Investigating the role of demography and selection in genome scale patterns of common and rare variant diversity in humans

Alexander Mörseburg

Department of Archaeology and Robinson College,
University of Cambridge
December 2018

Supervisor: Dr Toomas Kivisild
External Advisor: Dr Qasim Ayub

This thesis is submitted for the degree of Doctor of Philosophy

i. Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the introduction of each chapter and/or specified in the text.

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my thesis has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text.

It does not exceed the word limit of 80,000 words excluding references and appendices prescribed by the Archaeology and Anthropology Degree Committee.

Alexander Mörseburg

Cambridge, December 20th, 2018

ii. Abstract

Investigating the role of demography and selection in genome scale patterns of common and rare variant diversity in humans

Alexander Mörseburg, Department of Archaeology and Robinson College

In the last decade, an unprecedented increase in the availability of whole genome sequence (WGS) data has reshaped the field of human evolutionary genomics. However, many earlier sequencing projects like the HapMap and 1000 Genomes panels focussed on a limited set of populations. Therefore, more research has been required to better characterise genetic diversity in understudied regions, such as Island Southeast Asia and Siberia. This thesis contributes to this ongoing effort in the form of three partially related subprojects.

Firstly, population structure and local adaptations in Southeast Asia were investigated using novel autosomal 730,000 SNP data from 146 individuals in the context of a larger worldwide panel of 1,825 humans. The Kankanaey Igorot from the highlands of the Philippine Mountain Province were highlighted as the closest living representatives of the source population that may have given rise to the Austronesian expansion. Furthermore, consistent with archaeological, cultural and linguistic evidence of Indian influence in Southeast Asia starting from 2.5 kya South Asian admixture in the region was estimated to date back to the last couple of thousand years.

To provide an unbiased high-resolution picture of the patterns of functional and rare variants worldwide high coverage WGS data from 483 individuals (including 379 novel genomes) were analysed. Ingenuity Variant Analysis and the Ensembl Variant Effect Predictor were applied to a subset of these genomes (n = 382) to create a repository of functional and deleterious variants. Evidence for purifying selection in genes involved in pigmentation and immune defence against viruses was detected in African populations. The most differentiated sites across continental groups were integrated with haplotype-based selection tests and annotations from functional databases to pinpoint disease and metabolism-related candidate loci.

A subset of the WGS dataset, designed to maximise coverage of diverse ethnic groups (n = 447), was screened for variants occurring exclusively in two individuals in a heterozygous state (f_2 variants). It was shown that f_2 sharing correlates well with the results of CHROMOPAINTER, a state-of-the-art method to detect recent gene flow, and, allows for the detection of cryptic relatedness among distant populations. This was demonstrated by an example of a previously undetected low-scale African ancestry component in the South American Calchaquíes putatively related to the transatlantic slave trade.

Table of Contents

iii. List of figures

iv. List of tables

v. Acknowledgements

This thesis could not have been written without the support of numerous people and institutions. I am particularly grateful to my supervisors, Dr Toomas Kivisild and Dr Qasim Ayub. I am very much indebted to Toomas for his mentorship, his kindness and being so generous with his time. He is an inspiration as a scientist as well as a human being. To Qasim, for lending his expertise, his invaluable suggestions have improved this work greatly.

There are various people at LCHES who deserve a particular mention here and contribute to making it a great place to work. I thank my departmental graduate tutor, Dr Robert Attenborough for his unwavering support. I would also like to thank Silvia Hogg, Joanna Osborn and Pippa Milligan for being the good souls of LCHES. Furthermore, I am grateful to Fabio Lahr for the IT support. To Robin Morrison, for being a great office mate for more than three years and her feedback on data visualisation. I would also like to thank all members of the former Kivisild group at Cambridge for their advice and companionship. In particular, I am indebted to Dr Alexia Cardona for her bioinformatics support in the early stages of this project. To Dr Luca Pagani, for being a role model and a great collaborator on various projects over the years. To Anthony Wilder Wohns, who proofread parts of this thesis and to Dr Christina Eichstaedt for sharing her expertise on Andean populations.

I would like to thank Dr Mait Metspalu from the University of Tartu and his team, especially for granting me access to their HPC clusters which was crucial for the memory intensive bioinformatic analyses. Dr Chris Tyler-Smith, Dr Yali Xue and Yuan Chen kindly shared their insights on the analysis of loss-of-function and rare variants. I learned a lot from Prof Bill Amos and Dr Anders Eriksson who discussed variant sharing patterns and the use of demographic simulations with me.

Furthermore, I am indebted to Robinson College for supporting my research with a bursary, to Dr Brian McCabe, my tutor at Robinson for his excellent clerical support and to the Robinson MCR for being a great community of some of the most fun people I have ever met.

Finally, I am more grateful than I can express here to my friends and family. In particular to Alexander, Max, Sebastian and Stephan, who endured the monologues about my dissertation or provided a welcome escape, depending on what I needed at the moment. To Mandy, for having my back in small as well as big matters. And last but certainly not least, to my parents Brunhilde and Horst, for their encouragement, patience and support in all kinds of ways throughout this intellectual marathon. You are best parents I could wish for and I dedicate this thesis to you.

1. Literature review and thesis rationale

This chapter begins by introducing the scope of genetic variation in modern humans and then proceeds to provide a summary of the forces that have shaped it at the molecular level. It is followed by a brief overview of the main methodological and analytical advances in the field in the last decade and how they have improved our understanding of the recent evolutionary history of our species. The chapter concludes with a more detailed review of previous work in specific fields relevant to this thesis in sections 1.5, 1.6 and 1.7. In section 1.8 questions addressed in this work are highlighted.

1.1 Human genetic variation

1.1.1 SNPs

Every human being alive today is biologically unique. Genomic individuality, which even extends to monozygotic twins due to postzygotic somatic mutations (Dal et al., 2014; Ouwens et al., 2018), underlies this phenotypic singularity. Explaining how patterns of genetic diversity arose during the history of our species and what role they play in adaptive processes are the main challenges of human evolutionary genetics today.

Genetic variation manifests itself in elements of various lengths. The shortest and structurally simplest are single nucleotide polymorphisms (SNPs). In a broad sense, a SNP can be defined as a genomic site at which one base of the DNA sequence is substituted by another or a single base is inserted or deleted (indels). Furthermore, the minor allele frequency (MAF) of this change should be at least 1% in a population (Vignal et al., 2002). It should however be noted that the mechanisms which generate single-bp subs and single-bp indels are quite distinct (see section 1.1.2). In consequence, some authors (e.g. Sachidanandam et al., 2001) have adopted a narrower definition of the term SNP confined to the former. Rare variants (MAF <1%) have only recently become accessible due to improvements in sequencing technologies. In this thesis "variant" will be used as a general term to describe all differences between genome copies irrespective of their frequency. The term SNP will be used in a narrow sense, i.e. interchangeably to single base-pair substitutions and in the context of this work is limited to them unless single-bp indels are included explicitly. The total number of SNPs known in

humans is currently estimated at ca. 661 million (based on dbSNP build 151 as of 12/11/18, https://www.ncbi.nlm.nih.gov/projects/SNP/snp_summary.cgi?view+summary=view+summary&build_id=151).

The average number of SNPs relative to the human reference sequence in a randomly chosen African genome ($4.31*10^6$) is higher than the average number of SNPs in an Eurasian genome ($3.53*10^6$) (The 1000 Genomes Project Consortium, 2015) because of larger long-term effective population size in Africa (for a definition of this concept, see section 1.2.2 and for a discussion of the population history generating the observed patterns, see section 1.4.2).

Despite our wide-spread geographic range and current population size of more than seven billion these numbers are largely determined by long-term processes of natural history. This is supported by the observation that many great ape subpopulations exhibit higher levels of genome-wide heterozygosity than African and non-African humans despite their reduced habitats and correspondingly small population sizes today (Prado-Martinez et al., 2013).

Broadly speaking the functional impact of a SNP is dependent on its genomic context. The most basic distinction can be made on whether a variant is located within a gene or in proximity to one (genic and gene-related) or not (intergenic). Another way of distinguishing SNPs is by regions that are subsequently translated into amino acid sequences, i.e. are part of the protein coding sequence and those that are not.

Less than 1% of all SNPs in an average genome lie in protein-coding sequences (The 1000 Genomes Project Consortium, 2015). The potential importance of non-protein-coding SNPs in a regulatory context has been appreciated more in the last decade as our knowledge of these processes has expanded. However, in most functional studies protein-coding SNPs have been the focus. There are two groups of mutations that have a direct impact on the amino acid sequence, collectively referred to as non-synonymous mutations. The first is known as missense mutations; they cause the substitution of the original amino acid by another. The second are nonsense/stop-gain mutations, in which a triplet coding for an amino acid is replaced by a stop codon. Another class are called synonymous mutations in which there is a change observed at the base pair level, but the encoded amino acid remains the same. The functional implications of these variant classes are described in more detail in section 1.6

1.1.2 Short indels and other types of genomic variation

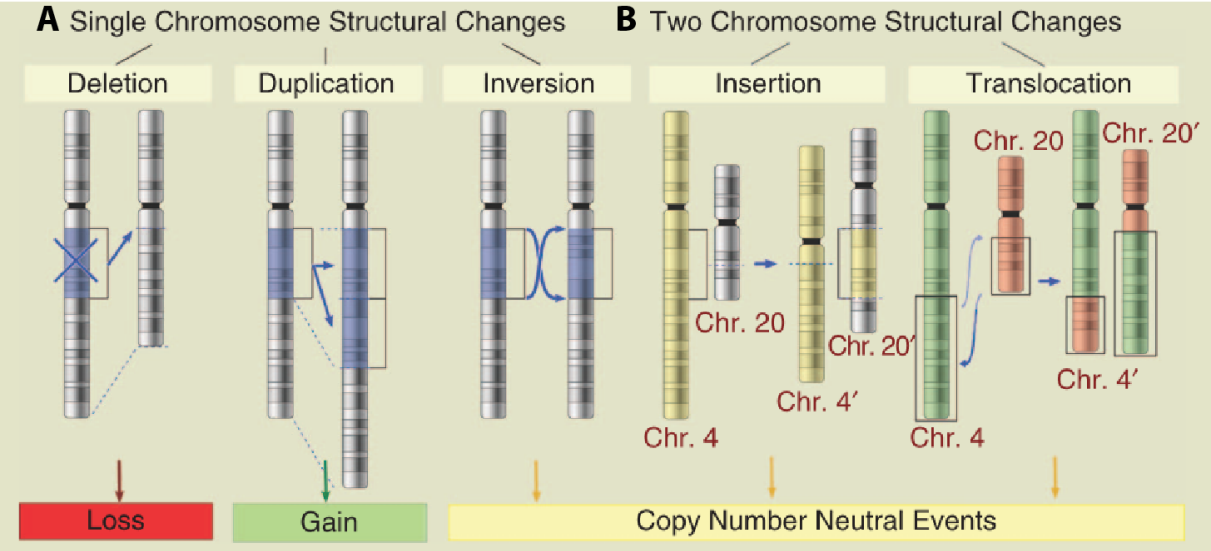


Figure 1.1: Different types of structural variation. A) A subset of these only affect one chromosome and can increase or decrease the number of copies of a certain region relative to the reference. B) Some of them are on a larger scale and involve multiple chromosomes, with often severe functional consequences (adapted from Alqallaf et al., 2013).

SNPs are the most studied type of sequence difference, however other types of genetic variation (Figure 1.1) can also be informative about human evolutionary history and can have a considerable impact on the phenotype.

Short indels are here defined as comprising all insertion/deletion polymorphisms up to 50 bp. As mentioned above (section 1.1.1), there are substantial differences in the mechanisms underlying single base-pair substitutions and indels which is also true for the estimated severity of their impact. The former, which are SNPs in a narrow sense, mainly arise due to misincorporation of nucleotides during replication and mutagenesis caused by various factors (see section 1.2.1). On the other hand, the majority of small indels are caused by DNA polymerase slippage, with other less well understood mechanisms also contributing (Montgomery et al., 2013).

On average, indels are assumed to have a more negative effective on phenotypical fitness than SNPs. Their impact is dependent on their position in the genome relative to other functional elements. The first possibility is that the indel is outside of the open reading frame (ORF), i.e. beyond all the regions of a gene contributing to the relevant protein product. The impact of this change is dependent on whether the relevant region plays a role in regulatory processes and can be assumed to be comparable to non-protein-coding SNPs. Indels which lie in an ORF can cause a frameshift in the translation machinery if they are not a multiple of three bases, which can

lead to devastating consequences on the protein product. There are however protein quality-control mechanisms which under certain conditions can lead to the degradation of these molecules (see section 1.6.2). Lastly, in-frame indels most likely have comparable impact to non-synonymous SNPs (Montgomery et al., 2013). The focus of this dissertation is on small scale variation which constitutes the vast majority (>99%) of all described variants.

1.2. Origins of human genetic diversity

1.2.1 Molecular forces creating genetic diversity

In a sexually reproducing species genetic variation arises because of two processes: *de novo* mutations and recombination events. As described in section 1.1., this variation manifests itself in the genome on different levels. Frequency changes of genomic variants mainly reflect demographic history and the impact of different kinds of directional selective forces, the latter often tightly linked to environmental factors.

While "mutation" is a summary term for any change in the DNA sequence, the class most widely used in demographic analyses and in this thesis are single base pair substitutions, also known as point mutations. Biochemically these are mainly the result of the mis-incorporation of nucleotides during replication while to a lesser extent they are also the outcome of spontaneous mutagenesis due to other endogenous processes and external chemical and physical mutagens (Cooper and Krawczak, 1990). Generally, DNA replication operates with very high precision and in case of a wrongly incorporated base there are stringent repair mechanisms; the error rates of DNA replication in different eukaryotes have been estimated as ranging from 10^{-9} to 10^{-11} bp^{-1} for each round of replication (Drake et al., 1998). The rate of germline mutations on the human lineage is a crucial parameter for theoretical studies in evolutionary genetics and more applied problems, e.g. understanding of the incidence rates of Mendelian diseases. In the last few decades a wide variety of approaches has been used to estimate germline mutations rates in the human lineage (reviewed in Scally, 2016).

One method is an indirect phylogenetic calibration using the known genetic divergence between two species and the time since they split according to the fossil record. From these calculations an approximate value of $1.0*10^{-9}$ bp^{-1} yr^{-1} for the yearly mutation rate on the hominin lineage was derived (Nachman and Crowell, 2000). Recently, advances in whole genome sequencing

(WGS) technologies have made new approaches to estimate mutation rates possible. The most direct of these is family sequencing where genomes from parent-child trios are compared and the new mutations in the offspring are divided by the whole length of the currently accessible genome to yield rate estimates for the last 1-2 generations (Genome of the Netherlands Consortium, 2014; Roach et al., 2010; Wong et al., 2016). Other approaches enable researchers to estimate mutation rates for a somewhat older time scale. An example of this is a method which compares chromosomes within an individual and then infers the mutation rate using the expected heterozygosity in diploid genomes and a fine-scaled recombination map of the human genome (Lipson et al., 2015). Where high quality ancient DNA (aDNA) data from an externally dated sample are available the mean number of additional mutations in present-day individuals can be divided by the time separating present-day humans and the ancient sample to yield a mutation rate (Fu et al., 2014a).

These estimates cover a range of approximately $0.4\text{-}0.6*10^{-9}$ bp^{-1} yr^{-1}, which is considerably lower than the rate obtained using the interspecies approach with fossil dates as calibration. One possible explanation for these differences is heterogeneity in germline mutation rates in space and time. Generally, it is assumed that the mutation rate has slowed down in more recent times on the primate tree. When using a low mutation rate to estimate ancient species splits based on molecular divergence this yields dates (e.g. for human and macaque) much older than the current fossil evidence seems to suggest (Scally and Durbin, 2012). These heterogeneities could be the result of variation in life history variables, e.g. generation time and of processes at the molecular level such as the efficiency of DNA repair. Both affect the number of germline mutations as the latter mostly result from replication errors during the cell divisions leading to gametes.

There is growing evidence for population-specific variation in modern human mutation rates, particularly for an increase in non-Africans vs Africans (Harris, 2015; Mallick et al., 2016; Mathieson and Reich, 2017). Irrespective of this, given the current state of evidence it seems reasonable to assume a lower mutation rate of ca. $0.5*10^{-9}$ bp^{-1} yr^{-1} for inferences about recent evolutionary events on the human lineage, i.e. within the last few hundred thousand years as suggested by Scally (2016).

The other main process generating genetic variation is recombination. Most recombination events are caused by a mechanism known as “crossing over”. This term describes the exchange

of genetic materials between paired maternal and paternal homologous chromosomes in the germ line during meiosis.

When studying two polymorphic loci on the same chromosome it can be measured to what extent certain allelic states at these two positions are inherited together. If the latter is the case the loci are said to be linked, i.e. few recombination events have occurred in the base pairs between them in the history of the respective population. This property can formally be described as linkage disequilibrium (LD) and quantified by various statistics such as D´ and r^2 (VanLiere and Rosenberg, 2008). Both can be considered normalised representations of the

Figure removed for copyright reasons. Copyright holder is Garland Science, Taylor & Francis Group, LLC.

Figure 1.2: Generation of genetic diversity through recombination: a) an example region in the genome contains four bi-allelic SNPs. Without recombination events this results in five possible haplotypes. b) when one recombination point is introduced this results in four additional haplotypes (adapted from Jobling et al., 2013).

correlation of particular alleles at adjacent loci, i.e. whether these alleles are more often passed on together than would be expected if they were inherited independently. A combination of

closely linked allelic states of polymorphisms along the same chromosome is described as a haplotype. In practice, these haplotypes are inferred by "phasing" algorithms (Browning and Browning, 2011), as shotgun sequencing typically yields genotypes without explicit phase information except for those located on the same short read. Recombination increases the number of haplotypes and therefore the observed genetic diversity (Figure 1.2). Generally loci which are physically close in the genome are also more often inherited together, however there is no a simple linear correlation of recombination probability and physical distance (Christiansen, 2008). The genetic map (of a species or a population) describes the distribution of linkage patterns and therefore implicitly also recombination probabilities across the genome. Understanding the variation in recombination patterns across the genome is relevant for a wide range of applications, including genome-wide association studies (GWAS).

It is also relevant for approaches using the length of shared genetic segments between individuals to infer the extent and time depth of shared ancestry (see section 1.3.3). The two main approaches by which a genetic map can be inferred are observing patterns of LD in a sample from a population or by tracking down recombination events in pedigrees.

The most well-known examples for the first approach are the HapMap (Frazer et al., 2007) and 1000 Genomes (The 1000 Genomes Project Consortium, 2010) projects which used high density SNP and whole genome sequence data respectively to estimate LD from worldwide population samples. Kong et al. (2010) reconstructed the first pedigree-based map based from over 15,000 parent-child pairs from Iceland, which had a resolution to a level of 10 kb and yielded more detailed information on sex-specific recombination events. Both studies confirmed that recombination probabilities are unevenly spread across the genome. The majority of recombination events (~60%, according to Frazer et al., 2007) occur within so-called recombination hotspots, most of which are only 1-2 kb long (Myers et al., 2006) and therefore cover only a relatively small fraction of the genome.

An important mechanism where a stretch of DNA is copied nonreciprocally from one chromosome onto the other while these are paired during meiosis is known as gene conversion. It can occur alongside crossing-overs or without them. This depends on how the heteroduplex characteristic for recombination events is resolved (Duret and Galtier, 2009). In this specific context, the term "heteroduplex" describes a structure which consists of one strand originating from the paternal copy of the chromosome and the one originating from the maternal copy, assuming that there are allelic differences between the two parental copies of a chromosome in

this region. The importance of gene conversion for understanding human evolution lies in a well-studied bias where GC alleles are favoured over AT alleles which leads to an increase in the frequency of the former over long evolutionary timescales (Glémin et al., 2015). This process is a relevant confounder for population genetic analyses at it can mimic or counteract selective processes (Capra et al., 2013)

1.2.2 The neutral theory and demographic forces that shape genetic variation

It first became possible to measure genetic diversity in humans and other species directly in the 1960s when small, functionally relatively well understood proteins were sequenced. The degree of polymorphism observed was many times greater than previously expected leading to the need for explanations such as offered by Kimura's (1968) neutral theory of molecular evolution. Its core statement is that the fate of most newly arising mutations is determined by stochastic processes, subsumed as genetic drift, and that selective forces mainly eliminate deleterious mutations from the gene pool. Genetic drift results from the observation that the genotypes in each generation of a species are a finite sample from the previous generation and that this sampling introduces randomness. The current weak version of the theory, arising from long debates between the "neutralists" and "selectionists" (reviewed in Nei et al., 2010) states that while the neutral theory applies to the majority of sites in the human genome at any given point in time in its strict version it is not sufficient to explain the observed patterns of genetic variation and adaptation in humans and other species (Jobling et al., 2013).

Assuming certain simplifying conditions of which the most important are constant population size, random mating and non-overlapping generations the expected patterns of genetic diversity under neutrality can be described by an approach known as the Wright-Fisher model (Fisher, 1930; Wright, 1931). It yields many quantitative predictions about genetic properties and therefore despite the abstractions involved can serve as a useful null hypothesis against which departures from these conditions can be detected. Therefore, it forms the ultimate theoretical basis underlying many of the analyses presented in this thesis.

Deviations from this simple model not only occur due to selective forces but also because of a range of demographic processes which violate the assumptions stated above and create distinct genetic patterns.

A way to quantify changes in population size and to understand their impact on genetic variation is the concept of effective population size (N_e). It was originally defined as the size of a population under the Wright-Fisher model which experiences the same amount of drift as the studied population (Crow and Kimura, 1970). Many factors are known to influence N_e (Charlesworth, 2009) and its relationship to the actual (census) population size N is complex. N_e is more than five orders of magnitude lower than N in humans, one reason being that there is high reproductive variance and therefore some individuals do not pass on their genetic material. Furthermore, the extent of genetic variation in a population is determined by its long-term N_e, which is the harmonic mean of the values of N_e over time, and, heavily affected by the lowest value of N_e across the whole time series. Therefore, in a recently expanded species such as humans the smaller ancestral population sizes largely determine the level of genetic diversity. It was first estimated as ca. 10,000 based on limited nucleotide diversity data in the 1990s (e.g. Takahata, 1993). In recent years WGS data have been used to infer human long-term N_e. Its value for the human ancestral population before the split of Africans and non-Africans was estimated at 9,000 (Gronau et al., 2011), which was subsequently corrected upwards to potentially as high as 14,000 as more genomes have become available (The 1000 Genomes Project Consortium, 2015).

Population size reductions that influence allele fixation rates are mainly caused by two processes: bottlenecks and founder effects. Both lead to a decline in the degree of polymorphism at many loci and consequently to a lowered overall diversity and heterozygosity. However, the underlying processes are different: the term bottleneck describes the sudden reduction of the size of an original population. A founder effect occurs when a subset of a larger population is spatially separated from its ancestral source group, e.g. during the colonisation of new territory.

The effect of geographical structure on N_e compared to random mating is complex. Under the assumption that there is some gene flow due to migration between the subpopulations and that humans as a whole are a continuous population the changes in genetic similarity over geographical distances can be modelled by an isolation by distance model (Wright, 1943). It states that mating choices are restricted by distance and that therefore the frequency of reproductive dispersals is negatively correlated to geographical distance.

1.2.3 Natural selection

Natural selection can be defined as all the deterministic forces leading to the differential propagation of an allele due to its impact on the phenotype (Vitti et al., 2013).

The origins of this concept lie in the seminal works of Charles Darwin and Alfred Russell Wallace and it was elaborated and mathematically formalised by Fisher (1930) and Haldane (1949) among others. In the following the main modes of selection and their effects on diversity are briefly outlined, computational methods for identifying selection signals will be described in more detail in section 1.3.4 and examples for selection at specific loci will be discussed in appropriate contexts.

When a variant increases the evolutionary fitness of its carrier it will likely undergo **positive selection** leading to an increase in frequency of the selected allele. This also affects surrounding neutral variation, which is referred to as "hitchhiking" effect. Especially when this process happens rapidly, which is also known as a selective sweep, this causes a reduction in genetic diversity surrounding the causal site (Smith and Haigh, 1974).

Conversely negative or **purifying selection** occurs when a variant has a negative effect on survival and/or reproduction and therefore declines in frequency. Most novel mutations are thought to be slightly deleterious on average and subject to this mode of selection (see section 1.6.). Therefore, genetic regions under strong purifying selection exhibit significantly reduced diversity as novel variants are removed from the gene pool before they can reach higher frequencies (e.g. Li, 1997).

Balancing selection is best understood as a process where multiple alleles are maintained at intermediate frequencies at a certain locus as this state is more beneficial than the fixation of any of these variants. There are two main mechanisms of balancing selection. The first is heterozygote advantage where the heterozygote state is actively selected for because of the positive effect it confers (Kwiatkowski, 2005; Poolman and Galvani, 2007).

Another possibility is frequency-dependent selection. In a simple case a rare allele provides a selective advantage and will lose this property over time once it reaches a higher frequency. The latter mechanism has been especially argued for in the coevolution of humans and their pathogens. In this scenario the carrier of a rare cell surface antigen will have an advantage because the pathogen will not yet have had the time to adapt to infecting the cell with this particular surface marker (Takahata and Nei, 1990).

1.3 Methods for genome-scale analysis of human population history

1.3.1 High coverage SNP arrays

Historically, processes shaping human genetic diversity were investigated using a subset of microsatellites, mtDNA and the non-recombining region of the Y-chromosome. These pioneering studies yielded valuable insights (e.g. Cann et al., 1987 for mtDNA). However, they presented only a limited perspective on overall variation. In the last two decades a more unbiased picture has gradually become available due to technological advances.

SNP array technology represents the first approach relevant in this context. The most important commercially available sets were developed by Affymetrix and Illumina. Both utilise a hybridization approach where fluorescence-marked fragments of single-stranded DNA bind to arrays containing a large set of unique nucleotide probe sequences representing the variable positions (reviewed in LaFramboise, 2009). The first incarnations of these technologies yielded information on thousands of markers on one chip while the resolution had increased to nearly one million by the end of the 2000s.

One of the first large scale projects to apply these technologies was the International HapMap project. Its main goal was to construct a common SNP-based genome-wide haplotype map to aid searches for disease susceptibility genes. Its first phase focussed on three groups representing Europeans, East Asians and West Africans respectively: Utah Residents with Northern and Western European Ancestry (CEU), Han Chinese from Beijing (CHB) and Yoruba from Ibadan (YRI). Its final phase included ca. 1,200 individuals from 11 European, African, East and South Asian populations genotyped for $1.6*10^6$ markers (International HapMap3 Consortium et al., 2010). HapMap provided a comprehensive picture of common variation in humans and a haplotype map indicating the extent of LD in their genomes (International HapMap Consortium et al., 2007). The latter was of great importance for medical research using SNP array data. One of the first large scale GWAS for a range of common diseases was undertaken by the Wellcome Trust Case Control Consortium (2007), it involved 17,000 individuals typed for 500k SNPs.

Another large-scale effort of a similar nature is the Human Genome Diversity Project (HGDP), which was launched in 1991. Its primary aims were anthropological with a focus of collecting data from isolated indigenous people around the globe. Lymphoblastoid cell lines were created

from these samples which were made available to investigators in 2002 (Cann et al., 2002). The most seminal application of SNP array technology on DNA from these was a paper by Li et al. (2008) that reported genotype data for 938 individuals from 51 indigenous populations for 650k SNPs. This allowed for the most comprehensive characterisation of human genetic variation up to this point in time and provided a reference still widely used for comparative purposes.

In conclusion, SNP arrays provide a cost-effective way to study common variants in a population, for making inferences about population structure, differentiation and admixture as well as haplotype homozygosity which is informative of recent selective processes.

One weakness of SNP arrays is ascertainment bias. Early genome-wide SNP array studies underestimated variation in non-European, particularly African, populations (e.g. Lachance and Tishkoff, 2013). Furthermore, SNP arrays by their nature have limited power to detect rare variants and do not allow for the discovery of novel variants. Another disadvantage is that commercially available arrays have been designed to include only one or a few SNPs from each of the LD blocks identified by projects such as HapMap. Additional SNPs from these LD blocks would not provide more information for a standard medical GWAS approach as they do not represent independent tests. This often results in too low (Clark et al., 2003a) and/or imprecise (Pengelly et al., 2015) estimates of the extent of population-specific LD from SNP array data.

1.3.2 Whole genome sequencing

The limitations of SNP array approaches make the need for a more unbiased way to access information from across the genome apparent. This problem has been addressed by technological breakthroughs, which have allowed the fast and relatively cost-effective sequencing of hundreds to thousands of human genomes (Figure 1.3). In the following, the basic approach, the most important platforms and some crucial pioneering studies are briefly described.

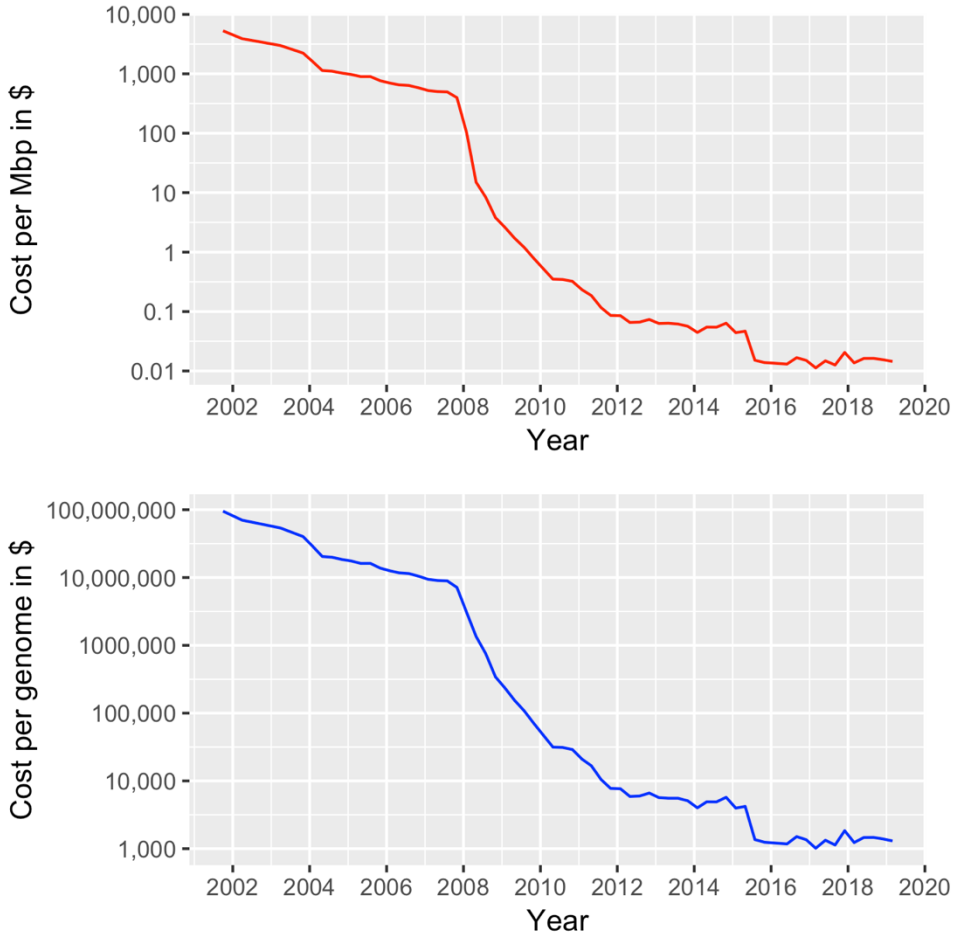


Figure 1.3: Decline of DNA sequencing costs at sequencing centres funded the by US National Human Genome Research Institute. The costs refer to sequencing to a Phred score-equivalent of 20 and to a human-sized genome respectively. The cost per Mbp refers only to the expenditures to generate the raw sequence data, whereas the cost per genome includes bioinformatic processing steps (data from Wetterstrand "DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP)" as of 12/09/2019, available from www.genome.gov/sequencingcostsdata).

The prevailing method is “shotgun sequencing”, which consists of four basic steps. First, nuclear genetic material from the individual of interest is broken down (either physically or enzymatically) into small fragments and a library is constructed from these fragments. These libraries are then sequenced in a massively parallel manner to obtain information about the sequence of nucleotides in these genome fragments. The resulting sequence reads are mapped to the reference sequence or assembled *de novo*. The length of the sequenced reads, which is highly variable between technology platforms, determines the length of accessible genome.

For the shotgun approach to be widely adapted, two conditions had to be fulfilled. Firstly, for many species, including humans, a reference sequence (presented by the International Human Genome Sequencing Consortium in 2001) was generated from scratch as an artificial composite

of DNA from many individuals. The reference genome is not static and is constantly updated incorporating our increased understanding of underlying genomic complexity. Secondly, there has been extraordinarily rapid progress in bringing down the costs and time required while increasing the output for each single step mentioned above. Currently, the next generation sequencing (NGS) market is dominated by the Illumina (Bentley et al., 2008) strategy. Each approach has different technical specifications, strengths and weaknesses (reviewed in Goodwin et al., 2016). The WGS sequences analysed in chapters 3 and 4 of this thesis were generated using technology from one of Illumina's main competitors, Complete Genomics (now belonging to BGI) (Drmanac et al., 2010).

These distinct features of the sequencing technologies and additional specifications in the bioinformatic protocols used to process the raw sequencing data result in considerable differences which need to be considered when interpreting data generated by different sequencing platforms. Generally, the Illumina approach is thought to operate with a very high sensitivity, however in its earlier versions, it was known to also have a relatively high false positive rate of 2.5% compared to micro-arrays (Bentley et al., 2008). Complete Genomics on the other hand reports a very low false positive rate, which can lead to true variants being missed (Bobo et al., 2016), this balance can be shifted by altering call quality thresholds (Drmanac et al., 2010). Follow-up studies indicate that the factors influencing error rates besides sequencing technologies are mainly related to the variant-calling algorithms used (Lam et al., 2011). In most cases the genotype quality scores obtained by downstream methods likely underestimate the true error rates for Illumina as well as Complete Genomics (Wall et al., 2014). Another contributing factor is that generally GC-content has been reported to result in higher sequencing coverage (Benjamini and Speed, 2012), even though decreased coverage for both AT-and GC-rich regions has been observed in some studies (Rieber et al., 2013).

Lam et al. (2011) systematically compared the genome of one individual sequenced by both Complete Genomics and Illumina technologies to a high coverage (76×). While ca. 88% of all SNP calls were concordant between the two platforms, the majority (60%) of the platform-specific inferences could be verified as true variants by independent approaches. Furthermore only 26.5% of all indels detected by the two approaches were overlapping. This very low concordance is most likely due to the higher complexity involved in calling these elements covering up to 200 bp by Lam et al.'s definition. One important effort to mitigate the problems

arising from these inconsistencies is the work of the Genome in a Bottle Consortium. They generated an integrated set of high confidence variant calls from different sequencing platforms and variant callers using a consensus approach that accounts for platform-specific biases for a small set of reference genomes (Zook et al. 2014, 2018). Thereby, they provided a benchmark against which sequencing and bioinformatics methods can be assessed.

NGS approaches have been applied to a vast range of questions of medical and anthropological relevance. The most important project, which greatly expanded the known range of human diversity and yielded insights into evolutionary history, is the **1000 Genomes Project**. In many ways it can be seen as a continuation and expansion of HapMap (The 1000 Genomes Project Consortium, 2010, 2012, 2015). Its primary aim was to provide a public resource of genetic variation for next-generation medical association studies, finding all accessible single nucleotide variants at a frequency of ≥ 1% across the genome in the studied populations. Its final phase yielded genomic information from 2,504 individuals hailing from Europe, East Asia, South Asia, some parts of the Americas and sub-Saharan Africa using a combination of low- and high-coverage full genome and exome sequencing. The term "exome" describes the totality of all exons across the genome, these are regions retained in the mature mRNA. It therefore encompasses all protein-coding genes including some flanking regulatory regions. It is analysed using a target enrichment approach where sequence-specific probes bind to single-stranded genomic DNA fragments which after purification and amplification are then sequenced (Coffey et al., 2011). Depending on the probes used it represents 1-2% of the total human genome. The **1000 Genomes Project** resulted in a total of ca. 88 million variants which were phased in haplotypes.

The **Exome Aggregation Consortium** (ExAC) (Lek et al., 2016) provided exomes from 60,706 individuals of diverse ancestries sequenced to a mean coverage of ~65×. Their contribution focussed especially on identifying genes for which the observed number of predicted protein-truncating variants is lower than would be theoretically expected and thereby identifying potential human knock-out genes (see section 1.6.2). As already mentioned, the coverage of the world in these datasets is not even, many indigenous populations from regions such as Siberia and Island Southeast Asia, but also from the Americas, Australia and Sub-Saharan Africa are not well represented. Several publications have attempted to close these gaps. Lachance et al. (2012) sequenced 15 genomes of African hunter gatherer populations, which considerably increased the known scope of human variation and uncovered potential adaptations relevant to

the pygmy phenotype. Subsequent studies have provided more whole genomes from these pygmy groups (Hsieh et al., 2016) as well as from across the whole African continent (Gurdasani et al., 2015).

On a global scale Mallick et al. (2016) recently reported the **Simons Genome Diversity Project** (SGDP) which provides high coverage sequences from 300 individuals representing a wide anthropological, linguistic and cultural diversity. Besides considerably improving the geographical spread of available human sequence data the authors found evidence for deep population splits of some African groups vs all other modern humans (>100 thousands of years ago, in the following abbreviated as kya). Malaspinas et al. (2016) presented whole genomes from 108 indigenous Australians and Papuans. They were the first to be able to give a deep picture of Australian population history and detected deep structure in populations of this region while generally supporting a scenario where all non-Africans descend from a single migration wave.

In the context of the previously unprecedented amounts of data which are now available, the developments in aDNA-sequencing can only be hinted at here (reviewed in Hofreiter et al., 2015 and Orlando et al., 2015). Together with the NGS techniques described above the improvements in enrichment capture have allowed the generation of high-quality whole genome aDNA data. This has offered researchers a window into the genomic past based on data from archaic hominins (Green et al., 2010; Meyer et al., 2012; Prüfer et al., 2013) and Palaeolithic modern humans (e.g. Raghavan et al., 2013) and from recent historical times (e.g. Rasmussen et al., 2011). Keeping in mind that the two main challenges in isolating and analysing aDNA are its degradation over time and the ubiquity of contamination the seminal studies mentioned above vary considerably in genomic coverage, but the obtained read lengths are very similar. For the former, the maximum achieved was 52× coverage for a Neanderthal individual from the Denisova cave in the Altai mountains in Siberia (Prüfer et al., 2013) compared to the lowest observed coverage of 1× observed for a 24 kya-old modern human individual from Mal'ta, also in Siberia (Raghavan et al., 2013). This range reflects differences in sample preservation as well as in the techniques used to retrieve and enrich DNA. The underlying aDNA fragment lengths fall mostly between 20-150 bp with the corresponding read lengths between 50-200 bp, i.e. in the lower range of most shotgun-sequencing analyses of modern DNA.

Overall, the last decade has seen major advances in the availability of WGS data, offering significant new insights into questions of evolutionary history and medical applications, especially in the rare variant spectrum. A limitation which affects many of the current high-throughput methods is their short read length of 35-700 bp depending on the specific approach used (reviewed in Goodwin et al., 2016). This results in a limited ability to correctly infer structural variation and to resolve information from repetitive regions of the genome. Current approaches to generate long reads are more expensive and slower but have a huge potential to expand our knowledge of the structure of the human genome by making previously unresolvable regions accessible (e.g. Seo et al., 2016).

The available datasets have increased in size and complexity during the last decade, which has brought new analytical and interpretative challenges. Some of the most frequently used approaches shall be summarised in the following sections.

1.3.3 Methods investigating population structure and neutral demographic processes

Since data from classical markers became available in the 1960s various statistical measures have been used to describe diversity within and among groups. These were mostly derived from the theoretical frameworks of F-statistics (Wright, 1951) and coalescent theory (Kingman, 1982). One commonly used example is π which describes the number of pairwise differences between two nucleotide sequences (Li and Sadler, 1991). Others include the heterozygosity H and the fixation index F_{ST}. The latter can be interpreted as the ratio of the variance of allele frequencies among subpopulations relative to the total variance, e.g. of humans from different continents relative to our species as a whole (Wright, 1965). In the last two decades these statistics have been partly superseded by approaches allowing a finer scale characterisation of population structure and inference of dynamic processes (reviewed by Schraiber and Akey, 2015).

When dealing with high coverage data consisting of hundreds of thousands or millions of markers typed across many individuals it is essential to get a first overview of the dataset to understand its basic characteristics and detect confounders. The most widely used approach for such purposes in population genomics is principal component analysis (PCA) (Price et al., 2006).

Assuming the dataset consists of n individuals an ($n \times n$) matrix can be constructed to describe the covariance of genotypes among individuals. PCA extracts the eigenvectors of this covariance matrix. Each of these vectors provides a coefficient for a linear combination of genotypes which most efficiently differentiates the various samples. The eigenvector explaining the highest proportion of variance is known as the first principal component (PC) and the following eigenvectors are ordered and named accordingly. The first two PCs are often plotted to obtain a PC space where individuals with similar genotypes will cluster. However, there are caveats to interpreting PC plots in terms of underlying population history. Novembre and Stephens (2008) used simulations to show that when spatial data are analysed and genetic similarity decays with geographic distance the application of PCA can result in highly structured patterns related to sinusoidal functions. Such a decay pattern can be produced by a homogenous short-range migration process and does not necessarily imply population expansions or long-distance migrations. More generally, PCA, like all unsupervised clustering methods, is highly dependent on sample sizes and the distribution of sampling locations.

In contrast to PCA there is a class of clustering approaches imposing an a priori stratification on the samples to find individuals sharing common underlying allele frequencies and to detect admixture events. They are summarised as STRUCTURE-like methods (named after the method published in 2000 by Pritchard et al. forming their conceptual basis). Apart from STRUCTURE itself FRAPPE (Tang et al., 2005) and ADMIXTURE (Alexander et al., 2009) are the most widely used of these approaches. Each individual (represented by a row in the input matrix) is analysed as the sum of k ancestral components. As mentioned above k is defined in advance and can be understood as the number of ancestral groups necessary to explain the observed diversity in each population. In the next step the genetic composition of each sample is described as a linear combination of these ancestral populations and the fraction coming from each ancestral group is estimated. The underlying technical details vary somewhat between approaches, for example ADMIXTURE, which is used in this thesis, utilises a likelihood method. The general goal of the latter class of approaches is to maximise the probability of the observed data occurring given the particular parameters of different models which are compared.

One potential error source are assumptions concerning the correct number of potential ancestral groups. Attempts have been made to optimise k, ADMIXTURE for example uses a method estimating the cross-validation error (Alexander and Lange, 2011). To calculate this quantity

a part of the genotypes in the dataset are masked/set to missing. ADMIXTURE analyses are run on the rest of the dataset and their output can be used to predict the missing genotypes. The goal is to minimise the error of this prediction. Furthermore, there are many evolutionary scenarios where the admixture of discrete ancestral groups that were (relatively) isolated for a previous time period is a misrepresentation of the true population history (Weiss and Lambert, 2014). The latter possibility needs to be considered for proper interpretation of the outcomes.

Furthermore, this class of approaches, like PCA, does not incorporate haplotype information and explicitly needs all entered polymorphisms to be independent of each other and therefore requires LD-pruning. This also means that it cannot make explicit inferences about which regions of the genome in an admixed individual derive from which putative ancestral group.

More recently, chromosome-painting methods addressing this need have been developed. These approaches are based on comparing each genome of interest (the recipient) to a set of phased chromosomes from a range of individuals (donors) serving as approximate matches for the putative ancestral sources. This comparison results in a probability for each segment of the recipient genome that this region derives from one of the donor individuals (Figure 1.4). These inferences require a model describing how ancestry changes along a genome. One of the most widely used approaches, which relies on a theoretical framework first proposed by Li and Stephens (2003), is ChromoPainter/fineSTRUCTURE (Lawson et al., 2012). As the results obtained with the latter will be incorporated in the analyses presented here, it shall be described briefly. The method consists of several subroutines. Firstly, ChromoPainter is used to construct a coancestry matrix x_{ij}, which displays the expected number of distinct genomic chunks shared by the recipient individual *i* with each possible donor individual *j*. The sum of these chunks over all chromosomes can be interpreted as a count of the number of recombination events leading to individual *i* being most closely related to *j* and is therefore an intuitive measure of ancestry sharing.

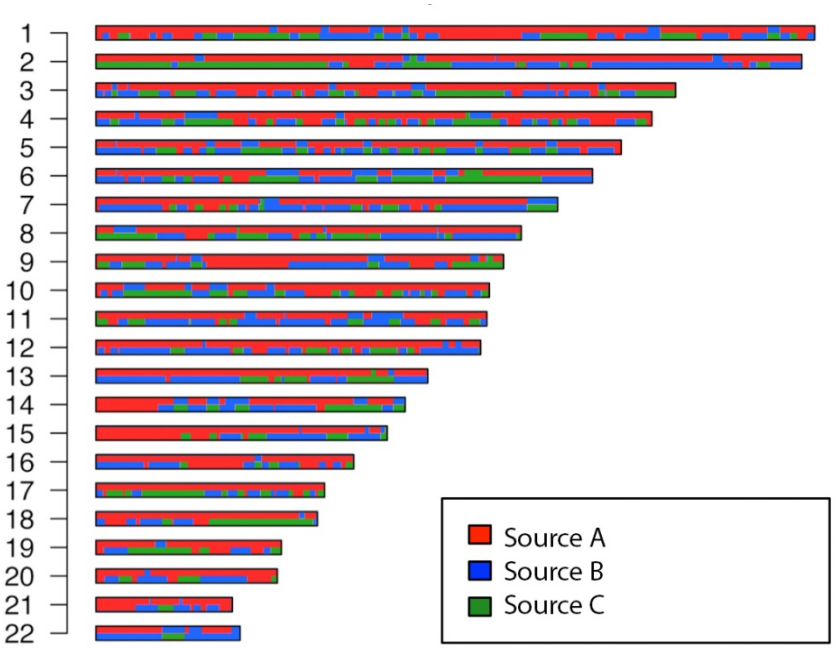


Figure 1.4: A possible visualisation of the autosomal local ancestry of a population derived from three source groups. In this case for most of the genome the maternal and paternal haplotypes in any given region are from different ancestry backgrounds. (adapted from Moreno-Estrada et al., 2013)

fineSTRUCTURE uses Bayesian inferences of populations to partition the dataset into *k* groups with indistinguishable genetic ancestry using a Markov chain Monte Carlo (MCMC) algorithm.

A population is defined such that a chunk from any recipient individual within this group has an identical donor and recipient distribution. This output can then be further analysed with the MCMC approach to yield a population tree.

One key limitation of ChromoPainter/fineSTRUCTURE is that it depends on the availability of good proxies for ancestral/donor groups which are not represented in the dataset. This has been partly addressed by the development of the GLOBETROTTER method (Hellenthal et al., 2014). It modifies and extends the fineSTRUCTURE approach accounting for ancestry from unsampled groups and uses the ancestral affinities and tract lengths to quantify and date the gene flow due to admixture events. One caveat with GLOBETROTTER and similar approaches is the underlying assumption that the admixture tract lengths are exponentially distributed and independent. This is a good approximation for low rates of admixture (Pool and Nielsen, 2008), however it has been shown to be less precise for recent strong gene flow (Liang and Nielsen, 2014).

Most approaches described until this point provide an understanding of the structure of the sample while working in an at least partly unsupervised manner. Other methods test complex

and parameter-rich models describing specific processes and properties such as population divergence times, migrations, gene flow events and changes in N_e. The majority of these are likelihood based and require the explicit specification of the model and starting estimates of relevant parameters.

These methods can be classified according to the type of data they require. A first class of approaches relies on the site frequency spectrum (SFS). It represents the distribution of variants present at each possible derived allele count in a sample of n chromosomes and has been shown to be a rich resource containing information on underlying neutral demographic history as well as selective events (Williamson et al., 2005). To enable quantitative estimates of the population history revealed by the SFS the methods described here use a Poisson random field model. The derived allele counts across the different frequencies are assumed to follow a Poisson distribution (Sawyer and Hartl, 1992), i.e. new mutations are presumed to occur in a population as a Poisson process, which is used in statistics to describe the incidence of high impact rare events.

The most widely used approach, which allows for up to three populations to be analysed relatively fast is known as "diffusion approximations for demographic inference" (*∂a∂i*) (Gutenkunst et al., 2009). It consists of the calculation of the expected SFS given a set of parameters specifying a demographic model under a diffusion approach which approximates the population genetics of a discrete set of individuals in discrete generations. The expected SFS is then compared to the observed SFS. The demographic model for which this similarity is maximised is chosen using a maximum likelihood framework.

Another approach which can yield information about the population history of multiple groups while being comparatively fast and requiring only a limited number of input parameters is TreeMix (Pickrell and Pritchard, 2012). It does not model the SFS as such but utilises a covariance matrix of allele frequencies between given populations while trying to account for LD by grouping neighbouring SNPs together. As in *∂a∂i* the model which yields the covariance matrix most similar to the observed one is chosen. TreeMix allows the estimation of the underlying population tree and admixture events between populations.

The F-statistics first introduced by Reich et al. (2009) also belong in this context. The F-statistics *sensu Reich* (as opposed to Wright's fixation indices) can be thought of as representing genetic drift shared between populations, i.e. the correlation of allele frequencies (Peter, 2016).

f_3 and f_4 statistics have interpretations as branch lengths in a tree describing the relationships of three and four taxa to each other respectively. The value of these statistics can be used to obtain the underlying tree topologies. If the assumed source groups are sufficiently diverged and the mixture proportions are high enough to have a detectable effect on the common allele frequency spectrum, admixture can be robustly inferred and quantified. TreeMix as well as the type of F-statistics *sensu Reich* described were used to analyse a part of the data presented here.

To analyse fine-scale population structure the focus should ideally be on haplotypes and on including all recombination events occurring across a genome in the models used to reconstruct demographic history. Applying the framework of coalescent theory, the genealogy underlying each region in the genome can be reconstructed, while recombination causes changes of these trees along the genome. The genealogies can be used to yield features of interest, e.g. N_e. The number of coalescent events detectable in a certain region of the genome is known to be inversely proportional to N_e. However, the full coalescent is very complex to model due to the long range correlations between sites along a genome and correspondingly an extremely large number of possible genealogies (Wiuf and Hein, 1999).

The sequential Markovian coalescent (SMC) framework proposed by McVean and Cardin (2005) addresses this problem by approximating the full coalescent under the assumption that the genealogy of the site where the recombination event occurs depends only on the genealogical tree underlying one neighbouring site.

The most widely used approach employing this framework is the Pairwise Sequentially Markovian Coalescent (PSMC) model developed by Li and Durbin (2011). In it each heterozygous site in a genome of interest is interpreted as resulting from the contributions of two haploid genomes. The density of heterozygous sites in each region can be used to infer the time to most recent common ancestor (TMRCA) between these two hypothetical genomes. A clustering of heterozygote sites indicates an older age of the common ancestor in a particular section of the genome. As mentioned above changes of N_e over time can then be inferred from the distribution of coalescent events throughout the past. PSMC has several advantages over other methods as it is non-parametric and no specific underlying (growth) dynamics are assumed and the input data do not require phasing. However, in its original version it is limited to a single individual. An extension of PSMC known as multiple SMC (MSMC) (Schiffels and Durbin, 2014), which requires phased data, can handle multiple individuals from different populations and yields estimates of the coalescence times within and between groups which can

in turn be used to approximate population split times. Besides the SMC-related there are a range of other methods to estimate N_e, each having different strengths and weaknesses, i.e. certain methods might give better estimates for certain timeframes.

Overall, the last two decades have seen great advances in our ability to infer human demographic history. Generally, there has been a move towards more sophisticated models often based on a maximum likelihood framework. While the exact likelihoods can in practice only be calculated for simple models there are robust frameworks to approximate these for many scenarios. Finally, due to the stochastic nature of coalescence processes we can never obtain a full picture of past demographic events solely based on present-day diversity.

1.3.4 Methods testing for signatures of natural selection

As described in section 1.2.3, selective processes can be subdivided according to their modalities and the resulting effects on genetic diversity. The focus of this section is on approaches detecting positive selection from genome-wide data; purifying selection will be considered in sections 1.6.3 and 1.6.4.

The approaches reviewed here infer events on a microevolutionary scale within our species as opposed to a macroevolutionary perspective, where the focus is on adaptations specific to modern humans. Methods to detect the latter include the K_a/K_S (also known as d_N/d_S) (Hurst, 2002), the McDonald-Kreitman test (McDonald and Kreitman, 1991) and the Hudson-Kreitman-Aguadé (HKA) test (Hudson et al., 1987). They are based on the comparison of the ratio of different site classes and/or types of polymorphism across various species to detect changes particular to one lineage. Generally, an excess of species divergence as opposed to intraspecies polymorphism could result from positive selection in a specific lineage whereas an excess of intraspecies polymorphism is thought to reflect balancing selection.

The earliest studies investigating the genetic basis of adaptations were based on a forward genetics approach. Phenotypes were hypothesised to be targets of natural selection and the underlying genetic basis was subsequently identified. One example of such a trait in humans is lactase persistence, for which positive selection was suggested as early as 1973 by Cavalli-Sforza and signatures of strong selection in the region of the lactase gene (*LCT*) were later detected by Bersaglieri et al. (2004). Another phenotype of this kind is skin pigmentation, its adaptive value was debated throughout the 20th century (e.g. Blum, 1961) and signals for

positive and purifying selection were subsequently found in multiple pigmentation related-genes (Harding et al., 2000; Norton et al., 2007). Technological advances allowed a shift towards approaches scanning the whole genome to identify candidate region targeted by natural selection without a priori knowledge of the underlying phenotypes (Sabeti et al., 2007; Voight et al., 2006). A plethora of methods have since then been applied to find candidates for positive selection on a genome-wide level (reviewed by Vitti et al., 2013; Wollstein and Stephan, 2015).

The basic rationale of most methods is to calculate a certain test statistic capturing one aspect of genetic diversity for one or multiple regions in the genome and to compare it to the expectations under a model of neutral evolution. It has been noted (e.g. Nielsen et al., 2007) that deviations from neutrality can also be caused by demographic events. One way to further evaluate the results of neutrality tests is to expand on the simple null hypothesis and instead use multiple non-neutral hypotheses for comparison. The demography of a specific population can also be explicitly modelled to identify loci with unusual properties given a particular demographic history.

A complementary approach is known as outlier analysis. It assumes that selection is locus-specific while demographic history should affect all regions of the genome more uniformly. Therefore, the empirical distribution of a statistic over all loci/regions of the genome should approximately reflect the null model. A threshold is then set and genomic units exhibiting a value of the test statistic more extreme than it are considered as candidates. One problem with outlier approaches it that it is very difficult to state with certainty what fraction of the genome has experienced a specific selective pressure. Therefore it has been recommended to consider outlier analysis as an enrichment tool for loci which require further analysis instead of providing conclusive evidence for selection on its own (Akey, 2009).

Another classification of tests depends on the different ways putative selective events affect the observed patterns of human diversity. Different tests detect a range of signals (Figure 1.5), which also has implications for the approximate underlying selection coefficients and time scales which can be investigated. Under the assumption that different populations are living in specific environments and are therefore subject to distinct selective pressures, frequency changes at particular loci might be caused by these different selective regimes. Accordingly, a locus with extreme frequency divergences between populations compared to the average difference could be the result of selection. A basic statistic to examine these patterns is known as ΔDAF. It is defined as the absolute value of the difference in the DAF between two

populations for a particular site (Colonna et al., 2014; The 1000 Genomes Project Consortium, 2012, 2015). It is considerably more powerful when unbiased WGS data are available. Another simple measurement of this particular effect selection has on genetic diversity is the F_{ST}. A high value at a particular locus would imply strong population differentiation potentially due to selection (Akey, 2002).

A wide range of statistics either derive from the F_{ST} or the ΔDAF. One example is the divergence statistic d_i introduced by Akey et al. (2010). It provides a measurement of locus-specific population structure (here F_{ST}) for a subgroup *i* of a species compared to all other subgroups normalised by the genome-wide average. This possibly indicates adaptive processes acting in a specific lineage. Some of the genomic data in this thesis will be analysed with the ΔDAF as well as d_i.

False positives are a challenge for the selection tests based on population differentiation, as strong population structure and drift can generate extreme allele frequency patterns.

As displayed in Figure 1.5 a selective sweep affects the SFS at neighbouring loci. In its early stages it causes a reduction of genetic diversity surrounding the locus. As time progresses novel mutations occur on this relatively homogeneous background, leading to an excess of rare alleles. The most commonly used test to detect this particular signal is Tajima's D (Tajima, 1989). It measures the number of pair-wise differences between individuals and compares it with the total number of segregating (i.e. polymorphic in an alignment) sites. As low frequency alleles contribute less to the number of pair-wise differences than intermediate frequency alleles this affects the difference between these statistics. Therefore, an excess of rare variants causes a negative D score which could either be due to positive selection or population expansion whereas balancing selection is thought to lead to an increase of intermediate frequency variants which would result in a positive D.

The genetic hitchhiking effect observed around the selected site can also be measured in terms of LD. As the causal allele and its neighbouring variants define a haplotype this particular combination of alleles will rapidly increase in frequency during the selective sweep as recombination acts much slower to break down the relevant associations. Therefore, a long (relative to the population average) high frequency haplotype is thought to be indicative of a selective sweep. This class of approaches is particularly powered to detect recent selective events (Vy and Kim, 2015). The integrated haplotype score (iHS) (Pickrell et al., 2009; Voight

et al., 2006) is a widely used test statistic to quantify these processes. It utilises the concept of extended haplotype homozygosity (EHH). The latter is defined based on a core allele (here the potential selection target) and an extended range spanning from it to a specified base. The EHH measures the probability than any two randomly sampled chromosomes within a group carrying the core allele are identical by descent (IBD) for the extended region. Given the molecular mechanisms underlying LD EHH decreases the further one moves away from the core region. For whole-genome data the iHS is calculated for each SNP and the area under the curve based

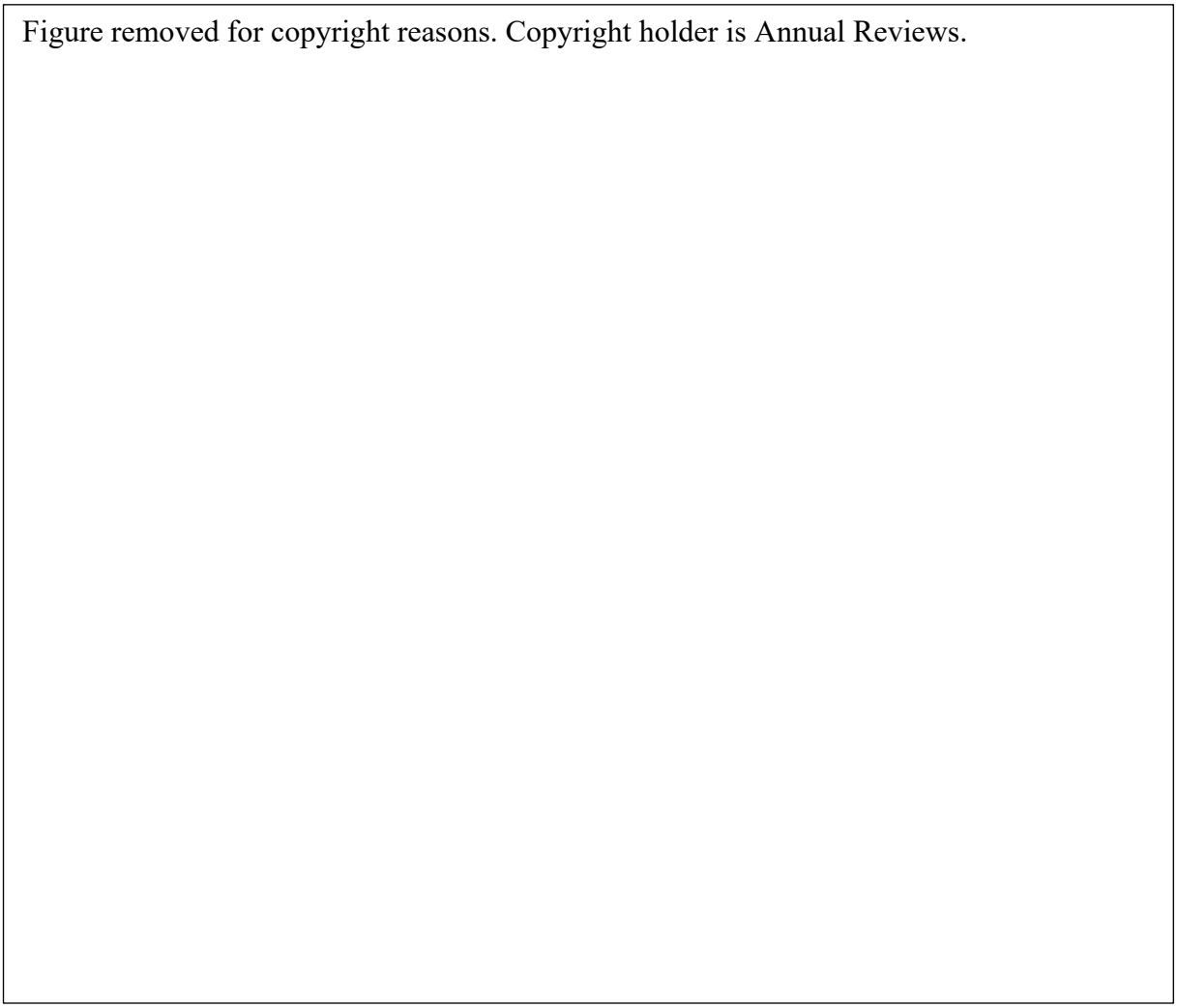


Figure 1.5: Methods detecting selective sweeps at the microevolutionary level. a) A beneficial mutation (red asterisk) rising in frequency also causes an increase in the frequency of derived variants surrounding the site. After completion of the sweep, novel mutations cause an excess of rare alleles. b) Extended haplotype homozygosity (EHH) is a measure of LD. It rises across the haplotype carrying the selected allele. After fixation novel mutations and recombination events lead to a breaking down of the EHH patterns. c) If selection acts in a specific population this will cause an increase in allele frequency differentiation as measured by the F_{ST} and other statistics. d) Composite methods integrate multiple signals of selection and can help pinpoint causal variants. The same can be achieved by WGS resequencing if the data were previously in genotype format (adapted from Vitti et al., 2013).

on the EHH is compared for the ancestral and derived alleles. This statistic is then standardised for the whole genome, empirical outliers with unusually long haplotypes surrounding either the ancestral or derived allele are considered sites of interest. The cross-population EHH (XP-EHH) (Sabeti et al., 2007) is another variation of the same framework. It relies on the comparison of haplotype lengths between populations and can therefore highlight localised population-specific selection. Simulations imply that the iHS and the XP-EHH are complementary: the former excels at detecting incomplete sweeps, the latter performs best for a sweep nearing completion in a particular population (Pickrell et al., 2009).

However, neither iHS nor XP-EHH were designed to pinpoint the causative selection targets at the variant level, they rather highlight regions with unusual haplotype homozygosity patterns. A tool known as derived intra-allelic nucleotide diversity (DIND) test (Barreiro et al., 2009) allows more precise inferences, at least for selection on derived variants. Under neutrality it is assumed that an allele with a high DAF has likely been segregating in the population for a long time. Therefore, it should be associated with high levels of diversity within the class of haplotypes defined by this core allele. For the DIND the standardised intra-allelic diversity is calculated for both the ancestral and the derived allele. These scores for the two alleles are then compared as a ratio statistic and plotted against the derived allele frequency (DAF) for all variants of interest. Outliers (either empirically or based on simulations) with high derived population frequency but low internal diversity at linked sites are then chosen as candidate variants.

Fagny et al. (2014) applied iHS and DIND to simulated and real low-coverage WGS data. Their outcomes indicate that both approaches are particularly powerful to detect selective sweeps targeting either a newly arisen allele or low frequency standing variation rising to high frequency in WGS data (80-100% at an allele frequency >0.4). They were also found to be robust to variation in recombination and mutation rates, negative background selection and coverage; however, they are sensitive to sample size.

One of the most widely used approaches which attempts to increase the power to detect selection by combining information from multiple tests is known as composite of multiple signals (CMS). It encompasses five selection tests including all three classes described above (Grossman et al., 2010, 2013). In its latest implementation a Bayesian framework is applied to calculate a factor for each test. This factor denotes the probability that a certain SNP with a particular score for a test statistic is under selection. The latter is inferred from comparing the

empirical test score to the bins of a spectrum of simulated scores for sites of which some are "known" selection targets (fixed in the simulation). These factors are then multiplied to obtain the composite score.

One major criticism of genome-wide selection scans has been the lack of reproducibility of selection signals. This has been attributed to the low power of individual tests to detect particular types of signals (e.g. Zhai et al., 2009) and signals being obscured and or mimicked by complex demographic histories which can be difficult to incorporate even in simulation approaches (Teshima et al., 2006).

In this aspect there have been some improvements in the last few years. Crisci et al. (2013) quantified the impact of particular demographic scenarios which deviate from a simple equilibrium model (e.g. bottlenecks) on the type I and type II errors for various haplotype homozygosity-based selection tests. Recent studies have indicated that more research is necessary to understand the statistical power of each test under different scenarios (Jacobs et al., 2016) and that caution should be taken against the over-interpretation of these results without additional functional follow-ups (Pavlidis et al., 2012). Nevertheless, there is a small but growing subset of loci mainly related to adaptations to agriculture, environmental variables and pathogen resistance (Fu and Akey, 2013; Wollstein and Stephan, 2015) which are consistently detected by selection screens.

A related question is the accuracy of the classic selective sweep model and its importance in human evolution. Generally, empirical data have revealed that the features which would be expected under the idealised strong and complete sweep scenario are relatively rare in human genomes. Among these are the low number of fixed differences between human populations and also the genome-wide absence of a pronounced decrease in diversity around human-specific missense mutations (Hernandez et al., 2011; Pritchard et al., 2010). However, work on the 1000 Genomes dataset implies a moderate but significant role for hard selective sweeps (Fagny et al., 2014). Soft selective sweeps and polygenic adaptations have been suggested as alternative modes of selection (e.g. Pritchard and Di Rienzo, 2010). While no definite consensus has been reached, it appears that multiple modes of selection have been important during recent human evolution and more work is required to understand the genetic signatures of those differing from the classic selective sweep model.

Another general trend has been relating selection candidates to functional information. This can be done a posteriori by using existing annotations from functional databases or by performing functional follow-up studies. Other approaches have explicitly incorporated information on environmental factors and have studied their correlation with allele frequency data (Coop et al., 2010; Fumagalli et al., 2011; Raj et al., 2013).

Finally, the explosive growth of genomic data has also opened new avenues in creating statistics to measure selection. For example, Field et al. (2016) used newly available whole genome data from the British population to infer very recent (beginning 2-3 kya) hard-sweep type selective events on derived alleles. The authors developed a new statistic, the singleton density score (SDS) which utilises the impact of recent selection on the rare variant spectrum. It exploits the lack of singletons and the distortion of the genealogy for haplotypes which are undergoing a selective sweep. To this aim, the distance between a test SNP and the nearest singletons is calculated to infer the ratio of terminal branch lengths (marked by singletons) for the derived vs ancestral alleles. Outliers with very short terminal branches are thought to be recent targets of selection.

A last aspect, which can only be briefly mentioned here, are the advances of aDNA studies. They provide an opportunity to study the temporal dynamics of selective processes (e.g. Mathieson et al., 2015 for Bronze Age Western Eurasia) and have yielded considerable improvements to estimates of selection coefficients.

1.4 The recent evolutionary history of humans: insights from modern genomics

1.4.1 African origins and early population splits

Unravelling the history of the human species requires integration of evidence from the fields of palaeoanthropology, archaeology, palaeoclimatology and evolutionary genetics. In the following the main insights gained from the recent progress of genetic studies into the development of our species in recent (primarily < 250 kya) evolutionary times will be reviewed.

For most of the 20th century there were two competing classes of models describing modern human origins. The first is known as the Out of Africa (OOA) model: according to it all living modern humans derive the vast majority of their ancestry from a relatively recent (Middle Pleistocene) African ancestral group that colonised the rest of the world and totally replaced

archaic hominins (Stringer and Andrews, 1988). In contrast multiregional models argued for an earlier dispersal and long-term continuity of Pleistocene hominins such as *Homo erectus* in different regions of the world (Weidenreich, 1947; Wolpoff et al., 1984). This would imply independent or partially independent, if gene flow was allowed, evolutionary lineages of these populations (Templeton, 2007).

The first studies of worldwide human genetic diversity which yielded phylogenetic trees based on mitochondrial (Cann et al., 1987; Ingman et al., 2000), Y-chromosomal (Thomson et al., 2000) and autosomal microsatellite (Bowcock et al., 1994) data identified Africa as the putative source of the human gene pool for these markers as it harbours the most diversity. Furthermore, it was hypothesised that all earlier hominins were completely replaced by the OOA wave and did not contribute to the modern human gene pool. The increase in the amount and complexity of genome-wide high coverage data has led to the development of more computationally intensive methods to estimate split times and other parameters of population histories (see section 1.3.3). In the following, the most important insights on human population history from evolutionary genetics will be summarised. The temporal subdivisions describing important transitions during the evolution of *Homo sapiens* are adapted from Mirazón Lahr (2016).

Much of the earliest history of our species and its origins remain unknown. Early fossil specimens exhibiting an unequivocally modern human morphology come from the sites of Omo Kibish and Herto, both in present-day Ethiopia, and are dated to 190-200 kya (McDougall et al., 2005) and 160-154 kya (Clark et al., 2003b) respectively. Very recently, these origins have been pushed back further by re-analysed as well as newly excavated specimens from the Moroccan site of Jebel Irhoud, situated in a layer dated to 350-280 kya (Richter et al., 2017). These fossil remains exhibit a mixture of archaic and derived traits with a facial morphology closer to modern *Homo sapiens* than any other currently known taxon (Hublin et al., 2017).

Genetic data cannot pinpoint the exact time at which a species arose because speciation is a gradual process, however it can be assumed that considerable substructure already existed in the hominin population which gave rise to *Homo sapiens* (e.g. Harding and McVean, 2004), which is consistent with archaeological evidence.

An African origin followed by a range expansion as suggested by uniparental marker analyses has been confirmed from world-wide samples of genome-wide SNP and WGS data (Sousa et al., 2014). Firstly, it has been shown that African genomes exhibit the highest variant numbers

(The 1000 Genomes Project Consortium, 2015). Furthermore, non-African diversity is mostly a subset of African variation (Campbell and Tishkoff, 2008) and genetic diversity between and within continents decreases along a presumed expansion axis which is compatible with multiple founder effects (Henn et al., 2012a; Prugnolle et al., 2005; Ramachandran et al., 2005). This is further supported by higher levels of LD (The 1000 Genomes Project Consortium, 2010) and longer runs of homozygosity observed in non-African populations (Frazer et al., 2007). It is debated in which part of Africa anatomically modern humans originated (reviewed by Lopez et al., 2016).

Some of the earliest known findings of *Homo sapiens* outside of Africa are from the sites of Skhul (Grün et al., 2005) and Qafzeh (Hovers, 2009) dating to 120 kya and 100-90 kya, respectively. The current interpretation is that these individuals are representatives of an early OOA movement which did not expand further and these populations either died out or returned to Africa. A recent finding consisting of anatomically modern human teeth from Southern China dated to at least 80 kya (Liu et al., 2015) corroborates earlier modern human dispersals. This is further confirmed by aDNA evidence from an at least 50 kya-old Neanderthal genome from the Altai mountains for which gene flow from an old modern human lineage was inferred (Kuhlwilm et al., 2016).

The deepest currently known split within extant humans is between the South African Khoisan and all other populations. Contingent on downward revisions of the genome-wide mutation rate (see section 1.2.1) this split could have occurred as long as 250-300 kya (Gronau et al., 2011; Scally and Durbin, 2012). Considering recent aDNA evidence from 2,000 year old South African Stone Age hunter-gatherer genomes this has even been pushed further back to 350-260 kya as recent admixture from an East African/Eurasian source into all modern Khoisan biases the divergence date estimated using the latter downwards (Schlebusch et al., 2017).

Figure removed for copyright reasons. Copyright holder for all three original underlying figures is Springer Nature.

Figure 1.6: Reconstruction of N_e based on WGS data using the PSMC method. A Li and Durbin (2011), B-C Schiffels and Durbin (2014), D-F Mallick et al. (2016)

The first inferences of N_e based on the PSMC methodology (Figure 1.6) indicated a shared increase for all modern humans populations until ca. 100 kya (Li and Durbin, 2011). This could either reflect a real increase in population size as suggested by the evidence for population expansion from the fossil and archaeological data or population structure involving separation and admixture. However, later studies (Mallick et al., 2016; Schiffels and Durbin, 2014) questioned these findings by showing a consistent decline of non-African N_e starting at 200 kya, long before any proposed OOA movement. These studies also found either a slight decline or constant population sizes for African groups during the period of interest.

1.4.2 Peopling the world: dispersal out of Africa

The next period of human evolutionary history can be described as a time of dispersals (70-50/45 kya) (Figure 1.7). From the fossil and archaeological record, we know that by its end modern humans coming from Africa had peopled most of the world excluding the Americas. They arrived in Europe (Fu et al., 2015; Nigst et al., 2014), East and Southeast Asia (Bae et al., 2017; Demeter et al., 2012) and Australasia (Clarkson et al., 2015) at least 55-45 kya. Generally, the climate throughout the world during this period was colder and drier than today, which would have made passages across straits easier for humans, however the sea barrier between Southeast Asia and Sahul (Australia and parts of Papua New Guinea) would still have posed a considerable obstacle (Davidson, 2010).

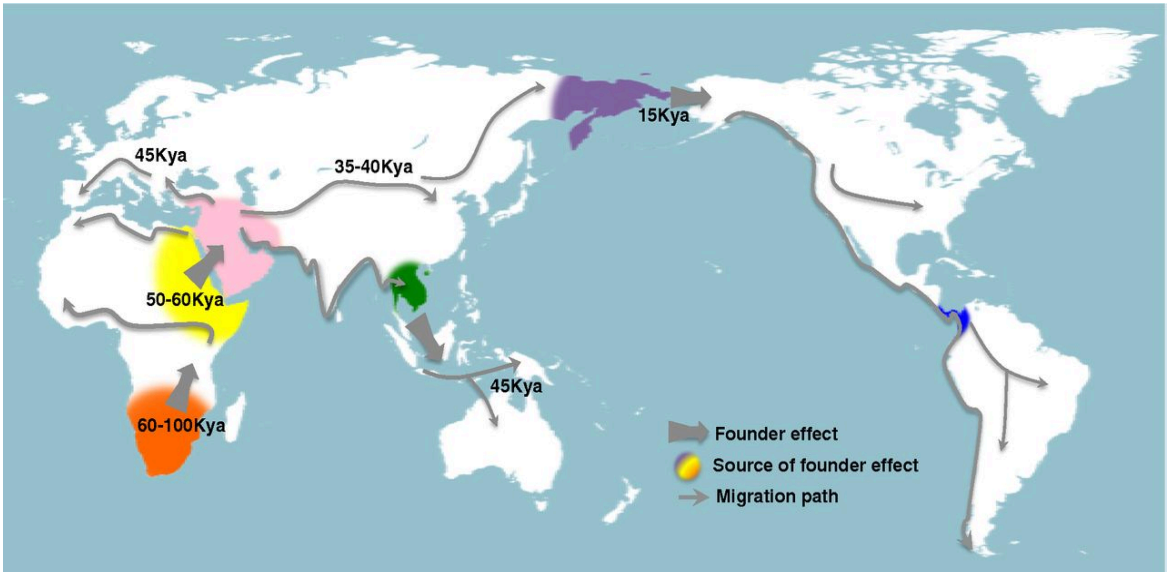


Figure 1.7: One possible scenario for the temporal and spatial framework of the OOA migrations. Wide arrows indicate major founder events while thin arrows symbolise migration paths. This depiction is a simplification of a more complex reality and some details displayed in this graphic are still very much debated, these include a) a southern African origin, b) a southern route exit from Africa, c) a relatively late arrival in Australasia. Figure taken from Henn et al. (2012a).

Recently, there has been much debate about the modalities of the OOA event. The first concerns the exact route taken when leaving Africa, the main two scenarios being a northern route (via Egypt and the Sinai) and a southern route (via Ethiopia and the Arabian Peninsula).

evidence consistent with a northern exodus comes from Pagani et al. (2015) who generated 225 mostly low-coverage whole genomes from modern-day Egyptians and Ethiopians. After the genomic components derived from recent Eurasian admixture in these populations were masked the Egyptian haplotypes showed a greater affinity to the non-African haplotypes than those from any other African population. This could be evidence that Egypt was the “last stop” on

the OOA migration, but this is dependent on underlying assumptions about population continuity and the geographical origins of the Egyptians' African component which are difficult to test without aDNA.

Another set of questions concerns the exact dating of the OOA event which gave rise to present-day non-Africans and the number of potential dispersals. One way to estimate these parameters is to develop spatially explicit simulations of human demography during the OOA expansion with different parameter combinations, generate summary statistics from these and then compare the latter to observed present-day genetic patterns (e.g. Eriksson et al., 2012).

Given certain assumptions about the geographical origin of particular haplogroups the mitochondrial evidence indicates an OOA event giving rise to modern humans at ca. 60-65 kya (Fernandes et al., 2012; Jobling et al., 2013). Time estimates based on WGS data are more variable depending on the methodology and the respective datasets used. Identity-by-state (IBS) (Harris and Nielsen, 2013) and coalescent-based (Gronau et al., 2011) approaches estimate the divergence times between Africans and Eurasians at 55 kya and 38–64 kya respectively. The PSMC and MSMC results (Li and Durbin, 2011; Schiffels and Durbin, 2014), as mentioned above, indicate a long continuous divergence process with considerable differences as early as 100 kya with continually ongoing gene flow until more recent times. The most comprehensive Markovian coalescent analyses suggest a divergence between Europeans and West Africans at ca. 63 kya and substantial differentiation between different non-Africans in terms of coalescence events and N_e starting ca. 50 kya (Mallick et al., 2016).

Another subject of debate is the number of OAA migration waves. The most parsimonious scenario would assume only one wave while others have proposed multiple migration events (reviewed by Groucutt et al., 2015). Most studies indicate that the genetic evidence for the relationships between different modern human groups are consistent with a single wave and a serial founder effect model (Deshpande et al., 2009). However, "weaker" versions of a multiple OOA model are still potentially viable. Firstly, recent simulations by Pickrell and Reich (2014) have shown that the characteristic decrease in heterozygosity with increasing distance from Africa can also result from past demographics diverging from this scenario with fewer bottlenecks and varying levels of admixture between different populations.

A second issue is how the Australo-Papuans (term hereafter used for Australians and Papuans) relate to Eurasians. Because of specific aspects of their morphology, mainly a higher robusticity

across multiple cranial features that they share with early anatomically modern humans, and the archaeological record Lahr and Foley (1994) proposed that they might derive their ancestry from an earlier OOA dispersal at 50-100 kya from a structured African population. While the genetic data have made this model in its strong form unlikely Australo-Papuans could still I) be derived from an earlier dispersal from a deme which gave rise to all non-Africans and/or II) exhibit a minor genomic component from an earlier OOA event.

A study of high coverage SNP data which included 25 individuals from Papua New Guinea applied an approximate Bayesian computation framework to address this question (Wollstein et al., 2010). Bayesian methods have already been mentioned several times in this chapter, broadly speaking, their goal in a population genetic context is to obtain the probability of demographic parameter values for a specified evolutionary model given the observed genomic data. Following Bayes' theorem, the probabilities of all possible parameter values, also known as the posterior distribution, can be computed as the product of the prior, here the assumed a priori plausibility of different demographic parameter values, and the likelihood function, here the probability of observed genomic data given all possible demographic parameter values. However, for complex demographic models the true likelihood function is very challenging to solve analytically. The key innovation of approximate Bayesian computation is that sampling from the posterior distribution can be performed without requiring explicit likelihood calculations (Beaumont et al., 2002). While the highest posterior probability was obtained by Wollstein et al. for a single-OOA scenario in which the New Guineans split from a common Eurasian ancestor there was also some support for a model in which the Papuans branch of from an ancestral East Asian clade. Rasmussen et al. (2011) sequenced the genome of an aboriginal Australian living ca. 100 years ago. They found support for an early branching of Australian aborigines from other Eurasians which they dated to 75-62 kya. As mentioned above (section 1.3.2) Malaspinas et al. (2016) provided an in-depth study of Australian and Papuan genomic diversity. Their analyses indicated that Australo-Papuans and Eurasians descend from the same OOA event. By comparing SFS spectra simulated under different demographic models to the observed patterns and MSMC analyses they found that the ancestors of Aboriginal Australians and Papuans diverged from other Eurasians shortly after the OOA event at 58 kya (51–72 kya).

However, Mallick et al. (2016) as part of their worldwide WGS study analysed $3*10^6$ SNPs derived from WGS data with the f-statistic-based method ADMIXTUREGRAPH (Patterson et

al., 2012). While correcting for known archaic gene flow into Australo-Papuans the best-fitting model placed them in a clade together with mainland East Asians.

The debate about the phylogenetic relationship of Australo-Papuans and other Eurasians is still ongoing and different studies have yielded apparently incompatible results. It has been noted that the long-term physical isolation and low N_e especially of some Australian and Papuan groups make the correct estimation of demographic parameters challenging (Lopez et al., 2016). Furthermore, approaches such as MSMC are hampered by phasing issues with Australo-Papuan samples, as these groups were not part of the 1000 Genomes Project. More research is needed to address these concerns and reconcile the apparent contradictions.

Regardless of this ongoing debate, studies of worldwide modern human diversity detect a split between West and East Eurasians which given the recent revisions of the nuclear mutation rate has been estimated at ca. 40–45 kya (Mallick et al., 2016). aDNA evidence has allowed us to place its lower boundary at a minimum of 36.2 kya based on the greater affinities of a Palaeolithic individual from Kostenki (Russia) to present-day Europeans than to East Asians (Seguin-Orlando et al., 2014). Similar observations for genomic fragments from a modern human from Tianyuan Cave (China), who clusters with an East Eurasian clade, support this and argue for a date of at least 40 kya (Fu et al., 2013). Furthermore the genome of a 7,000 old farmer from Stuttgart (Germany) revealed that this individual derived part of its ancestry from a "Basal Eurasian" group which branched off from all other non-Africans before the West-East-Eurasian split (Lazaridis et al., 2014).

These early divergences were followed by complex localised histories of population splits and gene flow events. For example modern Europeans are thought to trace their ancestry to at least three distinct groups: Western European hunter-gatherers, ancient north Eurasians (similar to ancient Siberian populations from the Upper Palaeolithic) and early European farmers with Middle Eastern affinities (Lazaridis et al., 2014). This model was further refined by Haak et al. (2015), who highlighted that these contributions were not necessarily direct. This complex ancestry of Europeans highlights that aDNA evidence often does not fit the most parsimonious demographic scenario constructed based on modern day genetic data (Haber et al., 2016).

1.4.3 Gene flow from archaic hominins

The unprecedented deluge of aDNA data has also affected our understanding of the relationship between archaic hominin species and modern humans. Genomic data have been generated for several Neanderthals (Green et al., 2010; Hajdinjak et al., 2018; Prüfer et al., 2013) and three individuals forming a distinct hominin lineage found at Denisova cave in Siberia (Meyer et al., 2012; Reich et al., 2010; Sawyer et al., 2015). These Denisovans are currently only represented by a phalanx and several teeth. Denisovans and Neanderthals are thought to form a clade which diverged from the ancestors of modern humans ca. 550–765 kya whereas the split between the two archaic hominins is estimated to have occurred ca. 381–373 kya (Prüfer et al., 2013). Based on D and f_4 statistics it was estimated that the genome of all modern non-Africans contains 1.5–2.1% Neanderthal ancestry (Green et al., 2010; Prüfer et al., 2013) that was proposed to derive from a single admixture event affecting the subpopulation all non-Africans trace their ancestry to.

This interpretation was challenged on several grounds. Firstly, it was suggested based on simulations that the observed affinity between Neanderthals and non-Africans can also be generated under scenarios involving only population substructure in Africa (Eriksson and Manica, 2012) which was disputed by other authors (Yang et al., 2012). For the second objection, it is important to recall that African populations are more diverse than non-Africans. A correlation between heterogeneity and mutation rate has been shown for microsatellites by Amos et al. (2008), but remains unproven for SNPs. If the latter could be demonstrated the lower Neanderthal-sharing with Africans could possibly be due to novel mutations in the latter and not because of shared polymorphisms between non-Africans and the archaic hominins (Amos, 2013, 2016). Despite these controversies the majority view in the field is that the wide-spread Neanderthal signature reflects genuine admixture. However, it is unclear how frequently it occurred and how many individuals were involved. Additional aDNA evidence from prehistoric non-Africans generally indicates longer shared tracts of ancestry with Neanderthals (Fu et al., 2014a, 2015; Seguin-Orlando et al., 2014) consistent with much more recent Neanderthal ancestry in these individuals and hinting towards multiple admixture events.

Gene flow from Denisovans into modern Australo-Papuans is thought to account for 4-6% of their genome, while East Asians and Native American populations appear to exhibit 0.2% of this ancestry component (Prüfer et al., 2013; Reich et al., 2011). The modalities of this putative

admixture are currently under debate and depend on assumptions about the geographical distribution of Denisovan-like hominins in the past.

Sankararaman et al. (2016) used a machine-learning approach to detect archaic ancestry based on WGS data and according to their own estimates they localised ca. 75% of all Neanderthal and 25% of all Denisovan ancestry. Consistent with other studies they showed that these regions are located less frequently than expected near genes implying that archaic ancestry on a modern human background has generally been detrimental to reproductive fitness.

However, there are a few regions in the genome which have been identified as introgressed, based on local phylogenetic trees, long range LD and/or probabilistic modelling that have also been inferred as targets of selection after entering the modern human lineage (reviewed by Racimo et al., 2015). Perhaps the most prominent example for adaptive introgression is *EPAS1,* a gene encoding a transcription factor which is important in response to altitude-induced hypoxia. A particular set of intronic mutations define a haplotype almost exclusively found in Tibetans thought to be advantageous at high altitudes and to have originated in a Denisovan-like population (Huerta-Sánchez et al., 2014).

1.4.4 The spread of agriculture and the recent impact of natural selection

The most recent phase in human evolutionary history largely corresponds to the Holocene, the current interglacial phase starting ca. 12 kya. Even before the Holocene the initial warming after the last glacial maximum led to an expansion of habitable niches in most parts of the world accompanied by an expansion of hunter-gatherers attested by the archaeological record (e.g. Housley et al., 1997; Sutton, 1977). This was followed by one of the most fundamental socioeconomic transformations in the history of our species: the adoption of farming and animal husbandry (Pinhasi and Stock, 2011). Its earliest centre was the Fertile Crescent with evidence for cattle herding and crop cultivation from 12 kya onwards (Zeder, 2008); during the first half of the Holocene domestication of animals and/or plants also originated in several other regions of the world (Bellwood, 2005; Diamond, 2002; Smith, 1998). These processes led to an increase in population sizes of the agriculturalist groups, which is also confirmed by the PSMC graphs of N_e for their modern descendants, especially for non-Africans (Figure 1.6), and to range expansions. It has long been debated to what extent these farmers replaced or assimilated the hunter-gatherers (Bellwood and Oxenham, 2008) and the available evidence from aDNA shows

that our understanding of the relationship of these ancient groups to contemporary genetic diversity is still evolving.

The adoption of agriculture also subjected our ancestors to a variety of new selective pressures. Together with the novel environments after the OOA and infectious disease in general these are thought to be the most important selective influences in our recent evolutionary past. Even for well-supported cases of selection (reviewed by Scheinfeldt and Tishkoff, 2013) the last few years have still yielded novel insights into their temporal and spatial dynamics.

For example, large scale aDNA studies have demonstrated that rs4988235, the most common SNP upregulating the expression of the *LCT* gene and therefore strongly associated with lactase persistence in modern Europeans, was still only at 5% during the Bronze Age implying a very high selection coefficient in recent evolutionary times (Allentoft et al., 2015). Based on re-analyses of worldwide modern human data a strong selection coefficient has also been estimated for the recessive (Hamblin et al., 2002) FY*O allele of *ACKR1,* which is protective against infection of erythrocytes by *Plasmodium vivax*, a major cause of malaria (McManus et al., 2017). Another phenomenon we have obtained first insights into is the adaptive role of structural variation. Inchley et al. (2016) demonstrated that the increased number of copies of an amylase gene (*AMY1*) in all modern humans relative to the great apes (Perry et al., 2007) is likely due to a shared selective sweep in the modern human lineage and potentially reflects the adaptation of a more starch-rich diet long predating agriculture.

Many conclusions about the broader outlines of human history inferred from the originally sparse datasets of genetic markers in conjunction with non-genetic evidence have been confirmed by genomics. However, the rapid development during the last 5-10 years, which have been hailed as the "age of genomic discovery" (Slatkin and Racimo, 2016), has highlighted the remarkable complexity underlying the history of modern human populations. The emerging consensus might be called a leaky replacement model (Pääbo, 2015) followed by complex regionalised demographic transitions. There are limitations inherent in genetic data besides the stochasticity already mentioned. On their own they only have a limited ability to inform us about human behaviour and cognition. Nevertheless, the general picture painted in the last four sections demonstrates that human genomic data in conjunction with other lines of evidence constitute a very powerful tool for inferences about our past.

1.5 Population history of Southeast Asia

Mainland (MSEA) and Island Southeast Asia (ISEA) are home to hundreds of different ethno-linguistic groups each displaying a complex demographic history (Lewis et al., 2015). Previous studies have revealed strong genetic correlations between populations which are geographically and linguistically close and suggested a common origin of all Southeast Asian and East Asian populations from a single migration wave (HUGO Pan-Asian SNP Consortium et al., 2009). It is well known, however, that in the more recent past the populations living in this region have undergone major demographic changes. One particular process, known as the Austronesian expansion, which occurred during the last five thousand years is thought to be associated with the spread of the Austronesian languages (Bellwood, 2007) and the Neolithic cultural complex. The latter consists of a core package defined by pottery and certain shell artefacts, though it has been stressed that this assemblage was most likely polythetic, i.e. while there is a typical combination of material culture traits defining the Neolithic package in the region, no single artefact type alone is sufficient to define group membership and not all types occur at all sites (Spriggs, 2011).

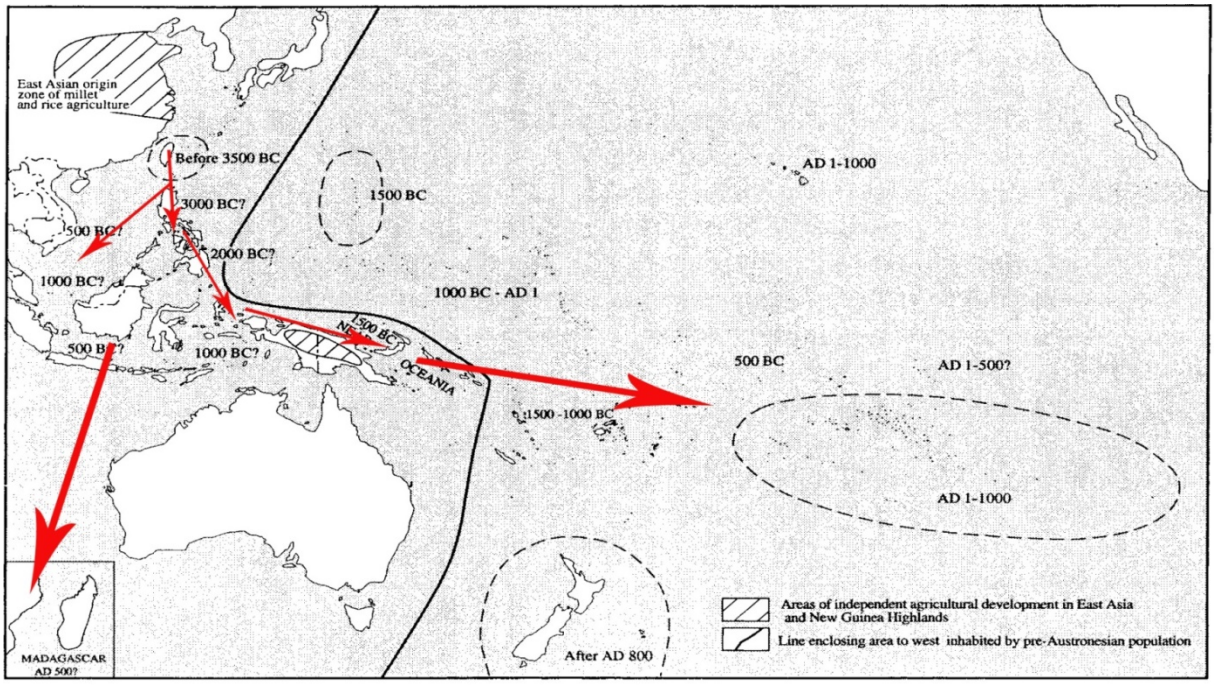


Figure 1.8: This map displays the approximate dates for the Austronesian colonisation events throughout ISEA and the more remote islands of the Pacific and the Indian Ocean. Putative dispersal routes are marked in red (adapted from Bellwood, 2006).

While the modality of this expansion is complex and still debated (Bulbeck, 2008), the current majority view holds that it began in Taiwan and spread into the Philippines from ca. 5 kya onwards. From there it proceeded west into present-day Indonesia and east to Near Oceania before extending to Far Oceania (Figure 1.8) (Bellwood, 2006). The latter movement has been associated archaeologically with the Lapita Cultural Complex and its distinct ceramics tradition (Gray et al., 2009; Kirch, 2000; Pawley and Ross, 1993). It first appeared in the Bismarck Archipelago ca. 3.4 kya and then rapidly expanded into the previously uninhabited islands of the Pacific, reaching Tonga and Samoa by ca. 2.9 kya and spreading as far as Madagascar and the Easter Islands between 1-2 kya.

The Austronesian expansion encompasses a period of dramatic increase in the movement of artefacts, peoples, ideas and languages across islands over vast geographical distances (Spriggs, 2011). Even before genetic evidence was available based on the integration of artefacts and practices from already resident groups in ISEA and Near Oceania it seemed plausible that the Austronesians admixed with locals.

One method to study these dynamics is the commensal approach where phylogeographic patterns observed for culturally and economically important animals and plants are used as proxies for reconstructing the pathways of the colonising canoes throughout the region (Matisoo-Smith, 2015). For example, domestic pigs in ISEA and Polynesia are thought to originally derive from MSEA stock (Larson et al., 2007) while many domestic plants in the region originated in New Guinea (Denham, 2011). Finally, recent research by Chang et al. (2015) on chloroplast DNA of the paper mulberry plant (an important resource of bark cloth for clothing and other purposes) indicates a Taiwanese origin of the lineages dominant in Polynesia. Taken together the archaeological evidence supports the notion of a complex history for the various components of Austronesian and Lapita cultures.

Wollstein et al. (2010) analysed the genetic make-up of the Malayo-Polynesians (speakers of Austronesian languages outside of Taiwan) from Near and Remote Oceania and they reported them as containing genetic contributions from people currently inhabiting Borneo (used as a proxy for Asian influence) and Papua New Guinea. These admixture events were dated to approximately 3 kya, consistent with similar population movements involving people of Asian ancestry eastwards through ISEA dated around 4-3 kya (Xu et al., 2012). More recent studies (Lipson et al., 2014; Pierron et al., 2014) have distinguished at least three major ancestral

components in MSEA and ISEA in association with Papuan-, Austro-Asiatic- and Austronesian-speaking populations.

aDNA analyses of samples from Remote Oceania suggest that, not dissimilar to what has become apparent for Europe (see section 1.4.2), the real settlement history of the region is more complex than the simplest possible scenario consistent with modern DNA and included probably at least three dispersals. Skoglund et al. (2016) presented aDNA from three individuals associated with the Lapita culture from Vanuatu, dated to ca. 3.1-2.7 kya, and one from Tonga, dated to ca. 2.7-2.3 kya. Comparative analyses indicated that these First Remote Oceanians were exclusively of East Asian-Austronesian ancestry and did not have any Papuan affinities. Two subsequent studies (Lipson et al., 2018; Posth et al., 2018) extended the aDNA record from Vanuatu and other Pacific islands considerably and revealed that only a few hundred years later the original Lapita people in Vanuatu had apparently been replaced by individuals of Papuan ancestry. This replacement was probably specific to Vanuatu, as the Papuan-like component in Tongans probably originates from a different source.

Analyses aiming to identify the precise source regions of these dispersals are still confounded by recent admixture in most modern ISEA populations with groups originating from other regions including MSEA (HUGO Pan-Asian SNP Consortium et al., 2009; Trejaut et al., 2014) (more details on the included candidate populations can be found in Appendix B.1).

In addition to the migratory events involving South East Asian sources, more recent South Asian influences in forms of cultural and trading networks, starting more than 2 kya, in ISEA and MSEA have been well established from historical and archaeological data (Ardika et al., 1997; Ardika and Bellwood, 1991; Lawler, 2014; Manguin et al., 2011). Exemplary for these developments are the sites of Khao Sam Kaeo and Phu Khao Thong from Peninsular Thailand yielding archaeological evidence dating to 2.3-1.2 kya. They confirm the earliest trade networks with India, which include rouletted ware, semi-precious stone beads and artefacts, and Indian crops (Castillo, 2013). In ISEA, one finds evidence of Indian trade either directly or via peninsular Thailand. Coastal sites located in Northern Bali dating to 2.1 kya yielded pottery of East Indian or Sri Lankan production, gold and carnelian objects from North India and mung bean (Calo et al., 2015). Furthermore epigraphy, i.e. evidence from ancient inscriptions indicates a strong Indian impact on the nascent political structures of the region (Mabbett, 1977) and provides records of Brahmanic rituals and animal sacrifices (Guy, 2011).

Linguistic evidence also supports early interethnic contact between Indian and Southeast Asian populations. Apart from the ubiquitous influence of Sanskrit (Gonda, 1973) where it is difficult to distinguish ancient from more recent borrowings, analyses of the earliest Maritime Southeast Asian literature demonstrate that it already exhibits signs of Tamil influence from South India, much of which most likely spread across the region through pre-existing local networks (Hoogervorst, 2015).

Traces of paternal (Y chromosomes) and maternal (mtDNA) Indian ancestry have been detected across several Indonesian islands at low frequency (<5%) (Chaubey and Endicott, 2013; Karafet et al., 2005, 2010; Kusuma et al., 2015). The influx of Indian ancestry is detectable in some genome-wide analyses of low density autosomal SNP data (HUGO Pan-Asian SNP Consortium et al., 2009) while being restricted to just a few populations from western Indonesia (Sumatra). Contrary to that, a more recent study (Pugach et al., 2013) using medium density SNP data could not find a South Asian genetic signature in South East Asia. The same authors, however, inferred gene flow from the Indian sub-continent to Aboriginal Australian populations and dated it at around 4 kya. In the absence of a similar South Asian component in SEA this finding was interpreted to require a direct sea route bypassing Southeast Asia.

1.6. Functional and deleterious variation

1.6.1 Missense variants: worldwide patterns and functional implications

Recent high-coverage exome and WGS projects detected 12.2–13.6k non-synonymous variants in each African genome compared to 10.2–12.4k in non-African groups (Table 1.1). This difference is thought to be another feature attributable to the generally higher genomic diversity in African populations. The vast majority of non-synonymous mutations (>99%) belong to the missense category (see section 1.1.1). Therefore, each human carries more than 10,000 missense mutations. To understand their biological importance, it is crucial to estimate the effects they have on phenotypic fitness. On average, missense mutations are thought to be slightly deleterious. This notion is supported by interspecies comparisons. The d_N/d_S ratio of sites which are different between humans and chimpanzees (0.24) is considerably lower than the one observed for intraspecies polymorphisms (0.38) (Bustamante et al., 2005).

Table 1.1: Per individual counts of non-synonymous sites for different world regions based on two recent large-scale sequencing projects. Data from the 1000 Genomes Project Consortium, (2015) and Lek et al. (2016). Abbreviations: ns…non-synonymous.

Source	Statistic	Africa	East Asia	Europe	Americas (admixed groups)	South Asia
1000 Genomes Project	Median autosomal ns variants	~12,200	~10,200	~10,200	~10,400	~10,300
ExAC project	Mean number of ns variants	13586.4	12426.3	11927.8	12211.1	12203.1

This indicates that most novel missense variants do not reach high enough frequencies to become fixed interspecies differences, likely due to purifying selection acting on them. Furthermore, it has been demonstrated that they are overrepresented in the rare frequency classes compared to synonymous alleles in empirical data.

After the first human and chimpanzee genomes had been sequenced comparative studies aimed to identify missense mutations uniquely shared by all modern humans. The first large scale systematic studies of this kind used the d_N/d_S and related statistics to identify genomic regions with an excess of species-specific missense mutations. Classes of genes which were overrepresented included those related to the immune response, chemosensory perception and gametogenesis (Bustamante et al., 2005; Nielsen et al., 2005). Nielsen et al. (2005) also categorised them according to the tissue where their expression levels are highest. Genes expressed in testes appeared to be the most differentiated with regards to non-synonymous mutations. In contrast those primarily transcribed in the brain appeared to be among the most conserved gene classes.

Subsequent analyses have corroborated this picture: Prüfer et al (2013) found that for regions of the human genome mappable to those of the great apes and archaic hominins there were only 96 fixed *Homo sapiens*-specific missense variants. Therefore, it seems plausible that non-coding changes playing a regulatory role, e.g. influencing gene expression are also an important source of phenotypic differences between humans and great apes (King and Wilson, 1975).

It has been suggested that some of the noncoding regions which show accelerated sequence evolution on the human lineage have an important role as enhancers (Franchini and Pollard, 2015). For example Boyd et al. (2015) demonstrated that such a region overlaps with the enhancer HARE5. This enhancer is physically associated with the promoters of genes important

for brain development and transgenic mice with the human variant of HARE5 exhibited increased neocortical size compared to those with the chimpanzee variant and wild types.

This would indicate that when analysing the results from selection tests in an inter- as well as intraspecies context the overlap of the output with databases describing regulatory variation should be considered. An example for the latter is the Genotype-Tissue Expression (GTEx) (Lonsdale et al., 2013) project which contains information on expression quantitative trait loci (eQTL), i.e. sites whose polymorphisms influence the expression of nearby genes.

This also has implications for the diversity at missense sites between different populations of *Homo sapiens.* When extreme frequency differences between groups sampled for the 1000 Genomes Project were considered, relatively few population-specific fixed sites were found (The 1000 Genomes Project Consortium, 2012). For continental groups, in this case Africa, Europe and East Asia, loci with a DAF difference of ≥ 0.9 included missense mutations in genes which had previously been detected as targets of geographically localised selection. These included SNPs in such well-known genes as *DARC/ARCKR1* (malaria infection resistance in Sub-Saharan Africans, Oliveira et al., 2012, see section 1.4.4), *SLC24A5* (light skin pigmentation in Europeans, Voight et al., 2006) and *EDAR* (influences multiple phenotypes among them scalp hair thickness and the number of eccrine sweat glands, the exact trait under selection is still contested, Kamberov et al., 2013; Sabeti et al., 2007). Within different populations from the same continental group there was a small fraction of loci exhibiting DAF differences of ≥ 0.25. The group which was most distinct from other Europeans were the Finns.

This highlights another important factor explaining differentiation in missense variants in modern human populations. In groups such as Ashkenazi Jews or the aforementioned Finns population-specific missense variants are among the causes for the elevated incidence of autosomal recessive disorders, e.g. Tay-Sachs disease (Kaback and Desnick, 2011). In both groups genetic drift, which has had more impact due to documented bottlenecks (Carmi et al., 2014; Liu and Fu, 2015; Palo et al., 2009) and the subsequent small N_e, is likely the underlying cause. A detailed discussion on the differences in the overall number of missense mutations and those classified as deleterious is presented in section 1.6.4.

Irrespective of the specific histories of modern human populations which can modulate them, the debate about the distribution of fitness effects (DFE) of *de novo* missense mutations is still ongoing. For example Zuk et al. (2014), who aimed at establishing an analytical framework for

rare variant association studies, assumed the respective impacts of missense mutations (based on the literature) to be as follows: 25% were thought to be strongly deleterious and 25% to be truly neutral, while the other half were classified as weakly deleterious.

Further modelling work also suggests that this distribution is dynamic. Changes in N_e and the strength of environmental fluctuations were shown to be important factors (Razeto-Barry et al., 2012). Empirical work investigating the DFE of missense (and other) mutations in well-studied genes in model organisms also indicates a strong dependence on the respective gene and stresses the role of complex phenomena such as epistasis which depend on the interaction of multiple mutations (Firnberg et al., 2014; Sarkisyan et al., 2016).

It is still technically challenging to determine the fitness effects for all possible human missense mutations in an experimental setting. Therefore, bioinformatic classification approaches to estimate them (reviewed in section 1.6.3) are required.

1.6.2 Nonsense variants: worldwide patterns and functional implications

Nonsense mutations, while considerably rarer than missense mutations, are nevertheless of great evolutionary and medical importance as the most prominent cause for the loss of gene function. In a strict sense, these are mutations that result in the introduction of a premature stop codon into the genetic code. Recent genome-wide studies have adopted a broader definition of protein-truncating variants (PTV) or loss-of-function (LoF) variants (both terms are treated as interchangeable in the literature, LoF shall be used in this thesis).

Besides nonsense variants, these also include short indels leading to a shift of the genetic reading frame and splice-site mutations where an exon with a count of nucleotides not divisible by three is skipped (Figure 1.9). While some studies (Narasimhan et al., 2016) focussed on these single position or only on small-scale changes, others also included larger deletions (>50 bp) removing either the first exon or more than 50% of the protein-coding portion of the transcript (MacArthur et al., 2012). The truncated polypeptides resulting from these changes are often either non-functional or have phenotypically harmful properties. An important mechanism to prevent their accumulation is nonsense mediated mRNA decay (NMD) (Maquat, 2004). Its typical targets contain termination codons positioned >50-55 nucleotides upstream of the last exon-exon boundary which has led to the formulation a somewhat heuristic 50/55 bp rule (Nagy and Maquat, 1998).

Figure removed for copyright reasons. Copyright holder is the American Association for the Advancement of Science.

Figure 1.9: Overview of the genomic changes which can cause a loss of gene function. SNPs (A) or small frameshift indels (B) resulting from de-novo mutations can result in a codon which leads to the termination of translation. Large deletions (C) and SNPs affecting splicing patterns (D) can have similarly severe effects (adapted from Rivas et al., 2015).

The master regulator of the molecular machinery underlying this process is thought to be UPF1, a protein with ATPase and helicase activities. However, UPF1 binding to the mRNA alone is not sufficient to initiate the decay of the molecule. Other specific factors, including UPF2 and UPF3 are required for the formation of an NMD-activating complex (reviewed by Lykke-Andersen and Jensen, 2015 and Hug et al., 2016).

Furthermore, there is increasing evidence that NMD also operates on mRNAs encoding full-length proteins. It has been shown to play a role in the regulation of genes related to cellular stress (Karam et al., 2015) and the generation of mature T cells as part of the immune response (Weischenfeldt et al., 2008). These pathways require a tight regulation of NMD activity by buffering mechanisms dependent on the genetic and environmental context, which are only partly understood.

This is underlined by the observation that not all transcripts containing LoFs that fulfil the 50/55-bp rule are measurably affected by NMD. This has been demonstrated by two studies analysing tissue-specific RNA sequence data collected as part of the Geuvadis RNA-seq and GTEx projects together with high quality exomes for 421 and 173 individuals respectively. Rivas et al. (2015) found that 69.5% of rare nonsense variants for which NMD is expected exhibit allele-specific expression, i.e. lower mRNA levels compared to the haplotype

surrounding the respective locus without the nonsense mutation. Following up on this finding the GTEx Consortium (2015) showed that only 38.4% of all high quality LoF variants cause lower expression across all transcripts which include them. It was also highlighted that certain LoF sites only affect gene expression in a specific tissue.

There are two more variant consequences that can lead to a loss of gene function and that shall be mentioned briefly here. Note that these are not included in most published surveys of LoF variants and will also not be further analysed in this thesis. The first are stop-lost variants where a mutation occurs in a triplet for a stop codon that is the ancestral type resulting in an elongated transcript. The second includes non-coding or synonymous variants that may affect mRNA stability.

Many *de novo* LoF mutations are associated with disease in humans, e.g. cystic fibrosis, an autosomal recessive Mendelian disease characterised by multiple organ system dysfunction and severely reduced life expectancy.

There is also some evidence that the total amount of rare LoF variants per genome might be a contributing risk factor for complex disease. Bellenguez et al. (2017) analysed whole exomes from 3,052 individuals of French ancestry. The cohort contained cases of late onset Alzheimer's disease (LOAD), early-onset Alzheimer's disease (EOAD) and controls. They found that the combined loads of rare (<1% MAF) LoF and strictly deleterious (all applied prediction algorithms agreed on the deleterious status) variants in the *TREM2, SORL1* and *ABCA7* genes were significantly associated with EOAD. Each gene explains ca. 1.1-1.5% of the heritability for this form of Alzheimer's.

The most extreme instances of these mutations lead to embryonal lethality, which has been studied as a recessive phenotype in consanguineous families (Shamseldin et al., 2015). LoF mutations leading to total gene inactivation, which is the case for at least a part of all LoF variants in humans, should cause lethal phenotypes for a considerable fraction of human genes. This is supported by evidence from animal models: 30% of all mouse genes lead to a lethal phenotype if both copies are knocked out (Ayadi et al., 2012).

However, there are also cases in which these mutations are neutral or evolutionarily advantageous. Olson (1999) proposed the "*less-is-more"* hypothesis according to which gene loss can act as a plausible mechanism of adaptive change. For example, a truncated version of the *CASP12* gene has been associated with higher resistance against severe sepsis (Xue et al.,

2006; Yeretssian et al., 2009). This variant is almost fixed in non-Africans, which is thought to be the result of natural selection. Examples of much rarer beneficial LoF variation are two nonsense mutations in *PCSK9* associated with lower LDL cholesterol plasma levels. They exhibit a combined heterozygote frequency of up to 6.9% in West Africans and are almost totally absent from all other worldwide populations (Cohen et al., 2005). Subsequently, drugs consisting of antibodies against the PCSK9 protein that mimic the physiological effect of these LoF variants have been developed. They are effective in reducing LDL cholesterol levels and cardiovascular mortality (Navarese et al., 2015). It is unclear if the inactivation of *PCSK9* was evolutionarily beneficial because of some other effect, e.g. it might interfere with the lifecycle of the malaria parasite (Horton et al., 2007), or if these frequency patterns result from genetic drift.

Before large-scale WGS datasets became available the most extensive survey of nonsense mutations comprised 805 SNPs genotyped in 1,151 individuals representing 56 worldwide populations (Yngvadottir et al., 2009). The main findings were that nonsense SNPs are on average under slightly purifying selection, while some cases of potentially advantageous variants were highlighted. Yngvadottir et al. furthermore observed 99 genes of which both copies were inactivated in at least one presumably phenotypically healthy individual, i.e. the LoF mutation was homozygous.

MacArthur et al. (2012) conducted perhaps the most influential study in the field. They analysed 185 low coverage genomes from four well-studied reference groups (CHB, CEU, JPT, YRI) generated in phase 1 of the 1000 Genomes Project. Using a broad definition of LoF (see above), which has since been widely adopted, the authors established a bioinformatic filtering pipeline and provided verification for every candidate site through additional genotyping assays. This resulted in a catalogue of 1,285 high-confidence LoF variants. The authors estimated that each human genome contains about ~100-120 genuine LoF-type variants, of which ~20 occur in a homozygous state (for a comparison to other studies see Table 1.2). In terms of the underlying frequency distribution MacArthur et al.'s (2012) results suggest that the majority of LoF variants found in an individual genome are common variants in nonessential genes. However, there are likely many LoF alleles with large fitness effects segregating at low frequencies in the human population, for which their study was underpowered due to the relatively low sample sizes.

Recently, large samples from individual populations have considerably extended the catalogue of genes that appear to be LoF-tolerant. Lim et al. (2014) extracted LoF information from ca.

Table 1.2: Statistics on the average number of LoF variants overall and in homozygous state per genome compiled from a range of genome-wide studies. Unless explicitly stated otherwise, the term knockout gene is defined as a gene containing a LoF variant for which at least one phenotypically healthy individual from the respective dataset is homozygous.

Source	Count of LoF sites per individual	Count of homozygous LoF per individual	Total number of genes inactivated without severe effects
1151 individuals genotyped for 805 SNPs (Yngvadottir et al., 2009)	32 (only nonsense in strict sense)	14	99
185 low coverage genomes from three major worldwide groups (MacArthur et al., 2012)	CHB + JPT: 103.5 CEU: 103.9 YRI: 121.5	CHB + JPT: 24.3 CEU: 22.5 YRI: 21.7	253
2,636 whole genomes of Icelanders and imputation of sequence variants from 101,584 chip-genotyped individuals (Sulem et al., 2015)	151.1	21.1	1,171
3,222 exomes of Pakistanis living in Britain (Narasimhan et al., 2016)	140.3	40.9	781
2,504 individuals of diverse ancestries, mixture of exome and low coverage WGS data (The 1000 Genomes Project Consortium, 2015)	Africa: 182 Americas (admixed): 152 East Asia: 153 Europe: 149 South Asia: 151 (median values)	NA	NA
300 individuals of diverse ancestries, high (43× on average) coverage WGS data (Mallick et al., 2016) (only stop-gain and frameshift SNPs)	136.9	NA	NA
60,706 exomes from different groups worldwide, down sampled to 3000 individuals per population (Lek et al., 2016)	Africa/African American: 140.7 Latino: 118.8 East Asian: 124 Finnish: 113.9 European (non-Finnish): 116.5 South Asian: 121.75	Africa/African American: 38.5 Latino: 43.2 East Asian: 42.9 Finnish: 39.7 European (non-Finnish): 39.8 South Asian: 40.15	10,374 (LoF-tolerant based on a maximum-likelihood classifier; the observed amount of truncating variation equals neutral expectations)

3,000 Finnish exomes and compared these to the same number of non-Finnish European controls. They demonstrated that Finns have a relative excess of rare and low frequency

deleterious variants as well as more complete gene knockouts, which was interpreted as resulting from a population-specific bottleneck. With a similar design Sulem et al. (2015) analysed WGS data from 2,636 Icelanders while imputing more information into >100,000 genotyped Icelandic individuals. Their most important finding was an expansion of the known rare variant spectrum, which showed that there are 1,171 genes which can contain homozygous LoF variants/"knockouts" in different individuals and still result in a generally healthy adult phenotype. Finally, Narasimhan et al. (2016) focussed on the exomes of 3222 British Pakistani individuals with high parental relatedness. Based on the observed deficit of homozygous knockouts they estimated an average load of 1.6 heterozygous recessive-lethal-equivalent LoF variants per adult. Where full health records were available, no correlation between the presence of a homozygous rare LoF variant and health status was found. The authors attributed this to a combination of reduced penetrance/mild phenotypes and technical issues such as alternative splicing or faulty genotype-phenotype associations.

On a global scale, the presence of LoF-causing variation was assessed in phase 3 of the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2015). The excess of LoF sites in African genomes compared to non-African genomes was attributed to the generally higher African diversity, while non-Africans appeared to show very little differentiation. The ExAC project comprising > 60,000 exomes (see section 1.3.2) is the most extensive survey of protein-coding variation in the human genome currently available. The authors classified genes according to the observed number of rare (MAF < 0.1%) LoF variants compared to a number expected from a sequence-context based mutational model. A total of 3,230 genes were classified as LoF-intolerant, of which 72% lacked any association with disease phenotypes, whereas 10,374 were identified as LoF-tolerant. Using a reduced subset consisting of 3,000 individuals per subgroup the conclusions from earlier studies on differences in LoF burden across populations were replicated, which were however less pronounced for LoF-intolerant genes.

Two aspects are worth highlighting from the review of the literature on this subject. Firstly, some regions of the world, such as Southeast Asia, Siberia and Oceania remain underrepresented in studies assessing LoF variation. Secondly, LoF variants alone are not sufficient to study potential interpopulation differences regarding the total load of potentially damaging variation and its phenotypic effects. Approaches assessing other exonic and ideally

also non-coding variation are needed to provide a more comprehensive picture of deleterious variation.

1.6.3 The genetic load and methods predicting variant deleteriousness

The theoretical concept of "genetic load" was defined by early genetic studies as the reduction of the evolutionary fitness of a population compared to a hypothetical comparative group with only the fittest genotypes (Haldane, 1937; Crow, 1958). One of the first empirical approaches to capture a fraction of this load by Morton et al. (1956) compared the mortality of the offspring of consanguineous marriages to those of non-related parents. From this the authors inferred the average number of mutations that would be lethal or lead to complete sterility if they occurred in a homozygous state as 3-5 per zygote, even though this particular number has been recently revised downwards to as low as 0.29 (Gao et al., 2015).

Potentially deleterious variation is continually introduced into the gene pool due to *de novo* mutations. Their persistence depends on the intensity of drift and purifying selection. Both are related to N_e, generally speaking with higher N_e drift becomes less important to the spectrum of deleterious variants and the impact of selective forces increases as empirically shown, e.g. for different *Drosophila* species by Petit and Barbadilla (2009). Kimura et al. (1963) suggested that in the long term mildly deleterious alleles might have the strongest impact on the genetic load of a population. Compared to very deleterious alleles they remain in the population for much longer and can rise to high frequencies.

To generate empirical estimates of the genetic load it is important to describe and analyse the phenomena contributing to it. In a recent review on the subject Henn et al. (2015) defined four components of the genetic load.

Firstly, the mutation load which is the fraction of the genetic load due the reduction in fitness because of recent deleterious mutations. Secondly, the inbreeding load occurring when homozygous recessive deleterious alleles are present more frequently in the population than would be expected if the loci were in Hardy-Weinberg equilibrium. The final two components are the segregation and the transitory load. These concepts respectively refer to scenarios where both homozygotes for a specific genotype exhibit a lower fitness than the heterozygote or when environmental changes cause previously optimal allelic configurations to become suboptimal. The focus of the following paragraphs will be on the mutation and the inbreeding loads. Before

describing the analyses comparing the global distribution and properties of potentially deleterious site classes contributing to the genetic load it is important to consider which classification approaches are used to determine the fitness impact of a novel mutation.

The latter are not only relevant for evolutionary analyses but have recently been shown to be potentially important in clinical practice. For one statistic, the CADD score (Kircher et al., 2014), it was demonstrated that its scoring scale correlated very well with the pathogenicity classification made by experts for variants underlying a hereditary type of colorectal cancer (van der Velde et al., 2015). However, it should be cautioned that functional deleteriousness does not necessarily result in a clinical phenotype. This reduces the practical predictive value of these approaches that might also vary by genomic regions and disease mechanisms (Mather et al., 2016).

Approaches to estimate variant deleteriousness can be subdivided into three categories. Firstly, there are methods using evolutionary conservation on the nucleotide level as a metric to identify regions where *de novo* mutations are most likely to have a disruptive impact such as GERP (Cooper, 2005) and PhyloP (Siepel et al., 2006). Another group of algorithms works on the protein level and assesses evolutionary conservation of amino acid sequences and/or the impact a particular missense mutation will have on protein conformation based on the properties of the amino acids involved. The widely used tools PolyPhen-2 (Adzhubei et al., 2010) and SIFT (Hu and Ng, 2012, 2013; Kumar et al., 2009; Ng and Henikoff, 2001) belong to this category.

More recently, machine-learning approaches have been developed. They integrate a wide range of predictions from other tools and annotations from functional databases to generate a score for each SNP or small scale indel in the genome. The most prominent of these is the CADD score mentioned above. A more detailed, although far from exhaustive, list of variant annotations algorithms and their underlying methodologies can be found in Appendix A.1.

Miosge et al. (2015) tested the accuracy of these inferences by a mutagenesis approach. They induced *de novo* mutations in 23 essential immunity genes in mice. Approximately 20% and 15% of these *de novo* mutations inferred to be deleterious by the predictive tools PolyPhen-2 and CADD respectively led to a measurable loss-of-function phenotype in homozygote individuals. A more detailed analysis of the protein encoded by the tumour suppressor gene *TP53* revealed that ca. 50% of all deleterious mutations caused a clear reduction of its transcription-enhancing activity. The other deleterious mutations while having no impact on

this measurable phenotype were, however, still depleted in empirical data and therefore presumably under purifying selection. This study underlines that more experimental data from high-throughput mutagenesis approaches and specific human phenotyping are crucial to improve our ability to interpret the consequences of deleterious variation. This information then would be fed back into the training datasets used by machine-learning algorithms to improve their prediction accuracy.

However, the *in silico* methods will always have limitations as demonstrated by the HGMD database which contains information on genetic variants related to human disease. Cooper et al. (2013) described the widely observed phenomenon of "reduced penetrance" where individuals with a disease-causing genotype do not exhibit any or only a few symptoms. Among the factors causing it are interindividual variation in gene expression patterns, copy number variations, the impact of adjacent sites and sex- or age-specific effects.

Many of the novel annotation tools are reliant on training datasets which are necessarily imperfect for the reasons already mentioned. There is also some evidence for negative epistatic effects from model organisms, i.e. the viability decline with the accumulation of deleterious variation appears to be non-linear (Fry, 2004). Additionally, some methods have shown a reference bias meaning that polymorphisms at sites where the reference is derived are more likely to be called as benign (Simons et al., 2014). Finally, if we assume that classic selective sweeps took place at least at a small fraction of loci in recent human evolution some high-frequency missense variants will be erroneously annotated as deleterious.

1.6.4 Potential population differentiation in the genetic load

The first systematic genome-wide study on the potential differences in the patterns of deleterious variation between human populations was conducted by Lohmueller et al. (2008). They presented exome data from 15 African and 20 European Americans and compared the ratios of heterozygous non-synonymous and homozygous non-synonymous genotypes between these groups. The main outcome was an excess of homozygous derived alleles at non-synonymous sites and of damaging homozygous alleles at sites classified as probably damaging by PolyPhen in European Americans. Furthermore, the authors ran forward simulations of the population histories of the respective groups with a severe bottleneck for non-Africans and subsequent growth for both populations. A comparison between empirical and simulated

genetic data demonstrated that differences in population history can have a considerable effect on the patterns of deleterious variation observed. The authors hypothesised that this could be indicative of the lowered efficacy of purifying selection in non-Africans in the past.

Peischl et al. (2013) explored the effect of particular demographic scenarios on deleterious variation, primarily of extreme drift at the front of the wave of population expansion occurring during serial founder events. They inferred an accumulation of deleterious variants and decreased mean fitness in groups at the periphery of an expansion wave vs those at the geographical centre using spatially explicit forward simulations.

Fu et al. (2014b) replicated Lohmueller et al.'s (2008) results regarding the discrepancies of patterns of deleterious mutation between European Americans and African Americans on a sample of >6,500 individuals from the ESP (Exome Sequencing Project) (Tennessen et al., 2012). Furthermore, they observed a significantly higher total number of derived deleterious (as indicated by PhyloP) alleles in the former. The alleles contributing to this excess were inferred by simulations to be mildly deleterious and often fixed in the population of European ancestry.

Recently, Henn et al. (2016) generated high-coverage exome and medium-coverage WGS data from 54 individuals belonging to seven worldwide populations of the HGDP representing a more even coverage of worldwide genetic diversity compared to previous work. Using the GERP score they detected an increase in the number of putatively deleterious alleles in non-African populations which was positively correlated with distance from Sub-Saharan Africa. This observation was much more pronounced for homozygous derived genotypes. However, the separation between populations was less clear for variants inferred to be highly deleterious, possibly indicating uniformly strong purifying selection. The authors observed a shift towards common variants in the deleterious spectrum in non-African populations and no excess of rare variants in the latter as reported by Casals et al. (2013) for the recently bottlenecked and expanded Quebecois vs French from France.

The most likely reasons for this discrepancy (besides different time scales) are that the sample sizes in Henn et al.'s study from each population are very small (n = 7-8) so that the excess of rare and intermediate frequency variants cannot be distinguished and some of the groups sampled by them (Yakuts) have not undergone a recent expansion like many other Eurasians. They demonstrated that a simple 2D range expansion model to simulate genomic outcomes

represents a better fit to the empirical data than the output generated from a single bottleneck and subsequent population growth model.

The above studies all found disparities in the number and/or frequency of deleterious mutations implying differences in the strength of selection and the resulting genetic loads. However, other studies arrived at fundamentally different conclusions.

Simons et al., (2014) analysed the same dataset as Fu et al.(2014b) and demonstrated that if PolyPhen-2 was used to measure the deleteriousness of variants the number of deleterious derived alleles per individual was the same in African Americans and European Americans. They also ran simulations indicating that for additive mutations and sufficiently strong selection coefficients the load of deleterious alleles is not affected by bottleneck or population growth scenarios. This led the authors to conclude that an individual's mutational burden is largely unaffected by demography. Do et al. (2015) reached similar outcomes mainly working on exomes from phase 1 of the 1000 Genomes Project and the ESP. They designed the new statistic $R_{X/Y}$, which compares the sums of DAFs over a particular site class between different populations while using neutral variants as reference (see section 3.1.4). Using this approach, they found no measurable differences in the load of derived non-synonymous or putatively deleterious (measured by a reference-independent version of PolyPhen-2) alleles between a range of diverse African as well as non-African populations. Applying this method to simulated data generated from fitted West African and European demographic models supported this observation and confirmed Simons et al.'s (2014) statement that under an additive model there are no appreciable differences in the number of potentially deleterious derived alleles between different "superpopulations". Their simulations also indicated that paradoxically selection might even be more effective in non-Africans insofar as it is measured as the rate at which non-synonymous mutations are removed from the population. However, the effects of prior drift due to the bottleneck and novel mutations because of recent explosive population growth counteract this.

Several authors have attempted to synthesise the apparently contradictory evidence presented above and to highlight the current gaps in our understanding of the subject (Gravel, 2016; Henn et al., 2015; Lohmueller, 2014a).

The first is how the genetic load should be calculated, as it cannot be directly measured but rather depends on parameters and assumptions in a population genetic model. These include

the underlying DFE for each mutation, which is currently unknown for humans and at least partly population-specific due to its relationship with environmental factors. Crucially, the genetic load also depends on the distribution of dominance coefficients h. Studies stating that the genetic load is not significantly different between modern human populations acknowledged that this only held true if all deleterious mutations were assumed to be additive, i.e. h = 0.5. Given that all studies using empirical data agreed that homozygous and heterozygous deleterious genotypes had population-specific distributions this assumption is crucial.

When the genetic load was calculated assuming total (h = 0) or partial (h = 0.25) recessiveness there were significant differences between populations (Henn et al., 2015, 2016). It is also important to keep in mind that a considerable fraction of loci are recessive. Antonarakis (2019) recently estimated that 9,000-10,000 protein-coding genes might be autosomal recessive for recognisable deleterious phenotypes extrapolating from known disease-associated genes and empirically observed patterns of mutation tolerance in the ExAC data (see sections 1.6.2 and 1.6.3). This would imply that roughly half of all protein-coding genes could be considered recessive loci. Experimental approaches where mutations were induced and their potential phenotypic consequences measured indicated an inverse relationship between dominance and severity of novel mutations (Agrawal and Whitlock, 2011) and that the average dominance of mildly deleterious mutations across a panel of non-human species is ca. h = 0.25 (Manna et al., 2011). Further evidence for this comes from mouse phenotyping efforts, e.g. Ayadi et al. (2012) found that a third of the mouse genes they surveyed were lethal if both copies were knocked out. Finally, many severe recessive genetic diseases, defined by their inclusion in new-born screening panels, are frequent enough that approximately half of all healthy individuals from the ESP dataset carried at least one risk allele in their exome (Tabor et al., 2014).

Secondly, it matters how the “efficacy of selection” is described. This is a contentious point because there is currently no generally accepted definition of the term and it is debated how the non-equilibrium setting in which human populations evolve impacts the statistics used to measure it (Brandvain and Wright, 2016; Gravel, 2016).

In conclusion, theoretical quantitative predictions about differential genetic load and the efficacy of selection require more reliable models of demographic history and the distributions of dominance and selection coefficients to improve their accuracy. From surveying empirical work on the subject it becomes clear that most analyses have focussed on some well-studied

reference populations (Henn et al., 2015) which has only very recently begun to change. Furthermore, most of these studies have used exome-sequencing data, which while containing the majority of all currently known disease-related and deleterious variants, is not a representative picture of full genetic diversity.

1.7 Rare variants and their application to problems of past demography

One of the main results of the first analyses of large-scale exome and WGS data, which were almost exclusively based on populations of European ancestry, was an excess of rare variants relative to older empirical studies and predictions derived from earlier models of population history (Coventry et al., 2010; Nelson et al., 2012; Tennessen et al., 2012). It has been demonstrated that this observation best fits a population history involving recent, exponential (Keinan and Clark, 2012), potentially even super-exponential (Reppell et al., 2014) growth starting < 5-10 kya (reviewed by Gao and Keinan, 2016).

Similar patterns have been observed for the (West) African component of African American genomes (Chen et al., 2015; Tennessen et al., 2012). Originally, estimates for East Asian populations pointed towards more mild growth (Gravel et al., 2011), however these earlier studies predicted a similar trajectory for Europeans. Most likely this is because the low sample sizes previously available did not allow for the observation of very rare variants. Also, population estimates for China based on historical census data, while containing considerable uncertainties, indicate population sizes of ca. 40-70 million as early as 2 kya (Durand, 1960).

The non-parametric PSMC and MSMC methods have recently been applied to a wide range of populations. By their nature they do not yield explicit estimates for growth rates, as opposed to the studies cited above that fitted models to summary statistics of population history, and their accuracy decreases for recent time frames. However, they can still indicate general trends, which mostly point towards recent growth for African and non-African groups with considerable intergroup variation (Mallick et al., 2016; Romero-Hidalgo et al., 2017; Schiffels and Durbin, 2014). However, irrespective of population-specific histories, even in groups that experienced only moderate or no recent growth the majority of the SFS consists of rare variants.

While there is no universally accepted definition of the term “rare” when applied to genetic variation, a commonly used threshold is a MAF of <1% or <0.1% in the total sample (Gao and

Keinan, 2016). Theoretical considerations imply that on average rare variants are younger than common variants (Kimura and Ohta, 1973) and in consequence particularly informative about recent population history. Therefore, rare variants should both improve the accuracy of existing approaches as well as allow for the development of new methodologies.

O'Connor et al (2015) demonstrated that for population clusters that split recently rare variants have a higher cumulative information content than common variants. They supported this by simulations where the former had a higher accuracy for assigning individuals to FRAPPE clusters corresponding to their true ancestries based on population splits. Empirically, a PCA limited to rare variants highlighted a previously undetected sub-cluster in European American individuals from the ESP dataset, which likely reflects individuals of Ashkenazi Jewish ancestry. More generally, methods that rely on the accurate inference of local ancestry such as ChromoPainter/fineSTRUCTURE and approaches detecting genomic runs that are IBD (Browning and Browning, 2013) should benefit greatly from incorporating rare variation as the sharing of one or multiple such variants provides strong support for locally shared ancestry (Gao and Keinan, 2016).

The 1000 Genomes Project Consortium (2012) defined variants present only twice in a heterozygous state in the whole dataset as f_2 sites and analysed their sharing patterns across all populations assembled for phase 1 of the project. The main outcomes of this study and subsequent applications of this statistic (e.g. Genome of the Netherlands Consortium, 2014) confirmed theoretical expectation. The majority of all f_2 variants were shared between individuals from the same population and elevated between-group sharing highlighted recent links consistent with hypotheses based on non-genetic evidence.

Mathieson and McVean (2014) utilised the 1000 Genomes Project phase 1 dataset to develop an approach to estimate the lower age boundary of f_2 variants from the age of f_2 haplotypes. They defined the latter as a region where two chromosomes from different individuals in a dataset are each other's closest genetic relative. To achieve this, they detected f_2 haplotypes based on unphased genotypes from high coverage SNP arrays and WGS data using relatively permissive criteria. The ages of these were then estimated using a maximum likelihood approach formalising the intuition that older haplotypes should be shorter and carry more singletons. Simulations showed that while there was great uncertainty concerning the age estimate for each individual run collectively, they can be informative about demographic history, e.g. the median f_2 haplotype age within Eurasian populations (50-160 generations) was

systematically lower than for internally shared runs in Africans (170-320 generations). In turn haplotypes shared across continents were much older than within.

Another independent analysis of the same dataset (Al-Khudhair et al., 2015) introduced the related concept of very rare genetic variants (vrGVs) with a MAF threshold of 0.2% in the total dataset. Al-Khudhair et al. (2015) in concordance with earlier works found considerable variation in the per-individual vrGV counts between different populations. Furthermore, they showed that known close relatives from the 1000 Genomes Project share thousands of vrGVs and that in principle even a very simple counting method of vrGVs should be informative about more distant genetic relationships. A subsequent paper by the same group (Fedorova et al., 2016) defined the concept of rare variant clusters (RVCs) consisting of five or more adjacent vrGVs. Sharing of RVCs between two or more individuals is very unlikely to result from coincidence and therefore indicative of IBD. The authors furthermore presented evidence that if rare variants were present in a region of the genome, they were able to detect at least a fraction of true very old short IBD segments.

1.8 Aims and Objectives

The main motivation for this thesis stems from the observation that the early studies on WGS datasets focussed on a limited subset of reference populations. To gain a more complete and granular understanding of the genomic history of our species, there is a need to better characterise genetic diversity in understudied regions, such as Island Southeast Asia and Siberia. This thesis contributes towards this goal in form of three studies. The first focusses on the population history of Southeast Asia, the second and third explore patterns of functional and rare variation in a worldwide genomic dataset. The literature on these subject areas has been reviewed in sections 1.5, 1.6 and 1.7 respectively and motivates the specific questions addressed in this work, which will be described in more detail in the following.

For the first project, in order to refine the current understanding on the source of the Austronesian expansion and to further explore potential South Asian genetic contributions in MSEA and ISEA, high density (730K) SNP Chip data were generated for 196 individuals from 10 populations. Of these, 50 from the Bajo and Lebbo populations are published already (Pierron et al., 2014) and 146 new (Burmese and Vietnamese from MSEA, Ilocano, Tagalog and Kankanaey from the Philippines, Murut, Malay and Dusun from ISEA plus four Australian

Aborigines). The newly generated dataset was merged with data from the literature and analysed with an emphasis on the following aspects. Firstly, the current knowledge on the putative source of the Austronesian expansion was re-examined. Secondly, the existence of signs of South Asian admixture in the newly available SEA populations was investigated. Finally, the extent to which signs of local adaptation are shared across local populations, as function of their common demographic history was determined. These results are presented in chapter 2.

As part of an ongoing effort to improve our understanding of the patterns of genomic diversity in non-reference populations a set of 483 whole genomes, of which 379 were novel, known as the Estonian Biocentre Human Genome Diversity Panel (EGDP) was generated (data published in Clemente et al., 2014 and Pagani et al., 2016). The author of this thesis was a member of the consortium providing the first comprehensive analysis of this dataset.

The second project is an investigation of the patterns of functional variation as well some measures of positive selection on a subset of the EGDP (n = 382) (see Appendix C.1 for samples, for details on the dataset see section 3.1.1). Particular attention will be given to the following aspects. Firstly, both the distribution of different classes of functional variants in non-reference populations and the extent of sharing across regions will be examined. Secondly, using this more complete picture of genomic diversity, it will be investigated whether statements made about purifying selection based on reference data must be reassessed. In this context, the potential differential impact of purifying selection on various gene classes will also be examined. Finally, improvements to the present understanding of potential target loci of positive selection based on unbiased WGS data using variant-based approaches and integration with information from functional databases will be explored.

The third project focusses on rare variant-based approaches, the EGDP dataset again represents an interest target as it consists of high-coverage full sequences and exhibits a broad geographical spread encompassing previously understudied regions of the world. Its main disadvantage for this type of analysis is the comparatively low number of individuals from any specific subpopulation. Therefore, the sharing of f_2 variants in a subset of the EGDP designated as the Diversity Set (n = 447) (Appendix C.1, for more details on the dataset see section 3.1.1) will be analysed in chapter 4 with a focus on describing potential cryptic interpopulation-relationships.

2. Insights into Southeast Asian population history from high coverage SNP data

This chapter describes analyses of high-coverage genotype data from 196 individuals representing a variety of populations spread across Southeast Asia (SEA). This chapter's content is based on a collaborative paper of which I was the first author. The text has been revised and expanded.

Mörseburg A, Pagani L, Ricaut F-X, Yngvadottir B, Harney E, Castillo C...Metspalu M, Kivisild T. 2016. Multi-layered population structure in Island Southeast Asians. European Journal of Human Genetics 24:1605–1611.

My contribution to this work builds on exploratory analyses by Dr Bryndis Yngvadottir (University of Cambridge) and Eadaoin Harney (University of Cambridge, present affiliation: Harvard University) who together with my supervisor Dr Toomas Kivisild first introduced me to the then unpublished dataset. I ran all analyses reported in this chapter unless explicitly stated otherwise and wrote the manuscript integrating comments and suggestions from my co-authors.

The history of human settlement in SEA has been complex and involved several distinct dispersal events. Previous genetic studies of the region have distinguished at least three major ancestral components roughly corresponding to the predominant ancestries in the genomes of present-day Papuan-, Austro-Asiatic- and Austronesian-speaking populations (HUGO Pan-Asian SNP Consortium et al., 2009; Lipson et al., 2014; Xu et al., 2012).

The chapter is mainly motivated by three questions, the background of which is described below. They will be addressed under consideration of the novel data together with previously published datasets.

1) A demographic event of great importance for Island SEA (ISEA) is the dispersal known as the Austronesian expansion. The current majority view in the field is that it began in Taiwan and spread across the ISEA region extending as far as Far Oceania and Madagascar. It is thought to be associated with the spread of the Austronesian languages and the Neolithic cultural complex (Bellwood, 2007; Spriggs, 2011). Therefore, Taiwanese aboriginals have been seen as

the best extant proxy for the source population giving rise to the Austronesian expansion. However, it is known that they have been subject to recent admixture from the Chinese mainland (HUGO Pan-Asian SNP Consortium et al., 2009). Several population groups in the Philippines and Indonesia are considered either to be direct descendants of the early Austronesian settlers or to derive the great majority of their ancestry from them. Some of these have been relatively isolated at least in historic times (for more details see Appendix B.1). Could these groups be the best living representatives for the Austronesian colonists and help us refine our knowledge of the early stages of the Austronesian expansion?

2) There are multiple lines of evidence for South Asian influences in ISEA and Mainland SEA (MSEA) starting more than 2 kya. These manifest in the form of cultural and trading networks, which have been inferred from historical records as well as archaeological excavations (Ardika et al., 1997; Ardika and Bellwood, 1991; Lawler, 2014; Manguin et al., 2011) and are reinforced by linguistic data (Gonda, 1973; Hoogervorst, 2015). Corresponding to the cultural exchange there is some indication of gene flow coming mainly from uniparental markers suggesting low but detectable levels of Indian ancestry throughout Indonesia (Chaubey and Endicott, 2013; Karafet et al., 2005, 2010; Kusuma et al., 2015) Do autosomal data support these findings of genomic signatures from South Asia in the region?

3) A wide range of methods has been developed to scan genomes for traces of positive selection. The features examined can be allelic differentiation, the SFS or LD patterns. The measures themselves vary from simple summary statistics to more complex approaches based on maximum likelihood or machine-learning frameworks. There is an ongoing debate about the power of these tests to a) detect patterns resulting from true selective events and b) distinguish these from neutral processes, especially under non-equilibrium demographies (Pavlidis et al., 2012; Jacobs et al., 2016; Xiang-Yu et al., 2016). Therefore, hypotheses concerning specific causal loci and populations should be formulated with great caution. These problems are exacerbated with SNP array data and when phenotypic information from the genotyped individuals is lacking, as is the case for the data presented here. In consequence, only the general sharing of outlier regions detected by selection statistics across SEA is examined. To what extent are the top signals from selection statistics common across Southeast Asia and how well does their similarity correlate with genetic distance measures?

2.1 Material and Methods

2.1.1 Newly generated data and quality control

The dataset this chapter focusses on consists of 196 individuals from 12 SEA (Figure 2.1 and Appendix B.2) and four individuals from one Australian population. DNA was extracted from saliva samples collected from healthy adult donors who signed an informed consent form. The field research to collect the samples was led by Dr Syafiq Abdullah (RIPAS Hospital, Brunei) (for the Dusun and Murut populations), Dr Francois-Xavier Ricaut (University of Toulouse) (Bajo, Lebbo) and Dr Toomas Kivisild together with Dr Joseph Wee (National Cancer Centre, Singapore) (Burmese, Kankanaey, Ilocano, Malay, Pangasinense, Tagalog, Vietnamese, Visayan). The study was approved by local Research Ethics Committees (SingHealth Centralised Institutional Review Board and the Medical and Health Research Ethics Committee of the National Cancer Centre, Brunei Darussalam), the Cambridge Ethics Committee (HBREC.2011.01) and the ERC Ethics Panel. All SEA samples were genotyped at Cambridge Genomic Services using Illumina OmniExpress Bead Chips for 730,525 SNPs. The data are accessible under the GEO accession number GSE77508. For the three Australian samples the Illumina Human 660K Quad Bead Chip yielded 655,215 SNPs, while for one Australian the 610K version of the latter chip gave 616,795 variants. These four samples are available under the accession number EGAS00001001738 in the European Genome-phenome Archive.

Data filtering and quality checks were performed using PLINK 1.07 (Purcell et al., 2007). Firstly, only autosomal SNPs with a genotyping success rate greater than 98% were included. PLINK was also utilised to detect all individuals from the same population more closely related than first cousins. This was done by estimating IBD iteratively within populations; individuals with an IBD > 0.125 were excluded from all analyses except the initial F_{ST} and Refined IBD approaches. Following these quality controls haplotypes were inferred from genotype data with SHAPEIT (Delaneau et al., 2013b). For all analyses presented in this chapter, unless explicitly stated otherwise, the four non-Kankanaey groups (Ilocano, Pangasinense, Tagalog, Visayan) from the Philippines are pooled as “Filipino” due to their low sample sizes and them all representing genetically similar urbanised lowland groups.

All pre-processing was done by Dr Tiago Antão (University of Cambridge, present affiliation: Embark Veterinary). Furthermore, eight full mitochondrial Kankanaey genomes were

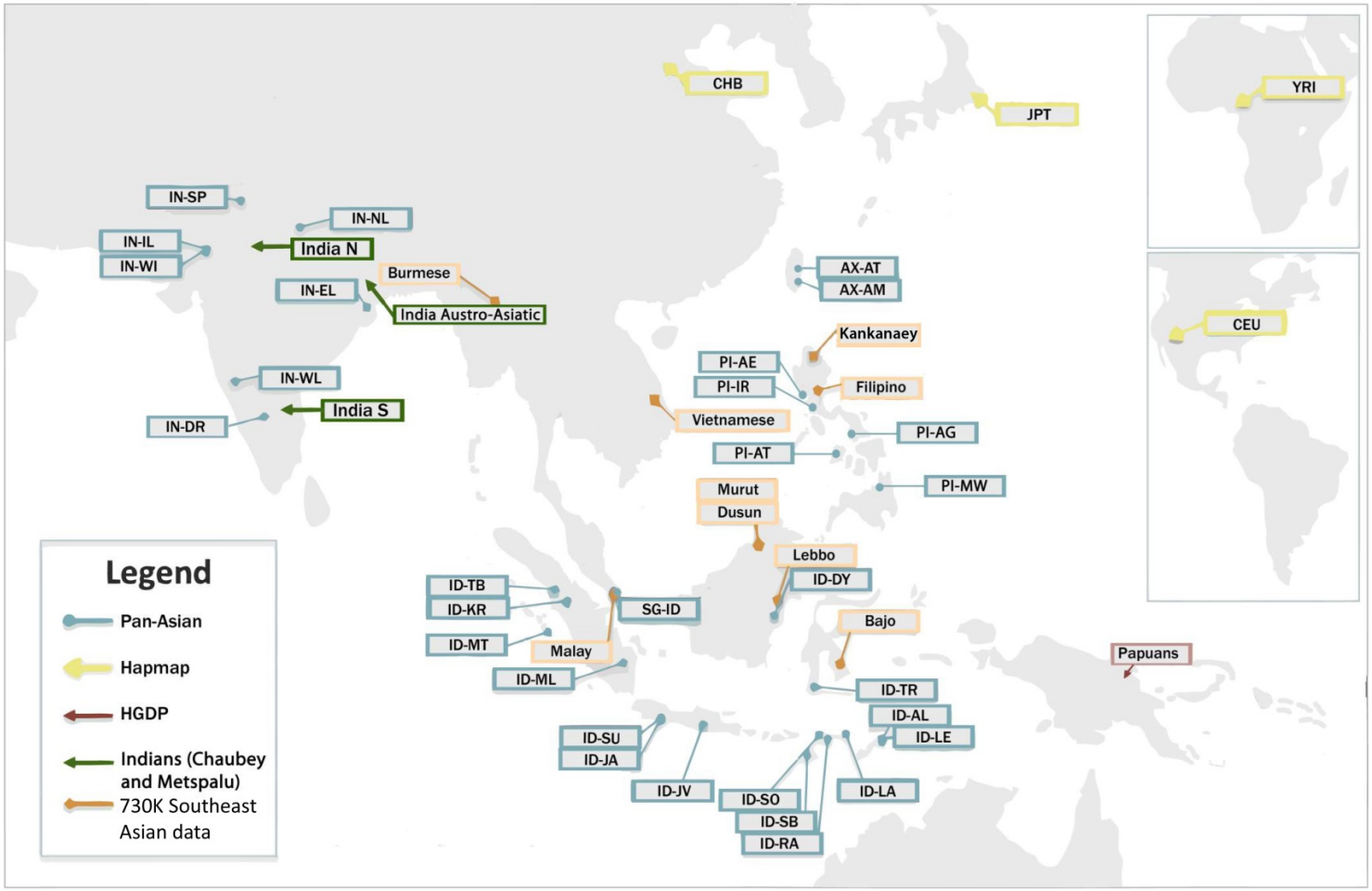


Figure 2.1: A map of Asia, highlighting the populations assessed in this chapter and showing all other sampling locations which were included in the ADMIXTURE analysis displayed in Figures 2.2-2.3. It is based on the HUGO Pan-Asian SNP Consortium (2009) population map. The population abbreviations for the Pan-Asia panel data are as follows: AX-AM = Ami, AX-AT = Atayal, ID-AL = Alorese, ID-DY = Dayak, ID-JA = Javanese I, ID-JV = Javanese II, ID-KR = Batak Karo, ID-LA = Lamaholot, ID-LE = Lembata, ID-ML = Malay (Indonesia), ID-MT = Mentawai, ID-RA = Manggarai I, ID-SB = Kambera, ID-SO = Manggarai II, ID-SU = Sundanese, ID-TB = Batak Toba, ID-TR = Toraja, IN-DR = Telugu, IN-EL = Bengali, IN-IL = Hindi (Upper Caste) I, IN-NL = Hindi (Upper Caste) II, IN-SP = Hindi (Upper Caste) III, IN-WI = Bhil, PI-AE = Aeta, PI-AG = Agta, PI-AT = Ati, PI-IR = Iraya, PI-MW = Mamanwa, SG-ID = Tamil (Singapore).

sequenced by Complete Genomics (Mountain View, California, USA) using CG software version 2.4. Access to the sequences is provided under the GenBank accession numbers KU752558 to KU752565.

2.1.2 Computational methodology

To get a first overview of the data, the SEA 730k SNP dataset was merged with four reference populations from the HapMap 3 panel (Frazer et al., 2007) , the HGDP Papuans (Li et al., 2008) and individuals from diverse South Indian populations (Metspalu et al., 2011) to obtain a set of 307,625 common SNPs. Runs of homozygosity (rOH) and average observed heterozygosity were obtained using PLINK default parameters. Pairwise F_{ST} was calculated using an *ad hoc* Perl script implementing an estimator for Wright's formula (Wright, 1931).

Furthermore, IBD scores were computed for all 22 autosomes (n_{SNP} = 306,198) with the Refined IBD algorithm (Browning and Browning, 2013) after they had been separately phased using SHAPEIT2 (Delaneau et al., 2013b). For details on IBD calculation using Refined IBD (see Appendix B.3), note that for the published paper IBD had been calculated using PLINK (results can be found in Appendix B.6). All these analyses except the Refined IBD were run by Dr Luca Pagani (University of Cambridge, present affiliations: University of Padua and University of Tartu).

ADMIXTURE analyses

To address more specific questions regarding the ancestries of the SEA populations, two distinct ADMIXTURE (Alexander et al., 2009; Alexander and Lange, 2011) analyses were run. A more detailed description of all samples can be found in Appendix B2. It lists the ethnic origins of the individuals, the number of genotyped markers, the literature source (if applicable) and whether the sample was included in either or both ADMIXTURE runs.

For comparative purposes publicly available genotype data from the HapMap (International HapMap Consortium et al., 2007) , HDGP (Li et al., 2008) and the Pan-Asian Consortium (HUGO Pan-Asian SNP Consortium et al., 2009) projects were added to 185 individuals from nine SEA populations (the lowered total results from the removal of 11 close relatives). Additionally, SNP data from studies focussed on Indian populations were used (Chaubey et al., 2011; Metspalu et al., 2011). This resulted in a dataset consisting of 1099 individuals.

For further verification of the inferences from the first ADMIXTURE analysis a second panel of 1010 samples was created, including 187 individuals from the nine SEA population and four Australian Aborigines. The minor discrepancies (185 vs 187) in the number of SEA individuals between the two ADMIXTURE analyses (see also Appendix B.2) can be explained by subtle differences in the pre-analysis filtering schemes. A detailed description of the merging and data curation for admixture can be found in Appendix B.4.The analyses on the verification panel were conducted by Dr Mait Metspalu (University of Tartu).

Additional demographic analyses

N_e for the nine SEA populations was estimated by analysing LD patterns with the NeON R package (Mezzavilla and Ghirotto, 2015). To further investigate genetic structure and gene flow between populations the TreeMix v1.1 software (Pickrell and Pritchard, 2012) was used. To measure how well the trees with different numbers of migration events reflect the relationship between population groups the fraction f of explained variance was calculated as described by the original authors of the method. MEGA v6.0.6 (Tamura et al., 2013) was used to create a graphic representation of the TreeMix output. For specific admixture events of interest suggested by the ADMIXTURE plots the respective sets of recipient and source populations were tested with the f_3 test (Pickrell and Pritchard, 2012). The population trios yielding a z-score smaller than -2 were considered significantly admixed. ALDER v1.03 was used to examine LD patterns potentially resulting from admixture with weighted LD curves and to date these putative admixture events (Loh et al., 2013). Furthermore, the f_4 ratio test (Moorjani et al., 2011) was applied to obtain a quantitative estimate of admixture percentages of interest.

For the analysis of the mtDNA data the haplogroup affiliation of each sample was assigned using HaploGrep 2.0 (Kloss-Brandstätter et al., 2011) and PhyloTree build 16 (as of 19/02/2014) (http://www.phylotree.org) (van Oven and Kayser, 2009). The variants are described relative to the rCRS (GenBank Accession Number NC_012920.10) (Andrews et al., 1999).

Selection tests

To capture haplotype homozygosity based signals the Integrated Haplotype Score (iHS) (Voight et al., 2006) and Cross Population Extended Haplotype Homozygosity (XP-EHH) (Sabeti et al., 2007) tests were used. Furthermore, the allele frequency-based Population Branch

Statistic (PBS) was calculated (Yi et al., 2010). Details on the implementation of selection tests can be found in Appendix B.5.

2.2 Results

2.2.1 F_{ST} and interpopulation Refined IBD-sharing

The mean average pairwise F_{ST} (Table 2.1) between the SEA groups described in this chapter is 0.020 with the lowest differentiation observed between the Bornean Murut and the Vietnamese (F_{ST} = 0.014) and the greatest difference exhibited by the Kankanaey from the Philippines and the Vietnamese (F_{ST} = 0.026). In an interregional comparison the East Asian groups are about as different from the SEA populations (average F_{ST} = 0.021) as the latter are from each other. This finding underlines the shared ancestry of the Chinese/Japanese and SEA groups. Furthermore, it indicates that the SEA populations exhibit considerable heterogeneity and do not form a tight cluster.

Relative to other macro-groups the SEA populations are somewhat closer to South Indians (average F_{ST} = 0.042) than to Western Europeans (average F_{ST} = 0.057) and less distant from Oceanians (Papuans) (average F_{ST} = 0.085) than from West Africans (average F_{ST} = 0.090). The relatively shorter genetic distances to the South Indians and the Papuans indicate either more recent population splits and/or subsequent gene flow, e.g. due to admixture events (the latter possibility is explored in detail in sections 2.2.2 and 2.2.4).

On average the mean rate of IBD sharing between any pair of SEA populations is $4.4*10^{-3}$ (Table 2.1). The highest IBD sharing values were obtained for the composite Filipino group and the Lebbo from Borneo at 0.0278 and the Kankanaey and the Dusun at 0.0192. The former value is partly driven by four individuals (Luz1, Luz4, lebbo5 and lebbo19) whose average cross-population IBD is 0.2791, comparable to second-degree relatives. It cannot be stated definitely whether this reflects a real biological relationship or as is more likely, an artefact, especially given the relatively low marker densities and confounding background relatedness. However, even if these individuals are removed the IBD sharing between the Filipinos and the Lebbo remains the highest for all interpopulation comparisons at 0.0210.

On a more general level, the Austronesian speakers from ISEA exhibit elevated levels of IBD sharing with each other and also with the Burmese. Notably, the Vietnamese (average IBD =.

Table 2.1: **Pairwise Refined IBD and F_{ST} values of SEA populations together with a set of reference populations**.

IBD/F_{ST}	Bajo	Burmese	CEU	CHB	Dusun	Filipino	Kankanaey	JPT	Lebbo	Malay	Murut	Papuans	South Indian	Vietnamese	YRI
Bajo		0.016	0.057	0.02	0.017	0.021	0.023	0.023	0.022	0.019	0.019	0.078	0.041	0.022	0.089
Burmese	0.0088		0.056	0.022	0.018	0.023	0.024	0.025	0.024	0.021	0.021	0.075	0.041	0.024	0.088
CEU	0	0		0.053	0.055	0.061	0.065	0.055	0.063	0.059	0.048	0.091	0.027	0.052	0.073
CHB	0.0003	0.0003	0		0.018	0.019	0.023	0.009	0.023	0.019	0.014	0.084	0.038	0.014	0.086
Dusun	0.0038	0.0034	0.0001	0.0003		0.019	0.015	0.022	0.02	0.017	0.017	0.085	0.039	0.02	0.088
Filipino	0.0034	0.0026	0	0.0004	0.0032		0.023	0.023	0.015	0.016	0.019	0.091	0.046	0.02	0.094
Kankanaey	0.0044	0.0041	0	0.0004	0.0192	0.0039		0.026	0.022	0.02	0.024	0.095	0.05	0.026	0.097
JPT	0.0002	0.0002	0	0.0007	0.0002	0.0002	0.0002		0.027	0.023	0.017	0.086	0.037	0.018	0.088
Lebbo	0.0048	0.0040	0	0.0003	0.0045	0.0278	0.0056	0.0002		0.019	0.022	0.093	0.048	0.025	0.096
Malay	0.0043	0.0037	0	0.0003	0.0042	0.0094	0.0054	0.0002	0.0068		0.017	0.088	0.043	0.02	0.092
Murut	0.0019	0.0014	0.0001	0.0005	0.0022	0.0017	0.0017	0.0003	0.0020	0.0019		0.079	0.032	0.014	0.083
Papuans	0.0012	0.0014	0	0	0.0002	0.0001	0.0001	0	0.0002	0.0002	0.0001		0.078	0.083	0.118
South Indian	0.0001	0.0001	0.0006	0	0.0001	0	0	0	0	0.0001	0.0002	0.0001		0.037	0.068
Vietnamese	0.0009	0.0007	0.0001	0.0006	0.0009	0.0010	0.0005	0.0004	0.0008	0.0009	0.0011	0	0.0002		0.087
YRI	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

$9*10^{-4}$) and the Murut (average IBD = $1.7*10^{-3}$) are more distant from other groups. For the Vietnamese, an MSEA population, this conceivably reflects a lack of recent shared ancestry. It is intriguing that the Murut are notably different from the Dusun, a neighbouring population from northern Borneo, which is thought to be closely related. The SEA populations have an average IBD sharing of $4*10^{-4}$ with the Papuans. This is mainly driven by the Bajo (IBD = 0.0012) and the Burmese (IBD = 0.0014) and could be suggestive of recent gene flow.

The sharing of the SEA pops with East Asians is on average somewhat lower at ca. $3*10^{-4}$ with the highest values exhibited by the Vietnamese. With regards to the hypothetical South Asian genetic impact in the SEA region there is an average IBD-sharing of $1*10^{-4}$ between South Indians and individuals from SEA groups with slightly elevated values in the Bajo, Burmese, Dusun, Murut and Malays and most likely related traces of European affinities in the Dusun, Murut and Vietnamese. Using an IBD threshold of $1*10^{-4}$ there is no detectable sharing of SEA groups with West Africans. Compared to the $F_{ST,}$ the IBD values indicate a more coherent clustering of SEA groups which is indicative of recent shared ancestry with each other, particularly relative to East Asian groups from whom they are not distinguishable by the $F_{ST.}$

Potential relationships between populations that can be hypothesised based on F_{ST} and IBD will be explored in more detail in the following section. F_{ST} is a relatively crude summary statistic and can only be interpreted as a very general measure of population differentiation resulting from neutral processes over large time scales, mainly drift and gene flow.

2.2.2 ADMIXTURE analyses

To further investigate general patterns of population structure in the SEA data two distinct ADMIXTURE analyses were performed. The first was mainly focussed on populations from SEA and South Asia while the second provided the context of a broader, worldwide genetic landscape and additional validation for inferences from the first analysis.

According to the cross-validation scores for both analyses K = 9 admixture fractions provide the best fit (for the local plot additional Ks are provided in Figure 2.3, for the global plot Ks from 3 to 15 are shown in Figure 2.4). The ADMIXTURE analyses of the newly generated data recapitulate the main ancestral components associated with Austronesian (k6), Austro-Asiatic (k5) and Papuan (k3) populations (Figures 2.2-2.3) described in the area by previous studies (Xu et al., 2012; Lipson et al., 2014). At lower K values the component associated with the Papuans is highly prevalent in Eastern Indonesia and the Mamanwa (a Negrito group from

the Philippines). However, from the groups displayed in Figure 2.2 at higher Ks it continues to persist only in the Alorese and Bajo from Indonesia (Figure 2.2B, Figure 2.3).

Burmese and Vietnamese exhibit significant proportions of the k2 component indicating shared ancestry with East Asian populations. The k4 component associated with South Asian ancestry is also consistently visible in Burmese and Malays and some Indonesian populations, mainly the Batak of Sumatra (HUGO Pan-Asian SNP Consortium et al., 2009). However, at lower Ks this component is also present in the Javanese and the Mamanwa Negritos,

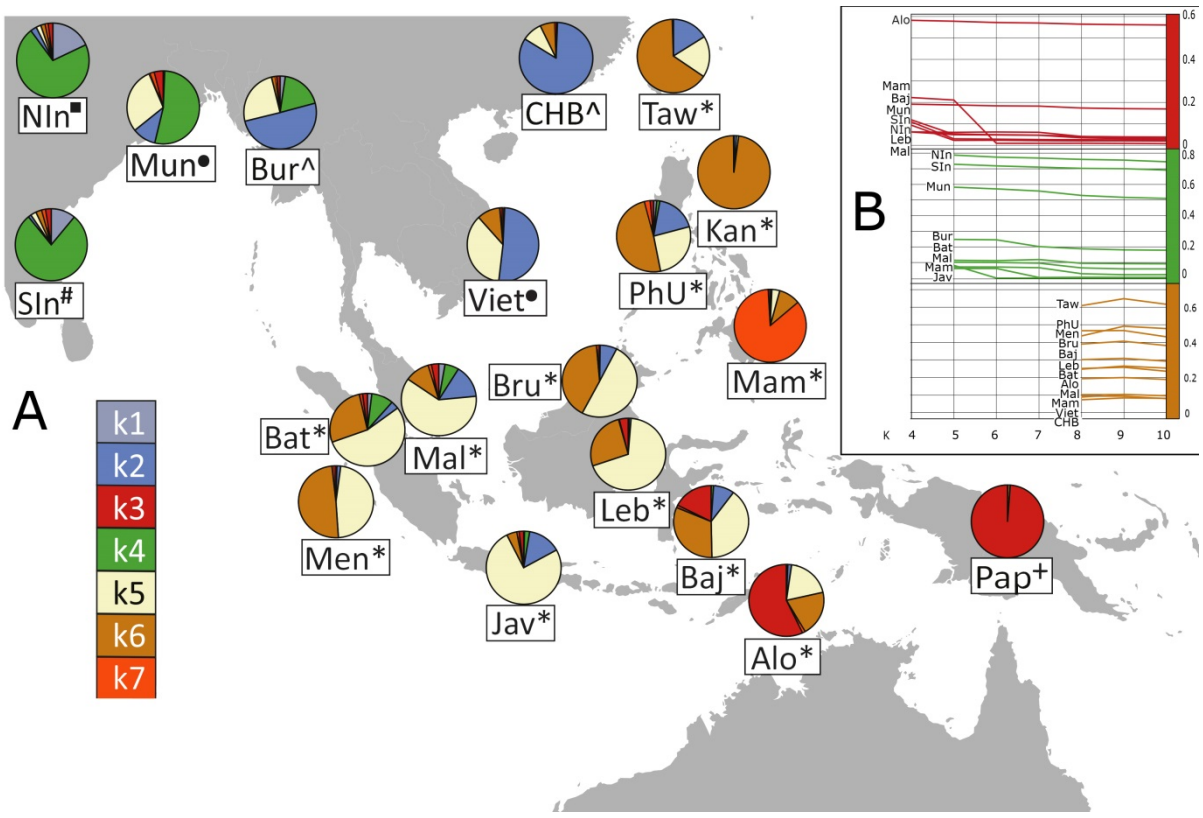


Figure 2.2: (A) A map of Southeast Asia, displaying a subset of populations assessed in this study and the distribution of ancestry components based on the local ADMIXTURE run with the optimal number of ancestry components (K = 9). The figure legend on the lower left section shows the list of genetic ancestry components whose colour codes correspond to those on the pie charts. Components k8 and k9 are mainly present in the Yoruba and Ati Negritos respectively and do not significantly contribute to the genetic diversity of the groups displayed in Figure 2.2 and are, therefore, not shown.
The population abbreviations are as follows: Alo-Alorese, Baj-Bajo, Bat-Batak, Bru-Brunei (Dusun, Murut), Bur- Burmese, CHB-Chinese from Beijing, Jav-Javanese, Kan-Kankanaey Igorot, Leb-Lebbo, Mal-Malay, Mam-Mamanwa Negritos, Men-Mentawai, Mun-Mundari, NIn-North Indians, Pap-Papuans, PhU-Philippine Urban, SIn-South Indians, Taw-Ami and Atayal from Taiwan, Viet-Vietnamese.
The symbols next to the population names reflect the linguistic affiliations. Austro-Asiatic languages: circle, Austronesian languages: asterisk, Indo-European languages: square, Dravidian languages: hash, Papuan languages: cross, Tibeto-Burman languages: caret.
(B) Three graphs showing the proportions of ancestry components k3, k4 and k6 from their emergence as independent components in the Papuans (k3, red), Indian populations (k4, green) and the Kankanaey Igorot (k6, brown) across multiple higher K values. All populations displayed show a percentage of at least 5% of the respective ancestry when it is first detected as a distinct component.

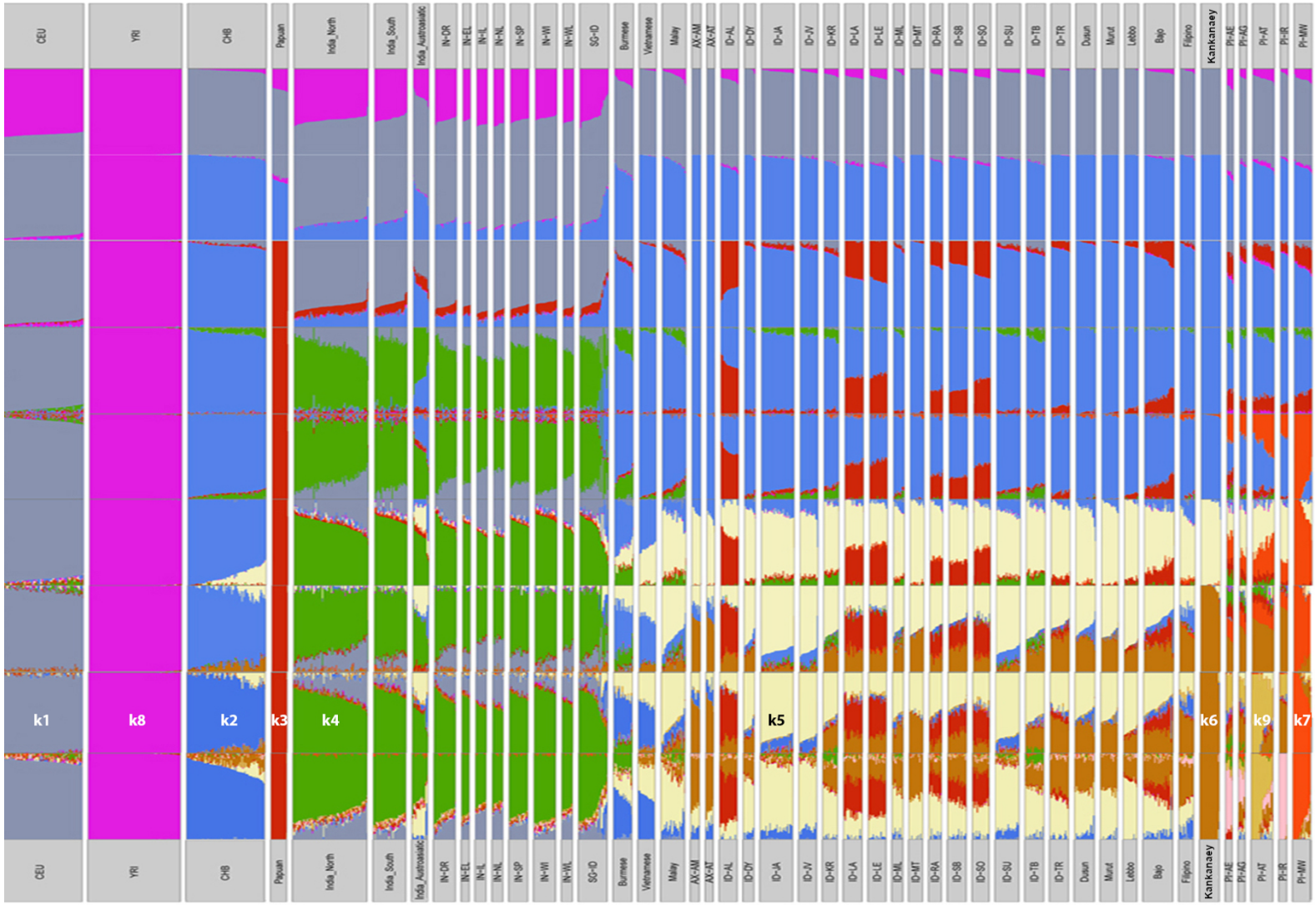


Figure 2.3: ADMIXTURE plot (K = 2-10) based on a panel of regional populations mainly from SEA. For population abbreviations for groups from the Pan-Asia panel data see the legend for Figure 2.1.

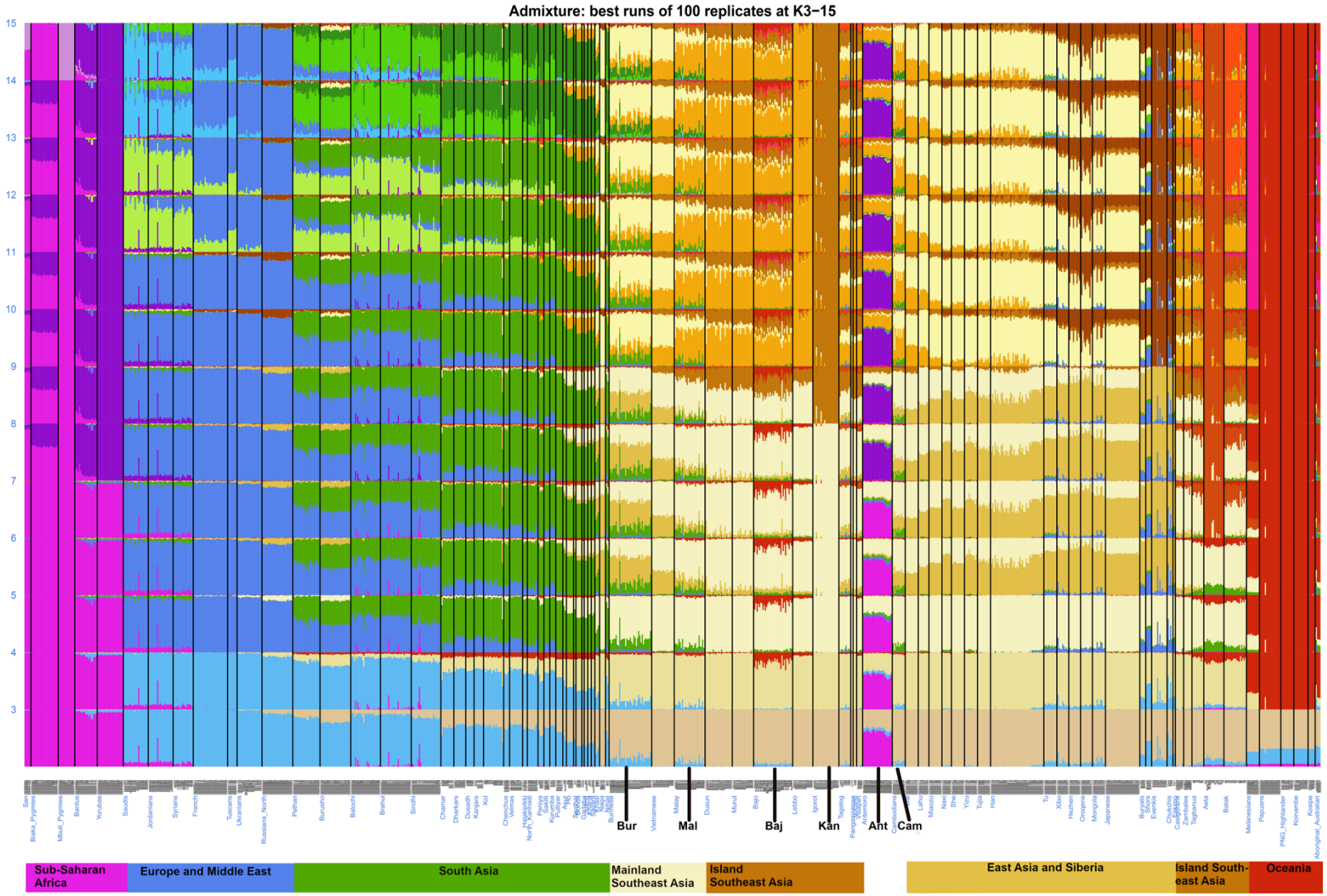


Figure 2.4: ADMIXTURE plot (K = 3-15) including a panel of worldwide populations. Six populations explicitly labelled, these include SEA groups analysed here and mentioned in the text, namely Baj, Bur, Kan, Mal – for population abbreviations see the legend of figure 2.1. and the Antemoro (Ant) and Cambodians (Cam).

suggesting affinities which, however, decline with higher Ks (Figure 2.2B, Figure 2.3). Notably, in the extended worldwide analysis (Figure 2.4) the Papuan-related component (red) in the Bajo and the South Asian signal (green) in the Burmese and Malays were also clearly detectable. The SEA groups described here exhibit a remarkable diversity from very heterogeneous groups such as the Malays to the Kankanaey who appear homogenous in their ancestry composition by the ADMIXTURE analyses (Figure 2.2B, Figure 2.3). The Kankanaey are an indigenous population of northern Luzon, belonging to the broader "Igorot" group. At K = 9, the majority of Kankanaey ancestry is in the k6 component, which they share with the Ami (AX-AM) and Atayal (AX-AT) from Taiwan and, hence, is putatively associated with the Austronesian expansion (Figure 2.2A, Figure 2.3).

When it emerges as distinct from the other Asian components, the k6 ancestry is spread throughout ISEA and remains stable in all these groups from K = 8-10 (Figure 2.2B, Figure 2.3). Remarkably in the regional admixture plots the Kankanaey remain unadmixed throughout all Ks from 2-10, even though at lower Ks they do not yet have their own distinct component. These findings are consistent with the Kankanaey's geographic location, the Mountain Province in the Northern Philippines (Figures 2.1-2.2) close to Taiwan, the likely homeland of the Austronesian peoples (Bellwood, 2007; Lipson et al., 2014).

2.2.3 Patterns of homozygosity, LD and reconstruction of N_e

Table 2.2: Summary statistics for each of the nine SEA populations along with a set of reference populations. Given are sample size (N), average amount of genome covered by runs of homozygosity (rOH), average number of rOH >1Mbp, average heterozygosity (Het) and average within-population IBD.

Population	N	Average genome in rOH (Mbp)	Average N of rOH > 1Mbp	Het	Average within-population IBD
Bajo	32	119.53	28	0.291	0.0338
Burmese	20	97.87	15	0.299	0.0249
CEU	86	85.18	5	0.325	0.0059
CHB	80	81.76	5	0.298	0.0010
Dusun	22	66.17	22	0.296	0.0139
Filipino	16	93.97	17	0.294	0.0164
Kankanaey	22	96.54	34	0.284	0.0392
JPT	86	86.62	7	0.297	0.0042
Lebbo	18	72.24	23	0.291	0.0383
Malay	25	113.63	22	0.29	0.0234
Murut	23	65.92	6	0.308	0.0057

Population	N	Average genome in rOH (Mbp)	Average N of rOH > 1Mbp	Het	Average within-population IBD
Papuans	17	97.63	42	0.24	0.0455
South Indian	40	88.42	19	0.314	0.0049
Vietnamese	18	48.05	4	0.306	0.0013
YRI	94	82.78	6	0.307	0.0082

IBD and related measures can also be used to characterise the general amount of inbreeding throughout the SEA populations. The four statistics in Table 2.2 describe related aspects of intrapopulation homozygosity.

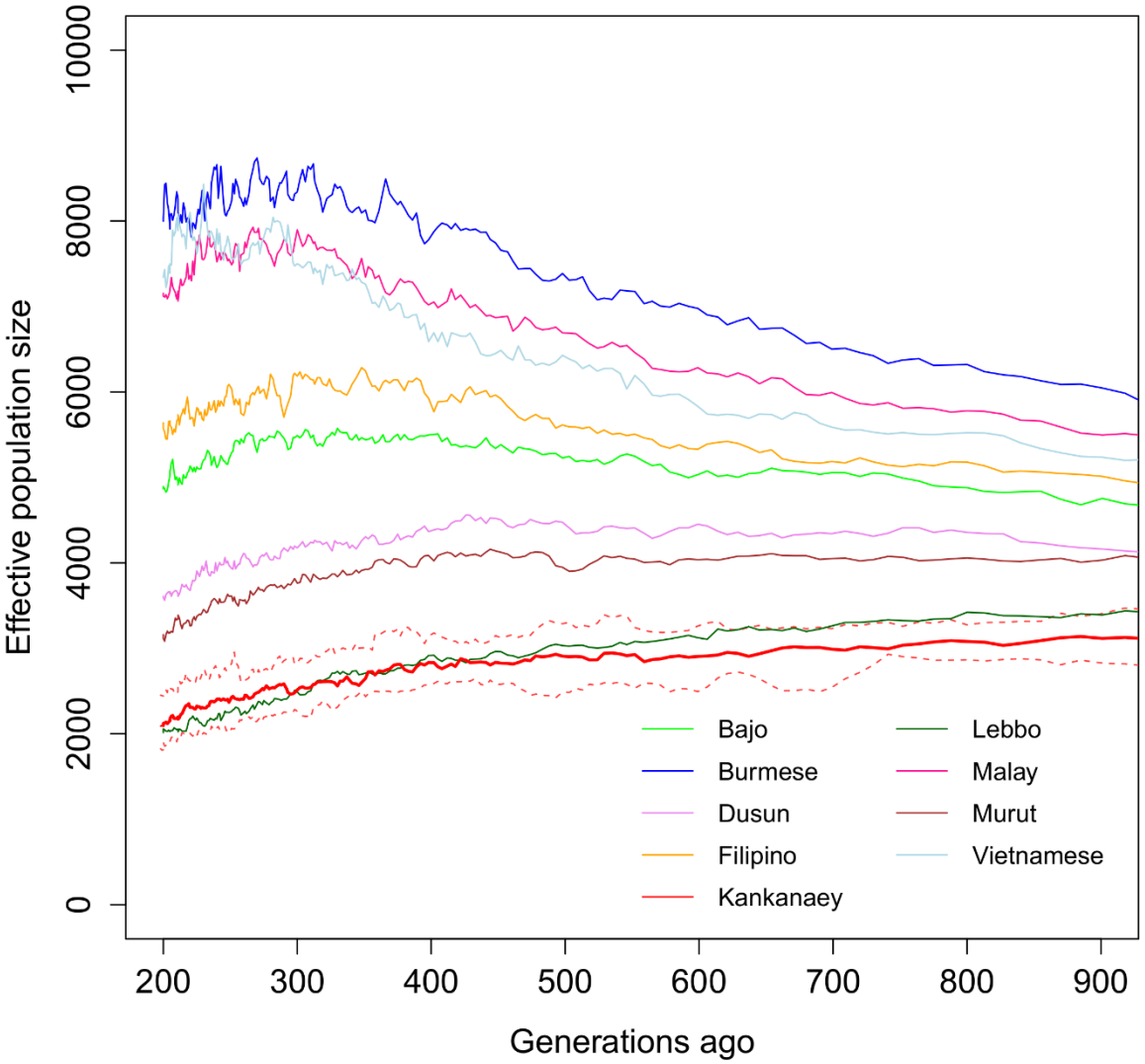


Figure 2.5: Plot of N_e over time for nine SEA populations estimated from LD. The red line describing the Kankanaey is thickened and surrounded by confidence intervals (dotted lines). The latter are the 5th and 95th percentiles of the distribution of N_e obtained from all bins (the bins contain pairs of markers by different recombination distances) across all chromosomes.

Generally, the Murut and the Vietnamese appear to be the most outbred groups whereas some of the Austronesian-speaking ISEA groups, notably the Bajo, Kankanaey and Lebbo exhibit

relatively elevated intra-group sharing. However, it should be cautioned that given the small sample sizes for the SEA groups (16-32 individuals), relative oversampling of a particular subpopulation can skew these measures.

With regards to the unusually homogeneous ancestry observed for the Kankanaey in the ADMIXTURE analyses these results demonstrate that they are comparable to other ISEA populations and that this pattern cannot solely have been caused by extreme inbreeding and/or genetic drift.

To further explore the potential effect of demographic history on population structure the N_e of the nine SEA populations presented here was estimated based on the development of LD patterns over time (Figure 2.5) (Mezzavilla and Ghirotto, 2015). The mainland Burmese and Vietnamese groups exhibit comparatively high N_e values and signs of recent expansion. This is also in line with their recent history of admixture with neighbouring populations, whereas there is more variation in the ISEA populations. Notably, the Kankanaey have one of the lowest values oscillating between 2,000 and 3,000 (6,000-27,000 kya). However, they are not an extreme outlier and are comparable to the Lebbo from Borneo (Mann–Whitney U test, p = 0.7938), who instead do not show such a homogeneous ADMIXTURE profile. Under the assumption that the brown k6 component reflects ancestry connected to the Austronesian expansion, the Kankanaey displayed a higher percentage than even Austronesian Taiwanese populations (AX-AT, AX-AM, Figures 2.2-2.3).

2.2.4 Admixture events detected and dated with f_3, f_4 and ALDER tests

While the Kankanaey exhibit a homogeneous ancestry profile when ADMIXTURE is applied, the method by itself is not a formal test for the absence of admixture. The f_3 method (Reich et al., 2009) considers the correlations of allele frequencies between groups (see section 1.3.3) to yield such a statistic, contingent on the magnitude of admixture fractions and other parameters (see section 4.3.2) . It shows that the Kankanaey cannot be modelled as any kind of mixture of pairs from 45 populations (Appendix B.8).

The notion that they are potentially a good proxy for the source group of the Austronesian expansion is supported by the ancestry composition of the Bajo, Filipino and Malays according to the f_3. The statistic describes their ancestry as being consistent with a mixture of the Kankanaey with either Papuans and/or an older local Asian substrate. The latter is here called

"Austroasiatic" even though populations like the Javanese, where it is the majority component, are today culturally and linguistically Austronesian (see section 2.3.3). The attempt to date these admixture events using ALDER (Loh et al., 2013) highlighted a clear admixture pattern between a "Kankanaey-like" people and earlier substrates, dated to at least 2.2 kya in the Bajo (Table 2.3). The emerging picture from f_3,ALDER and ADMIXTURE seems to be compatible with a scenario of local "Austroasiatic" and Papuan components influenced by the incoming Austronesian (brown k6, Figures 2.2-2.3) wave 4-3 kya which originated from a population living in Taiwan and, perhaps, in the North Philippines (Lipson et al., 2014).

f_3 together with ALDER was also used to further contextualise the potential South Asian connections of some SEA groups. Both statistics (Table 2.3) suggest the presence of variable degrees of South Asian-related ancestry in the MSEA and ISEA populations (Bajo, Burmese, Filipino and Malay). Assuming a generation time of 30 years (Fenner, 2005) the earliest possible midpoint of the South Asian admixture is estimated at 2.4 kya. The overall proportion of South Asian ancestry was further estimated by applying the f_4 statistic (Moorjani et al., 2011) (Table 2.4) according to the tree presented in Figure 2.6.

The estimated values were 24.9% for the Burmese, 8.3% for the Malays and 5.3% for the Bajo. One limitation of this approach is its dependence on shared genetic drift. As the Papuans and South Indians have a similar position in the phylogenetic tree relative to the other groups, Papuan ancestry could be mistaken as South Indian. This has probably no effect in the Burmese and Malay, who do not show Papuan admixture as inferred from ADMIXTURE (Figure 2.2A, Figure 2.3) but could contribute to the South Indian ancestry detected in the Bajo.

Table 2.3: ALDER admixture dates of newly typed populations. Admixture dates for combinations of Javanese (ID-JA), Kankanaey, South Asians and Papuans were reported only when the f_3 statistic yielded significant z-scores (Z <= -2). Tests involving ID-JA as source population were run on 12k SNPs, while the remaining tests were run on the higher resolution 300k SNPs dataset. Standard errors (SE) estimated from a jackknifing procedure over the 22 autosomes are given. "na" denotes combinations of populations where a one-reference and/or a two-reference ALDER exponential decay could not be fit to the LD patterns, i.e. the latter do not resemble patterns expected under admixture LD.

Recipient	Source1	Source2	f_3 z-score	ALDER Date (generations)+/SE	years+/SE
Filipino	Kankanaey	ID-JA	-4.1	35+/-17	1050+/-510
Malay	Kankanaey	ID-JA	-2.8	Na	Na
Bajo	Papuan	ID-JA	-14.1	61+/-10	1830+/-300
Malay	Papuan	ID-JA	-7.4	12+/-7	360+/-210
Burmese	South Indian	ID-JA	-13.1	49+/-5	1470+/-150
Malay	South Indian	ID-JA	-11.6	36+/-13	1080+/-390
Bajo	Papuan	Kankanaey	-18.5	62+/-10	1860+/-300

Recipient	Source1	Source2	f_3 z-score	ALDER Date (generations)+/SE	years+/SE
Burmese	Papuan	Kankanaey	-2.2	52+/-4	1560+/-120
Filipino	Papuan	Kankanaey	-6.5	Na	Na
Malay	Papuan	Kankanaey	-6.3	58+/-10	1740+/-300
Bajo	South Indian	Kankanaey	-4.8	66+/-14	1980+/-420
Burmese	South Indian	Kankanaey	-14.0	53+/-6	1590+/-180
Filipino	South Indian	Kankanaey	-10.4	Na	Na
Malay	South Indian	Kankanaey	-10.8	45+/-12	1350+/-360

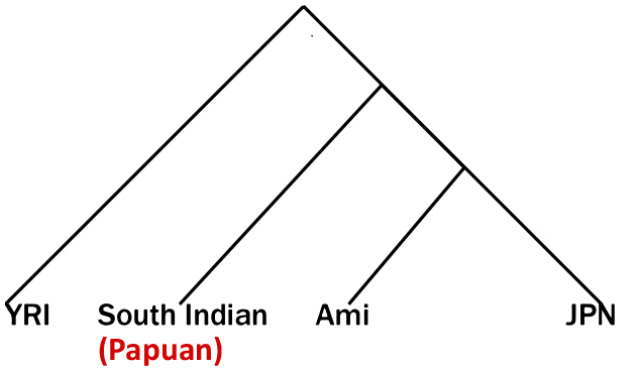


Figure 2.6: Proposed phylogenetic relationship between populations analysed with the f_4 test.

True Indian ancestry in this population still seems conceivable given the presence of South Asian lineages in uniparental marker analyses (Kusuma et al., 2015). Taken together, these analyses indicate a South-Asian related component in the genetic make-up of at least some SEA groups which entered their gene pool as early as 2.4 kya, being supported by ADMIXTURE, f_3 and f_4 analyses for the Burmese and the Malay and by f-statistics for the Bajo (f_3, f_4) and the lowland Filipinos (f_3).

Table 2.4: Proportion of South-Indian related ancestry inferred using the f_4 ratio statistic.

Population	South-Indian-related ancestry
Burmese	~24.90%
Malay	~8.30%
Bajo	~5.30%
Vietnamese	~0% (f_4 ratio -0.060)
Lebbo	~0% (f_4 ratio of -0.117)
Filipino	~0% (f_4 ratio of -0.143)
Dusun	~0% (f_4 ratio -0.180)
Murut	~0% (f_4 ratio of -0.208)
Kankanaey	~0% (f_4 ratio -0.324)

2.2.5 Inference of tree topologies and mixtures with TreeMix

TreeMix (Pickrell and Pritchard, 2012) was applied to estimate admixture in context of an explicitly reconstructed tree of the SEA populations. Four migration events were incorporated into the inferred tree as the amount of variation of allele frequency patterns explained reached a first local maximum at this stage (Appendix B.7B). The general relationships outlined by the TreeMix graph (Figure 2.7) mostly recapitulate the population splits inferred from other global datasets (Li et al., 2008; Pickrell and Pritchard, 2012; Mallick et al., 2016).

The nine novel SEA groups form a cluster with the Han Chinese and ISEA groups from the Pan-Asia panel. The affinity of the Kankanaey to the Taiwanese aboriginals is supported by them forming a distinct subclade in this cluster. Of note are the Philippine Negrito groups whose ancestry to varying degrees consists of a very distinct component, which branches off and here clusters with the Papuans, and sources closer to other ISEA groups (see also Migliano et al., 2013).

Another gene flow event already highlighted by the ADMIXTURE analyses (Figures 2.2-2.4) is from a source close to the basal split of East and West Eurasians into the Burmese. As inferred with other approaches it is most likely South Asian. Finally, the Papuan component in the Bajo could be confirmed. It should be noted that the relatively low marker density (n_{SNPs} = 9,808), due to the reduced overlap of several datasets, limits the power of TreeMix. Furthermore, for datasets with complex population relationships there are often multiple graphs which are equally compatible with the data (Pickrell and Pritchard, 2012).

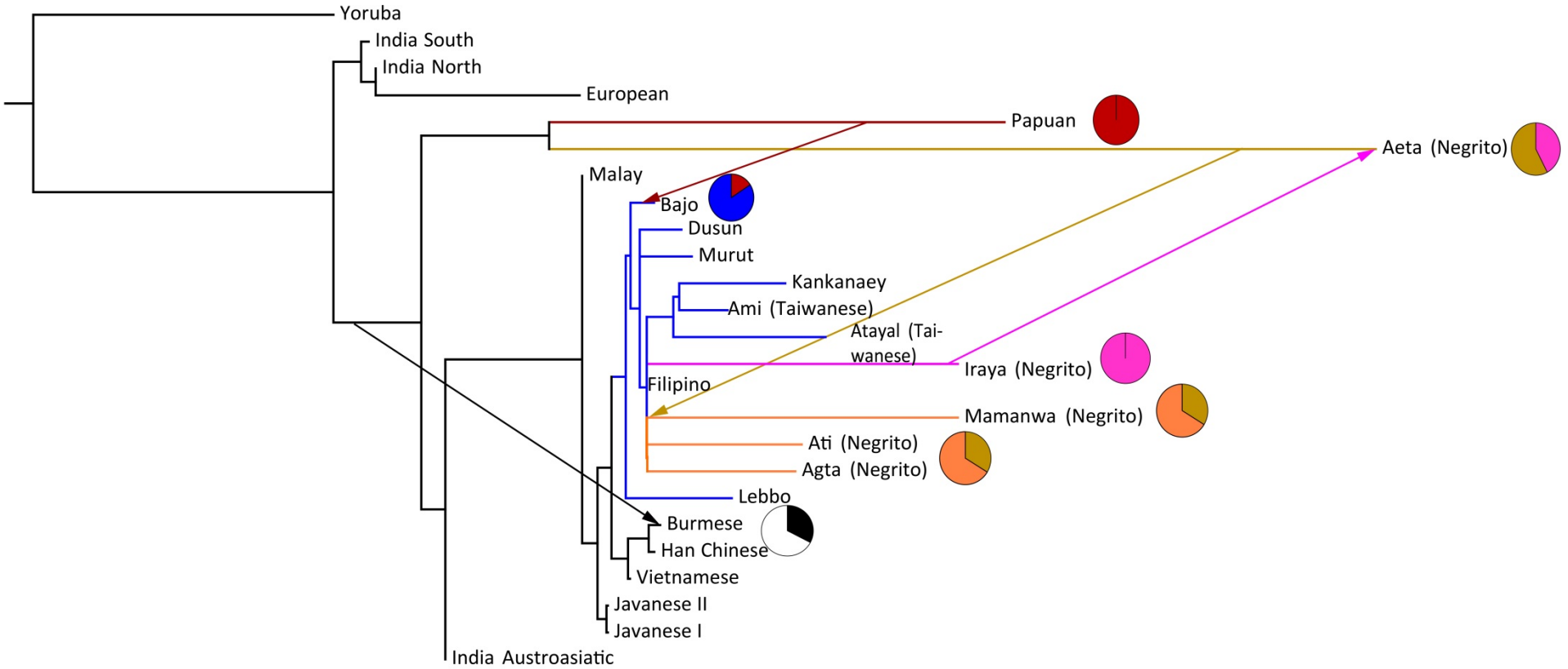


Figure 2.7: TreeMix analysis involving 25 populations. Four migration edges (indicated by arrows) were allowed. The pie charts describe the admixture proportions for populations resulting from inferred migration events with the branches who are the closest matches to the putative admixture sources coloured accordingly. The Kankanaey cluster most closely with the Taiwanese aboriginals and that there is no detectable gene flow from other populations to them.

2.2.6 Mitochondrial DNA lineages in the Kankanaey Igorot

The affinities of the Kankanaey with the Ami and Atayal from Taiwan and their potential role as a good proxy for the Austronesian expansion are further highlighted when analysing uniparental markers. The eight available Kankanaey mtDNA sequences (Appendix B.9) exhibit lineages (B4a1a;M7b1a2a1) which are typical markers of Malayo-Polynesian speaking populations (Trejaut et al., 2005; Soares et al., 2011). Taken together, the evidence from this approach and ADMIXTURE, f_3 and TreeMix on autosomal data suggests that the Kankanaey could potentially represent an unadmixed remnant population close to the source that may have given rise to the Austronesian expansion.

2.2.7 Selection signal sharing compared to overall genomic distance

As an additional tool to explore relationships among populations patterns of haplotype homozygosity and allelic differentiation were examined using the test statistics IHS (Voight et al., 2006), XP-EHH (Sabeti et al., 2007), and PBS (Yi et al., 2010) (windows with population-

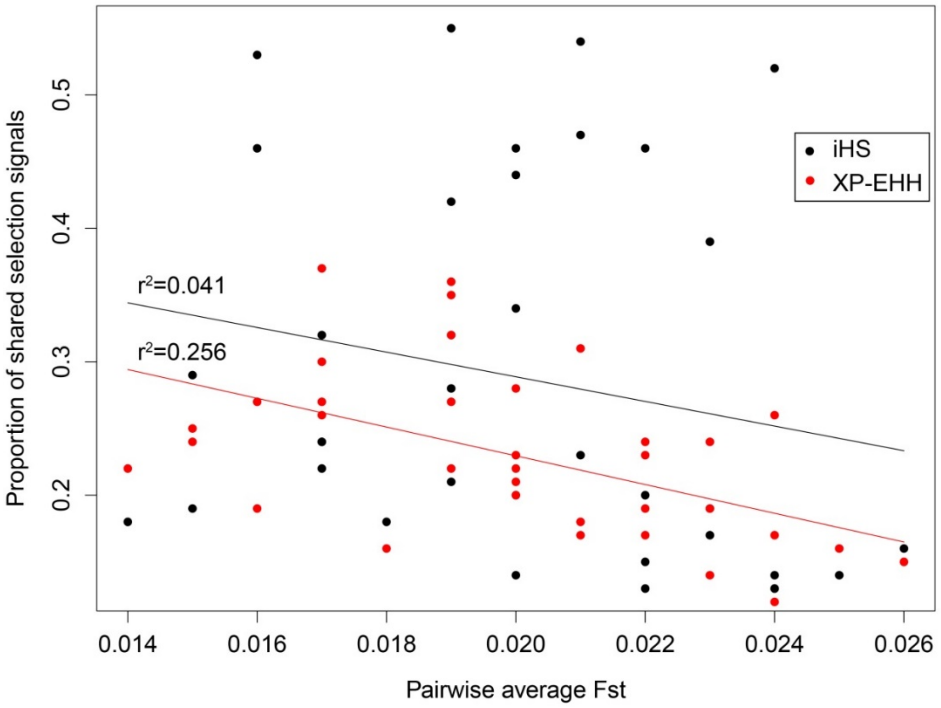


Figure 2.8: Relationship between F_{ST} and selection signal sharing. The proportion of 200-kb windows that were detected as 1% selection outliers in one population and also found in the 5% of signals in the other population is shown for the iHS (red dots) and the XP-EHH (black dots). Linear regression lines were fitted to the data, the respective coefficients of determination are shown.

specific top hits for all tests can be found in Appendices B.10A-B.10C). The focus here is not on specific putative selection signals, but rather on signal sharing, which is defined as follows for the haplotype homozygosity statistics iHS and XP-EHH. A shared signal is one that falls in the 1% selection outliers of the test for one population and is also found in the top 5% of signals in other population(s). For the iHS the amount of signal sharing between two groups is not significantly correlated (r = -0.2033, p = 0.2344) to overall genetic similarity as expressed by F_{ST} (Figure 2.8, underlying data in Appendix B.11). However, the MSEA groups and the Han Chinese (included as a reference) who share a considerable proportion of East-Asian ancestry (Figure 2.2A, Figure 2.3) also show a great affinity to each other regarding haplotype homozygosity patterns (Appendix B.11A). In ISEA, those groups with at least three significant ancestry components at K = 9 (Bajo, Filipino, Malay, Figure 2.2A) exhibit more signal sharing. In contrast, Kankanaey, Lebbo and Murut show reduced sharing with all other populations, which perhaps highlights phenomena of deep population splits and separate demographic histories in recent times when the haplotype homozygosities have accumulated. However, these inferences are highly dependent upon the approach utilised.

A different picture presents itself for the XP-EHH, which considers both haplotype homozygosity and allelic differentiation, with the Han Chinese used as outgroup. The average fraction of signal sharing declines from 0.31 to 0.22, while the correlation with F_{ST} becomes significant (r = -0.5071, p =0.0016) (Figure 2.8).

This is probably because signals connected to shared ancestry with East Asians are excluded. As a result, the Burmese, who exhibit a large fraction of the k2 East Asian-related component (Figure 2.2A, Figure 2.3) become an outlier with respect to the uniqueness of their top 1% XP-EHH signals, only 15% of which are shared with other groups on average (Appendix B.11B).

2.3 Discussion

2.3.1 New insights into the dynamics of the Austronesian expansion

In this chapter the Kankanaey from the northern Philippines were identified as the population harbouring the highest proportion of the Austronesian genomic component, even higher than the ones detectable in modern aboriginal Taiwanese (Figure 2.2A, Figure 2.3). This conclusion rests on evidence from several independent analyses including ADMIXTURE, f_3, rOH, reconstructed N_e, TreeMix and uniparental markers.

The Kankanaey belong to the broader group of populations collectively known as Igorot (Appendix B.1). Various studies exist on the Kankanaey language (Allen, 2014) and customs (Kohnen, 1986), although works on their prehistory are lacking. Genetic data from 30 Kankanaey-speakers were included in a study of mtDNA haplotype diversity in the Philippines (Delfin et al., 2014). They were shown to share many lineages with two other Igorot groups (Ibaloi and Ifugao) from Northern Luzon. These results are broadly consistent with the uniparental data presented here (Appendix B.9), where the Kankanaey show haplotypes also found in Taiwanese aboriginals (Ko et al., 2014) and generally associated with the Austronesian expansion (Trejaut et al., 2005; Soares et al., 2011). Therefore, it can be concluded that the Kankanaey are either the best-preserved descendants of the original source population of the Austronesian expansion, or a case of total replacement that followed it. The dominant model suggests a southward diffusion of Austronesians from Taiwan around 4 kya, which impacted the Philippines, the north of Borneo and Sulawesi between 3.8-3.6 kya, and later spread into the Pacific (Bellwood, 2007). Even if the modality of this expansion is complex and still debated (Bulbeck, 2008), the location of the Kankanaey in the northern Philippines, close to Taiwan, suggests that they may be considered as one of the least admixed living groups tracing their ancestry from the source populations of the Austronesian expansion.

This conclusion has been strengthened by independent analyses since the publication of the paper this chapter is based on (published online on 15/06/2016). Genomic data from some Kankanaey were included in the analyses of aDNA from Lapita individuals by Skoglund et al. (2016) (published online on 03/10/2016) (see section 1.5). Using TreeMix and qpGraph (Patterson et al., 2012), which assesses the fit of different admixture graphs to allele frequency correlation patterns, the Kankanaey were identified as the closest outgroup to the ancient First Remote Oceanians among all analysed ISEA populations.

As a note of caution, it should be added that while the Kankanaey appear to be a genetic relict population this does not imply cultural isolation. Recent archaeological evidence from the linguistically and genetically (Delfin et al., 2014) closely related Ifugao indicates economic intensification and increases in political complexity coinciding with the start of the Spanish presence in the Philippines (Acabado, 2017) (Appendix B.1). It can therefore be speculated that the Kankanaey were at least indirectly impacted by colonialism and underwent somewhat similar developments to their geographic neighbours.

2.3.2 Indian impact in SEA

A minor South Asian component was detected by the ADMIXTURE analyses of MSEA and ISEA populations (green k4, Figures 2.2A-2.3; green, Figure 2.4). It was further confirmed by f_3, f_4 and ALDER results and dated to have entered SEA from 2.4 kya (Table 2.3). While this component is more widespread at lower value of K (Figure 2.2B, Figure 2.3), at the best K = 9 (Figure 2.2A) the evidence is strongest for the Burmese and the Malay and somewhat weaker for Bajo and Filipinos, where it is limited to f_3/f_4 (Tables 2.3-2.4). The Refined IBD method (Browning and Browning, 2013) further supports these outcomes. Each SEA population shares about twice as many IBD segments with South Indians (excluding the Austro-Asiatic speakers) as with Europeans. This indicates more recent contacts of the SEA groups with the former.

It is important to explore how these results relate to the linguistic and archaeological evidence, attesting a continuous presence of South Asian cultures in SEA since 2.5 kya (Gonda, 1973; Manguin et al., 2011; Bellina et al., 2014; Calo, 2014). In most SEA populations analysed here the Indian component is absent or below the scale of detectability. Firstly, it is most likely that the "carriers" of South Asian culture were traders, artisans (Bellina et al., 2014) and at a later date, religious scholars (Brahmins) who were influential as advisers to Southeast Asian rulers. Some of these might have been locals educated in India who brought home Sanskrit texts and Brahmanic rituals (Bronkhorst, 2011). So, this rather small group would not have left a major genetic signature. Secondly, the epigraphic record and evidence from monumental archaeology from 1.5-1 kya attest that the Indian presence is biased towards courts and generally higher social strata, which can lead us to overestimate the impact on the majority of the population (Bronkhorst, 2011). More generally speaking there are a wide range of scenarios relating to the spread of cultural elements and gene flow and the patterns of this relationship are highly complex to model, e.g. for the Neolithisation in Europe (Fort, 2015). So, except for the

Burmese, who are also geographically very close to the Indian subcontinent, the evidence points to rather minor Indian gene flow. This is a notable contrast to the documented cultural influence which, however, temporally overlaps with the dates obtained for admixture with ALDER (Table 2.2). A related signal of minor South Asian gene flow was also detected in some other populations across ISEA (Karafet et al., 2005, 2010; HUGO Pan-Asian SNP Consortium et al., 2009; Chaubey and Endicott, 2013; Kusuma et al., 2015).

Taken together these findings suggest SEA as a potential waypoint for the South Asian migration into Australasia detected by Pugach et al. (2013). However, Pugach et al. themselves disputed a scenario where Indian ancestry would reach Australia indirectly via ISEA. Furthermore, the date obtained using ALDER (2.4 kya) is at least 1500 years after the proposed South Indian migration into Australasia. A preliminary conclusion would envisage the SEA and Australasia migrations as two separate events. Besides the fact that the dating methods differed between this work and Pugach et al. (they used a method based on wavelet transform analysis) at least two caveats can be brought up to reconcile this fragmented scenario. Given the evidence presented here, it seems reasonable to assume a constant gene flow from South Asia into SEA via land, with Australasia being only a sporadic endpoint. In this case the 4 kya estimate provided by Pugach et al. would be a point estimate of the sparse arrival into Australasia, while the ALDER estimate given here should be interpreted as the midpoint (Loh et al., 2013) of such a flow between 4 kya and more recent times. Secondly, given the concordance of lines of evidence from linguistics and archaeology for a South Asian presence in SEA around 2.5 kya, it is possible to imagine a particularly intense corresponding gene flow during that time further biasing the ALDER estimate toward this period. It should be noted that a re-analysis of the signal reported by Pugach et al. using a larger panel of genomes from Sahul failed to reproduce it and concluded that it is most plausibly explained by technical artefacts (Bergström, 2018).

2.3.3 Other general findings

The comparison of haplotype-based scans of positive selection revealed that as opposed to earlier studies on a continental level (Metspalu et al., 2011) the correlation between haplotype sharing patterns and genetic distance (F_{ST}) is relatively weak in ISEA (Figure 2.8).

Populations showing more diversity in the admixture plots also exhibit higher levels of shared signals with other groups. Additionally, the sharing patterns proved to be dependent on the kind

of test utilised. Notably, when the XP-EHH, which uses the Han Chinese as outgroup, is applied, all signals shared with East Asians are excluded, causing the Burmese to become an outlier (Appendix B.11B). This potentially reflects haplotype homozygosity signals unique to their South Asian-related ancestry component (Figure 2.2A, Figure 2.3).

The finding of an Austro-Asiatic-related component in ISEA populations first reported by Lipson et al. in 2014 was confirmed. Support for the potential existence of a pre-Austronesian farming culture of Austro-Asiatic origin at least on Borneo, which would have later been assimilated by the incoming Austronesian speakers, comes in fact from multiple sources. These include the presence of Austro-Asiatic loanwords in Bornean Austronesian languages, the transfer of taro cultivation from the mainland and similarities in synchronous archaeological assemblages from Vietnam and Borneo (Blench, 2011). Given its wide spread in linguistically diverse groups in MSEA and ISEA, the explicit association of k5 (Figure 2.2A, Figure 2.3) with a specific language family should be taken with caution. However, it is worthwhile noting that in India this component was specifically found in Munda-speaking populations, whose languages belong to the Austro-Asiatic family. The k5 component could represent an ancestral substrate, which was once distributed widely throughout SEA and was encountered by the Austronesians when they spread from Taiwan. Another possibility is that there was an early split into several subgroups during the Austronesian expansion and that this component belongs to the ancestral make-up of a subgroup of Malayo-Polynesians who expanded into western Indonesia. The latter hypothesis however does not satisfactorily account for the aforementioned Indian affinities.

3. Functional and deleterious variants in worldwide WGS data

The second results chapter of this thesis consists of a survey of the global distribution of functional and/or deleterious variants in a dataset of 382 whole genomes. In this context, potential signals of positive and purifying selection are also examined. A considerable portion of these results were first reported in a paper by a consortium of more than 90 researchers of which I was a member.

Pagani L, Lawson DJ, Jagoda E, Mörseburg A, Eriksson A, Mitt M, Clemente F, ...Kivisild T, Metspalu M. 2016. Genomic analyses inform on migration events during the peopling of Eurasia. Nature 538:238–242.

My main contribution to this work lies in the analyses of exonic variation, purifying selection and some tests for positive selection. I ran the respective analyses and wrote the corresponding sections of the paper and the supplementary materials with support from my collaborators. All analyses described in this chapter were conducted by me unless explicitly stated otherwise.

In recent years, the amount of WGS data available for studies of neutral and functional variation has increased exponentially thanks to large projects focussed on the global patterns of variation such as the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2015) and exonic data from >60,000 samples presented by ExAC (Lek et al., 2016). However, the coverage of the world in these datasets is uneven, with regions such as Siberia and ISEA being underrepresented. Furthermore, many of these studies either provided just exomes or low-coverage data, which do not include non-coding and rare variation respectively.

These gaps have only started to be addressed very recently. One example is the SGDP (Mallick et al., 2016) providing data from 300 individuals spanning a wide geographical range. However, the first publications on this dataset concentrated mainly on reconstructing demographic history.

Most of what we know about the global patterns of functional variation and related phenomena such as potential differences in purifying selection is still based on a relatively small set of reference populations. As these distributions are known to be influenced by demographic

history a more even geographic coverage is essential. A range of questions related to this subject will be investigated using the new dataset of 382 whole genomes described in this chapter. They fall in three broad categories briefly contextualised below.

1) Each human genome carries more than 10,000 missense mutations and 70-180 variants thought to cause a loss of gene function if present in a homozygous state (Lek et al., 2016; The 1000 Genomes Project Consortium, 2015). On average, the former are thought to be mildly and the latter moderately deleterious when appearing *de novo*. In terms of geography, there is a higher number of polymorphic sites from all classes in Africans than in non-Africans; this pattern is reversed for derived homozygous genotypes (Fu et al., 2014b; Lohmueller et al., 2008) However, relatively few fixed differences of missense and nonsense mutations are observed between continental groups (The 1000 Genomes Project Consortium, 2012). While the differences between Africans and all non-Africans are well established, considerably less is known about variation between non-African continental groups. The analyses on this dataset of 382 individuals aim to explore this subject under particular consideration of the following questions. How are different classes of functional variants distributed throughout the genome across worldwide populations and to what extent are they shared? Furthermore, which factors can explain these differences? Are there any particularly unusual patterns in previously understudied groups?

2) The concept of genetic load, defined as the reduction of the evolutionary fitness of a population compared to a hypothetical comparative group with only the fittest genotypes (Haldane, 1937; Crow, 1958), predates modern molecular genetics. There is a long-standing debate on the extent to which human populations vary with regards to this quantity. Early theoretical work demonstrated that the efficiency of selective processes, including purifying selection, relative to drift should increase with higher N_e (Kimura et al., 1963) which could be an important contributing factor to this potential interpopulation variation. The relevance of this is underlined by more recent evidence for deep divergences and long standing differences in N_e among modern human populations (Mallick et al., 2016). Support for significant differences in the strength of selection and corresponding genetic loads comes from multiple studies. As mentioned above, there was a reported excess of homozygous derived deleterious variants in non-Africans and in some studies (Fu et al., 2014) also a higher total of derived deleterious alleles in them. Often these empirical outcomes were supported by simulations exploring the impact of demographic history on the number of deleterious variants carried by an individual.

However, other studies using ratio statistics of particular allelic classes and corresponding simulations arrived at contradictory results (Simons et al., 2014; Do et al., 2015). They found that the mutational burden was unaffected by demography, particularly if it was assumed that all mutations are additive. Several authors have recently attempted a synthesis of the growing literature on the subject (Lohmueller, 2014a; Henn et al., 2015; Gravel, 2016). They have highlighted the two main reasons for these discrepancies: a) conclusions about the genetic load are highly sensitive to underlying assumptions about dominance effects and selection coefficients and b) how the relevant statistics are calculated. Furthermore, there is considerable disagreement about how biologically meaningful these differences in mutational load would be on a population level. Notwithstanding this ongoing debate it is important to gather more empirical evidence to improve our understanding of the forces affecting genetic load. With increasing global population coverage, do earlier statements made about purifying selection based on a few reference groups have to be revised? Furthermore, is there evidence for differential purifying selection on particular gene classes?

3) One possible cause for extreme frequency differences between populations at a particular locus is regionally specific positive selection. Strongly supported selection candidates include polymorphisms in the *DARC/ACKR1*, *SLC24A5* and *EDAR* genes. As already mentioned, there are concerns about the power of selection tests and their ability to distinguish patterns resulting from neutral and adaptive processes. Some of these are less relevant with WGS as opposed to the SNP data discussed in the previous chapter, e.g. it is less likely that true causal variants are missed. The dataset presented here represents a considerable expansion of geographical coverage compared to the 1000 Genomes project for which highly differentiated loci were examined for its first phase. It also provides another opportunity to re-examine the "*less is more*" hypothesis (Olson, 1999) according to which a small fraction of nonsense mutations should be under positive selection due to their advantageous effects in particular environments which could lead to strong allelic differentiation. Keeping the relevant caveats in mind, can we improve our understanding of the potential targets of positive selection by applying variant-based approaches on high coverage WGS data from a diverse panel of worldwide populations?

3.1 Material and Methods

3.1.1 The EGDP dataset and demographic analyses

The dataset analysed in this chapter is a subset (n = 382) of a larger panel of 483 genomes sequenced with Complete Genomics technology and analysed by the EGDP consortium Pagani et al. (2016). Before the criteria for the selection of this subset and the respective analyses are described it is important to outline a) the general pipeline for the creation of VCF files and b) some demographic analyses. Outcomes from a) and b) were later used as either a starting point or as supporting evidence to investigate patterns of functional variation.

379 of the 483 genomes were first published by Pagani et al. (2016), whereas the other samples were taken from previously published sources (Drmanac et al., 2010; Lachance et al., 2012; Clemente et al., 2014). In addition, publicly available Personal Genomes Project (PGP, http://www.personalgenomes.org/) data, available at the time of the study, was used, bringing the total number of high coverage genomes used to 730. Appendix C.1 is a table containing information on all genomes including the collaborator/institution that collected the samples for the EGDP consortium. All samples were sequenced, mapped and called by Complete Genomics (Mountain View, California, USA) using Complete Genomics software versions 1.5, 2.0, 2.2 and 2.4.

Samples for the genomes reported by Pagani et al. (2016) were collected either from blood or saliva with informed consent and sequenced to >40× and >80× coverage, respectively.

Before further describing the initial filtering and phasing of the data it should be noted that, primarily to increase the power for phasing, the full set of 730 genomes was used for these upstream filtering procedures. Most PGP genomes, however, were discarded for downstream analyses due to the lack of specific information on their ethnic affiliation and origin.

Data filtering and phasing for creation of VCF files

The starting point for these pre-processing steps were the VAR files provided by Complete Genomics (details on the output of the Complete Genomics pipeline can be found in www.completegenomics.com/documents/DataFileFormats_Standard_Pipeline_2.2.pdf).
These files contain all short genomic variants (SNPs in a narrow sense, i.e. single base-pair

substitutions see section 1.1.1, multiple base-pair substitutions, small indels) in each genome mapped against the human reference genome hg19 build 37.

For this purpose, the 730 VAR files were combined into chromosome-based variation matrices with each row representing a variant and each column representing an individual. This was done with the CGA-tools modules *listvariants* and *testvariants* with presence, absence and no-call for each variant being coded as 1, 0 and N.

These matrices were then further refined by only keeping SNPs of high calling quality using the *Varfilter* module from CGA-tools with the VAF varscore (which is a confidence score obtained in the CG pipeline by a maximum likelihood scoring model) of 40 and 20 for heterozygous and homozygous variants respectively. The output of this step was transformed into PLINK format (Purcell et al., 2007). Three further filters were then applied on these files.

Firstly, for each polymorphic SNP in the PLINK files for the 730 individuals it was tested whether the allele frequencies at this site fulfil the conditions of the Hardy-Weinberg Equilibrium (HWE) using the Wigginton et al. (2005) method that controls for type 1 error. Furthermore, a one-tailed exact test for the detection of an excess of heterozygotes was performed and SNPs showing a $p_{high} < 0.0001$, i.e. sites where the count of heterozygotes is so high that the probability of observing it under the assumption of HWE is below 0.0001 were discarded as likely sequencing artefacts and genotyping errors. In total 19,458 autosomal SNPs were excluded reducing the dataset to 46,881,865 variants. Following the test for HWE all SNPs which were not biallelic and those that had more than 5% no-call/low quality calls were removed. The files were converted from CG format to PLINK as described above by Mario Mitt (University of Tartu), Dr Lauri Saag (University of Tartu) and Dr Alexia Cardona (University of Cambridge).

SHAPEIT2 (Delaneau et al., 2013b) was applied to phase the filtered PLINK files. To speed up the calculations, singletons were removed before phasing. Subsequently singletons were re-introduced into the phased data, randomly picking one of the two available haplotypes for assigning the derived allele at each singleton position. Chromosome 20 was phased twice to estimate phasing accuracy, yielding an average per individual switch error rate of 2-4%. The file conversions and the phasing were performed by Mario Mitt and Dr Luca Pagani.

Definition of the Diversity Set for demographic analyses

For different downstream analyses subsets of relevant samples were defined. The first subset was designated as the Diversity Set, it was used to reconstruct past demographic events.

For this Diversity Set only samples with location information and self-reported unmixed ethnicity were considered (Appendix C.1). Therefore only 17 of the samples from the Personal Genomes Project were kept. Furthermore, close relatives were filtered by removing one sample from all pairs which were related to second degree or closer. The degree of relatedness was calculated using the KING (Manichaikul et al., 2010) software package.

To avoid biases in the results the sample sizes of overrepresented populations in the dataset were reduced. This had no effect on the newly sequenced data as the overall strategy was to include a low number (3-4) of individuals from a wide range of populations. Preferably, samples carrying the "Unusual metrics" flag as indicated by the CG analysis pipeline were excluded.

The final Diversity Set, which also formed the basis for the rare variant analyses presented in chapter 4 of this thesis, contained 447 samples (Appendix C.1). For further downstream analyses groupings and sub-lists were formed depending on what was required for the particular task.

Relevant demographic analyses

ChromoPainter (CP) was used (Dr Alexia Cardona, Cambridge, and Dr Daniel Lawson, University of Bristol) in combination with fineSTRUCTURE (Lawson and Falush, 2012; Lawson et al., 2012) (see section 1.3.3) to investigate structure among the different populations. The CP linked model was applied on all haplotypes extracted from the phased data created for the individuals from the Diversity Set. In the context of this thesis the ChromoPainter/fineSTRUCTURE is of importance for two follow-up analyses. Firstly, they inform the definition of larger groups for the analyses of functional variation and selection scans. Secondly, they provide a useful summary of the coancestry patterns across the whole variant spectrum which will be contrasted with the rare variant patterns in chapter 4.

Temporal dynamics of N_e over population histories were inferred using MSMC (Schiffels and Durbin, 2014). This approach was applied to two partially overlapping sets of populations. The first consisted of 22 groups chosen as representatives of the sampled human diversity. For these N_e over time was reconstructed based one or two phased genomes. To increase resolution at

more recent time scales MSMC was furthermore applied to all populations (n = 42) for which four genomes were available (detailed information on the composition of these sets can be found in Appendix C.2 and the raw data are given in Appendix C.3). The relevant parameters were set to a mutation rate of $\mu = 1.25*10^{-8}$ bp^{-1} per generation and a generation time of 30 years (these parameter values are identical to those used by Schiffels and Durbin, 2014). The MSMC analyses were conducted by Dr Luca Pagani.

3.1.2 Definition of subsets for downstream analyses

Definition of the Selection Set and the Variant-Based Analysis Set

The coancestry matrices resulting from the CP approach formed the basis for the larger groupings chosen to analyse patterns of natural selection. A minimum sample size > 15 was set

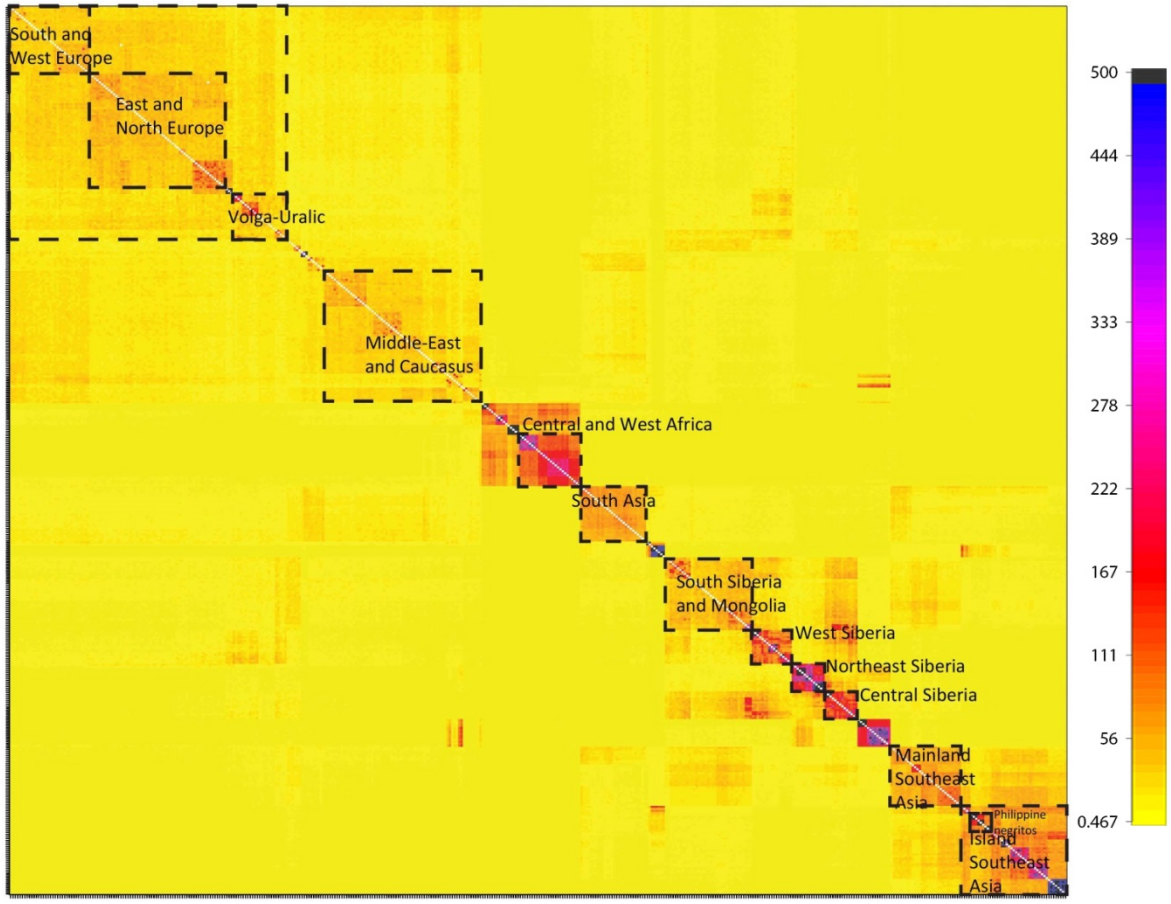


Figure 3.1: A fineSTRUCTURE-based coancestry matrix was used to define twelve groups of populations for the downstream selection scans. The quantity used to measure similarity is total length of genomic sharing in cM. The respective groups are highlighted in the plot by boxes with broken line edges. The number of individuals in each group is reported in Appendix C.4. Please note that the unlabelled cluster situated between Central Siberians and Mainland Southeast Asians at the bottom right of the figure consists mainly of Native Americans from the Andes.

as a threshold because prior research (Pickrell et al., 2009) indicates that the statistical power of haplotype homozygosity-based selection tests rapidly decreases with lower sample sizes once a threshold of 10-20 individuals (depending on the exact method applied) is reached. A very similar set of individuals was chosen for the analyses of functional variation (for the exact differences see below) to ensure consistency for publication purposes. The selection analyses based on haplotype patterns were run by Dr Tiago Antão, Dr Georgi Hudjashov (Massey University), Evelyn Jagoda (University of Cambridge, present affiliation: Harvard University), Dr Charlotte Inchley (University of Cambridge, present affiliation: University of Oxford), Sarah Kaewert (University of Cambridge, present affiliation: University of Colorado) and Dr Florian Clemente (University of Cambridge, present affiliation: University of Lausanne). They are not the subject of this thesis; however, the selection groups are described here as they also were the groupings chosen for other population-differentiation based selection tests (di, DIND, see section 3.1.4) that are included as supporting evidence.

Although the coancestry matrices produced by CP and the derived trees inferred by fineSTRUCTURE capture the population structure expected from previously analysed global datasets well (data not shown) they are very sensitive to balanced sample sizes and not based on an explicit underlying model of evolutionary history. A matrix containing the total length of genomic sequence shared between the pairs of donors and recipients was used to determine the groupings for the selection scans (Figure 3.1). For the selection scans in total 12 groups (all listed in Table 3.1, excluding Australo-Papuans and South Americans due to sample size requirements) consisting of 369 individuals were formed and designated as the Selection Set (Appendix C.4, Figure 3.1). While the variant-based analyses are still sensitive to sample size as they rely on allele frequency patterns, additional comparisons between global groups can be made if this factor is considered as a caveat. Therefore, for the variant based analyses two groups with smaller sample sizes from outside (mainland) Eurasia were added. The first was a summary group of Oceanians (here used interchangeably with the term “Australo-Papuans”) (n = 8) consisting of individuals from Papua New Guinea and Aboriginal Australians. The second was a specific subset of Native American diversity, a very regionally distinct cluster of populations from the Andes (n = 13). Eight individuals that were part of the Selection Subset were removed from the Variant-Based Analysis Subset. These included five African-Americans, one Hungarian, one Pole and one Mari. The final Variant-Based Analysis Set contained 382 individuals from 14 (sub)-continental macro-groups (Table 3.1).

Table 3.1: Macro-groups used for i) the Variant Based Analysis Set on which patterns of functional and deleterious variation were investigated and ii) the Diversity Set on which patterns of rare variation were investigated Appendix C.1 contains a more detailed description of the samples. Please note that "Assyrians" is a self-designated name of Aramaic-speaking Syriac Christians from the Middle East.

Macro-group	Abbreviation	N(functional)	N(rare)	In both datasets	Functional analyses only	Rare variant analyses only
(Central and West) Africa	Afr	21	39	4 Luhya, 8 Pygmies, 9 Yoruba		5 African-Americans, 5 Hadza, 3 Maasai, 5 Sandawe
Middle East (and Caucasus)	MiE	26	59	6 Armenians, 7 Arabs, 3 Assyrians, 3 Druze, 4 Iranians, 2 Jordanians, 1 Lebanese		3 Abkhazians, 3 Avars, 3 Azerbaijanis, 3 Balkars, 3 Circassians, 2 Georgians, 4 Kabardians, 3 Kumyks, 4 Lezgins, 2 North Ossetians, 3 Tabasarans
South and West Europe	WEu	31	36	3 Albanians, 2 British, 9 CEPH Europeans, 4 Croats, 1 French, 4 Germans, 2 Hungarians, 4 Italians, 2 Moldavians		1 Hungarian,1 Portuguese, 3 Roma
East and North Europe	EEu	53	63	4 Belarussians, 3 Cossacks, 6 Estonians, 3 Finns, 3 Ingrians, 3 Karelians, 3 Latvians, 3 Lithuanians, 1 Mishar Tatar, 4 Poles, 7 Russians, 2 Swedes, 7 Ukrainians, 4 Vepsians		1 Cossack, 2 Komis, 3 Mordvins, 1 Pole, 3 Saami
Volga-Uralic	Vol	21	22	5 Bashkirs, 3 Chuvashes, 3 Maris, 6 Tatars, 4 Udmurts		1 Mari
South Asia	SoA	28	29	2 Andhra Pradesh, 3 Bangladeshi, 1 Bengali,4 Gujarati, 4 Jharkhand, 1 Kerala (Malayalam), 2 Madhya Pradesh, 1 Nepali Brahmin, 1 Orissa, 1 Punjabi, 1 Rajasthani, 7 Uttar Pradesh		1 Tamang

Macro-group	Abbreviation	N(functional)	N(rare)	In both datasets	Functional analyses only	Rare variant analyses only
Central Asia	CeA	-	24	not part of Variant-Based Analysis Set		2 Ishkashim, 3 Kazakhs, 7 Kyrgyz, 2 Rushan-Vanch, 1 Shugnan, 1 Tajik, 3 Turkmens, 1 Uyghur, 3 Uzbeks, 1 Yaghnobi
West Siberia	WSi	17	18	3 Forest Nenets, 3 Kets, 3 Khantys, 3 Mansis, 3 Selkups, 3 Tundra Nenets		1 Mansi
South Siberia and Mongolia	SSi	34	23	6 Altaians, 6 Buryats, 6 Mongolians, 2 Shors, 3 Tuvins	11 Buryats	
Central Siberia	CSi	31	17	6 Evenks, 3 Evens, 2 Nganasans, 6 Yakuts	10 Evenks, 2 Evens, 2 Yakuts	
Northeast Siberia	NSi	25	14	5 Chukchi, 4 Eskimos, 5 Koryaks	11 Koryaks	
Mainland East and Southeast Asia	SeM	29	29	8 Burmese, 7 Chinese, 4 Japanese, 10 Vietnamese		
Island Southeast Asia	SeI	45	45	3 Aeta, 3 Agta, 3 Batak (Philippine Negritos, Batak mentioned in chapter 2 are Indonesian Austronesians) 4 Bajo, 8 Dusun, 8 Igorot, 4 Lebbo, 2 Luzon, 8 Murut, 2 Visayans		
Americas	Ame	13	21	5 Calchaquíes, 4 Colla, 4 Wichi		5 Mexicans, 3 Puerto Ricans
Oceania	Oce	8	8	3 Koinanbe, 3 Kosipe, 2 Aboriginal Australians		

3.1.3 Annotation of functional and deleterious variants

The first step in analysing genomic data for functional variation was to compare the composite loads of different classes of functional variants globally. This task was performed using two main approaches which complement each other: QIAGEN's Ingenuity® Variant Analysis™ (IVA) (www.qiagen.com/ingenuity) and the web application and Perl scripts provided as part of the Ensembl Variant Effect Predictor (VEP) (McLaren et al., 2016) toolset.

Unless stated explicitly otherwise, all data analysis steps described in this subchapter and section 3.1.4 and 3.1.5 were run using R (R Core Team, 2017) and UNIX shell scripts custom-written for the respective purposes by the author.

Extraction and analysis of functional variation using IVA

Functional annotations of the exonic sites were explored firstly with IVA version 3.0.20140417 (as of 19/05/2014) with the masterVarBeta files from the CG pipeline as the starting input. The potential translation impacts of a particular variant were inferred from its location relative to the set of transcripts provided by the RefSeq (O'Leary et al., 2016) release 63 database.

All samples were pooled in one analysis to reduce the effect of biases particular to certain groups. The first step was a pre-filtering option to narrow down the analysis to exonic regions and 20 bp upstream and downstream from those regions. This filter was based on a bed file describing exonic boundaries as given in Ensembl release 75 (27/02/2014) which was provided by Dr Alexia Cardona. The cut-offs chosen were call quality (this quantity is equivalent to the VAF varscore mentioned in section 3.1.1 as the mastervarBeta files are derived from the VAR files) of 30 and a read depth of 20×. Furthermore, the 0.2% most exonically variable 100-base windows in healthy public genomes were excluded. Also, only biallelic autosomal sites were considered for further analyses.

To ensure consistency with the results of selection scans run by collaborators, additional filtering steps were taken. Firstly, a more sophisticated coverage filter was applied. After the average coverage for all windows was calculated and the immediate removal of windows with a coverage of zero, any windows which had a coverage value above or below two standard deviations (SD) from the genome-wide average were also excluded. This left 13,127 200-kb windows with a coverage between 32× and 72×. The windows derived from this filter were provided by Dr Florian Clemente.

Secondly, only those SNP sites which had passed the general quality filters (see section 3.1.1) and were included in the VCF files (for the original larger panel of n = 730 as described above) were kept. This reduced the number of sites stepwise from initially 334,857 to 294,399 (Table 3.2). The final number of exonic variants from IVA was 294,360, as 39 sites that were fixed for the reference allele had passed the quality filters (due to an originally larger sample size of 495 in the IVA analysis).

Table 3.2: Filtering scheme for variant-based analyses. The abbreviations used are as follows: syn - synonymous, mis - missense, non - nonsense. Among the excluded sites missense variants are overrepresented relative to their proportion in the overall dataset.

	Included				Excluded			
	all	**syn**	**mis**	**non**	**all**	**syn**	**mis**	**non**
Autosomal biallelic	334,857	128,789	182,351	2,896	-	-	-	-
VCF filters	296,407	116,830	158,858	2,381	38,450	11,959	23,493	515
Coverage filter	294,399	116,096	157,694	2,357	2,008	734	1,164	24

The functional annotations provided by IVA were complemented by other databases. Information on the ancestral states of variants was imported based on an alignment of genomic data from six primate species (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/ancestral_alignments/). This allowed inferences about the polarity of 266,993 exonic variants for which high-confidence ancestral calls were available. Also, the most likely calls for Neanderthal (Prüfer et al., 2013) and Denisovan (Meyer et al., 2012) alleles were retrieved from online repositories. The bulk data on all exonic variants from IVA including the additional annotations described here are presented as Appendix C.5.

Furthermore, information on NMD for the stop-gain variants was retrieved from Ensembl release 87 (08/12/2016) using the Ensembl VEP Perl scripts (McLaren et al., 2016). Before any further annotations the total number of stop-gain variants analysed was reduced from 2,357 to 2,096 as sites without high confidence calls indicating that the reference matches the ancestral allele were removed. Out of these 2,096 ca. 30% (n = 628) were part of at least one transcript affected by NMD. For the filtering steps described below and the annotation of the data it should be stressed that a) a variant can be part of many different transcription contexts and b) multiple nonsense variants can be present in an individual in a transcript which will potentially undergo NMD.

However, the focus here was on stop-gain variants potentially causative of NMD, even though this relationship cannot be established definitively from the extracted data. Therefore, the set of nonsense mutations was further narrowed down to those annotated as

"stop_gained&NMD_transcript". This excludes variants which are part of an NMD transcript but not stop-gain in the respective context.

Applying this criterion led to a reduction to a final set of 229 nonsense mutations affecting a total of 345 transcripts. For these the count statistics were generated. The only exceptions were two adjacent stop-gain mutations on chromosome 14 (14:93022121, 14:93022130) that were both found in a Han Chinese individual (NA18558-200-37-ASM). They lie in the same NMD-affected transcript (Ensembl Feature ID ENST00000555589.1) resulting from the *RIN3* gene. As each of those mutations has a 50% chance of being causative for the NMD effect they were counted together as one heterozygous variant instead of two. Information on the variants potentially causative of NMD in their respective transcripts is reported separately in a VCF file as Appendix C.6.

Downsampling to make population-level summary statistics comparable

For every statistic whose properties are such that it cannot be directly compared between macro-groups with different sample sizes, e.g. the DAF spectrum, a multiple random subsampling approach similar to that suggested by Leberg (2002) was applied. Let n be the original sample size of a macro-group of interest and m the sample size of the smallest group involved in the comparison. From each group with n individuals a subset of m was drawn without replacement 100 times to account for population substructure in the larger sample. The statistic of interest was calculated independently for each of the 100 replicates and averaged over all of them.

Comparison of annotations obtained using IVA and the VEP

Exonic variants in 382 genomes were extracted from the masterVarBeta files using IVA as described above with an exome filter based on the exonic boundaries as defined by Ensembl 75 and RefSeq annotations to predict translation impacts.

Further information on variant context was obtained by applying VEP with the Ensembl transcript set. These analyses were performed on genomic data in the VCF format, based on those originally generated for 730 individuals. To narrow down the VCF records to the subset of 382 individuals relevant to this chapter, and accordingly remove variant sites that had become invariant due to the reduction in sample size, VCFtools v0.1.12b (Danecek et al., 2011) was utilised. All releases after Ensembl 75 (27/02/2014) are based on the human genome assembly GRCh38.1 or higher. As the data presented in this chapter were mapped to GRCh37, a GRCh37-compatible version of Ensembl release 87 (08/12/2016) was used for variant

annotation. While the information on population frequencies and potential pathological consequences has been updated since then, the transcript set is still the latest GRCh37-compatible Ensembl transcript set, i.e. an extension of GENCODE19 (July 2013) (Harrow et al., 2012). This also means that the predictions concerning the translation impact are equivalent to those in Ensembl 75.

Matching categories for variant annotation

In order to meaningfully present and interpret the comparisons of variant annotations obtained with different methods, functional consequences were matched as closely as possible. Specifics on category matching can be found in Appendix C.7. In case of multiple annotations for a specific site from IVA as well as VEP, multiple outcomes were prioritised for each method separately using a severity ranking (Appendix C.8) following McCarthy et al (2014) and the European Bioinformatics Institute (https://www.ensembl.org/info/genome/variation /prediction/predicted_data.html). For visualisation purposes the outcomes obtained when comparing the annotations assigned by IVA and VEP were z-score transformed. Please note that the variant annotation "splice_severe_variant" is the result of merging the Ensembl variant consequences „splice_acceptor_variant" and „splice_donor_variant" and effectively describes a variant that either changes the first two or last two bases of an intron.

Annotation of deleterious variants in VCF files

Multiple approaches to infer deleterious and potentially pathogenic variants were applied to the VCF files of the 382 individuals forming the Variant-Based Analysis Set. These are complementary to the earlier inferences regarding stop-gain variants and those annotated as deleterious by SIFT and PolyPhen-2 based on the IVA output.

To infer deleteriousness of SNPs the CADD score (Kircher et al., 2014) was used. This machine-learning-based approach reports a phred-scaled C Score for each variant. It indicates how dissimilar the annotations it received from a wide range of approaches are from those reported for selectively neutral variants. The pre-computed data assigning scores to all possible SNPs of GRCh37/hg19 obtained with CADD v1.3 (11/07/2015) was retrieved from http://cadd.gs.washington.edu/download. The scores reported here are phred-like scaled based on their rank relative to all n possible SNPs in the human reference genome as follows: $-10 * \log_{10}(\text{rank}/n)$.

LoF variants were annotated using the LOFTEE (version 0.2.2beta) plugin (https://github.com/konradjk/loftee) (Karczewski, 2016) for the VEP based on the translation consequences inferred by the GRCh37-compatible version of Ensembl release 87 (08/12/2016). Its LoF definition is broadly based on the one proposed by MacArthur et al. (2012) and includes variants with the following effects: I) stop-gained, II) splice-site disrupting, III) small indels leading to a frameshift. The analyses here were limited to I) and II). Complete Genomics provides information on small indels of up to 50 bp, however, without extensive further curation, these calls appear to be unreliable (demonstrated on a well-studied example in Appendix C.9). The LOFTEE plugin considers additional filters to determine putative high confidence (HC) LoF variants. These filters exclude variants a) for which the proposed LoF allele is ancestral, b) occurring in non-canonical splicing contexts, or c) falling in the last 5% of the processed transcript (the latter criterion applies only to stop-gain variants).

In addition, the inferences of stop-gain mutation annotations by VEP and IVA were compared by a repeated measures correlation test using the R package rmcorr (Bakdash and Marusich, 2017) to detect whether there are any macro-group-specific patterns in the relationship between VEP and IVA annotations.

Subsets of genes containing HC LoF variants were analysed for the enrichment of particular biological processes, molecular functions and the cellular locations of particular gene products as given by the Gene Ontology (GO) database (The Gene Ontology Consortium, 2015) (Release 14/08/2017). These analyses were conducted using the PANTHER Overrepresentation Test (Release 20170413) with the Bonferroni correction for multiple testing (Mi et al., 2017). Information on tandem duplicates observed throughout the genome was extracted from the Duplicated Genes Database (last update: 25/02/2015) (Ouedraogo et al., 2012).

Multiple regression analyses were run to infer how well the number of derived homozygous alleles for different putatively deleterious variant classes can be predicted by a) distance from Africa and b) long term N_e since the OOA event. The great-circle distance was considered from either Eastern Africa (Addis Ababa) or Southwestern Africa (Windhoek) with the Sinai added as a waypoint for all non-Africans and Beringia for the Native Americans. The waypoints were chosen to reflect the current palaeoanthropological evidence concerning modern human origins and possible migrations routes out of Africa (see sections 1.4.1 and 1.4.2) and to be consistent with earlier studies investigating the relationship of different genome-wide statistics vs distance from Africa (Henn et al., 2012a; Prugnolle et al., 2005; Ramachandran et al., 2005).

The MSMC data (see section 3.1.1) provided information on changes in N_e during the timespan from 5-65 kya for 215 individuals from the Variant-Based Analysis Set. As the data points were unevenly distributed over it, the data were binned into 5,000-year intervals and one data point was randomly sampled from each interval. The long-term N_e(5-65 kya) was then calculated as the harmonic mean of these 13 data points. This procedure was repeated 1000 times and the final value used in the regression analyses for each subpopulation was the average N_e(5-65 kya) across all replicates.

The Human Gene Mutation Database (HGMD) (Stenson et al., 2017) was mined to highlight potentially disease-related variants. A VCF file based on the HGMD Professional version (HGMD_PRO_2015.3 as of 30/10/2015), was kindly provided by Prof David Cooper's group (University of Cardiff). This version of HGMD covers a total of 153,883 sites.

These analyses were supplemented by a core allele panel of 674 mutations identified as being highly penetrant for severe, mostly autosomal-recessive, childhood Mendelian disorders constructed by Chen et al. (2016).

3.1.4 The $R_{X/Y}$ statistic

To quantify purifying selection and the distribution of the mutational loads of certain variants classes the $R_{X/Y}$-statistic (Do et al., 2015) was used. This statistic consists of two terms. The first indicates the relative load of derived mutations seen in population X but not in population Y for the site class of interest A relative to a reference site class B, defined as $L_{X,\ not\ Y}$. Analogously, $L_{Y,\ not\ X}$ describes the sum of the frequencies of derived mutations in class A compared to class B observed in Y but not in X.

The ratio of these two terms is then the $R_{X/Y}$-statistic:

$$R_{x/y} = L_{X,\ not\ Y} / L_{Y,\ not\ X} \tag{3.3}$$

It indicates whether one population shows an excess of derived mutations of class A relative to class B compared to the other population. If selection on two lineages has been equally effective since they split, assuming that mutation rates have been the same, $R_{x/y}$ should be equal to 1.

To assess the significance of these scores a bootstrap procedure was performed, where 90% of all sites included in the calculation of the original score were resampled. One thousand replicates of these analyses provided the basis for the calculation of an empirical p-value against

the null hypothesis of $R_{x/y} = 1$, a significance threshold of 0.05 was chosen. The original Perl scripts for these analyses were kindly provided by Ms Yuan Chen (Wellcome Sanger Institute) and then modified by the author. $R_{x/y}$ was calculated on all missense mutations and on missense variants from seven subsets of genes related to phenotypes of interest. Four of these were based on GO phenotypic annotations (The Gene Ontology Consortium, 2015) (Release 01/01/2015) (Appendix C.10) and these were originally retrieved by Dr Charlotte Inchley and Sarah Kaewert. Furthermore, the Top 50 significant genes associated with susceptibility to malaria were extracted from the MalaCards database (Rappaport et al., 2013) (as of 18/03/2015). Additionally, a list of 810 genes containing at least one SNP whose frequency distribution displays a strong correlation with the diversity of helminth species transmitted in different geographic areas was used to define genomic elements with a potential role in the immune response against these parasitic worms (Fumagalli et al., 2010). Finally, a manually curated list of 35 genes significantly associated with any aspect of normal pigmentation variation in humans based on a strict choice from GWAS (as of 18/03/2015) was kindly prepared by Dr Mircea Iliescu (University of Cambridge) (Appendix C.11).

3.1.5 Tests for positive selection

The ΔDAF approach was based on a cross matrix of DAFs in 12 macro-groups (n = 361) from the Variant-Based Analysis Set (Table 3.1), Native Americans and Oceanians were excluded due to small sample sizes. The most differentiated pairs of macro-groups were selected based on the maximum allele frequency difference. To account for LD if several highly differentiated SNPs were in the same 200-kb window only the most extreme signal was reported. The subsets of interest were the top 20 ΔDAF missense and nonsense mutations a) among all macro-groups and b) among non-Africans only.

To visualise the macro-group sharing of highly differentiated mutations in relation to demographic history a phylogenetic tree was constructed using the TreeMix v1.1 software (Pickrell and Pritchard, 2012). The input data were based on a subset of the VCF files of the Diversity Set (see section 3.1.1). All groups were contained in this analysis except Central Asians and Native Americans. Singletons and doubletons were removed, and the data were polarised based on the ancestral alleles extracted from a six-primate alignment (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/ancestral_align

ments/). TreeMix analyses were conducted by Dr Mait Metspalu. MEGA v7.0.14 (Kumar et al., 2016) was used by the author to create a graphic representation of the TreeMix output.

The DIND (Barreiro et al., 2009) statistic served as an additional measure to explore whether variants highlighted by ΔDAF also lie on haplotypes of unusually low diversity. DIND scores were generated for the macro-group with the highest DAF respectively for all variants in the top 20 of the several ΔDAF analyses. Details on the how the statistical significance of the DIND scores was assessed can be found in Appendix C.12 and the R scripts to obtain them were kindly provided by Dr Florian Clemente.

3.1.6 Integration of functional databases with selection results

The outcomes of the ΔDAF, DIND and d_i were merged with databases providing further functional annotations to support their interpretation. This candidate set consisted of the top 20 ΔDAF for different comparisons and subsets of the top candidates for the other two approaches briefly described in the following. The top 12 of the most highly divergent SNPs by the d_i score were chosen in each of the twelve population groups from the Selection Set. If two highly differentiated SNPs were above an LD threshold ($r^2 > 0.05$) the one with the higher d_i score was reported.

The two highest ranking windows in each of the twelve population groups were identified based on a composite signal from the empirical p-values of three selection tests (iHS, nSL and Tajima's D) and subjected to further DIND analyses. The results for iHS, nSL and Tajma's D were generated by Evelyn Jagoda, Dr Guy Jacobs (University of Southampton, present affiliation: Nanyang University) and Dr Charlotte Inchley, respectively. They are not further described here as they were not relevant for any other analyses in this thesis. GWAS results were obtained from I) the NHGRI-EBI Catalog of published GWAS studies (MacArthur et al., 2017) (www.ebi.ac.uk/gwas, bulk data v1.0.1 downloaded on 07/09/2017, data last updated on 31/08/2017) and II) GWAS Central (Beck et al., 2014) (www.gwascentral.org, data accessed using GWAS Mart on 07/09/2017). Where possible the effect allele was inferred either from the databases or manually from the source papers.

The gene expression data used for the analyses described here were taken from the GTEx Portal (Lonsdale et al., 2013) (https://www.gtexportal.org/home/, GTEx Analysis V7, bulk data on cis-eQTLs downloaded on 14/09/2017). This dataset contains information from 11,688

transcriptomes of 714 donors across 53 tissues; because of sample size requirements, eQTLs were determined for 620 individuals and 48 tissues. GENCODE 27 (released August 2017) (Harrow et al., 2012) was used as an additional resource to infer the strength of support and the biotype of transcripts affected by eQTLs. In a similar manner as described for the genes containing HC LoF (see section 3.1.3) the protein-coding genes targeted by the eQTL were tested for enrichment of functional categories using PANTHER. LD statistics for the region surrounding the potential selection target rs11227639 were calculated using PLINK 1.9 (Chang et al., 2015a). BED files containing information about DNase hypersensitivity sites, which are considered markers of regulatory regions (Tsompana and Buck, 2014), generated by the Roadmap Epigenomics consortium (Kundaje et al., 2015) were retrieved (ftp://ftp.ncbi.nlm.nih.gov/pub/geo/DATA/roadmapepigenomics/by_experiment/DNase_hypersensitivity/ downloaded on 28/09/2017).

3.2 Results

3.2.1 Distribution of functional and deleterious variation inferred with Ingenuity Variant Analysis

Firstly, the distribution of functional variation was explored using IVA by considering the differences in exonic SNPs among 14 regional pools of populations (Table 3.1). Among the 382 samples considered 294,360 exonic single nucleotide variants are observed (Table 3.3), of these 51.4% are singletons.

Table 3.3: Total exonic SNP counts by macro-group extracted using Ingenuity Variant Analysis. Abbreviations: DAF - derived allele frequency, short codes for the macro-groups taken from Table 3.1.

	all	synonymous	missense	nonsense
Afr	96,223	45,723	43,398	413
MiE	70,293	31,218	33,931	328
WEu	67,637	29,786	32,969	350
EEu	80,203	33,944	40,406	488
Vol	59,601	27,144	28,059	315
SoA	71,163	31,492	34,510	374
WSi	51,369	23,404	24,129	230
SSi	68,109	30,105	33,080	390
CSi	53,904	24,366	25,567	281
NSi	47,516	21,978	22,041	253

	all	synonymous	missense	nonsense
SeM	63,658	28,532	30,499	359
SAm	40,210	18,365	18,867	159
Oce	39,155	18,666	17,534	163
Whole dataset	294,360	116,069	157,683	2,357
average DAF	0.056	0.074	0.041	0.016

In a subset of 266,993 variants for which the ancestral state was known a significant shift of the DAF spectrum for missense and nonsense mutations towards low global frequencies (<10%) compared to synonymous variants is observed (Figure 3.3; Table 3.4) (missense vs synonymous $X^2 = 2550.6$, $p < 10^{-15}$; nonsense vs synonymous $X^2 = 238.66$, $p < 10^{-15}$).

For each individual genome on average ~9,300 synonymous and ~7,800 missense sites are observed compared to only 56 nonsense (here limited to stop-gain) variants (Table 3.5). Of these stop-gain variants on average 3.2 are part of at least one transcript affected by NMD. As

Table 3.4: Proportions of variants in different functional classes by DAF. The first row contains information on all samples pooled together. To make the regional patterns comparable the frequency spectrum for each macro-group was normalised to the sample size of the smallest group (the Oceanians at n = 8). To account for population structure in each regional group this procedure was repeated 100 times for all groups except the Oceanians. The proportions given in this table represent the mean fractions obtained from these 100 runs. Abbreviations: DAF - derived allele frequency, short codes for the macro-groups taken from Table 3.1.

	synonymous			missense			Nonsense		
DAF	**<10%**	**10-50%**	**>50%**	**<10%**	**10-50%**	**>50%**	**<10%**	**10-50%**	**>50%**
All	0.861	0.079	0.059	0.923	0.046	0.030	0.977	0.019	0.003
Afr	0.397	0.432	0.170	0.472	0.389	0.138	0.561	0.379	0.060
MiE	0.288	0.416	0.296	0.400	0.375	0.225	0.523	0.382	0.095
WEu	0.266	0.421	0.313	0.370	0.387	0.242	0.519	0.380	0.101
EEu	0.259	0.426	0.315	0.369	0.387	0.244	0.534	0.370	0.096
Vol	0.269	0.427	0.304	0.370	0.392	0.238	0.568	0.337	0.095
SoA	0.282	0.424	0.294	0.394	0.378	0.228	0.550	0.358	0.091
WSi	0.282	0.424	0.294	0.341	0.405	0.254	0.505	0.379	0.116
SSi	0.262	0.415	0.323	0.363	0.382	0.255	0.541	0.352	0.107
CSi	0.216	0.427	0.358	0.303	0.408	0.289	0.501	0.376	0.123
NSi	0.198	0.432	0.370	0.279	0.419	0.302	0.466	0.405	0.128
SeM	0.261	0.407	0.332	0.359	0.375	0.266	0.542	0.346	0.112
SeI	0.247	0.417	0.335	0.345	0.390	0.265	0.492	0.384	0.124
SAm	0.196	0.397	0.407	0.306	0.386	0.308	0.403	0.395	0.202
Oce	0.243	0.403	0.354	0.330	0.391	0.280	0.511	0.364	0.126

they are almost always the only stop-gain variants detected in their respective transcript in the individuals carrying them it they are likely causative for the NMD. In Africans on average 5.2

Table 3.5: Per individual exonic SNP and/or allele counts for different functional variant classes for each of the 14 macro-groups used in the variant-based analyses. Abbreviations: del - deleterious (SIFT and PolyPhen-2 combined), der – derived, mis- missense, NMD - stop gain variants which are part of transcripts predicted to be affected by nonsense-mediated mRNA decay as inferred from Ensembl Release 87, non – nonsense, syn – synonymous, hom – homozygous genotypes, het - heterozygous genotypes, short codes for the macro-groups taken from Table 3.1. The divergences between the "hom" column and the subsequent columns giving the number of derived homozygous missense and synonymous variants occur because for a subset of alleles the reference allele was derived and therefore the allelic states were reversed before counting the variant.

	All	syn	mis	del	non	NMD	hom	Het	hom syn der	hom mis der	hom non der	hom del der
Afr	22855.9	11622.9	9487.2	647.6	67.2	5.3	7454.2	15401.7	4236.5	3017.3	5.2	70
MiE	18663.2	9360.1	7876.2	566.3	54.1	3.3	6746.2	11917	5098.8	3626.5	6.9	89.1
WEu	18346	9200.6	7739.1	548	54.7	3.5	6663.4	11682.6	5150	3679.1	6.7	91.8
EEu	18361.4	9206.4	7748.1	559.2	53.8	3.3	6688.6	11672.8	5163.4	3698.8	5.8	90.4
Vol	18529.2	9299.1	7800.5	546	54.7	2.8	6736.5	11792.7	5144.4	3650.9	7.1	88.6
SoA	18816.5	9457.6	7919.6	567.7	57.1	2.2	6842.3	11974.2	5071.6	3613	7.4	88.6
WSi	18357.8	9189.1	7748.6	558.2	53.9	2.1	7200.9	11156.9	5313.9	3778.8	7.2	97.2
SSi	18459.4	9254.8	7772	544.4	56.5	2.8	7109.5	11349.9	5237.9	3726.1	6.6	91.8
CSi	18055.3	9071.2	7593.3	523.3	53	3.1	7561.7	10493.6	5421.6	3857	6.1	103.2
NSi	17995.3	9048.8	7541.3	514	54.6	2.9	7676.6	10318.8	5484.8	3899	7.8	102.3
SeM	18323.6	9217.6	7683	516.6	57	3	7297.5	11026.1	5284.7	3764.9	7.3	91.5
SeI	18326.2	9200.9	7718.7	536.5	58.3	3	7631.8	10694.4	5391.4	3845.6	8.9	106.7
SAm	17677.4	8837.5	7482.4	563.8	54.1	3.3	8207.5	9469.9	5763.4	4106.4	11.2	122.9
Oce	18559.5	9312.5	7826.8	534.3	59.6	3.2	8265.6	10293.9	5478.9	3917.6	8.4	113.3
Whole dataset	18604	9344.9	7827.4	550	56.1	3.2	7184.3	11419.7	5217.6	3718.6	7.1	94.9

Table 3.6: Average per individual ratios of different types of non-synonymous variants compared to synonymous variants for each of the 14 macro-groups used in the variant-based analyses. Outcomes of Tukey's HSD for each pairwise comparisons are given in a matrix, significant p-values are bolded. The upper half contains the values for the ratio of missense to synonymous variants, the lower half for nonsense to synonymous. Abbreviations: mis- missense, non – nonsense, ns – non-significant syn – synonymous. Short codes for the macro-groups taken from Table 3.1.

	mis/syn	non/syn	Afr	MiE	WEu	EEu	Vol	SoA	WSi	SSi	CSi	NSi	SeM	SeI	SAm	Oce
Afr	0.8163	0.0058		**$< 10^{-15}$**	**$< 10^{-15}$**	**$< 10^{-15}$**	**$< 10^{-15}$**	**$< 10^{-15}$**	**$< 10^{-15}$**	**$< 10^{-15}$**	**$< 10^{-15}$**	**$< 10^{-15}$**	**$< 10^{-15}$**	**$< 10^{-15}$**	**$< 10^{-15}$**	**$< 10^{-15}$**
MiE	0.8415	0.0058	ns		ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
WEu	0.8412	0.0059	ns	ns		ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
EEu	0.8416	0.0058	ns	ns	ns		ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
Vol	0.8388	0.0059	ns	ns	ns	ns		ns	ns	ns	ns	ns	ns	ns	ns	ns
SoA	0.8374	0.0060	ns	ns	ns	ns	ns		ns	ns	ns	ns	ns	ns	ns	ns
WSi	0.8432	0.0059	ns	ns	ns	ns	ns	ns		ns	ns	ns	ns	ns	ns	ns
SSi	0.8398	0.0061	ns	ns	ns	ns	ns	ns	ns		ns	ns	ns	ns	ns	ns
CSi	0.8371	0.0058	ns	ns	ns	ns	ns	ns	ns	ns		ns	ns	ns	ns	ns
NSi	0.8334	0.0060	ns	ns	ns	ns	ns	ns	ns	ns	ns		ns	ns	**0.028**	ns
SeM	0.8335	0.0062	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns		ns	**0.022**	ns
SeI	0.8389	0.0063	**0.018**	**0.006**	ns	**0.002**	ns	ns	ns	ns	**0.013**	ns	ns		ns	ns
SAm	0.8467	0.0061	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns		ns
Oce	0.8405	0.0064	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	
Whole data	0.8376	0.0060														

of these variants were detected, whereas only 3.1 were found in non-Africans. It is unclear whether the excess observed in Africans with regards to these variants is of functional importance because a) the absolute numbers involved are relatively small and b) experimental verification of variants predicted to be causative for NMD *in silico* indicates that only ca. 40-70% of them actually lead to a measurable decrease in gene expression (GTEx Consortium et al., 2015; Rivas et al., 2015, see section 1.6.2). When the ratio of total missense vs synonymous variants per individual was compared by population groups an ANOVA showed that there are significant differences between the population means regarding this statistic ($p < 10^{-15}$). To further investigate the pairwise comparisons Tukey's HSD test, which corrects for multiple comparisons, was applied. It indicates that all non-African population groups have a significantly higher number of missense variants relative to (almost) neutral synonymous variants compared to the Africans ($p < 10^{-4}$ for all tests) (Table 3.6). Furthermore, the Native American group exhibits the highest missense/synonymous ratio of all macro-groups. However, this excess of missense variants is only significant in comparison to two of the other non-African groups: North Siberians and mainland East/Southeast Asians ($p < 0.03$ for both tests).

When applying the same framework to the number of nonsense variants relative to synonymous variants the ANOVA is again significant ($p < 2*10^{-4}$). Among all macro-groups the Oceanians and the ISEA group have the highest value for this statistic. Only the latter exhibit a significant excess of nonsense variants compared to the Africans and three other non-African groups ($p < 0.012$ for all tests). The higher fraction of non-synonymous variation in non-Africans could have been caused by multiple mechanisms. Simulation studies suggest that recent population growth leads to an increase in the proportion of non-synonymous (i.e. mainly missense) SNPs relative to that expected in non-expanded populations (Lohmueller, 2014b).

To empirically test this idea the harmonic mean of the subpopulation-specific N_e for the last 5,000 years was correlated with the per-individual missense/synonymous ratio for all non-Africans. For the 186 individuals for whom such data were available there is no significant linear relationship between recent N_e and this ratio ($r = -0.0595$, $p = 0.4171$).

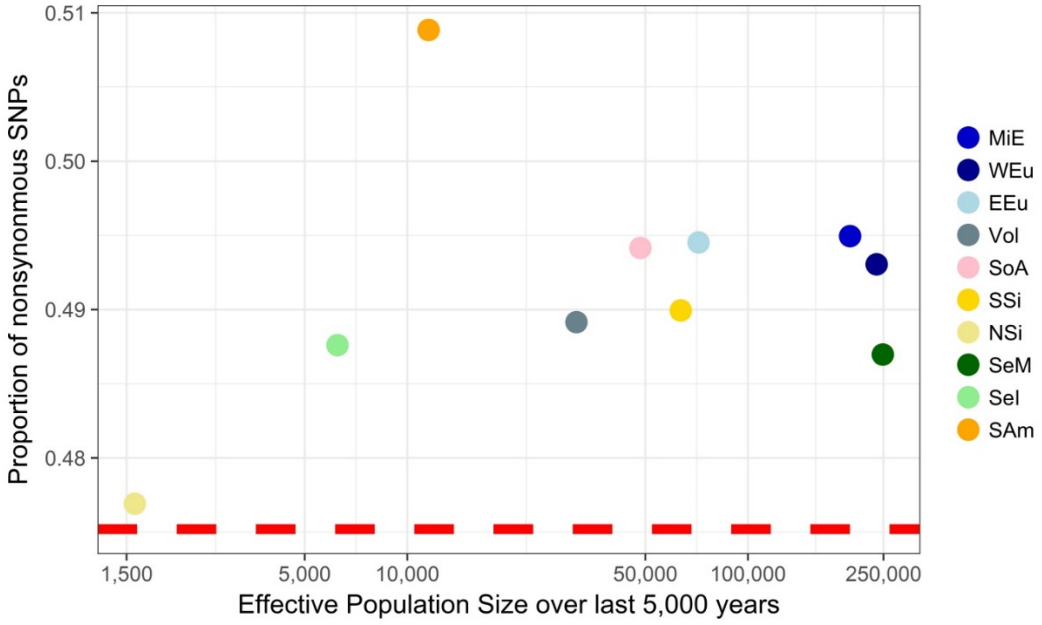


Figure 3.2: Scatter plot of the fraction of non-synonymous SNPs in non-African macro-groups relative to the harmonic mean of N_e obtained with MSMC over the last 5,000 years. The dashed horizontal line indicates the proportion of non-synonymous SNPs in Africans. For all groups except the South Americans fractions were calculated from the mean (rounded to full numbers) counts obtained from repeated downsampling to n = 13. This graph only includes continental groups containing at least one population for which 4 genomes were sequenced so that MSMC could be run with good resolution. If there were multiple such populations, N_e values were averaged. Short codes for the macro-groups taken from Table 3.1.

Germans and Burmese constitute extreme outliers, their population sizes are beyond the "upper fence" of N_e = 775,888, defined as the upper quartile plus three times the interquartile range, following the standards suggested by the US National Institute of Standards and Technology (NIST, 2013). If they are removed the results remain essentially unchanged (r = -0.0704, p = 0.353).

Additional support for this result comes from analysing the data in exactly the same manner as proposed by Lohmueller (2014b). There the fraction of non-synonymous variants was calculated relative to all exonic sites and on a per-macro-group basis (above it was done per-individual). Given the unequal sample sizes involved this was achieved by using a strategy described in section 3.1.3 to downsample all groups except the South Americans (Oceanians were excluded due to their small sample size) to n = 13 and to obtain the fraction of non-synonymous SNPs. The main outcomes are very similar: all non-Africans have a higher fraction of non-synonymous variants compared to Africans and for non-Africans this fraction does not correlate (r = 0.0217, p = 0.9525) with the harmonic mean of N_e for the macro-groups over the last 5,000 years (Figure 3.2, Appendix C.13). This also holds true if the statistic correlated to the non-synonymous proportion is $N_{e5,000\_harmonic\_mean}$/ $N_{e20,000}$, i.e. the relative increase in N_e

over the last 20,000 years (r = 0.0269, p = 0.9412). Notably, the Native Americans exhibit the highest fraction of non-synonymous SNPs of all non-Africans, whereas the Northeast Siberians have the lowest while the other non-Africans cannot be distinguished by this metric.

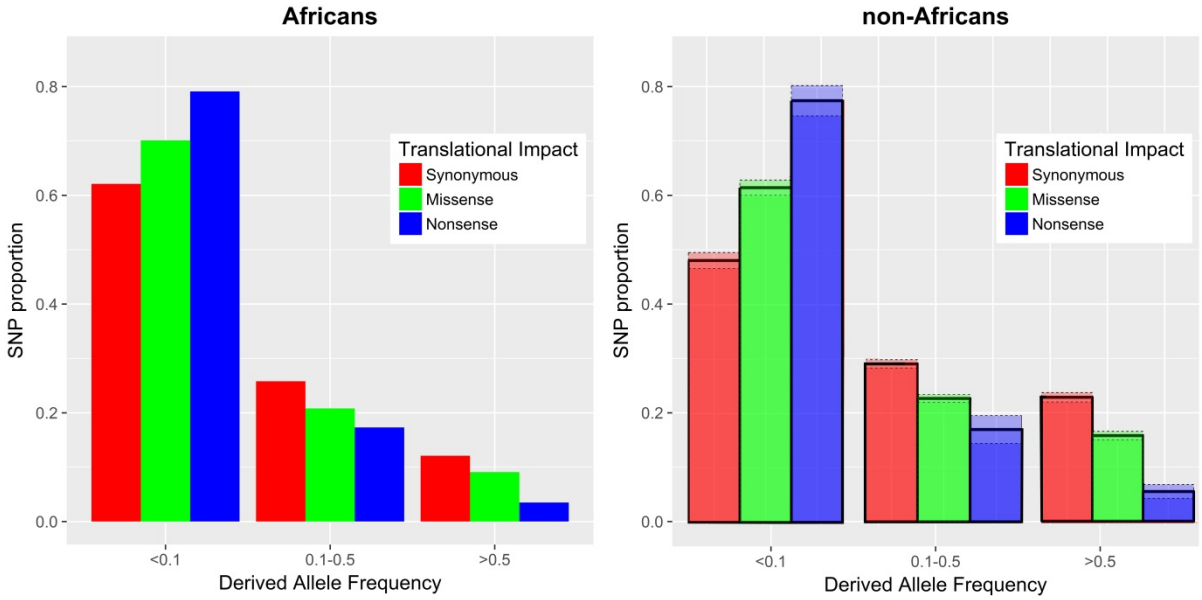


Figure 3.3: Bar plot comparing the binned DAF distributions of different classes of exonic variants in Africans and non-Africans. As Africans have a pooled sample size of n = 21, non-Africans (n = 361) were downsampled. In the right plot the bolded bars indicate the respective mean of the SNP proportions obtained from all sampling replicates, the dotted lines mark 2 SD below and above this mean. For synonymous and missense variants Africans show a significant excess of low frequency sites based on the underlying count data (African synonymous vs non-African synonymous $X^2 = 159.3$, $p < 10^{-15}$; African missense vs non-African missense $X^2 = 836.59$, $p < 10^{-15}$).

Additional complexity emerges when DAFs are considered. Non-Africans were downsampled to n = 21 to make them directly comparable to Africans. There is a significant ($p < 10^{-15}$ for both chi-square tests) shift towards the rare end of the frequency spectrum for synonymous and missense variants in the latter (Figure 3.3). While there is a similar trend for nonsense variants it did not reach the thresholds required for statistical significance. The downsampling approach described above was applied to compare the 14 macro-groups to each other. For synonymous and missense variants, the majority of the DAF spectra are significantly different between macro-groups (for these 91 comparisons the Bonferroni correction was applied, leading to a more stringent significance threshold of $p = 5.495*10^{-4}$) (Figure 3.4, raw data in Appendix C.14). Groups for which the DAF spectra are not differentiated are either phylogenetically close and/or would have experienced similar population histories. The former applies to two clusters, a West Eurasian one comprising the two European groups and the Volga-Uralic group and an

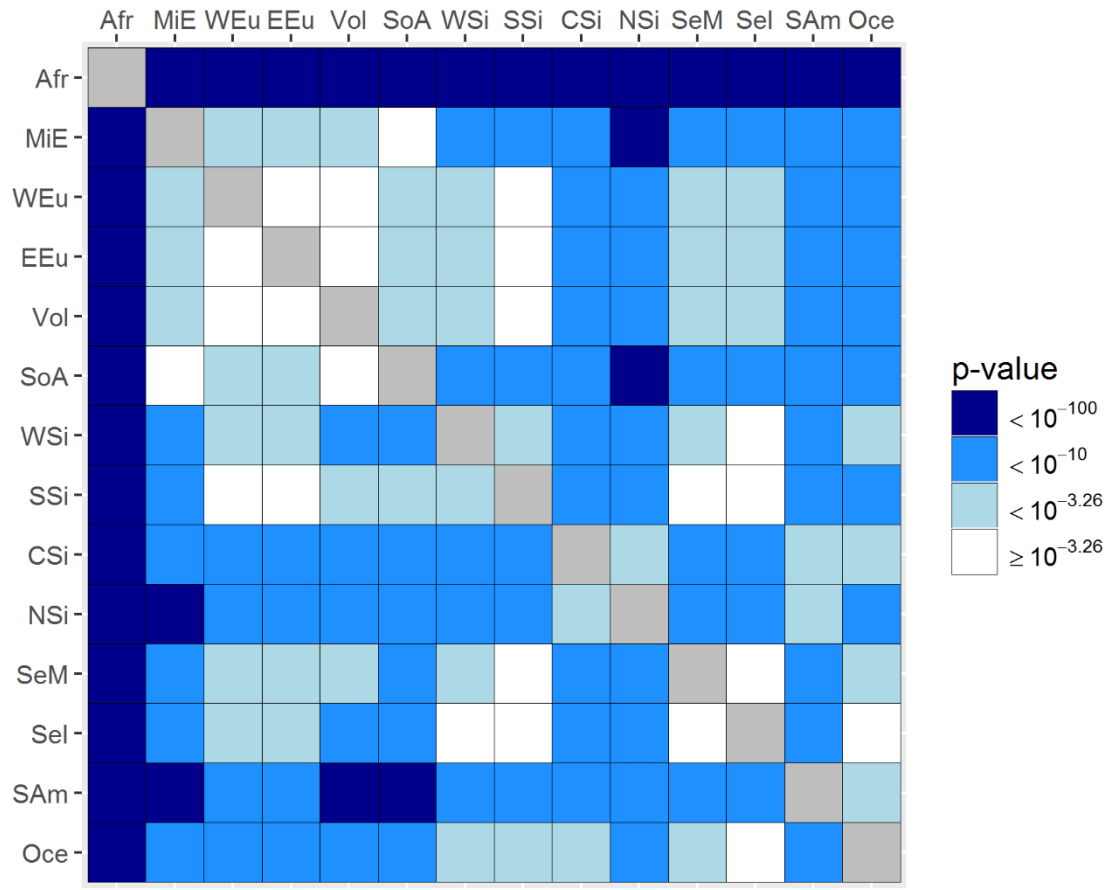


Figure 3.4: Heatmap displaying the outcomes of pairwise chi-square tests comparing the binned (see Table 3.4) DAF spectra between each of the 14 macro-groups. For all groups except the Oceanians these spectra represent the mean (rounded to full numbers) counts obtained from repeated downsampling to n = 8. The lower half contains the values for the DAF spectra of synonymous, the upper half for those of missense variants. Cells containing the p-values for comparisons that are not significant according to the Bonferroni-corrected threshold ($p = 10^{-3.26}$, as there are 91 comparisons each) are white.

East Eurasian macro-grouping consisting of the island and the mainland groups from East and Southeast Asia plus the South Siberians. The latter group's DAF spectra are also similar to those of West Eurasian groups, most likely indicating higher N_e resulting from a recent expansion compared to the other Siberians.

The picture for nonsense variants is very different. For most of the macro-group comparisons there are no significant differences (Appendix C.15). The only notable exception are Native Americans (Andeans) who show a significant excess of high frequency nonsense variants compared to the Africans ($p = 1*10^{-4}$) and a similar trend compared to seven other non-African groups ($p < 0.05$ for all these comparisons, however they were not significant according to more strict criteria). The non-Africans to which Native Americans are similar are Siberian groups and island Southeast Asians.

Finally, the patterns of regional sharing of exonic variants were examined. When the focus is on variants from across the whole frequency spectrum in general the more severe the functional consequence of a variant the less likely it is to be shared across regions (Table 3.7).

Table 3.7: Average relative sharing of exonic SNPs from each macro-group with all others. Singletons on a global level were excluded for the calculation of these ratios. Furthermore, the results of the "all sites" column and those based on DAF filters are not directly comparable as the ancestral state could only be inferred for a subset of sites and the polarity for the synonymous and missense variants was changed in some cases (see section 3.1.2). Abbreviations: DAF- derived allele frequency, mis- missense, non – nonsense, syn – synonymous Short codes for the macro-groups taken from Table 3.1.

	all sites (shared non-reference alleles)			>50% DAF			<10% DAF		
	syn	mis	non	syn	mis	non	syn	mis	non
Afr	0.55	0.48	0.41	0.46	0.45	0.44	0.14	0.13	0.12
MiE	0.70	0.62	0.54	0.66	0.66	0.58	0.29	0.28	0.26
WEu	0.71	0.62	0.53	0.67	0.67	0.59	0.30	0.29	0.26
EEu	0.70	0.61	0.52	0.69	0.68	0.59	0.32	0.31	0.29
Vol	0.73	0.65	0.54	0.70	0.70	0.63	0.33	0.30	0.27
SoA	0.72	0.64	0.54	0.68	0.68	0.63	0.30	0.29	0.26
WSi	0.73	0.64	0.55	0.68	0.68	0.65	0.29	0.27	0.26
SSi	0.72	0.64	0.54	0.71	0.71	0.66	0.33	0.31	0.29
CSi	0.71	0.63	0.52	0.67	0.66	0.65	0.28	0.26	0.28
NSi	0.71	0.62	0.53	0.64	0.65	0.60	0.24	0.22	0.23
SeM	0.71	0.63	0.53	0.68	0.68	0.61	0.27	0.24	0.22
SeI	0.68	0.59	0.49	0.67	0.66	0.62	0.25	0.23	0.21
SAm	0.66	0.57	0.50	0.55	0.54	0.42	0.16	0.16	0.13
Oce	0.64	0.55	0.44	0.50	0.50	0.44	0.13	0.12	0.11

It seems plausible that this is a consequence of the differences in the global allele frequency spectra for different site classes, i.e. the variants which are more often deleterious appear to be more constrained and shifted towards lower frequencies.

As an additional criterion, variants were stratified by their DAF in such a manner that a variant was counted as shared if it was above or below a certain threshold in both populations of interest. Unsurprisingly, in all three classes of exonic variants, including missense and nonsense mutations, those with a higher DAF (>50%) show considerably more sharing among macro-groups than those at a low (<10%) frequency. These analyses also confirm that the lower sharing of nonsense variants is mostly a function of their lower global frequency as the relative differences in sharing between different functional classes are less pronounced if allele frequency is adjusted for.

Across the different filtering categories generally macro-groups from mainland Eurasia who have multiple geographical borders with other clusters such as the Volga-Uralic and the South Siberian regional groups exhibit the highest sharing. Africans share the least functional variants with the others reflecting their greater genetic diversity and the depth of the African/non-African split. Other groups exhibiting relatively little overlap in exonic variation are Oceanians and Southern Americans from the Andes. The latter two groups have been genetically mostly isolated from the other non-Africans examined here since at least ca. 15 kya (Raghavan et al., 2015; Malaspinas et al., 2016). They share as few rare variants, which are generally of recent origin, with other macro-groups as Africans. Rare variant sharing patterns will be explored in more detail in chapter 4.

3.2.2 Ingenuity Variant Analysis compared to Ensembl's Variant Effect Predictor

In total, 354,396 SNPs were assigned exonic and splice site-related annotations by either IVA or VEP (Table 3.8). 276,388 variants are either exonic or splice-site altering according to the RefSeq release 63 transcript set which was the basis for the annotations by IVA. A larger total

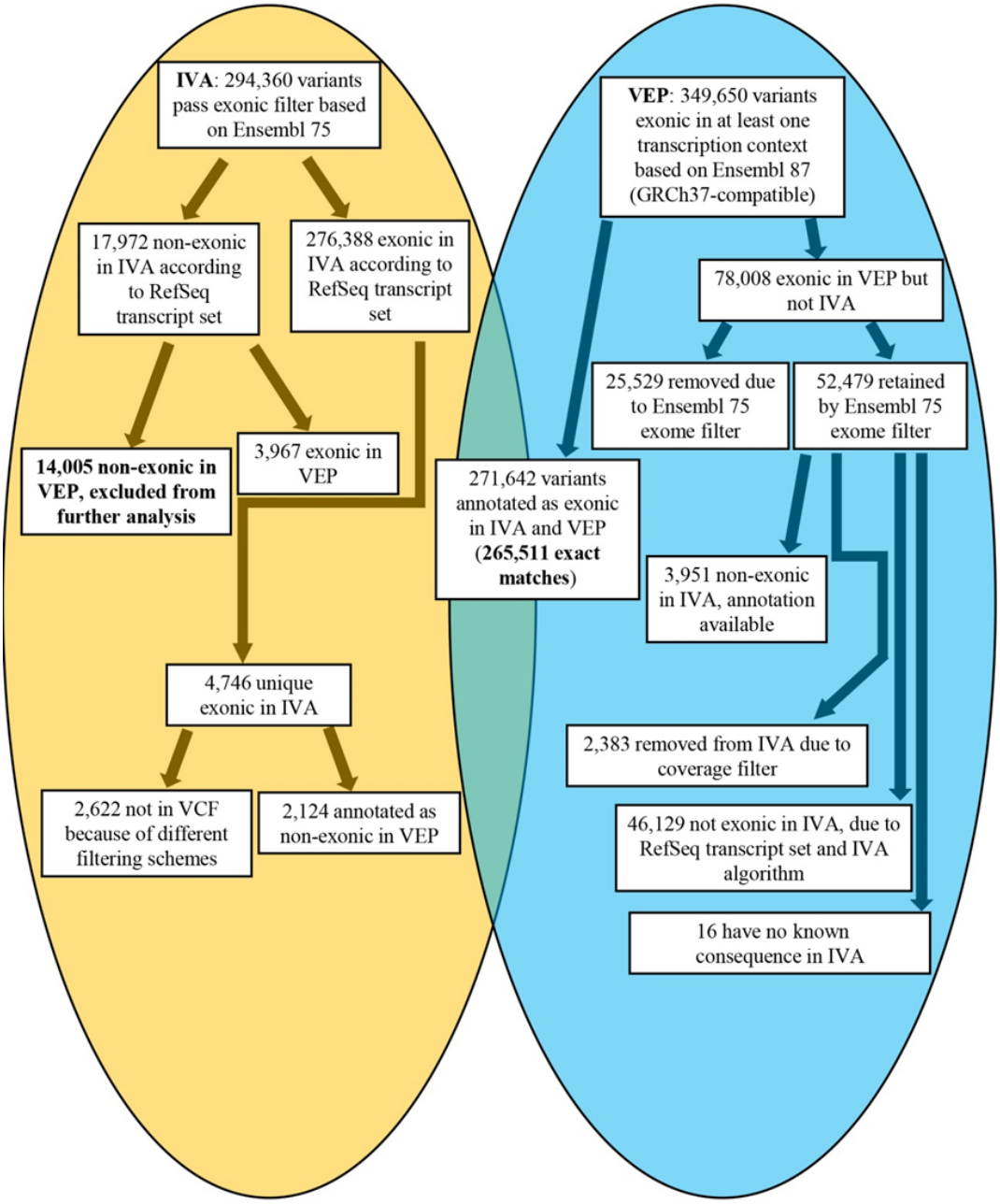


Figure 3.5: Filtering scheme displaying the steps of the comparison between IVA and VEP and the amount of overlap between sites called as exonic from both approaches. The absence of 2,622 sites in the VCF containing 382 individuals occurred because the IVA data were filtered based on sites being variable in a larger initial VCF of 730 individuals. For this subset of variant sites, mostly consisting of singletons, the calls reported in the IVA output and the 382-individual VCF were inconsistent due to a slightly more permissive call quality filter (threshold of 30 for IVA and 40 for VCF files) applied to heterozygous variants for IVA.

of 294,360 sites passed the exome filter applied to the IVA output, however this filter was based on Ensembl 75. The VEP annotated a total of 349,650 sites to either be exonic or to influence splicing patterns in at least one transcription context. Overall, 271,642 variants are categorised as exonic by both tools, of which 265,511 (74.9%) are exact matches, while the other 6,131 are assigned divergent exonic annotations. If splice region variants are excluded, the total match rate improves to 83.8%. For the other sites (n = 82,754), only one tool yields an exonic annotation, the majority of these are exclusively called by the VEP (Figure 3.5).

Table 3.8: Summary of the number of annotations matching between IVA and VEP for each category of exonic or splice-site altering annotations. The first column contains "IVA+VEP", i.e. it represents the union of all sites assigned to each consequence type based on IVA and VEP. The exonic and splice sites total for the "IVA+VEP" column is smaller than the sum of the rows giving the numbers for each annotation separately. This is because there are 6,131 sites categorised as belonging to different exonic categories by the two approaches and therefore counted twice when adding them up. The columns "IVA" and "VEP" hold the number of variants categorised by each approach separately. The match column is the logical intersection, i.e. the number of sites that receive the same annotation from both tools. The match rates in columns five to seven are reported as the proportion of these matches relative to the total number of annotations in the category either from IVA, VEP or both.
106 sites belonging to the "stop_retained_variant" category for IVA were manually re-annotated. The original IVA annotation for these variants was "synonymous". However, when considering additional information (Appendix C.5) given by IVA in the "Protein Variant" column it became apparent that in these cases the respective mutation caused a stop codon to be replaced by another stop codon.

	IVA +VEP	IVA	VEP	Match	IVA match rate (%)	VEP match rate (%)	Total match rate
stop_gained	3,479	2,354	3,406	2,281	96.9	66.97	65.56
stop_lost	549	233	522	206	88.41	39.46	37.52
splice_severe_variant	4,920	553	4,899	532	96.2	10.86	10.81
start_lost	630	328	607	305	92.99	50.25	48.41
missense	185,152	157,492	181,353	153,693	97.59	84.75	83.01
splice_region_variant	37,790	-	37,790	-	-	-	-
stop_retained_variant	182	106	178	102	96.23	57.3	56.04
synonymous	127,825	115,322	120,895	108,392	93.99	89.66	84.80
LoF_equivalent	8,948	3,140	8,827	3,019	96.15	34.20	33.74
all_exonic_and_splice_site	354,396	276,388	349,650	265,511	96.06	75.94	74.92

While this excess of exonic annotations by the VEP is consistently observed across all categories of exonic/split-site altering variants match rates vary considerably. Synonymous variants show the best concordance, as 84.8% of them are called consistently by IVA and VEP.

On the other end of the spectrum, severe splice variants are only assigned the same status in 10.8% of all cases. Recall that splice sites are the genomic regions which correspond to exon-intron boundaries and where the primary RNA transcript is cleaved before the excised introns are degraded and the exon sequences are rejoined. Therefore, even small differences in the genomic coordinates of the exon-intron boundaries between the Ensembl and RefSeq transcript sets should contribute to these marked differences in annotations. Furthermore, putative LoF variants that have a potentially more severe effect than missense and synonymous variants, show only a relatively low concordance of 33.7%.

In order to further study these patterns, the data can be visualised as a heatmap (raw data in Appendix C.16). It gives a detailed breakdown for each annotation with one method displaying to which categories assigned by the complementary approach this annotation corresponds.

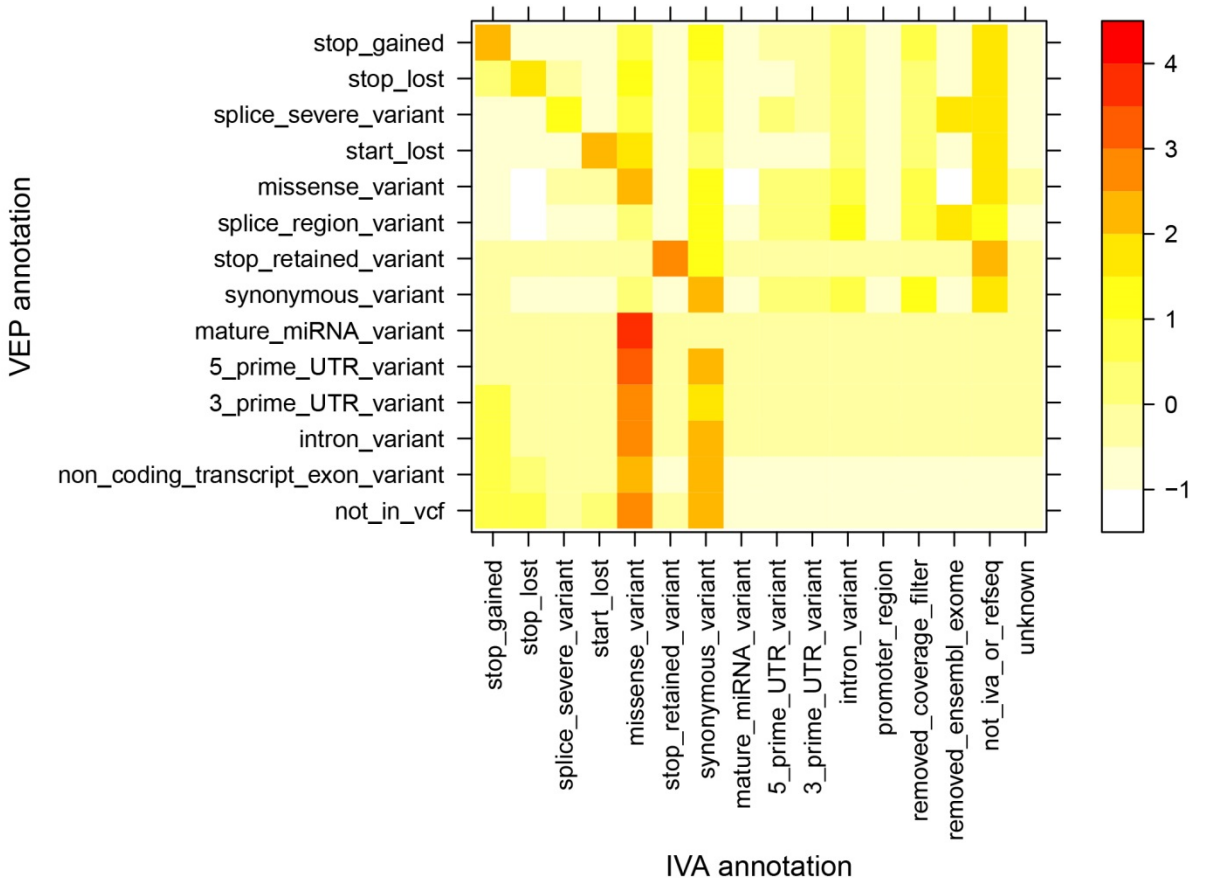


Figure 3.6: This heatmap contains the number of variants for all different combinations of annotations derived from VEP (rows) and IVA (columns). Z-scores from <-1 to >4 are indicated by a colour bar with increasing intensity from white to yellow to red for higher values. High z-scores indicate that a particular IVA annotation is over-represented across all other IVA annotations inferred for sites with a particular VEP annotation.

In general, many variants assigned to a specific exonic category by the VEP are either given a different exonic annotation by IVA or are called as non-exonic (Figure 3.6). These non-exonic calls are spread across several categories reflecting different factors contributing to this divergence. A significant fraction is explicable by the additional filters applied to IVA data, namely the exome filter derived from Ensembl version 75. It led to the exclusion of 25,529 sites (Figure 3.5). All of these are either categorised by VEP as being located in a splice region or as severe splicing variants. From manually investigating a few examples (data not shown) it became apparent that these sites are intronic but located very close to an exon/intron boundary where splicing takes place. Another large (n = 46,129) (Figure 3.5) fraction of these sites is removed either because a) they are not part of the RefSeq transcript set or b) while they are part of the RefSeq set the IVA annotation algorithm does not report them as exonic.

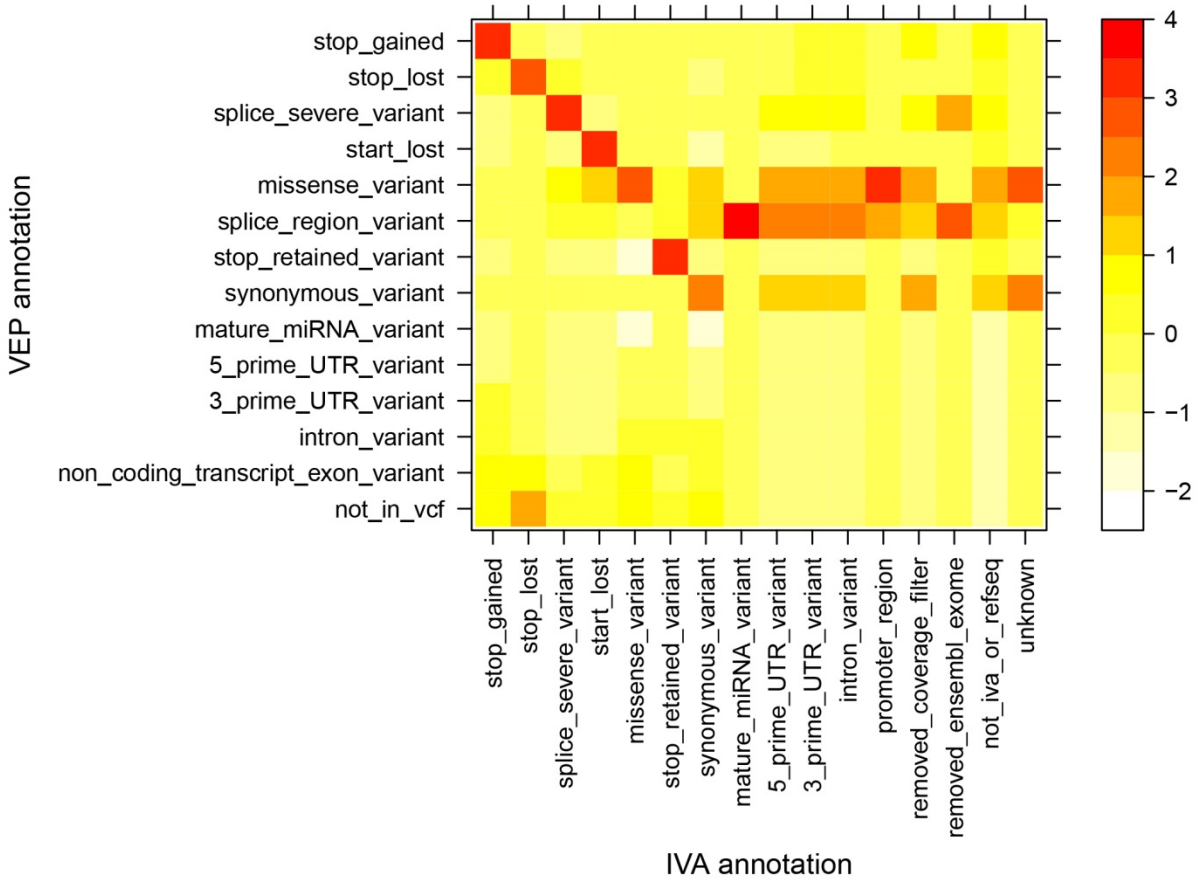


Figure 3.7: This heatmap contains the number of variants for all different combinations of annotations derived from VEP (rows) and IVA (columns). Z-scores from <-2 to 4 are indicated by a colour bar with increasing intensity from white to yellow to red for higher values. High z-scores show that a specific VEP annotation is over-represented across all other VEP annotations inferred for sites with a particular IVA annotation. The clearly visible diagonal of high z-scores for exonic annotations indicates a very high concordance of variants called as exonic by IVA with the corresponding VEP calls.

One way to learn more about the contributions of a) and b) is to use the RefSeq and Ensembl transcripts sets with the same annotation algorithm. Therefore, chromosome 22 was analysed using the GRCh37-compatible version of VEP with the curated RefSeq transcripts as source. The overall number of matches between RefSeq and Ensembl for this chromosome is 7,046/8,675 equalling a total match rate of 81.2%. Even when leaving splice region variants aside, that are mostly not recorded by IVA even if their impact is estimated to be severe (Table 3.8), the Ensembl transcript set recognises 7,686 variants as exonic, of which 6,531 receive the exact same annotation when using the RefSeq transcript set from VEP (84.9%). This match rate is similar to that observed for the whole genome with IVA when splice-related sites are excluded. Therefore, the use of the RefSeq transcript set probably is responsible for most of the divergences regarding the calling of exonic variants in the strict sense whereas the algorithm applied by IVA significantly contributes to the particularly low number of matches for splice-related variants.

On the other hand, the chance of a variant in IVA exactly matching to VEP is very high for all categories, particularly synonymous variants and missense mutations (Table 3.8 and Figure 3.7). This underlines a general asymmetry between VEP and IVA with regards to the number of variants annotated as exonic. As described above, even excluding splicing related variants, for which the divergences are particularly large, there is an excess of more than 1,000 stop-gain and more than 23,000 missense variants when using VEP as opposed to IVA for annotating the genomic data.

3.2.3 Patterns of deleterious variation

The CADD score was chosen as a measure of deleteriousness for every SNP recorded in the VCF files. The cut-offs for the analyses presented here were set strictly to reduce the number of miscategorised variants, at ≥ 20 and ≥ 30, respectively representing the most deleterious 1% and 0.1% of the genome. One of the advantages of the CADD is its ability to assign scores to protein coding and non-coding variants.

Using a CADD cut-off of ≥ 20 (hereafter referred to as CADD20) a total of 178,893 potentially deleterious variants were identified in the 382 individuals analysed. Of these CADD20 sites 88,389 (49.4%) lie in exonic regions as defined by Ensembl Release 75 and 90,504 (50.6%) are non-exonic.

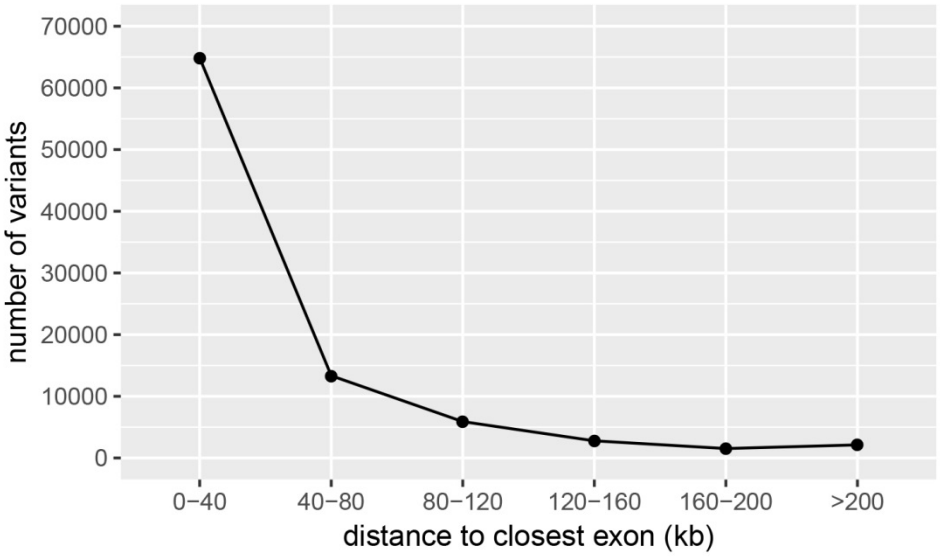


Figure 3.8: This graph shows the distance of non-exonic variants annotated with a CADD score of ≥20 to the next exon. The majority (71.4%) of these sites are within 40 kb or less of the next exon. The intervals up to 200 kb exclude their lower and include their upper boundary.

The variants that are annotated as non-exonic are in relative proximity to the next exon, with a median distance of 15,782 bp (Figure 3.8). This is consistent with theoretical expectations as they are likely to be a) splice site affecting variants on the intron side of an exon-intron boundary and b) different classes of regulatory sites. The former are by definition located nearby exons while important types of the latter such as regulatory eQTLs have been shown to cluster near exonic regions in general and transcription start points in particular (Bryois et al., 2014). For the strictest criterion, CADD30, all 12,747 highlighted variants are exonic. This is not unexpected, as it has been demonstrated previously (Kircher et al., 2014) that the variant types with the highest mean CADD score, such as stop-gain mutations, belong to this class.

Table 3.9: Per individual SNP counts for different classes of putatively deleterious derived variants for each of the 14 macro-groups in the Variant-Based Analysis Set. Abbreviations: CADD20 – variants with a CADD score of ≥ 20, CADD30 – variants with a CADD score of ≥ 30, hom – homozygous genotypes, LoF – variants classified as high confidence loss-of-function variants by the LOFTEE plugin to Ensembl's VEP, Short codes for the macro-groups taken from Table 3.1.

Macro-group	CADD20	CADD20_hom	CADD30	CADD30_hom	LoF	LoF_hom
Afr	7228.3	1116.4	185.2	18.4	116.6	15.5
MiE	5841.7	1340.5	159.5	22.3	97.4	21.7
WEu	5749.3	1337.0	151.6	20.9	95.5	21.6
EEu	5720.4	1339.7	151.0	21.6	96.6	20.5
Vol	5784.7	1333.4	153.2	22.2	97.4	22.3
SoA	5892.0	1349.0	156.4	20.4	98.0	22.6
WSi	5686.3	1447.1	150.0	23.6	96.2	24.0
SSi	5759.4	1441.6	155.8	23.0	98.6	22.6
CSi	5575.5	1587.4	145.4	24.4	95.2	24.3

Macro-group	CADD20	CADD20_hom	CADD30	CADD30_hom	LoF	LoF_hom
NSi	5521.3	1599.4	145.0	24.4	95.6	26.6
SeM	5735.1	1481.8	148.0	23.9	95.6	23.8
SeI	5693.0	1609.8	149.7	26.7	93.6	25.4
SAm	5382.2	1804.2	150.7	36.2	85.5	29.7
Oce	5765.9	1786.1	149.1	30.6	94.5	27.6
whole dataset	5794.5	1446.3	153.1	23.4	97.0	23.0

Only sites where the deleterious allele is derived were considered for further analyses. This reduced the dataset of CADD20 variants to 172,867 (96.6%) and of CADD30 variants to 12,561 (98.5%). Table 3.9 reports the average total and homozygous SNP counts for these two different thresholds.

For the total number of variable sites with a CADD20 annotation most comparisons between the macro-groups yielded statistically significant differences (Figure 3.9, raw data in Appendix C.17). The macro-groups which are not differentiated among each other belong to five clusters

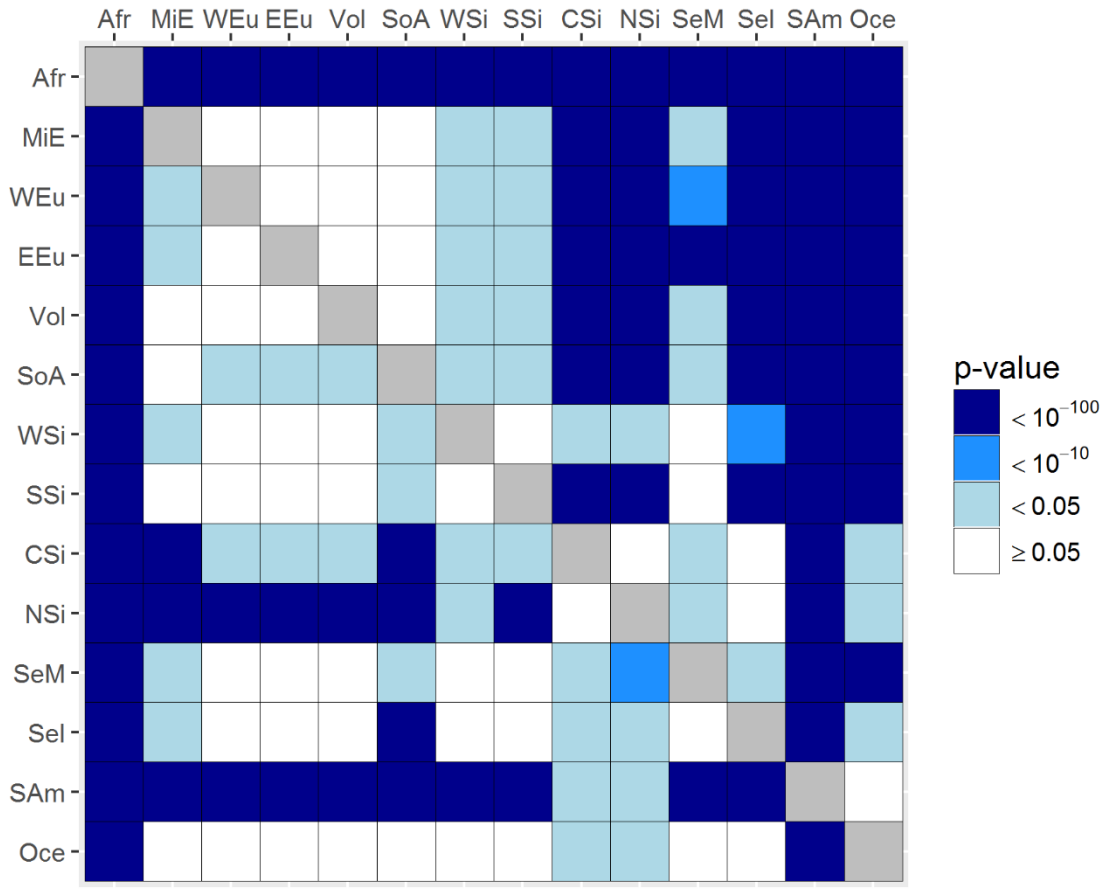


Figure 3.9: Heatmap displaying the outcomes of Tukey's HSD for each pairwise inter-macro-group comparison of the per individual derived SNP counts for CADD20 variants. The lower half contains the values for the total number of SNPs, the upper half for homozygous genotypes only. Cells containing the p-values for comparisons that are not significant are white.

from the highest to the lowest total number of variants as follows: i) Africans, ii) Middle Easterners /South Asians, iii) Western Eurasians and some Eastern Eurasian groups iv) Central and Northeast Siberians v) South Americans. The Oceanians fall in between clusters ii) and iii). If the two Australians, who were shown to be admixed (Pagani et al., 2016), are excluded the remaining Papuans are somewhat less diverse ($\overline{\text{CADD20}}_{\text{Papua}} = 5658.2$) but are still located within the range of most Eurasian groups.

This order of macro-groups seems to reflect the distance from Africa as well as population history after the OOA event. The relationship between these variables will be examined quantitatively below. For the stricter CADD30 cut-off these differences become much less pronounced with only the African group exhibiting a constant excess of CADD30 variants (Figure 3.10, Table 3.9).

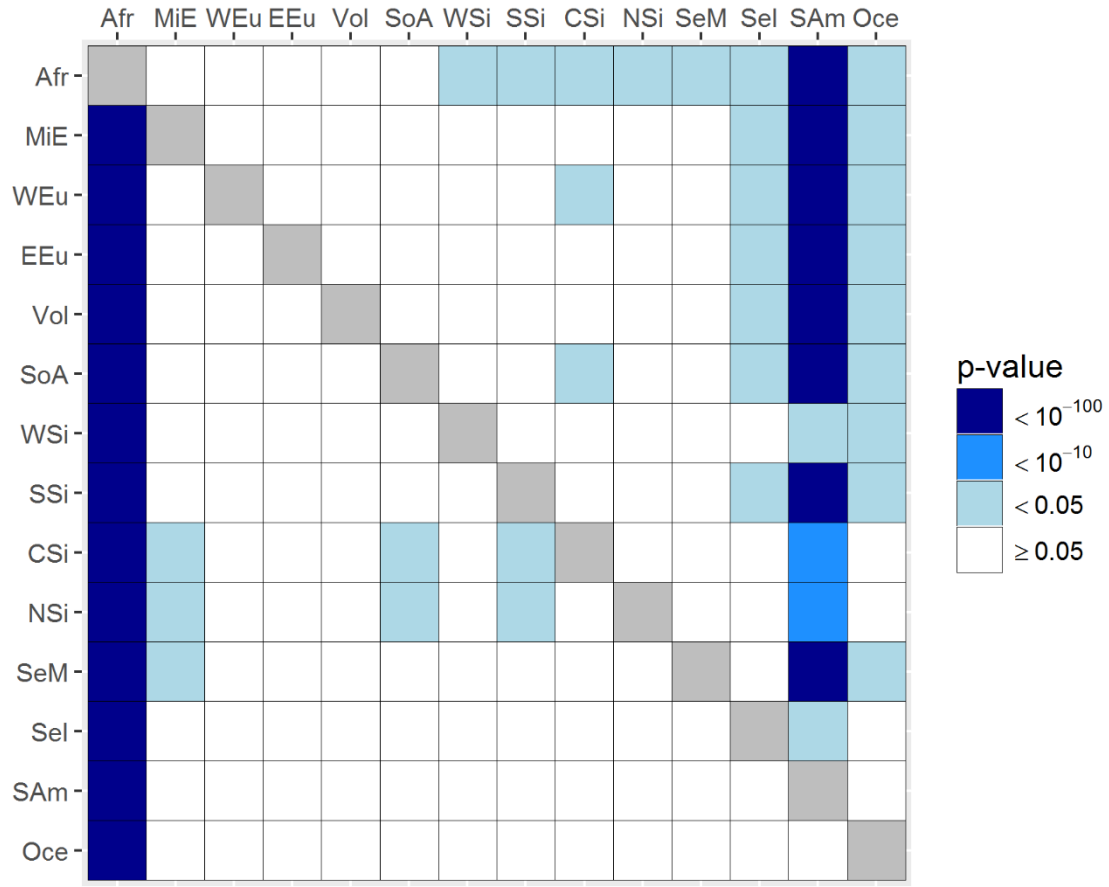


Figure 3.10: Heatmap displaying the outcomes of Tukey's HSD for each pairwise inter-macro-group comparison of the per individual derived SNP counts for CADD30 variants. The lower half contains the values for the total number of SNPs, the upper half for homozygous genotypes only. Cells containing the p-values for comparisons that are not significant are white.

Concerning the number of homozygous CADD20 sites most macro-groups are statistically distinct. Compared to the total CADD20 derived SNP counts the order of macro-groups is reversed, with the Oceanians (if the Papuans are treated separately, this rises to $\overline{\text{CADD20}}_{\text{Papua\_Hom}} = 1942$) and South Americans exhibiting the highest average totals and the Africans the lowest (Figure 3.9, Table 3.9). Furthermore, Western Eurasians and South Asians form a homogeneous cluster. Again, for the putatively more severely deleterious CADD30 variants groups show less differentiation.

However, the Africans have particularly low numbers of homozygous variants whereas the Native Americans are an extreme outlier with a total about twice as high as that observed for the former (Figure 3.10 [raw data in Appendix C.18], Table 3.9). Oceanian and ISEA groups have an excess of CADD30 homozygous variants relative to the other Eurasians (effect still significant for the latter if the Philippine Negrito populations are excluded, data not shown).

Another approach to quantify deleterious variation in the human genome is to focus on variants predicted to lead to a loss of function of the respective gene they are located in. According to the LOFTEE plugin for the VEP an average human genome in the dataset described here contains 97 high confidence (HC) LoF SNPs, of which 23 occur in a homozygous state (Table 19, Appendix C.19 reports the numbers of for different LoF types).

Like the patterns observed for the stricter CADD30 cut-off the total counts of HC LoF SNPs are much less differentiated between macro-groups compared to CADD20 sites (Table 3.9, Figure 3.11 [raw data in Appendix C.20]). Only the African group has a significantly higher number of HC LoF sites than all other non-Africans whereas the Andeans have a significantly lower number. Consistent with the patterns described for the homozygous CADD variants the Africans have a lower number of homozygous LoF sites compared to all other groups. The South Americans exhibit a significant excess of these sites compared to all other groups except the Northeast Siberians, who are also phylogenetically closest to them.

Furthermore, Eastern Eurasian groups that have comparatively low N_e values, such as the Northeast Siberians and ISEA groups, together with the Oceanians, also appear to carry significantly more homozygous HC LoF variants relative to the West Eurasian cluster. As described above, the match rates between VEP and IVA are comparatively poor for LoF variants. However, when only the overlap of HC LoF variants with IVA is considered these numbers improve. 2,073 (75.3%) stop gained and 490 (32.2%) severe splice site disrupting

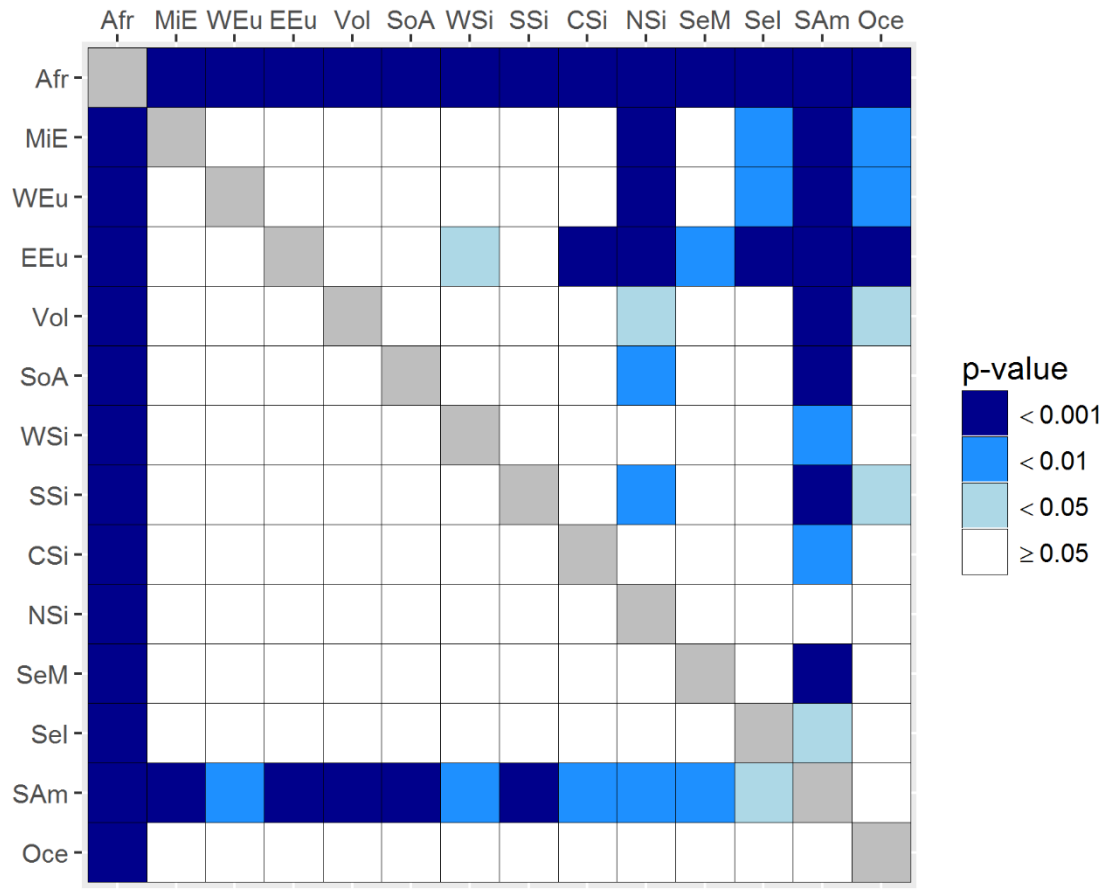


Figure 3.11: Heatmap displaying the outcomes of Tukey's HSD for each pairwise inter-macro-group comparison of the per individual derived SNP counts for HC LoF variants. The lower half contains the values for the total number of SNPs, the upper half for homozygous genotypes only. Cells containing the p-values for comparisons that are not significant are white.

variants from the HC LoF category received the same annotation with IVA. This leads to a higher total match rate of 59.9% relative to only 33.7% for the unfiltered VEP LoF variants. This can be interpreted as evidence that the additional filtering steps applied by the LOFTEE plugin cause an enrichment of LoF inferences that are more robust to differences in transcript sets and annotation algorithms and accordingly more likely to represent a true LoF event.

To further investigate how sensitive these results are to the applied methodology, correlation coefficients between the number of stop-gain variants detected by IVA and those classified as HC stop-gain by LOFTEE/VEP were calculated. The approaches were as follows: i) aggregate/pooled correlation across all individuals, ii) correlation based on macro-group averages, and iii) correlation on an intra-macro group level. The latter statistic is an adaptation of a repeated measures correlation approach (Bakdash and Marusich, 2017). Originally, it was used to compare multiple measurements taken from one individual, here it was utilised to assess

whether the relationship between the stop-gain variant counts inferred by both approaches is consistent in each macro-group.

Generally, there is a good individual-level correlation (r/ρ = 0.68-0.75, $P< 2.2*10^{-16}$) between the stop-gain variant counts obtained for SNPs and homozygous derived sites obtained from IVA and LOFTEE/VEP, regardless of whether a simple linear correlation or Spearman's ranked correlation were applied (Appendix C.21). There is little intra-macro-group deviation from this general pattern (Appendix C.22), even though the repeated measures correlation coefficients are somewhat lower ($r_{\text{stop-gain SNP}} = 0.72$, $r_{\text{stop-gain homozygous derived}} = 0.64$).

Notably, the correlation coefficients improve when average values for each macro-group are compared (r/ρ = 0.84-0.93). One exception is the ranked correlation coefficient for stop-gain SNP counts that was only reported as 0.53 and furthermore non-significant at $\alpha = 0.05$. However, this low correlation of the ranks should not have a strong impact on the relevant results as most macro-groups are not distinguishable by this metric (Figure 3.11). The only effect that is lost using the IVA vs the VEP approach is the significant depletion of the total number of stop-gain SNPs in the South Americans.

In the above paragraphs, differences between the numbers of putatively deleterious variants observed in different macro-groups have been contextualised in terms of geographical proximity, population history and/or distance from Africa. A simple linear regression approach was applied to quantitatively examine how well the distance from I) East Africa (Addis Ababa) and II) Southwest Africa (Windhoek) predicts the number of derived homozygous deleterious alleles with synonymous variants included as control group (Figure 3.12).

For moderately deleterious variants the distance from both African reference points is a good linear predictor ($r^2_{\text{CADD20\_Addis\_Ababa}} \sim 0.65$ and $r^2_{\text{CADD20\_Windhoek}} \sim 0.69$) of the total number of derived homozygous genotypes an individual carries. However, if the more rigorous CADD30 score and HC LoF variant subsets are analysed the relationship becomes much weaker ($r^2 \sim 0.27$-0.32 for all four regressions).

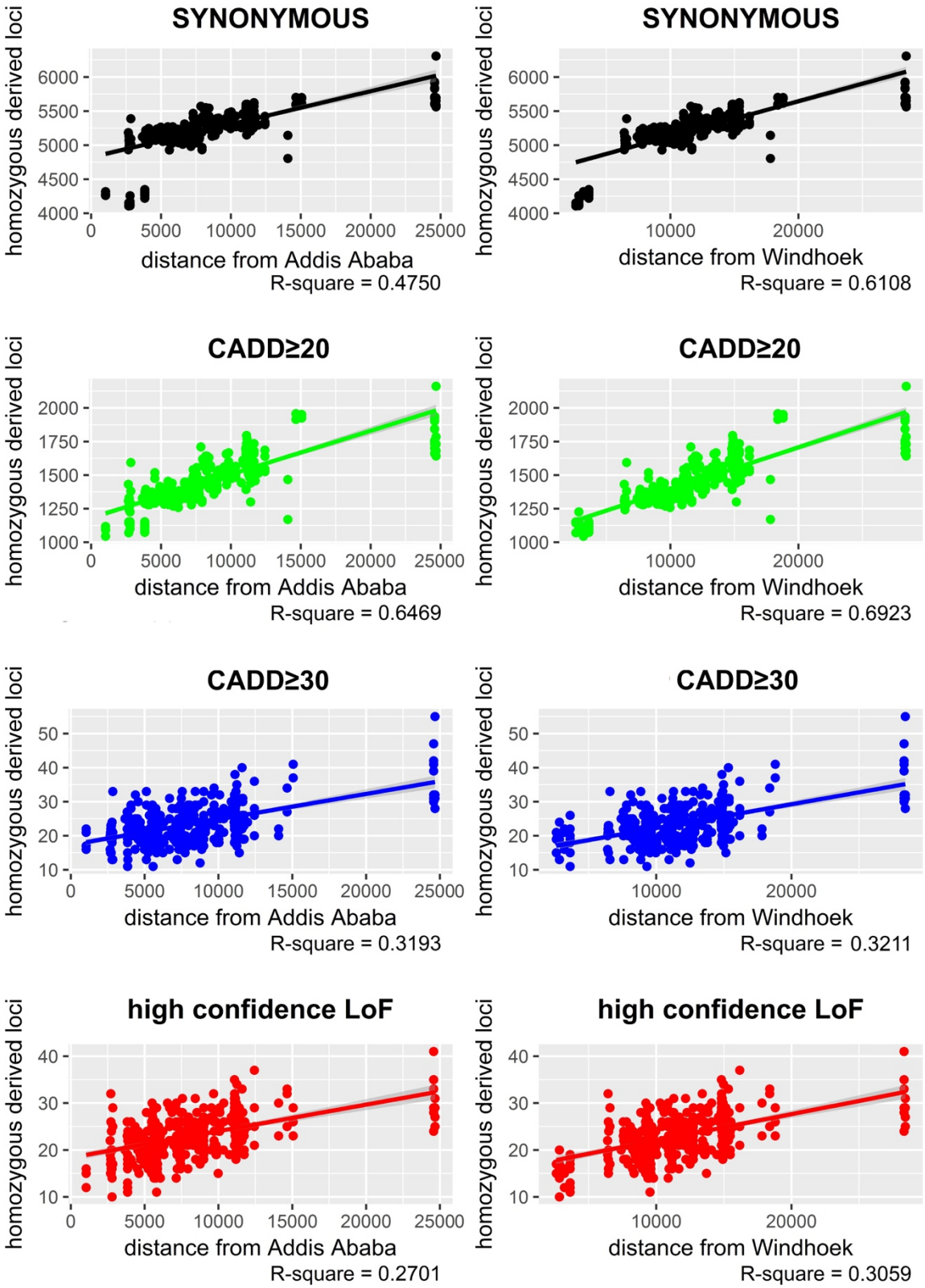


Figure 3.12: Number of derived homozygous sites per individual for different functional variant categories in relationship to the distance from East (Addis Ababa) and Southwest (Windhoek) Africa. For the deleterious variant classes, the coefficient of determination decreases with increasing severity of the variant type. The grey shaded areas surrounding the regression lines indicate the 95% confidence interval for predictions of the number of homozygous derived loci from this linear model.

A seemingly counterintuitive observation is that the correlation of the number of derived homozygotes with distance from Africa is worse for synonymous (annotations based on IVA) variants than for CADD20 variants. Several factors could be contributing to it. Firstly, for the synonymous subset the relationship appears to be partially non-linear. In particular, the Africans have a lower homozygote genotype count than would be expected based on the regression line inferred from the non-Africans, and therefore cannot be adequately captured by a simple linear regression (Figure 3.12).

Secondly, the synonymous variants were treated differently with regards to the inferences of derived homozygous genotypes. For the deleterious variant classes only sites at which the reference allele is equivalent to the ancestral state were recorded, as the focus here was on the accumulation of deleterious variants on the human lineage. However, at synonymous sites for which the ancestral allele was non-reference the genotype calls were adjusted accordingly, i.e. flipped such that the reference derived homozygotes were counted. In this context, it is important that the non-Africans have an overall higher DAF at such sites (0.58) than the Africans (0.5). Briefly, this is because non-Africans are on average more closely related to the human reference sequence and therefore more likely to share derived alleles with it (see also Appendix C.45). Therefore, the per-individual counts of derived reference homozygous loci are also lower in Africans which increases the observed differences between the two groups. This is supported by an additional linear regression with only synonymous variants for which the ancestral allele is the same as the reference were considered, it yielded coefficients of determination closer to those observed for CADD20 sites ($r^2_{SYN\_Addis\_Ababa}$ ~0.68 and $r^2_{SYN\_Windhoek}$ ~0.71).

Finally, all regressions using the distance from Windhoek instead of from Addis Ababa yielded consistently higher coefficients of determination. Using the distance of the sampling location from Africa as the only explanatory variable for the linear regression reflects the approximate impact of the OOA bottleneck and the subsequent initial range expansions; however, it does not incorporate subsequent differences in population history. Therefore, the long-term N_e based on the inferences of this parameter by MSMC across time was included as a predictor (which had been computed for 215 individuals in the Variant-Based Analysis Set). An ANOVA based on the chi-square statistic was used to compare whether the addition of N_e significantly improved the fit of the respective regression models. This is the case for 6/8 models with the exception of CADD30 variants (Table 3.10). Notably, most of the differences in derived

Table 3.10: Determination coefficients obtained from multiple regression analyses aiming to predict the number of homozygous derived sites for different functional classes based on the distance from Africa and the long-term N_e since the OAA event. The addition of N_e was evaluated by an ANOVA using the chi-square statistic; the respective p-value for the comparison of the model pairs is recorded. Note that N_e was only available for a subset of populations. Abbreviations: D_AA – distance from Addis Ababa, D_WH – distance from Windhoek.

	r^2 (D_AA)	r^2 (D_AA +N_e)	P	r^2 (D_WH)	r^2 (D_WH +N_e)	P
Full dataset						
synonymous (total)	0.4750			0.6108		
synonymous (ancestral is reference)	0.6771			0.7074		
CADD ≥20	0.6469			0.6923		
CADD ≥30	0.3193			0.3211		
LoF HC	0.2701			0.3059		
Reduced dataset (individuals for which N_e is available)						
Synonymous	0.5444	0.9009	$7.36*10^{-168}$	0.6754	0.9077	$5.42*10^{-118}$
CADD ≥20	0.6953	0.7667	$7.96*10^{-16}$	0.7489	0.7699	$1.10*10^{-5}$
CADD ≥30	0.4013	0.4027	0.4855	0.4029	0.4059	0.3006
LoF HC	0.3363	0.4062	$5.89*10^{-7}$	0.3786	0.4099	$7.95*10^{-4}$

synonymous homozygotes can be explained by the distance from Africa and the long-term N_e (r^2 ~0.90 for East and r^2 ~0.91 for Southwest Africa as reference points).

While the differences in the number of homozygous derived moderately (CADD20) deleterious variants are relatively well accounted for (r^2 ~0.77) by these two factors used as proxies for neutral processes in population history, the fit is considerably worse than for synonymous variants. Furthermore, the distance from the African reference points and long-term N_e are themselves moderately negatively correlated ($r_{Ne\_Addis\_Ababa}$ = -0.4194, $r_{Ne\_Windhoek}$ = -0.5440) ,

as expected under a serial founder model (Deshpande et al., 2009). However, since the long-term N_e is calculated as the harmonic mean it is less influenced by recent expansions compared to the arithmetic mean and as shown above, considering N_e yields additional value for the explanation of the observed patterns of genomic diversity.

Finally, when derived homozygous genotypes at LoF sites (stop-gain + splice donor/acceptor disrupting) inferred using IVA are considered, the amount of variation explained by the regression analyses is lower than for the HC LoF sites inferred with VEP. The coefficient of determination for the best model (distance from Windhoek + N_e) is only ~0.28 (details reported in Appendix C.23), lower than the maximum r^2 ~0.41 observed for the HC LoF. Most likely this occurs because the RefSeq transcript set on which the IVA annotations are based is more

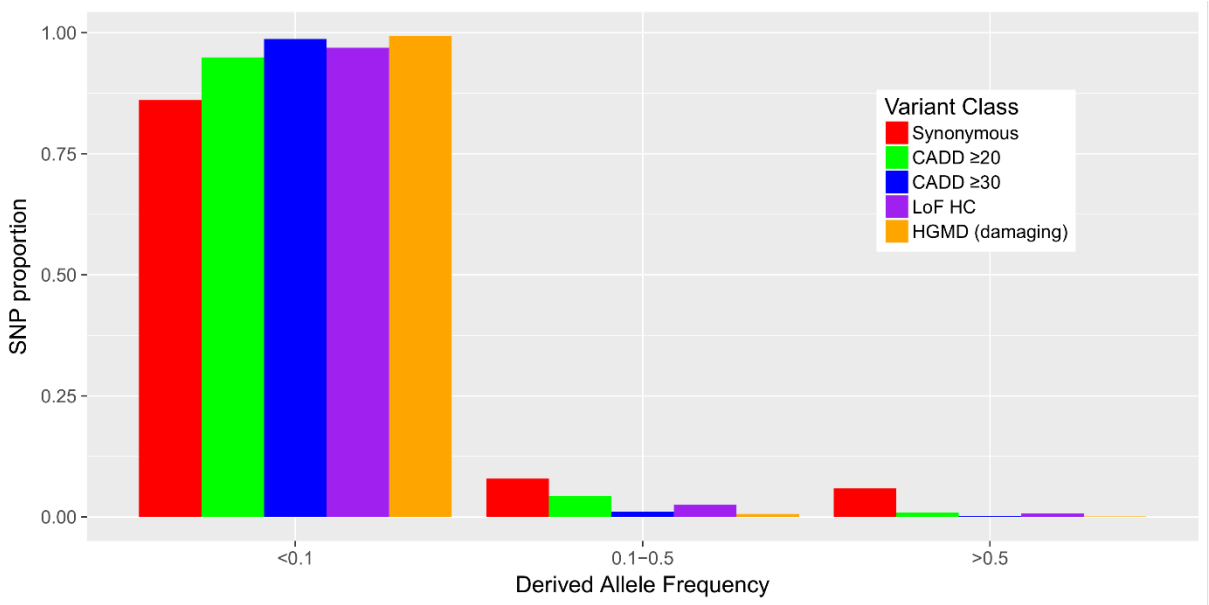


Figure 3.13: Proportions of variants in different putatively deleterious classes by DAF compared to synonymous variants as inferred by IVA.

conservative, whereas as VEP LoF results even in the filtered format that leads to a higher matching rate (see above) contain considerably more variants, of which only a fraction likely has a small or no impact on fitness. Similarly to the analyses of functional variants inferred by IVA by their DAF spectra, the binned spectra for the types of deleterious variation inferred in this subchapter were compared to those observed for synonymous variants (Figure 3.13).

As theoretically expected, all three deleterious variant classes exhibit a significant depletion of high frequency derived variants due to purifying selection. This pattern is most pronounced for the CADD30 variants, while LoF HC variants are in turn more shifted towards rare variation

than CADD20 sites (all differences are significant using the chi-square-test with a Bonferroni-corrected significance threshold of 0.01, $p < 10^{-13}$ for all comparisons).

To further explore inter-macro-group differences with regards to the DAF for deleterious variants the same downsampling approach as described for the outcomes of IVA (see section 3.2.1) was applied. CADD20 variants exhibit patterns that are similar to those observed for synonymous and missense variants inferred with IVA.

Table 3.11: Proportions of variants in different putatively deleterious classes. To make the regional patterns comparable the frequency spectrum for each macro-group was normalised to the sample size of the smallest group (the Oceanians at n=8) by downsampling. Abbreviations: DAF - derived allele frequency, Short codes for the macro-groups taken from Table 3.1

	CADD ≥20			CADD ≥30			LoF HC		
DAF	**<10%**	**10-50%**	**>50%**	**<10%**	**10-50%**	**>50%**	**<10%**	**10-50%**	**>50%**
Afr	0.553	0.409	0.038	0.714	0.268	0.018	0.580	0.389	0.031
MiE	0.480	0.433	0.087	0.691	0.276	0.032	0.526	0.395	0.079
WEu	0.453	0.452	0.096	0.687	0.273	0.040	0.527	0.391	0.082
EEu	0.444	0.459	0.096	0.674	0.284	0.042	0.512	0.407	0.082
Vol	0.450	0.458	0.092	0.678	0.282	0.040	0.536	0.381	0.084
SoA	0.477	0.435	0.088	0.699	0.268	0.033	0.555	0.367	0.079
WSi	0.409	0.485	0.106	0.625	0.332	0.043	0.482	0.419	0.099
SSi	0.445	0.448	0.107	0.673	0.284	0.043	0.515	0.395	0.090
CSi	0.375	0.490	0.135	0.591	0.355	0.054	0.444	0.437	0.119
NSi	0.346	0.512	0.142	0.572	0.379	0.049	0.425	0.447	0.128
SeM	0.444	0.442	0.115	0.689	0.266	0.045	0.532	0.373	0.095
SeI	0.417	0.466	0.117	0.646	0.306	0.048	0.484	0.410	0.106
SAm	0.360	0.472	0.168	0.582	0.328	0.090	0.446	0.395	0.159
Oce	0.403	0.454	0.143	0.581	0.353	0.066	0.518	0.357	0.125

Nearly all comparisons yield evidence of significant differences (chi-square tests with a Bonferroni-corrected significance threshold of $5.495*10^{-4}$ adjusted for 91 comparisons, the same threshold was applied for CADD30 and HC LoF sites) (Table 3.11, Figure 3.14 [raw data in Appendix C.24]).Some West Eurasian groups, e.g. the two European macro-groups are not significantly differentiated, possibly because of their shared population histories. Interestingly, the South Siberians also appear similar to the Europeans, even though they belong to the East Eurasian macro-clade.

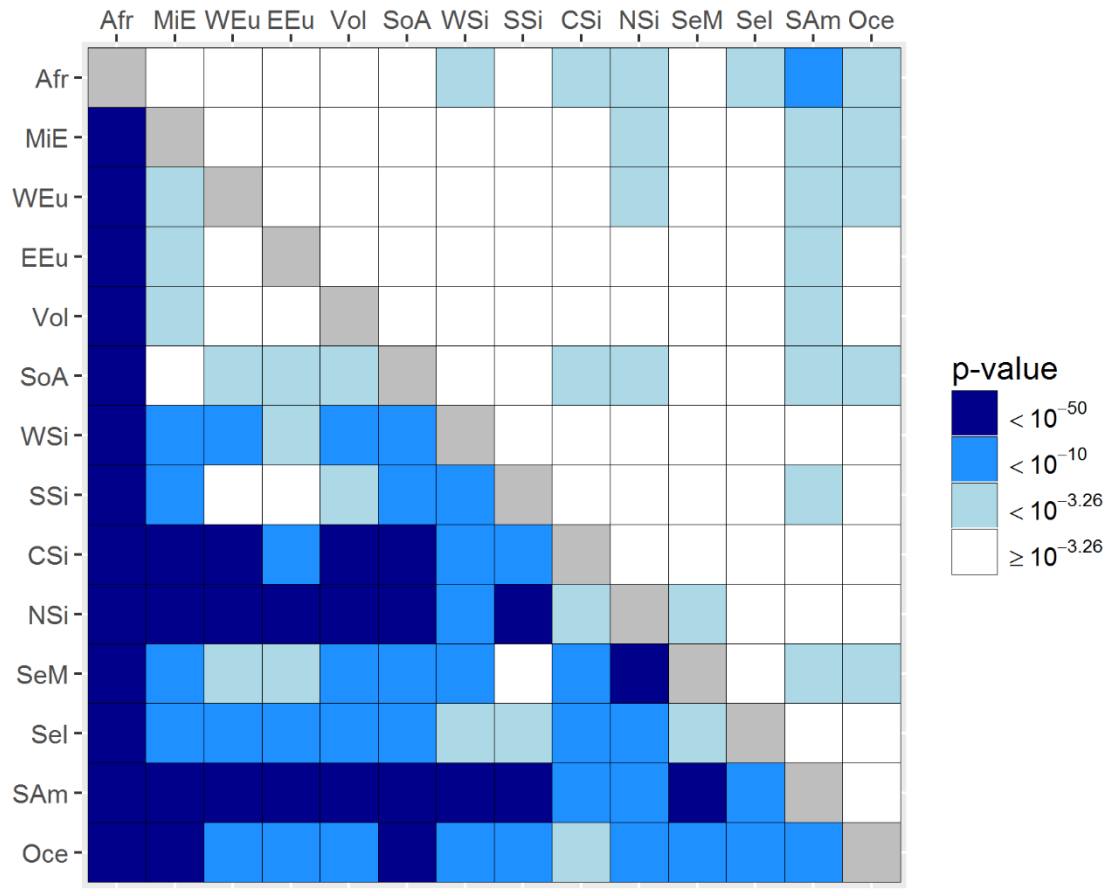


Figure 3.14: Heatmap displaying the outcomes of pairwise chi-square tests comparing the binned (see Table 3.11) DAF spectra between each of the 14 macro-groups. For all groups except the Oceanians these spectra represent the mean (rounded to full numbers) counts obtained from repeated downsampling to n = 8. The upper half contains the values for the DAF spectra of CADD30, the lower half for those of CAD20 variants. Cells containing the p-values for comparisons that are not significant according to the Bonferroni-corrected threshold ($p = 10^{-3.26}$, as there are 91 comparisons each) are white.

These comparisons are based on a relatively low number of bins; therefore, the more fine-grained full DAF spectra were calculated and for a subset of eight macro-groups are plotted in Figure 3.15. Generally, these are consistent with the patterns observed for the binned data. The higher resolution indicates an approximately tenfold increase (1.6% vs 0.16%) in the proportion of the derived fixed moderately deleterious CADD20 variants in the South American group compared to the Africans.

For the more severely deleterious CADD30 variants the binned allele count spectra appear to be more uniform across macro-groups (Table 3.11, Figure 3.14). Macro-groups that exhibit a noticeable shift towards more common derived CADD30 variants are South Americans (chi-square test significant for 8/13 comparisons), North Siberians and Oceanians (for each chi-square test significant for 5/13 comparisons).

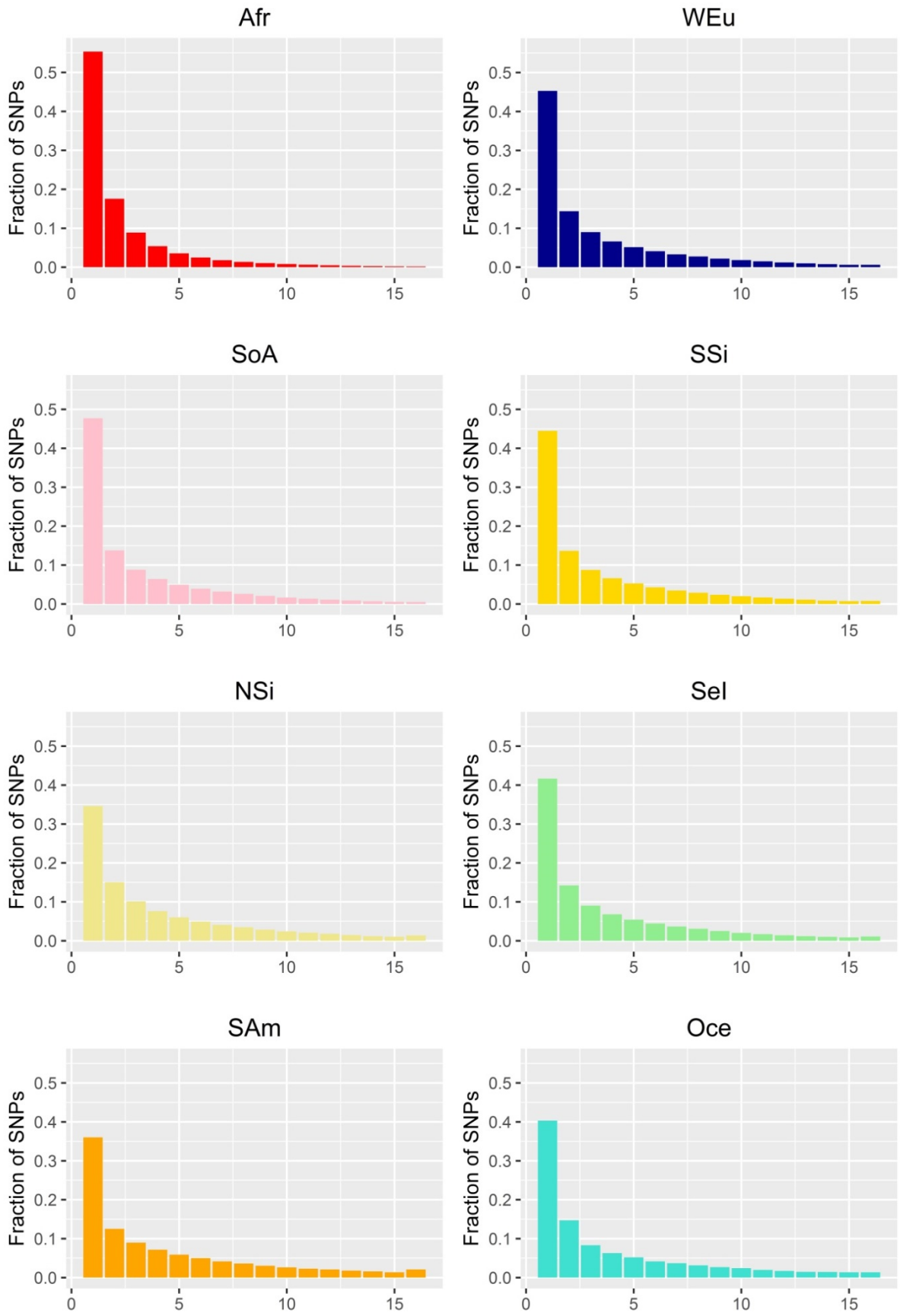


Figure 3.15: DAF spectra of CADD20 variants for eight global macro-groups. Note the shift towards more common deleterious variants in non-Africans. Short codes for the macro-groups taken from Table 3.1.

The frequency patterns for the HC LoF sites (Table 3.11, Appendix C.25) also appear to be relatively similar when compared between macro-groups. Africans display a trend towards more rare variants that is significant relative to all groups from East Eurasia (except South Siberians), the Americas and Oceania.

3.2.4 Application of the $R_{X/Y}$ statistic to infer purifying selection

The $R_{X/Y}$ statistic tests for the accumulation of derived mutations from a particular class of interest in a population compared to another population while using a reference site class as a control (note that the functional annotations for this subchapter were inferred using IVA).

Table 3.12: $R_{X/Y}$ scores of 13 non-African groups relative to Africans They were calculated for missense variants across the genome and from seven different phenotype-based subsets of genes relative to all synonymous variants. Unless indicated otherwise, the phenotype-based subsets of genes are based on GO annotations. A value below 1 indicates a relative excess of missense variants in the non-Africans, while a value above 1 shows a depletion of these. Results for which a bootstrapping procedure gives a p-value below 0.05 are highlighted with an asterisk; two asterisks indicate a p-value below 0.01.
Abbreviations: bact – [immune response against] bacteria, helminth –[immune response against] helminths, genes listed contain at least one SNP whose geographic distribution is strongly correlated with the diversity of helminth species, malaria - Top 50 most significant genes associated with susceptibility to malaria from the MalaCards database, olfac – olfaction, pigment – strict manually curated list of pigmentation genes from GWAS, thermo – thermoregulation, virus – [immune response against] viruses. Short codes for the macro-groups taken from Table 3.1.

Macro-group	all	bact	helminth	virus	malaria	Pigment	thermo	olfac
MiE	0.992	0.976	1.007	0.900*	1.083	0.678**	0.972	1.083
WEu	0.988	1.040	1.003	0.901*	0.926	0.719*	1.001	1.092
EEu	0.986	1.000	1.012	0.902*	0.988	0.682*	0.949	1.076
Vol	1.000	1.002	1.034	0.922	0.922	0.712*	1.012	1.122
SoA	0.993	1.035	1.033	0.895*	0.953	0.713*	1.014	1.130
WSi	0.990	1.034	1.030	0.908	0.878	0.696*	1.017	1.101
SSi	0.995	1.054	1.033	0.926	0.921	0.723*	1.019	1.106
CSi	1.007	1.055	1.040	0.918	0.935	0.713*	0.975	1.108
NSi	1.016	1.050	1.035	0.911	0.878	0.737*	1.093	1.137*
SeM	1.004	1.077	1.045	0.916	0.985	0.729*	1.051	1.134*
SeI	0.997	1.045	1.013	0.874*	1.086	0.722*	0.992	1.171**
SAm	0.998	0.996	1.012	0.879*	0.794	0.757*	1.060	1.134*
Oce	0.994	1.103	1.017	0.904	1.085	0.792	0.991	1.097

Statistically, loads of missense mutations (using synonymous variants as a reference) are indistinguishable across the whole genome in non-African macro-groups compared to Africans (Table 3.12). To test whether genes involved in the response to environmental factors, such as

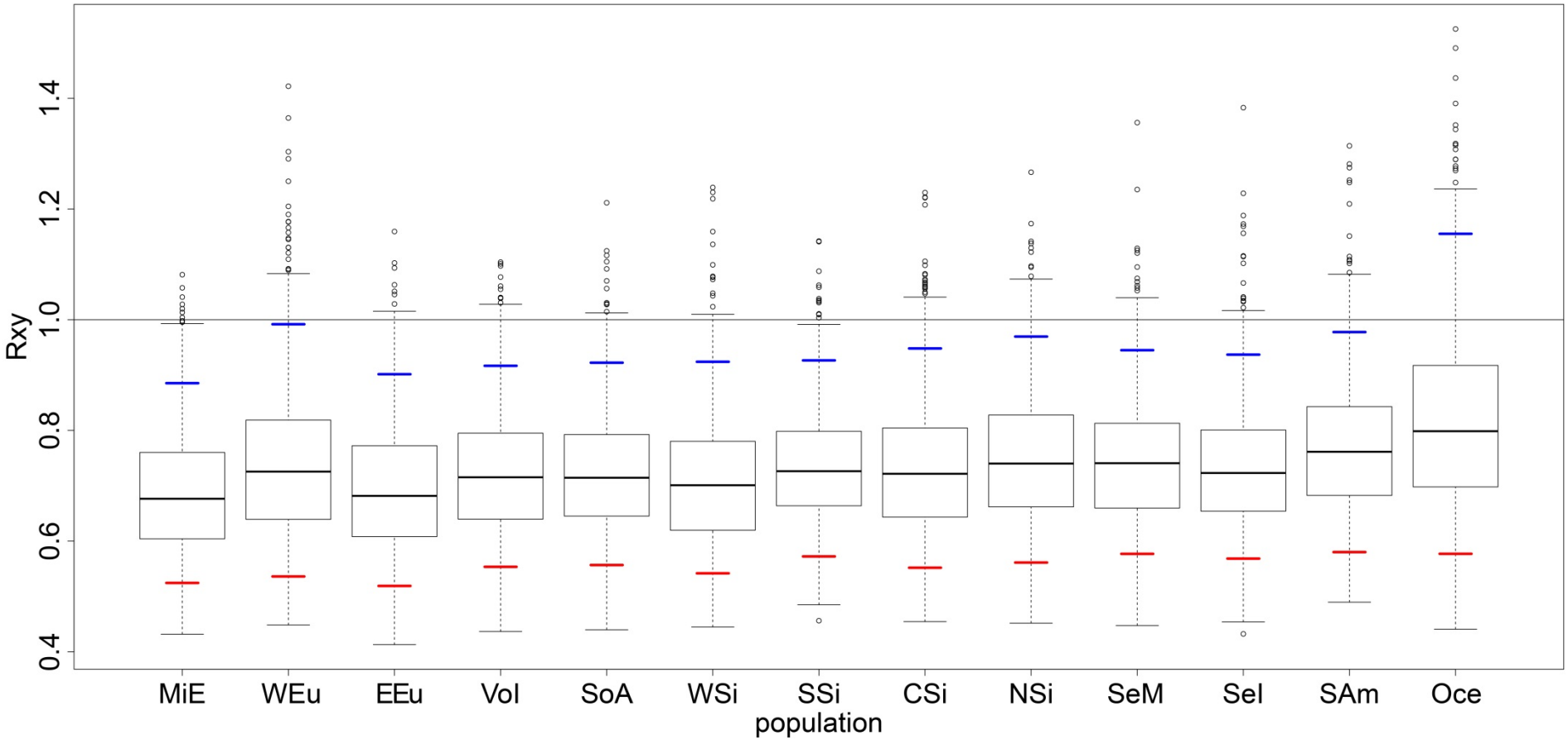


Figure 3.16: Results of 1,000 bootstrap replicates of the $R_{X/Y}$ test for a subset of pigmentation genes highlighted by GWAS (n = 35). The horizontal line provides the African reference (x = 1) against which all other groups are compared. The blue and red marks show the 95th and the 5th percentile of the bootstrap distributions respectively. If the 95th percentile is below 1, then the population shows a significant excess of missense variants in the pigmentation subset relative to the Africans. This is the case for all non-Africans except the Oceanians. Short codes for the macro-groups taken from Table 3.1.

variation in temperature and pathogen exposure, showed signals of purifying selection as a class phenotype-related gene lists were created (Appendices C.9-10).

For three of these classes there are significant differences between the Africans and non-African groups (Table 3.12). The former exhibit a signal of purifying selection in pigmentation genes; this observation is significant compared to all non-Africans except the Papuans (Figure 3.16). There is a quantitatively weaker signal of a similar nature with regards to viral immunity genes where six non-African groups have a significant excess of derived missense variants. In contrast, three East Eurasian groups and Native Americans have a significant depletion of derived missense variants in genes related to olfactory reception. Given that these genes are thought not to be strictly constrained in general the biological significance of this is unclear and it could also be a by-product of genetic drift in these groups that would cause a reduction in the overall diversity of derived missense mutations.

3.2.5 Homozygous LoF variants

LoF sites, particularly when occurring in a homozygous state, can add considerably to our knowledge about gene function. This applies even in the absence of detailed phenotypic information, as the donors were selected based on criteria that make it unlikely that they suffered from severe Mendelian diseases when the samples were taken.

Overall, 4,281 sites in the genome for which at least one individual in the dataset is heterozygous were annotated as HC LoF (reported in a VCF file as Appendix C.26). These LoF sites were found in 3,395 genes. Only sites for which at least one individual carries a homozygous LoF variant were selected for further analysis (n = 369 in 352 genes). According to theoretical expectations this subset should be enriched for redundant genomic elements. This is because if there exist multiple genes which are duplicates of each other, in many cases one intact copy of the duplicated gene would be enough to ensure the presence of the relevant gene product. Then, LoF mutations in the other copies could rise to detectable frequencies as they would not have a negative fitness impact (e.g. Wagner, 1998). To test this prediction information on duplicated genes was retrieved from the Duplicated Genes Database. A total of 102 genes with homozygous LoF variants have at least one tandem duplicate, with a median value of four tandem duplicates. Compared to all genomic elements (including pseudogenes

and non-coding RNAs) catalogued in Ensembl Release 71 that the Duplicated Genes Database is based on this represents a significant enrichment ($X^2 = 93.904$, $p < 10^{-15}$).

When the PANTHER overrepresentation test using annotations from the GO database is applied these genes containing homozygous LoF variants exhibit a significant depletion of genic elements contributing to developmental processes of any kind ($p = 0.0093$) and for the development of anatomical structures in particular ($p = 0.024$) (Appendix C.27 gives the full table of outcomes). The same also applies for genes coding for any type of protein that binds to other proteins ($p = 0.0276$). However, more of these genes than expected are involved in interactions with cations ($p = 0.0177$), especially metal ions ($p = 0.00664$).

One limitation of this overrepresentation approach is that it does not explicitly consider differences in length of the coding sequence (CDS) of each gene. However, there is little evidence that in this dataset CDS length is the main factor explaining whether a gene carries a homozygous LoF detected (for details see Appendix C.28). Recent large exome sequencing studies have focussed on creating a catalogue of rare homozygous LoF (rhLoF) variants in human populations. As naturally occurring knockouts they are important to understand which genes are tolerant for homozygous LoF variants and to assess the impact of rare variants on the phenotype more generally. The focus of the following paragraphs is on novel variants of this type in the data analysed here, as the latter comprises regions only sparsely covered by other major sequencing projects, ISEA and Siberia in particular.

To assess the occurrence of these variants in the 382 genomes of interest all sites with a DAF ≥2% were removed, which left 116 sites, each located in a different gene. In the next step variants in genes previously described as containing rhLoF were excluded. The set of genes was compared to an aggregated list reporting such genomic elements derived from four large scale exome sequencing studies comprising >170,000 individuals (Sulem et al., 2015; Lek et al., 2016; Narasimhan et al., 2016; Saleheen et al., 2017). This filtering step reduced the dataset to 56 variants. Subsequently, the remaining variants were compared to the global WGS data generated by the 1000 Genomes Project (n = 2,504) (The 1000 Genomes Project Consortium, 2015) and the SGDP (n = 279, publicly available subset) (Mallick et al., 2016). Variants were discarded if they were already reported as occurring in a homozygous state in either of these datasets. In a final step all these variants were manually checked against the ExAC Browser (Karczewski et al., 2017) that also records variants that were not used for the LoF-analyses in the main ExAC paper (Lek et al., 2016) as they are covered in less than 80% of all individuals

in this dataset. This led to the exclusion of three further variants (overlaps resulting from all filtering steps described here are displayed in Appendix C.29).

The resulting set of previously unreported rhLoF variants consists of 34 variants (Appendix C.30). If the two variants detected in African Pygmy samples first published by Lachance et al. (2012) and three variants in Papuans that would potentially have been reported if the data from Malaspinas et al. (2016) on the genomic diversity of Oceanians are included in the filters, this leaves 29 variants, 28 of which were detected in populations from Eurasia.

With regards to the distribution of variation between continental groups, almost two thirds (n = 18) of these Eurasian rhLoF variants are observed in the Central Siberian, Northeast Siberian and ISEA groups. Compared to all other Eurasians (n = 239) these continental populations taken together (n = 101) exhibit a significant excess of rhLoF carrying individuals (Fisher's exact test, $p = 5.251*10^{-4}$). The subpopulations within these clusters accounting for most of the novel homozygotes are the Kankanaey (see section 2.3.1) and the Evenks.

To evaluate the functional importance of these variants the LoF-tolerance of the genes they are located in was assessed based on scores derived from an algorithm comparing the gene-wise theoretically expected LoF counts to those observed in the >60,000 exomes from the ExAC project. These rhLoF-containing genes are neither enriched for LoF-tolerance ($X^2 = 0.54484$, $p = 0.4604$) nor depleted for LoF-intolerance ($X^2 = 1.2937$, $p = 0.2554$), even though these non-significant results are at least partially caused by the small sample size of the rhLoF subset. The CADD scores retrieved for the novel rhLoF variants have a median of 23.75 indicating that they belong to the ~0.42% most deleterious variants in the genome. This is considerably lower than the median CADD of 37 observed for nonsense variants in disease-causing genes (Kircher et al., 2014). The only variant in this subset reaching the latter threshold is located on chromosome 19 in the *ZNF568* gene encoding a ZNF protein. The severity of the resulting phenotypic effect is difficult to estimate, given that the carrier was healthy enough to serve as an adult donor of genetic information and that this gene is theoretically predicted to be LoF-tolerant and belongs to a family of functionally very similar genes (Appendix C.31).

Only one of the variants highlighted here has an entry in HGMD (CS982369). It is observed in a Dusun individual from Brunei and is situated in *SLC14A1*. The resulting protein product is known as the Kidd glycoprotein and is located on the erythrocyte membrane, where it plays a role in urea transport across the former and in the collecting duct system of the kidney that increases the concentration of urea in urine. However, its absence does not lead to a severe

disease phenotype, as another homozygous splice site mutation in *SLC14A1* leading to its loss on a protein level has been seen in Polynesians (Irshaid et al., 2000), which was not recorded in the LoF catalogue derived from existing exome sequencing projects. The affected individuals are unable to maximally concentrate urine but their erythrocytes have a normal morphology and lifespan (Dean, 2005).

3.2.6 Variants related to Mendelian disease

The final class of deleterious mutations that will be examined here are those documented in the HGMD. Mutations from this database are only reported if they were thought to be disease-causing by the HGMD authors, i.e. they received the DM flag and the deleterious allele was derived. These steps reduced the dataset to a total of 1,596 mutations (~21.1% of originally 7,851 variants). The DAF spectrum of these mutations is similar to that obtained for CADD30 variants (Table 3.12) ($X^2 = 4.818$, $p = 0.090$).

Each individual in the global dataset carries on average ca. 16.3 HGMD variants, of which 1.8 are present in a homozygous state (Table 3.13). Africans exhibit the highest number of SNPs of this type, likely driven by their higher genome-wide diversity compared to non-Africans. The total of homozygous site counts is very low as expected for disease-causing variants and has therefore to be interpreted with caution. However, generally West Eurasian groups exhibit the lowest number of HGMD homozygotes, elevated numbers are found in Eastern Eurasians and Native Americans.

Table 3.13: Per individual SNP counts of derived mutations that are thought to be causally related to a disease phenotype as indicated by the HGMD for the 14 macro-groups used in the variant-based analyses. This also includes polygenic diseases. Short codes for the macro-groups taken from Table 3.1.

Macro-group	HGMD	HGMD_hom	HGMD_allele_count
Afr	20.2	2	22.1
MiE	17.5	1	18.5
WEu	17.1	1	18
EEu	16.1	0.9	17
Vol	17.3	1.2	18.5
SoA	17.8	1	18.8
WSi	15.2	1.2	16.5
SSi	16.4	2.4	18.8
CSi	15.3	2.5	17.7
NSi	14.2	2.4	16.6
SeM	16.4	2.6	19

Macro-group	HGMD	HGMD_hom	HGMD_allele_count
SeI	15.5	2.7	18.2
Sam	13.8	2.6	16.4
Oce	13.8	2.9	16.6
whole dataset	16.3	1.8	18.1

It is difficult to evaluate the potential clinical significance of the HGMD-variants in absence of phenotypical data from the individuals analysed here as these include many loci putatively involved in complex polygenic diseases.

Therefore, a panel of 513 autosomal SNPs that have high penetrance and underlie dominant or recessive Mendelian childhood disorders was retrieved from Chen et al. (2016). A subset (n = 20) of these sites was detected in the Variant-Based Analysis Set. However, they almost exclusively occur heterozygously and as the relevant (mostly metabolic) diseases caused by mutations at these 20 sites are recessive the latter are very unlikely to lead to a disease phenotype (Appendix C.31). A total of 31 individuals are heterozygous for at least one of these sites: there is one homozygote finding described in more detail below. The resulting total carrier burden for such mutations in the Variant-Based Analysis Dataset is 0.086 alleles per individual. Half (n = 16) of the individuals carrying Mendelian disease alleles are of European ancestry, indicating that the carrier burden reported for the total dataset is most likely an underestimate. The roughly estimated prevalences (Appendix C.31) are generally lower than those previously reported. This is to be expected, as the panel of populations in the Variant-Based Analysis Dataset is in its composition very different from the datasets analysed in the medical literature. However, there is good agreement with regards to the general trends: recessive Mendelian diseases that are more common according to previous studies tend to have more heterozygous carriers in the dataset presented here.

The most relevant single finding occurred in a Murut individual (Murut13) who carries a homozygous G to A transition on chromosome 7 at position 117,199,683 (NM_000492.3:c.1558G>A) resulting in a change from valine to isoleucine. This variant lies in the *CFTR* gene, mutations in which are causative for cystic fibrosis. Given the severity of the expected phenotype and the recruitment of randomly chosen healthy donors, it seems unlikely that this individual suffered from overt genetic disease when the sampling was done.

3.2.7 Variant findings by maximum allelic differentiation and other positive selection tests

While the previous sections of this chapter were concerned with exploring interpopulation differences across classes of functional and deleterious sites these final paragraphs will focus on common variants showing strong allelic differentiation between ancestry groups.

For the ΔDAF and the other site-specific selection tests only samples overlapping the Selection Subset (Appendix C.4) were considered, i.e. Oceanians and South Americans were excluded. The rationale was twofold, a) these groups have low sample sizes, and b) to ensure consistency when the analyses are compared to selection results obtained by colleagues who only worked on the Selection Subset. It was ascertained that there were no fixed missense or nonsense allele frequency differences between those two groups and the other 12 macro-groups (data not shown).

When applying the ΔDAF on missense variants (functional annotations in sections 3.2.7 and 3.2.8 based on IVA unless explicitly stated otherwise) for all top 20 most differentiated sites the lowest DAF is observed in Africans (Table 3.14). Besides the known variants in pigmentation, ear wax and hair morphology genes, the sites showing high differentiation between Africans and non-African groups lie in genes involved in immunity (*ACKR1, DUOX2*), obesity and diabetes (*ALMS1*), ciliogenesis (*TMEM216*) and motor activity (*MYO18B*). Those that are most differentiated among non-African groups again included genes involved in ciliogenesis (*PCDH15, B9D1*), the formation of muscular tissue (*TTN, FLNB*) and

Table 3.14: Most differentiated missense variants by population groups sorted in descending order by maximum ΔDAF, a) top 20 sites for all comparisons, b) top 20 sites for all comparisons between non-African populations. To account for LD if two or more "most differentiated" SNPs were located in the same 200-kb window only one of the signals is reported here. * = The asterisk indicates that this site was highlighted as an outlier (>5 SD) by the DIND analyses. Short codes for the macro-groups taken from Table 3.1. Chr - Chromosome

Chr	Position	dbSNP ID	Max ΔDAF	Pair	Gene	Gene function
All populations						
16	48258198	rs17822931	1.00	Afr-CSi	*ABCC11*	ear wax type
15	48426484	rs1426654	0.98	Afr-EEu	*SLC24A5*	pigmentation
14	77843814	rs11844594	0.98	Afr-SeM	*SAMD15*	-
2	109513601	rs3827760	0.97	Afr-CSi	*EDAR*	hair thickness
2	73651967	rs3813227	0.96	Afr-CSi	*ALMS1*	insulin, obesity
5	33951693	rs16891982	0.95	Afr-EEu	*SLC45A2*	pigmentation
1	159175354	rs12075	0.94	Afr-SeI	*ACKR1*	malaria resistance
17	73565171	rs1671021	0.94	Afr-NSi	*LLGL2*	cell division
22	26422980	rs2236005	0.93	Afr-SeI	*MYO18B*	motor activity

Chr	Position	dbSNP ID	Max ΔDAF	Pair	Gene	Gene function
10	15145855	rs15772	0.93	Afr-CSi	*RPP38*	RNase
11	61165741	rs10897158	0.93	Afr-WSi	*TMEM216*	ciliogenesis
17	39659913	rs9891361	0.93	Afr-Vol	*KRT13*	cytokeratin
16	48122582	rs7193955	0.92	Afr-CSi	*ABCC12*	multi-drug resistance
16*	31099011*	rs11150606*	0.92	Afr-CSi	*PRSS53*	endopeptidase
1	16042766	rs12091750	0.90	Afr-SeM	*PLEKHM2*	lysosome localisation; salmonella pathogenesis
9	14722477	rs3747532	0.90	Afr-CSi	*CER1*	morphogen activity
9	127262802	rs1110061	0.90	Afr-rest	*NR5A1*	transcriptional activator (sexual development)
12	65269047	rs939875	0.90	Afr-rest	*TBC1D30*	GTPase activating protein
7	99081730	rs6962772	0.89	Afr-NSi	*ZNF789*	potential transcription factor
15	45392075	rs269868	0.89	Afr-MiE	*DUOX2*	oxidase (thyroid hormone synthesis and antimicrobial defence)
Non-African populations						
15	48426484	rs1426654	0.98	SeI-MiE	*SLC24A5*	pigmentation
2	109513601	rs3827760	0.97	WEu-CSi	*EDAR*	hair thickness
16	48258198	rs17822931	0.96	MiE-CSi	*ABCC11*	ear wax type
5	33951693	rs16891982	0.95	SeM-EEu	*SLC45A2*	pigmentation
16	31099011	rs11150606	0.92	MiE-CSi	*PRSS53*	Endopeptidase, hair morphology
4	38798648	rs5743618	0.87	SeM-EEu	*TLR1*	immunity
3	10302056	rs2241314	0.82	CSi-WEu	*TATDN2*	Putative DNase
10	55955444	rs4935502	0.82	WEu-CSi	*PCDH15*	stereocilia, cell adhesion
19	1003172	rs2240154	0.82	MiE-NSi	*GRIN3B*	glutamate receptor
16*	89986154*	rs885479*	0.82	MiE-NSi	*MC1R*	pigmentation
11	120107411	rs882856	0.81	SeI-WEu	*POU2F3*	keratinocyte differentiation
15	63937209	rs2229749	0.80	MiE-CSi	*HERC1*	membrane transport
15	40581543	rs936212	0.80	WEu-NSi	*PLCB2*	taste perception, ciliogenesis
3	58118555	rs12632456	0.80	Vol-SeI	*FLNB*	actin binding
17	19247075	rs4924987	0.79	SeI-EEu	*B9D1*	ciliogenesis, hedgehog signalling pathway
3	44692564	rs2272044	0.79	MiE-SeM	*ZNF35*	cell differentiation, spermatogenesis
3	130368069	rs7614116	0.79	NSi-EEu	*COL6A6*	collagen, cell binding
2	179644855	rs10497520	0.78	NSi-MiE	*TTN*	muscle assembly
3	133941320	rs1131262	0.78	NSi-WEu	*RYK*	transmembrane signalling

Chr	Position	dbSNP ID	Max ΔDAF	Pair	Gene	Gene function
16*	31088625*	rs749670*	0.78	CSi-SoA	*ZNF646*	lipid metabolism, bile synthesis (transcription factor, influences the expression of *HSD3B7* according to GTEx)

immunity (*TLR1*). Among the ciliogenesis genes, the rs4935502 variant in *PCDH15* has been previously detected as a selection candidate gene in East Asians (Sabeti et al., 2007). The gene is necessary for normal cochlear and vestibular functions in humans and mice and has also been associated with animal echolocation (Shen et al., 2012).

Notably, for the majority (18/20) of the most highly differentiated genes among non-Africans the variation in frequency patterns is consistent with the clustering of macro-groups by their reconstructed demographic history (Figure 3.17), one exception being the SNP rs7614116 in *COL6A6* that has low and high frequency populations both among East and West Eurasian regional groups (Appendix C.32). The placement of a gene on the branch preceding the population split of groups sharing a similar DAF does not imply that a putative selective sweep occurred in this common ancestral population, even though it provides an estimate for the upper boundary for such a hypothetical event.

However, the phylogeny displayed in Figure 3.17 does not contain information on intra-group diversity and underestimates the complexity of gene flow events as the number of migrations was limited to two for the sake of clarity. The former especially affects South Asians, whose different subgroups are known to exhibit varying proportions of deeply diverged ancestries (Reich et al., 2009) and the Australasians, whose relationship to the other Eurasians is still debated (see section 1.4.2).

Accordingly, the known selection target rs1426654 in *SLC24A5,* where the derived allele strongly correlates with a lighter skin tone (Basu Mallick et al., 2013), is shared between West Eurasians and South Asians. With a finer resolution it becomes clear that in individuals from Northern and Western India (known to carry more West Eurasian ancestry) the derived allele is almost fixed (93%), whereas in South Indians and Austro-Asiatic speakers its frequency is only 33% (Appendix C.33).

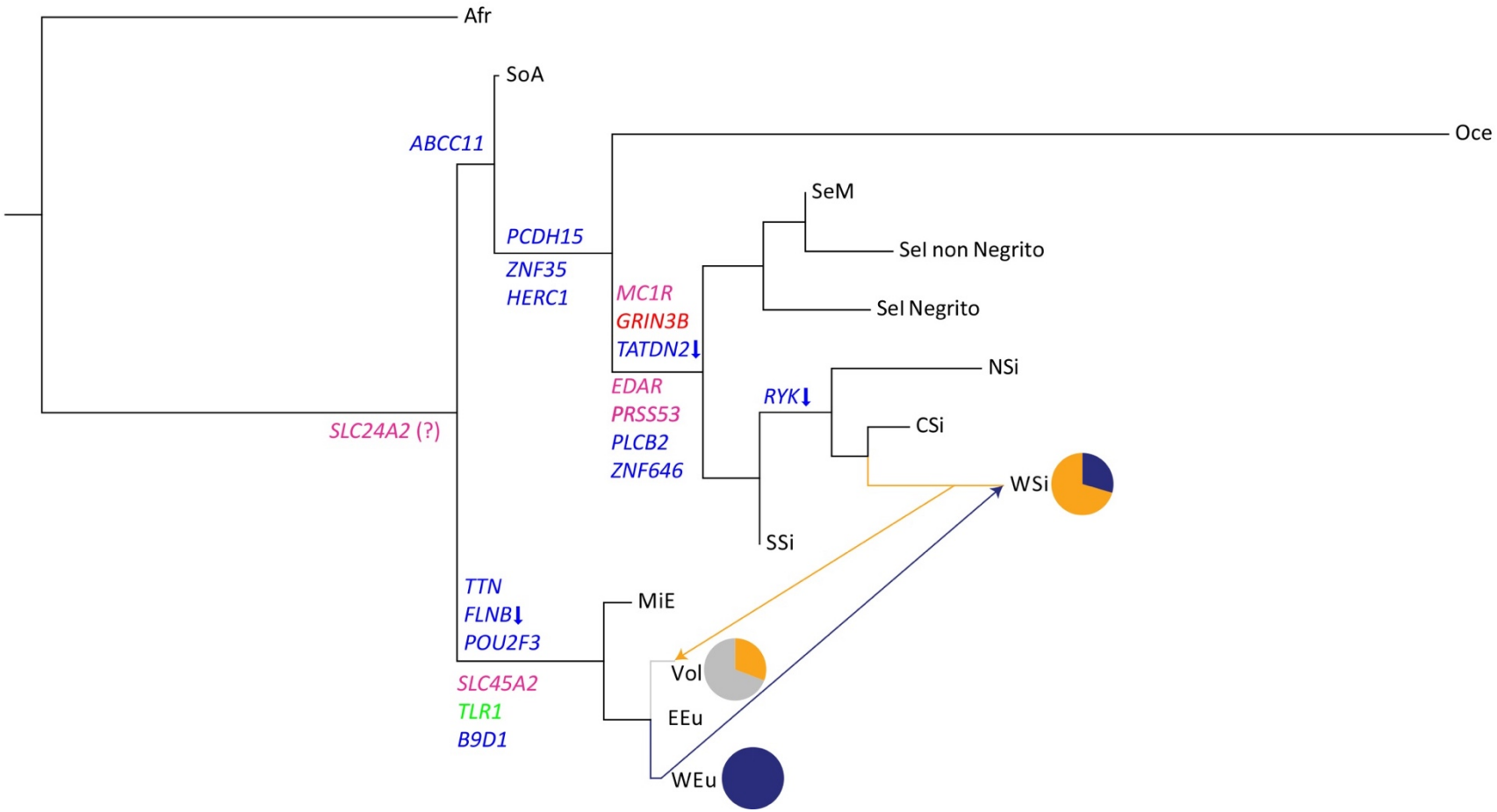


Figure 3.17: TreeMix analysis involving 14 macro-groups with two migration edges; 19 genes containing highly differentiated missense SNPs are highlighted. The branch labelled with a specific gene name is the one preceding the split of all the macro-groups sharing a particular ΔDAF signal due to a SNP in the respective gene. The colours chosen for the gene names reflect their putative function: blue – miscellaneous, green – immunity, pink – morphology and pigmentation, red – nervous system. Downward arrows indicate a high DAF in Africans and a DAF decrease characteristic for the respective non-African cluster. The Native Americans are not included and the ISEA group is separated into Negrito and non-Negrito groups. Short codes for the macro-groups taken from Table 3.1.

In the case of ten out of the twenty most differentiated missense mutations in non-Africans the derived allele is absent in the African group suggesting they are *de novo* mutations among non-Africans, as opposed to the other ten variants, the high ΔDAF of which can be interpreted as potentially resulting from selection on standing variation.

A further subcategory of functional variants that was investigated using the ΔDAF approach were stop-gain mutations. More than half (12/21) of the most differentiated stop-gain mutations segregate between Africans and any other non-African groups. In contrast to the missense mutations the majority of highly differentiated nonsense variants do not show allele frequency patterns consistent with the phylogenetic clustering of populations (Figure 3.17) regardless of whether Africans are excluded (Appendix C.34). Three of the nonsense variants are absent in the African samples analysed here (rs5758511, rs2270002 and rs16930998) and very rare in the African populations from phase 3 of the 1000 Genomes Project, their frequency increase therefore postdates the OOA event.

The results also include a well-documented (Balasubramanian et al., 2011) East Asian variant in the *ZAN* gene that codes for a protein known to play a role in the specificity of the sperm adhesion to the egg's zona pellucida (Tardif et al., 2010). Three out of the top 20 highly differentiated nonsense variants are in olfactory receptor genes, confirming the results of some earlier studies (MacArthur et al., 2012; Gudbjartsson et al., 2015), which reported an excess of common nonsense variants in this gene class.

Table 3.15: Most differentiated stop-gain variants by population groups sorted in descending order by maximum ΔDAF, a) top 20 sites for all comparisons, b) top 20 sites for all comparisons between non-African populations, these only include 4 sites, as all other sites have already been highlighted in a). To account for LD if two or more "most differentiated" SNPs were located in the same 200-kb window only one of the signals is reported here. * = The asterisk indicates that this site was highlighted as an outlier (>5 SD) by the DIND analyses. Short codes for the macro-groups taken from Table 3.1. Chr – Chromosome, NMD-variants are expected to lead to nonsense-mediated mRNA in at least one transcript they are part of.

Chr	Position	dbSNP ID	NMD	Max ΔDAF	Pair	Gene	Gene function
All populations							
11	56431216	rs11228710	No	0.62	Afr-NSi	*OR5AR1*	olfaction
7	100371358	rs2293766	No	0.61	Afr-SeI	*ZAN*	sperm binding to egg
9	125391241	rs1476860	No	0.60	Afr-NSi	*OR1B1*	olfaction
5	1240757	rs7447815	No	0.58	NSi-SeI	*SLC6A18*	neurotransmitter transport
17	4803711	rs35400274	No	0.54	Afr-WSi	*C17orf107*	-

Chr	Position	dbSNP ID	NMD	Max ΔDAF	Pair	Gene	Gene function
12	40834955	rs10784618	No	0.54	NSi-SeM	*MUC19*	ocular mucus homeostasis (?)
22	42336172	rs5758511	No	0.53	Afr-SeM	*CENPM*	cell division
19	35719020	rs541169	No	0.50	CSi-Vol	*FAM187B*	-
17	4461748	rs7215121	No	0.49	SeI-WEu	*GGT6*	glutathione (anti-oxidant) synthesis
11	5444136	rs2647574	No	0.48	SeI-WEu	*OR51Q1*	olfaction
11	63057925	rs1790218	Yes	0.46	CSi-Vol	*SLC22A10*	membrane transport
1	152323132	rs12568784	No	0.45	SeM-WSi	*FLG2*	epithelial cornification
17	72588806	rs545652	No	0.45	Afr-rest	*C17orf77*	-
1	55251314	rs2270002	No	0.44	Afr-NSi	*TTC22*	protein binding
11	5462702	rs16930998	No	0.41	Afr-SeM	*OR51I1*	olfaction
21	35334566	rs766425	No	0.41	Afr-WEu	*LOC400863*	-
5*	74965122*	rs34358*	No	0.40	NSi-SoA	*ANKDD1B*	protein binding/ signal transduction
17	74077797	rs1043149	Yes	0.39	Afr-NSi	*ZACN*	membrane transport
17	33772658	rs8072510	Yes	0.39	CSi-SoA	*SLFN13*	ATP-binding
1	20501582	rs12139100	No	0.38	Afr-NSi	*PLA2G2C*	inactive phospho-lipase (probable)
15	55722882	rs57809907	Yes	0.38	Afr-rest	*DNAAF4*	neuronal migration
Non-African populations							
11	62848487	rs11231341	No	0.38	Vol-WSi	*SLC22A24*	membrane transport
16	81183325	rs59980974	No	0.34	NSi-WEu	*PKD1L2*	membrane transport
4	15482360	rs1861050	No	0.33	MiE-SeM	*CC2D2A*	ciliogenesis, morphogenesis (?)
3	53899276	rs1043261	No	0.32	SeI-WEu	*IL17RB*	Immunity

Other prominent categories of genes highlighted in the non-African comparisons by the top 20 ΔDAF nonsense mutations are membrane transport (*SLC22A10, SLC22A24, ZACN* and *PKD1L2*), immunity (*IL17RB*) and the formation of ectodermal tissues, e.g. neuronal migration (*DNAAF4*) and epidermal cornification (*FLG2*). The nonsense variants in the two latter genes have been previously associated with disorders. The polymorphism rs57809907 in *DNAAF4* was shown to be correlated with the occurrence of dyslexia in a Finnish cohort (Taipale et al., 2003), however follow-up studies were unable to replicate this finding (Tran et al., 2013), while rs12568784 in *FLG2* was associated with more persistent atopic dermatitis in a cohort of African-American children suffering from the condition (Margolis et al., 2014). Furthermore,

the highly differentiated nonsense-variant containing genes are significantly enriched for having one or more tandem duplicates from the Duplicated Genes Database compared to all genomic elements (9/25, $X^2 = 11.87$, $p = 5.7*10^{-4}$).

DIND scores for all top 20 variants in the ΔDAF analyses were generated to test for cases of significantly lower diversity of the derived haplotype. Four loci were found in >5 SD range for the DIND. The first is a well-known missense variant (rs885479) in the *MC1R* gene associated with light skin pigmentation in East Asians (Pneuman et al., 2012) and found at a high DAF in all East Eurasian groups (Appendix C.32). It has been highlighted as a potential target of positive selection in previous studies (Coop et al., 2009). Variation at this site has also been linked to the diagnosis of depression (Wu et al., 2011). A similar pattern can be observed at rs749670, a missense SNP in the *ZNF646* gene that is most likely a transcription factor. This locus was previously reported as a selection target in East Asian populations (Xue et al., 2009), the data presented here indicate that this signal is shared with Siberians (Appendix C.32B). However, the potential selective pressure is currently unknown; the variant could be of regulatory importance (see section 3.2.8).

The other loci where the DIND and ΔDAF overlap are the nonsense variant rs34358 in *ANKDD1B* and the missense variant rs11150606 in the serine protease *PRSS53.* The derived allele (A) at the former was found to be associated with a lowered value of total blood cholesterol (Teslovich et al., 2010) and exhibited the highest observed DAF in the Northeast Siberians (Appendix C.34A). The latter signal is shared between all Eastern Eurasian groups. A recent GWAS has shown that the derived state (C) is correlated with straight hair in an admixed Latin American population, where the site was also highlighted as a selection target (Adhikari et al., 2016a). The authors also demonstrated that *PRSS53* is expressed in the hair follicle during active hair growth and that the rs11150606 polymorphism affects the processing and secretion of the resulting protein.

3.2.8 Overlap of selected variants and functional association databases

To further investigate the functional relevance of the variants highlighted by selection tests the annotations for the top 215 SNPs highlighted by ΔDAF (Tables 3.14-3.15), d_i (Appendix C.35A) and/or DIND (Appendix C.35B) were retrieved from databases containing information from GWAS and concerning gene expression across a range of tissues.

When querying the GWAS Catalog (MacArthur et al., 2017) and GWAS Central (Beck et al., 2014) for SNPs associated with variation in phenotypic traits ($p < 5*10^{-8}$) 11 loci overlapping with the top selection signals were found (Appendix C.36). Many of these have been highlighted by previous selection scans (e.g. rs3827760 in *EDAR,* rs4988235 in *MCM6*) while for others the functional relationship between the potential selection target and the GWAS signal is unclear. Interestingly, there are loci known to be under positive selection and well-supported on a mechanistic level where the selected derived allele is correlated with phenotypes that are either diseases or predispose to them. A good example is rs4988235; it resides in an enhancer region of the *LCT* gene and thereby confers lactase persistence. The derived state (A) at this SNP is associated with obesity indicators. Similarly, the derived G allele at rs16891982 in *SLC45A2* resulting in lighter pigmentation of the skin, hair and eyes correlates with higher risks of melanoma and squamous cell carcinoma. Both relationships are likely enhanced by very recent environmental changes. For the lactase persistence allele, it has been suggested that this relationship is modulated by higher dairy consumption (Manco et al. 2017), as these high energy density foods are abundantly available to many populations in the modern world. Similarly, behavioural changes are thought to have increased the exposure to UV radiation in light-skinned European populations over the last century (Parkin et al., 2011).

Subsequently, the GTEx database (V7) (Lonsdale et al., 2013) was mined for overlaps using a similar approach. For the most recent release of this resource only cis-eQTL data were available. This term was defined as an association between local genetic variation and gene expression within ≤ 1 Mb from the transcription start site (tss). The significance of this relationship was determined by the GTEx Consortium (Aguet et al., 2017) using a multi-step procedure. It accounts for a) multiple genetic variants in the 1-Mb cis-association window surrounding the tss of the genomic element whose expression levels were quantified and b) the high number of such gene expression phenotypes measured throughout the genome. Trans-eQTLs and longer range intra-chromosomal eQTLs were retrieved for the previous release (V6) of GTEx, none overlapped with the 215 SNPs from the selection scans.

However, this approach highlights a total of 1,458 cis-regulatory relationships (i.e. variant-gene-tissue trios) across a total of 47 tissues involving 97 unique SNPs and the expression levels of 283 genomic elements (Appendix C.37). A total of 210 (74.2%) of the genomic elements are classified as protein-coding genes. The median distance of each eQTL from the tss is 58.6 kb. The ubiquity of these associations is not surprising given that in a study on the previous version of the GTEx dataset (V6) 48.5% of all common SNPs included were shown to be associated significantly with at least one gene in at least one tissue (Aguet et al., 2017).

Given these conditions it was important to estimate whether the observed overlap of cis-expression data and highly differentiated sites represents an enrichment over neutral expectations. For the latter purpose 100 randomly sampled sets of 215 variants matched to the original potentially selected loci for frequency and chromosome were generated. Each of these sets was overlapped with the GTEx (V7) cis-eQTL dataset. On average, 1093.5 cis-regulatory relationships mapping to 84.1 unique SNPs were found. The empirical p-values resulting from the comparison of the observed data to the distribution of these neutral replicates are 0.02 for cis-regulatory trios and 0.04 for the total SNP count. While these are significant at $\alpha = 0.05$, the pattern observed for highly differentiated sites is not an extreme outlier relative to frequency-matched controls.

To explore whether particular biological terms from the GO database were enriched in the protein-coding genes affected by the eQTLs overlapping highly differentiated sites the PANTHER overrepresentation test was applied.

The resulting terms are clustered by two main categories. Firstly, there is a general enrichment of genes involved in the activation of the immune response ($p = 0.034$) (all results are listed in Appendix C.38). The subclass driving this outcome are genes coding for HLA class II molecules (p-values ranging from $5.2*10^{-10}$ to 0.0081 for various related functional and structural GO terms). These have long been known to be expressed primarily on antigen-presenting cells of the immune system thereby initiating the $CD4^+$-T cell-immune responses. Secondly, there are many results from the "cellular component" category, that are mainly related to the Endoplasmic Reticulum and the Golgi apparatus (all p-values < 0.046), indicating that this subset of genes encodes more proteins involved in the processing and the intracellular transport of proteins and lipids than expected. Given that the genomic architecture of the HLA region is characterised by high gene density and extreme sequence variation as well as extended runs of LD (Horton et al., 2004) the first result should be interpreted with caution. There are

only two out of 97 polymorphisms regulating at least one HLA gene and the overrepresentation signal is mainly caused by rs3135371, a polymorphism that is an eQTL for eight genes in the region. For this eQTL the directionality of the expression change correlating with the derived allelic state is variable between genes but constant for one gene across tissues.

One possible confounder for these analyses is that the allele frequency patterns detected by the d_i statistic (Appendix C.35A) can be biased as short-read WGS approaches are known to have an elevated error rate for SNPs in the HLA region. This bias leads to an overestimation of the reference allele frequency (Brandt et al., 2015). As both alleles observed at rs3135371 and rs60438747 at an unusually high population-specific frequency are non-reference it seems likely that the highly differentiated loci are genuine. Recent work has shown that clusters of eQTLs are part of HLA haplotypes influencing HLA-wide gene expression (Lam et al., 2017). It can be speculated that the non-reference alleles at rs3135371 and rs60438747 belong to HLA haplotypes that are much more common in the relatively understudied Southeast Asians and Siberians than in other parts of the world. The resulting differences in the expression of multiple HLA class II genes could represent the original selection target leading to the observed patterns.

A way to prioritise regulatory relationships from the large overlap found is to consider those eQTLs affecting genes that are highly expressed in the particular tissue where the correlation between the allelic state of the SNP and mRNA expression was detected. Therefore, a subset of eQTL-gene pairs was selected as follows: the gene had to be among the top 5% genes a) in terms of total expression in this tissue, and b) in relative terms of how much it is overexpressed in it relative to all other tissues using a z-score approach. These analyses were limited to protein-coding genes.

This led to a reduction of the dataset to seven variant-gene-tissue combinations (Table 3.16). They represent at least five independent underlying signals. This is because rs882856 and rs1941406 that both upregulate the expression of *POU2F3* are in moderately strong LD in the Variant-Based Analysis Set (r^2 ~0.58 when all populations were pooled). Furthermore, rs269868 influences both *DUOX1* and *DUOXA1* that are arranged head-to-head, i.e. on opposite DNA strands. These two genes are part of a system that is crucial for thyroid hormone synthesis and only functional if both are expressed (Grasberger and Refetoff, 2006). Therefore, they are best considered as a transcriptional unit.

Table 3.16: Overlap of variant-based selection tests and cis-eQTL from GTEx V7. For this subset of variants, the gene is strongly and specifically expressed in the tissue the gene-variant relationship was highlighted in (top 5% of protein coding genes in that tissue in absolute as well as relative terms). Furthermore, the raw p-value for a t-test between the respective variant and the expression of the target gene (nominal P) and the normalised regression coefficient (slope) for a simple linear regression of the genotype at a respective site and the expression of the target gene are given. For the latter a positive value indicates an increase in gene expression with the number of derived alleles an individual carries, a negative value designates a decrease respectively. Abbreviations: anc – ancestral, Chr – chromosome, der – derived, dist – distance, Pop – Population, Pos – Position, Tss – transcription start site. Translation impacts were inferred from Ensembl v75 and are abbreviated as follows: DOWN- downstream gene variant, IN – intronic, INTER- intergenic, MIS – missense, NC – non-coding transcript variant, NC.EX – non-coding transcript exon variant, STG – stop gained, UP – upstream variant. Short codes for the macro-groups taken from Table 3.1.

Chr:Pos	Anc/ der	rs_ID	Tests	Pop	Translation impact	Gene	Gene function	Tissue	Tss (dist from)	P (nominal)	Slope (der)
1:152323132	G/T	rs12568784	ΔDAF	SeM, SeI	MIS, STG, IN, NC	*CRNN*	mucosal/epithelial immune response (tumour suppressor for oesophageal cancer); epidermal differentiation	oesophagus mucosa	-63607	$3.06*10^{-13}$	-0.22
9:33469427	G/A	rs10813983	Di	WSi	NC.EX, IN, UP, DOWN	*AQP3*	water channel protein	oesophagus mucosa	21818	$9.67*10^{-12}$	-0.33
11:66753650	G/A	rs11227639	Di	NSi, WSi	INTER	*ACTN3*	part of sarcomeric Z line; affects athletic performance	skeletal muscle	439784	$3.61*10^{-10}$	0.40
11:120107411	G/A	rs882856	ΔDAF	diffused	MIS, UP	*POU2F3*	keratinocyte differentiation	skin (suprapubic, not sun-exposed)	62	$4.12*10^{-6}$	0.15
11:120141165	A/G	rs1941406	Di	WEu	IN, UP	*POU2F3*	keratinocyte differentiation	skin (suprapubic, not sun-exposed)	33816	$5.84*10^{-11}$	0.22
15:45392075	G/A	rs269868	ΔDAF	all non-Africans	MIS, NC.EX, DOWN	*DUOXA1*	activator of DUOX1	thyroid	-30055	$5.31*10^{-6}$	-0.29
15:45392075	G/A	rs269868	ΔDAF	all non-Africans	MIS, NC.EX, DOWN	*DUOX1*	oxidase (thyroid hormone synthesis and antimicrobial defence)	thyroid	-30062	$2.58*10^{-8}$	-0.37

Perhaps the most remarkable finding is rs11227639, an intergenic variant affecting the expression of *ACTN3* in skeletal muscle. It is located 440 kb downstream from the transcription start. The derived (A) allele at this site causes a higher expression of *ACTN3* in the GTEx samples (Figure 3.18). DAFs of more than 70% were observed in Northeast and West Siberians relative to a global DAF of 18% in non-Siberians (Appendix C.39A).

To understand the role of rs11227639 as a potential selection target it is necessary to consider the biological context. *ACTN3* encodes actinin-3, a structural muscle protein linking actin filaments that is only expressed in type II fast twitch muscle fibres. The role of this gene has been well characterised as it contains the wide-spread LoF mutation rs1815739 for which ca. 18% of the world population are homozygous (MacArthur et al., 2007). The functional ancestral allele at this site has been associated with increased explosive power performance in the general population as well as elite athletes (Moran et al., 2007; Eynon et al., 2013).

Studies on animal models and human muscle biopsies have linked this SNP to a variety of changes in muscles (Lee et al., 2016). These include differences in cross-sectional area, which is increased when the functional allele is present (Broos et al., 2016) and several metabolic

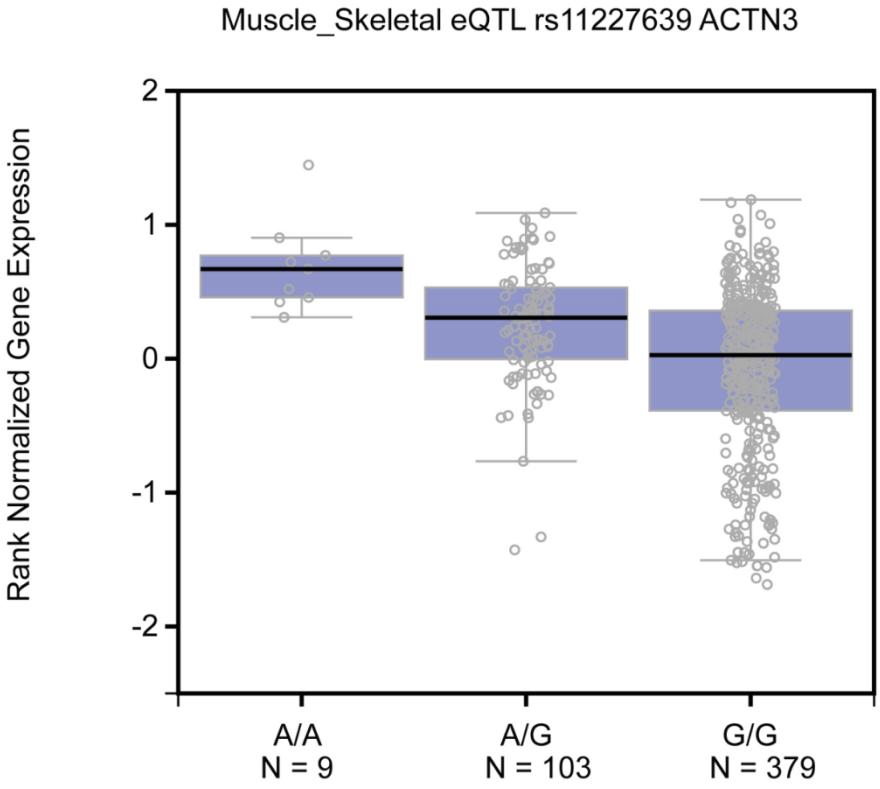


Figure 3.18: Normalised gene expression levels in GTEx V7 for all three possible genotypes at rs11227639.

properties such as the relative importance of anaerobic vs aerobic energy generation (MacArthur et al., 2007, 2008). Appendix C.40 gives an overview of the consequences of higher *ACTN3* expression, i.e. when the gene is functional and potentially influenced by additional up-regulatory variants such as rs11227639.

The LoF variant rs1815739 has in the past been highlighted as target of natural selection in non-Africans in general (MacArthur et al., 2007; Amorim et al., 2015) and in groups like the Kalash from Pakistan (Ayub et al., 2015) in particular with the main suggested pressures being scarcity of food resources and cold climate (Friedlander et al., 2013).

There should be an epistatic relationship between rs1815739 and rs11227639, i.e. the effects of the upregulating mutation can only manifest when at least one functional copy of *ACTN3* is present. The rs1815739 mutation was not highlighted in section 3.2.7 as it was excluded from the IVA analyses, however it was included in the VCF files. Intriguingly, the Northeast and West Siberian macro-groups have the highest non-African frequency of the functional allele at

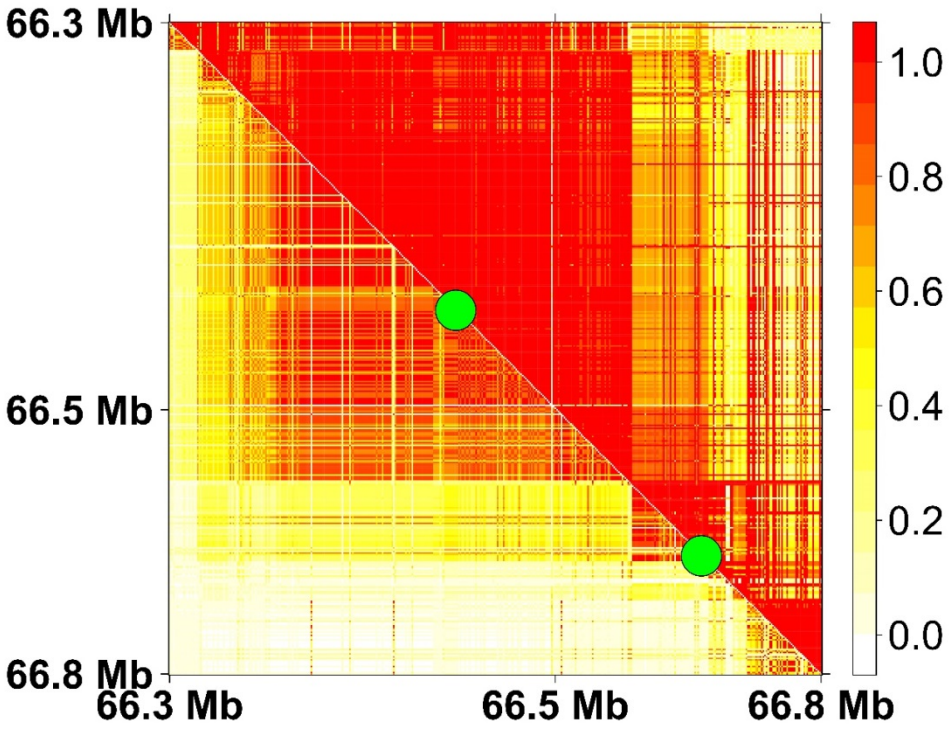


Figure 3.19: LD for Northeast Siberians between 66.3 and 66.8 Mb on chromosome 11. The upper half of the triangle shows D´, the lower half r^2. The upper dot marks rs1815739 and the lower rs11227639. Values of the LD metrics from 0 to 1 are indicated by a colour bar with increasing intensity from white to yellow to red for higher values.

rs1815739 (Appendix C.39A). For Siberians and South Americans, the frequency of the functional allele at rs1815739 at the macro-group level correlates well with the presence of the upregulating allele at rs11227639 (r = 0.94). This relationship is also detectable among Siberian subgroups (r = 0.8, groups with n<5 were excluded) (Appendix C.39B). To further examine this potential epistatic relationship LD patterns between rs1815739 and rs11227639 for Northeast and West Siberians were analysed in WGS data. LD is almost non-existent for West Siberians (r^2 = 0.03, D′ = 0.24) but high (r^2 = 0.63, D′ = 0.79) for Northeast Siberians (Figure 3.19).

To test whether this amount of LD is unusually strong for two SNPs with the respective frequencies of rs1815739 and rs11227639 that are ~426 kb apart in Northeast Siberians LD was calculated for this population across the whole genome binned into 1-Mb windows. From each window one pair of SNPs fulfilling these conditions was drawn randomly, yielding a total of 2,682 SNP pairs.

Compared to this distribution the r^2 between rs1815739 and rs11227639 had an empirical p-value of 0.016, i.e. the vast majority of SNP pairs from randomly chosen windows exhibit lower LD (Figure 3.20). The potential case for selection in Northeast Siberians is further supported by the iHS statistic, where the window rs11227639 lies in (chr11:66.6-8 Mb) also reaches genome-wide significance (p = 0.012) (Appendix C.41). Without further functional analyses and/or fine-mapping of the signal the evidence presented here is suggestive of positive selection on the cis-eQTL rs11227639 in Northeast Siberians but not conclusive. The main reason is the

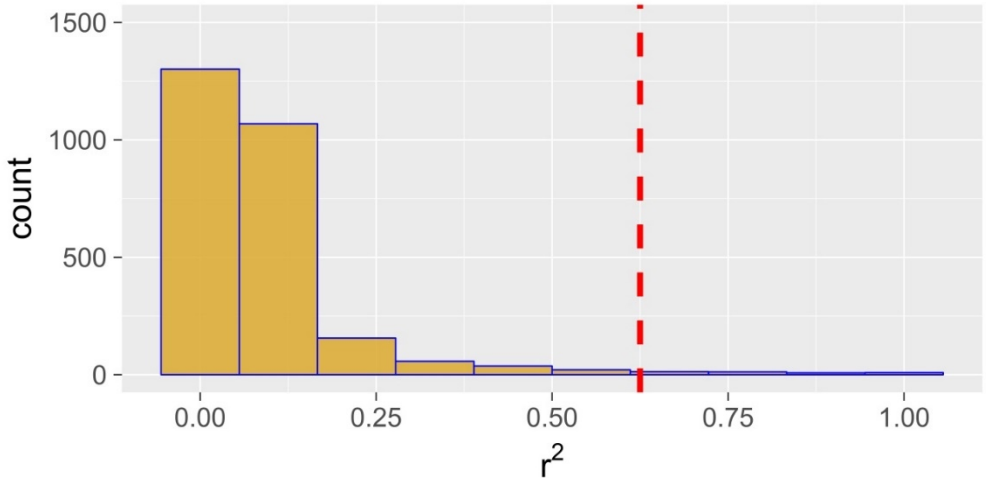


Figure 3.20: Genome-wide distribution of r^2 values for randomly drawn SNP pairs of the same frequency and distance as rs1815739 and rs11227639 in Northeast Siberians. A dotted red line shows the r2 between these two variants.

regulatory signal as well as the pattern of population differentiation are shared across tightly linked loci (Appendices C.42-C.43). In consequence, seven SNPs have an $r^2 \geq 0.5$ in GTEx genotype data with the potential selection candidate rs11227639; these loci are also tightly linked in the Northeast Siberians (Appendix C.44).

One approach to make this signal more precise is incorporating information on regulatory epigenomic markers. Regulatory DNA sequences are known to be co-located with open chromatin regions where nucleosomes are removed and/or destabilised (Tsompana and Buck, 2014). Therefore, these regions are accessible to DNase I. Information on such DNase I hypersensitive sites was obtained for psoas muscle based on data from 111 reference genomes assembled by the Roadmap Epigenomics Consortium. These were overlapped with rs11227639 and the surrounding candidate sites as intuitively the true causal locus should lie in an open-chromatin region. Evidence from fine-mapping analyses suggests that causal loci are more likely to be located in these regions (Aguet et al., 2017). This filter reduces the number of candidate loci to three including the original candidate rs11227639, the others being rs58462309 and rs6591228.

3.3 Discussion

3.3.1 Effect of methodology on functional variation annotation

Before the new empirical data presented in this chapter will be contextualised within the current understanding concerning the load of deleterious mutations and signals of positive selection it is important to consider how sensitive functional annotations, which are crucial for these inferences, are to the choice of software and transcript set.

A comparison of the IVA and VEP approaches applied to 382 worldwide WGS yielded two major results. Firstly, VEP classifies an excess of more than 73,000 variants as either exonic or splice-site altering that are not reported by IVA. This contributes to an asymmetry in functional annotations; most elements in IVA are given the same annotation by VEP but not vice versa. The analyses point to three causative factors. The IVA results were filtered more strictly, as the remaining sites effectively represent the overlap of Ensembl 75 and RefSeq release 63. Furthermore, the RefSeq transcript set, used as reference by IVA, is generally less likely to describe a particular variant as exonic or splice-site altering than Ensembl. The reason for the latter is that the Ensembl transcript set describes a larger fraction of the genome as protein-coding and therefore any randomly chosen position is more likely to have a more severe functional consequence (Wu et al., 2013). This effect was also observed controlling for the choice of annotation software when the VEP inferences using Ensembl and RefSeq were compared for chromosome 22. Finally, the IVA algorithm does not report most splice site variants, even if they are contained in the source transcript set. Secondly, the overall match rate between the two methodologies for exonic and splice-site altering variants is 74.9%. Underlying this figure is considerable variation between different variant classes, with the highest agreement rate observed for synonymous variants compared to a much lower concordance for LoF variants (Table 3.8).

Both outcomes reported here are mostly consistent with published studies of a similar nature. McCarthy et al. (2014) functionally annotated 276 25× coverage genomes and tested for the effect of either transcript set or software tool by fixing the other element. For both analyses they obtained exonic match rates of >80%, somewhat higher than reported here where both factors varied simultaneously. Furthermore, they note that the transcript set appears to have a larger impact on the match rates than the choice of software and that the VEP outperforms other

software tools in detecting splicing variants. One limitation of the comparative approach used here is that for simplicity only the most severe variant was reported (precedence ranking in Appendix C.8). This underrepresents the true biological complexity as many variants can be part of multiple transcripts.

There are also more general issues related to variant annotation, e.g. not all software tools use the same definitions for reporting particular variant classes and considerable gaps remain in our knowledge concerning the roles of many transcripts. As the terminology becomes more standardised due to such efforts as the Sequence Ontology Project (Eilbeck et al., 2005) and our representations of the true underlying genomic structure are continually improving, in the future the annotations might converge increasingly. This is supported by recent exploratory analyses on all possible SNPs and short indels up to a length of 3 bp in the exons and exon-flanking regions of the CTFR gene. The mutations were generated in silico and different variation annotation tools were used on the output. Even before normalisation of effect nomenclature the concordance exceeded 95% (Jesaitis, 2017).

Based on the outcomes described above, if a relatively low number of variants is of interest and false positives are potentially costly, a higher level of certainty can be reached if multiple tools assign the same annotation to these sites. However, for most of the analyses presented here the focus is on more general patterns, i.e. the relative abundance of different variant classes in various populations. For this case, as demonstrated for stop-gain variants, the results based on annotations obtained from different methods appear to be mostly consistent, especially if category definitions match and interpopulation differences are sufficiently strong.

3.3.2 Patterns of deleterious variation and purifying selection

A significant part of this chapter consists of analyses of the patterns of deleterious variation in 382 whole genomes from a world-wide sample. It provides a broader geographic coverage than most previous empirical explorations of the subject (see, however, Henn et al., 2016). This is especially important as following the initial non-African bottleneck changes in N_e (Figure 3.21) and various levels of gene flow have interacted to create unique population histories which impact the patterns of deleterious variation.

One of the disadvantages of the dataset analysed here are the sample sizes. Even if summarised on a continental level, they are comparatively low, which has been shown to impact the

behaviour of population-level summary statistics (Koch and Novembre, 2017) as a considerable fraction of rare variation is not captured.

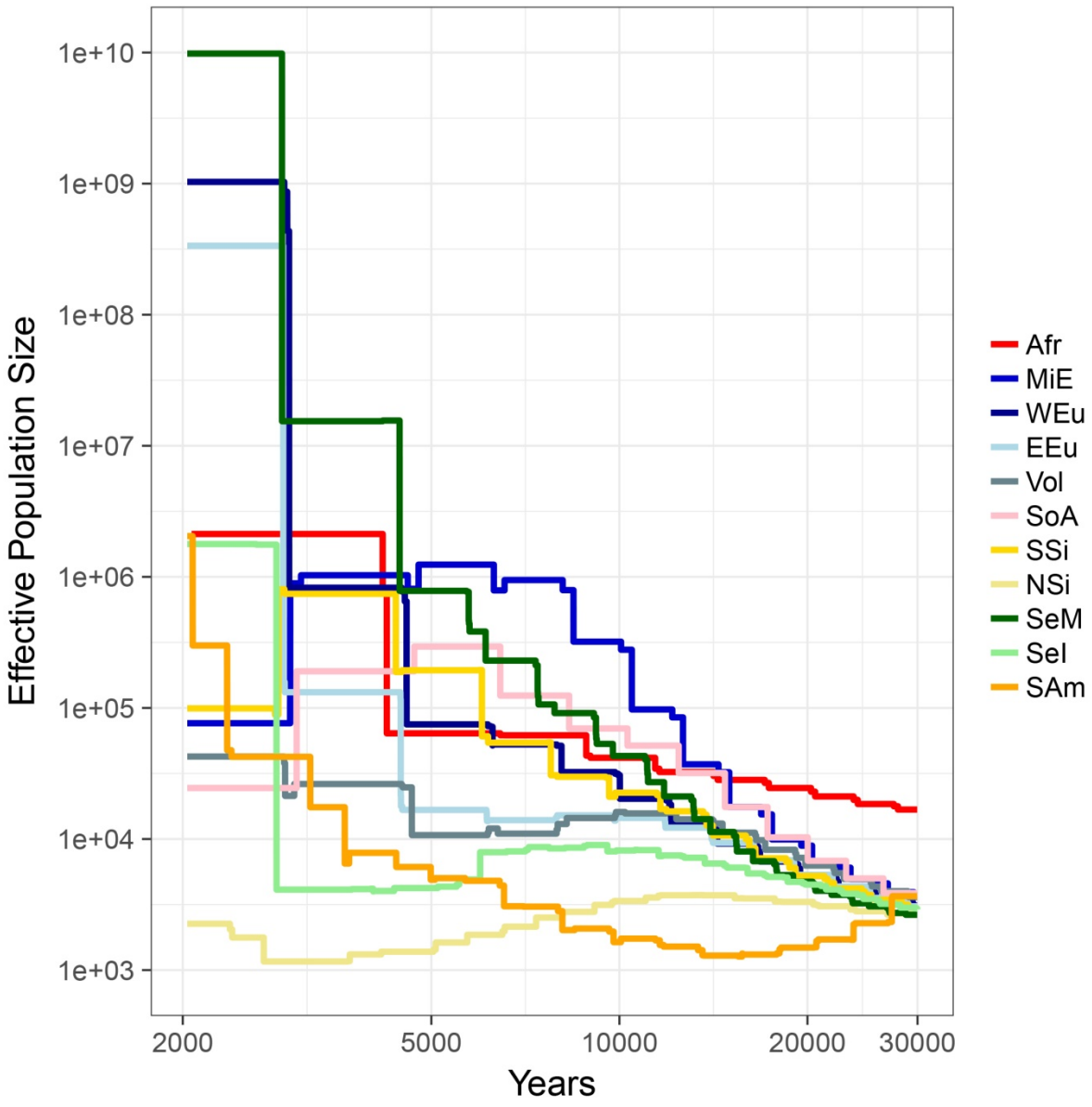


Figure 3.21: Plot of N_e over time for different groups in the Variant-Based Analysis Set estimated from 4 diploid genomes with MSMC. Both axes are logarithmic. This graph only includes continental groups containing at least one population for which 4 genomes were sequenced. If there were multiple such populations, the values were averaged for the continental group. The time scale was obtained assuming a generation time of 30 years and a mutation rate of μ = 1.25*10-8 bp^{-1} $generation^{-1}$. Short codes for the macro-groups taken from Table 3.1.

Here, many findings concerning the differences in patterns of genomic diversity between different variant classes and/or population groups are recapitulated. On a global level a significant shift towards rare variation compared to synonymous variants is observed for a wide range of putatively functional/deleterious mutations, including missense, nonsense, broader

LoF, different CADD score cut-offs and those contained in the HGMD database. This is consistent with previous studies (Hughes et al., 2003; The 1000 Genomes Project Consortium, 2010) and the predicted effect of purifying selection.

One key observation from previous studies on the subject is the higher proportion of non-synonymous and/or deleterious mutations among all exonic sites in non-Africans (Lohmueller et al., 2008; Peischl et al., 2013; Lohmueller, 2014b). This finding is reproduced here with two slightly different summary statistics (per-individual ratios averaged over a macro-group vs per-macro-group ratios calculated on total site counts) and is found to be mainly driven by missense variants (Table 3.6). Three non-exclusive explanatory mechanisms have been proposed. The first is exponential population growth in recent times causing an influx of new mutations. As there are more possibilities for non-synonymous changes compared to synonymous changes in coding regions the majority of these changes are non-synonymous and selection would not have enough time to purge these, despite its power increasing with higher N_e (Lohmueller, 2014b). In other scenarios, this increase is a consequence of the range expansions out of Africa. The extreme strength of drift on the wave front would allow either standing or novel deleterious mutations to behave approximately like neutral variants and therefore reach higher frequencies (Peischl et al., 2013; Peischl and Excoffier, 2015). A bottleneck could cause a similar outcome (Lohmueller et al., 2008). In the first case the differences would be mainly driven by rare variation, whereas in other models common variation would also contribute.

The summary statistics for the proportion of missense/non-synonymous mutations are not correlated with the harmonic mean of N_e over the last 5,000 years obtained with MSMC (Figure 3.2) or population growth over the last 20,000 years. This appears to contradict the simulation studies (Lohmueller, 2014b) that suggested recent population growth as the main cause for the relative excess of non-synonymous mutations in non-Africans. However, it is possible to reconcile these results. Firstly, as mentioned above the dataset analysed here is underpowered to fully assess the contribution of rare variants. Secondly, the accuracy of MSMC has been shown to decline for very recent population history. By the retrieval of N_e using MSMC on simulated data it was verified that when (as is the case here) eight haploid genomes are used the method is reliable until ca. 2 kya (Schiffels and Durbin, 2014).

Furthermore, the relationship between recent growth in N_e and the proportion of non-synonymous variants might follow a logarithmic model. This is supported by Lohmueller

(2014b) who found that this statistic is vastly different for simulated non-expansion vs 10-fold expansion scenarios but plateaus quickly when higher magnitudes of growth are assumed. In addition, Northeast Siberians, the only non-African group that did not experience rapid population growth in the timeframe that can be reliably reconstructed by MSMC (Figure 3.21) have a non-synonymous proportion that is essentially indistinguishable from that of Africans (Figure 3.2, Appendix C.13).

Overall, the analyses presented here indicate that the proportion of non-synonymous variants is not linearly related to recent population growth and that the latter is unlikely to be the only factor underlying global variation of this statistic. At least part of the effect appears to be driven by common variation which is further supported by simulations confirming that even for very small sample sizes non-Africans have a higher non-synonymous proportion (Koch and Novembre, 2017).

Another summary statistic for allele frequency patterns, the DAF spectrum, was generated for six different variants classes (synonymous, missense, stop-gain, CADD20, CADD30 and HC LoF from LOFTEE) and for all 14 macro-groups normalised to the sample size of the smallest group. The CADD score allows for stratification of variants by approximate mutational effect sizes. This is supported by the negative correlation of a variant's CADD score and its DAF (Kircher et al., 2014, also Table 3.11). Furthermore, assuming a CADD score of 0 represents neutrality a range of demographic models can be fit to the SFS for each subsequent CADD score bin using an ML approach. The best-fitting models for increasingly severe CADD scores include monotonically increasing negative selection coefficients (Racimo and Schraiber, 2014). Here, different CADD thresholds are uniformly applied across all genes. This is a simplification, as "true" biological deleteriousness thresholds likely vary considerably between genes as recently demonstrated by Itan et al. (2016) based on CADD scores assigned to disease-associated mutations in different HGMD-listed genes.

Two limitations of CADD should be briefly touched upon here. Firstly, as a general problem, also associated with selection detection methods, reduced diversity that correlates with high CADD scores can result from neutral forces, e.g. variation in local mutation rates and GC-biased gene conversion. However, it has been shown that the negative correlation of CADD with DAF holds when stratifying for related genomic features such as GC and CpG content (Kircher et al., 2014). Secondly, CADD exhibits a reference-bias, i.e. a site where the reference

is derived is more rarely called as deleterious compared to a site where the reference is ancestral. While for the CADD this bias is less severe than for SIFT and PolyPhen-2 it still results in an underestimation of the total number of deleterious variants in non-Africans, as they have a higher DAF at sites where the reference is derived (explored in more detail in Appendix C.45).

For missense and CADD20 variants, the majority of which are likely mildly to moderately deleterious, there is a significant rightward shift of the DAF spectra for non-Africans relative to Africans across all possible comparisons (Tables 3.4, 3.11, Figure 3.15). Among non-Africans most pairs exhibit significant differences in the spectra for both variant categories unless their population split occurred recently. However, it should again be stressed that phylogenetic relatedness as such is not causative but rather a similar N_e trajectory. The reported shift towards more common moderately deleterious variation in non-Africans confirms earlier studies (Lohmueller et al., 2008; Fu et al., 2014b; Henn et al., 2016). In the literature, this pattern has most often been explained as resulting from (serial) bottlenecks (Lohmueller, 2014a) and/or spatial range expansions (Peischl et al., 2013). This observation was also replicated when pooling all non-African macro-groups together (Figure 3.3), consistent with theoretical expectations there is still more population structure in the African than in the non-African sample.

Other variant classes, such as stop-gain (IVA), HC LoF and CADD30, exhibit considerably less differentiation with regards to the regional DAF spectra (Appendices C.24-C.25, Figure 3.14). Consistent to what is observed for more weakly deleterious variants there is a shift of Africans towards rare variation compared to most East Eurasian groups and an excess of high frequency derived deleterious variants in Andeans (Northeast Siberians exhibit a similar but quantitatively weaker pattern), especially relative to West Eurasians and Africans.

The most plausible explanation for this observation is that according to theoretical expectations the fate of new deleterious mutations in an idealised population of constant size depends on N_e and s: if $|4*N_e*s| >> 1$ then selection primarily determines the observed alleles frequencies, if the product $|4*N_e*s|$ becomes smaller genetic drift increases in importance (Kimura, 1994). Many mutations in the CADD30, HC LoF and stop-gain classes are expected to have high negative selection coefficients. They should intuitively be disruptive as they often lead to a loss of function (see section 3.3.3) and exhibit low global frequencies (Racimo and Schraiber, 2014). Therefore, most variants of this type should be uniformly purged in all populations leading to

the much more homogeneous patterns observed. A potential criticism of this argument is that this approximation only holds under equilibrium conditions. However, recent simulation work on non-equilibrium demographies demonstrates that for variants with high selection coefficients the associated genetic loads should become comparable for OOA- and non-OOA-like populations 2000-3000 generations after their split (Gravel, 2016).

The number of CADD20 variants observed in non-Africans, which declines with distance from Africa (Figure 3.9, Table 3.9), reflects the decreasing diversity when moving away from Africa also observed for neutral variation (Prugnolle et al., 2005; Henn et al., 2012a) and agrees with patterns reported previously for deleterious variants (Lohmueller et al., 2008; Fu et al., 2014b). The generally accepted interpretation of this pattern is that it results from the OOA bottleneck (Lohmueller, 2014a). However, the observed differences between East Eurasian groups approximately equidistant from Africa hint at the role of subsequent population history, most plausibly different trajectories of N_e. It should be noted that different processes can generate comparable values of N_e, i.e. the clustering of Middle Easterners and South Asians and their high genetic diversity relative to other Eurasians, as previously shown by Melé et al. (2012), does not necessarily indicate shared population history.

For homozygous derived CADD20 sites the sequence of macro-groups can broadly be described as reversed compared to that observed for the total number of sites (Table 3.9). Apart from the Papuans all-non-Africans show increasing numbers of homozygous derived CADD20 genotypes with distance from Sub-Saharan Africa (Figure 25), confirming previously reported observations (Lohmueller et al., 2008; Fu et al., 2014b; Henn et al., 2016). Most macro-groups appear to be significantly differentiated except for a West Eurasian/South Asian cluster and some East Eurasian groups respectively (Figure 3.10).

Intriguingly, Oceanians are very similar to other Eurasians concerning the total amount of variants in this class but for homozygous derived loci constitute a distinct group together with the South Americans. When the Australians, who are known to be recently admixed, are removed the Papuans (n = 6) retain this somewhat counterintuitive pattern. These remaining individuals have the highest fraction of homozygous CADD20 loci relative to the total number of sites (~0.343). However, they exhibit a higher total site count relative than South Americans and Central/Northeast Siberians (this pattern was also observed for all exonic sites, Table 3.5).

One possibility is that the well-attested additional pulse of Denisovan admixture into Papuans caused their higher diversity accompanied by more homozygous derived sites. The Denisovan alleles are estimated to have entered the genome of the ancestors of the Australo-Papuans ~44 kya (Malaspinas et al., 2016). As Denisovans had a higher baseline homozygosity than human modern populations (Meyer et al., 2012) these alleles could be reflected as additional homozygous derived variants unique to Papuans. However, observations from the Kosipe and Koinanbe might not be generalisable, as recent large-scale genomics papers (Malaspinas et al., 2016; Bergström et al., 2017) have found a high amount of diversity within Papua New Guinea.

The main focus in the following will be on homozygote CADD counts as there is a negative relationship between s and the dominance coefficient h and therefore on average deleterious mutations are likely to be at least partially recessive. Support for these statements comes from a) experimental studies examining the phenotypic consequences of deleterious mutations in model organisms (Agrawal and Whitlock, 2011; Manna et al., 2011; Ayadi et al., 2012), b) the observation that ~58% of known Mendelian diseases in humans are recessive (3,077 out of 5,317 listed in the online database OMIM as of April 2019, see Lee et al., 2019), and c) recent work using the SFS of mutations observed in outcrossing and selfing species of *Arabidopsis* to estimate h and s jointly (Huber et al., 2017).

Under a simple linear model, distance from different African locations explains ca. 2/3 of the global variance in per-individual derived homozygote counts for the moderately deleterious CADD20 variants (Table 3.10). This value is slightly lower than that reported here for synonymous variants (if only ancestral = reference sites are considered). This also applies when it is compared to broadly similar analyses that regressed quantities related to neutral genome-wide heterozygosity/homozygosity against geographic distance from Africa where r^2 was 0.75-0.85 (Prugnolle et al., 2005; Li et al., 2008; Pemberton et al., 2012). To provide a proxy for potential differences in post-OOA population history estimates of N_e (harmonic mean between 5 and 65 kya) from MSMC were included in the regression model. For 6/8 scenarios (synonymous, CADD20 and HC LoF variants) the inclusion of this parameter results in a significant improvement of the amount of explained variation. For the best fit (N_e + distance from Windhoek) 91% of all variation in synonymous homozygous derived genotypes could be explained.

For CADD30 as well as for HC LoF variants most macro-groups are not significantly differentiated from the others concerning derived homozygotes. Exceptions are the Africans and the Native Americans (to a lesser extent also the Oceanians, island Southeast Asians and North Siberians) with the directionality of effects as expected based on their distance from Sub-Saharan Africa (Figures 3.10-3.11, Table 3.9). Correspondingly, the fit of the regression models is worse and for both variant classes the best models explain ca. 40% of the global variation (Table 3.10).

One likely cause for this observation is again that on average the selection coefficients for CADD30 and HC LoF variants are considerably more negative than for CADD20 variants, reducing the importance of neutral forces. In contrast, the latter determine most of the diversity observed for CADD20 variants. Simulation studies suggest that in addition to the serial bottlenecks occurring during the OOA migrations explicitly incorporating range expansions (Peischl et al., 2013) helps to explain how moderately deleterious mutations behave approximately neutrally and therefore reach higher frequencies which contribute to the observed increase in homozygote counts away from Africa.

The lower degree of geographical stratification for CADD30 and HC LoF variants has here been interpreted in terms of purifying selection. However, it could also result from the lack of power to find differences between macro-groups due to lower total variant counts in these categories. A resampling procedure was applied to correct for this confounding effect (Appendix C.46). Its outcomes show that for CADD30 the slope of the increase in derived deleterious homozygotes with distance from Africa is considerably flatter than for CADD20, i.e. at least part of the effect is caused by uniformly strong purifying selection. Intriguingly, for HC LoF geographical differentiation is almost as marked as for CADD20 when differences in the size of variant classes are corrected for. However, the low per-genome totals still suggest that HC LoF variation is constrained by natural selection.

The $R_{X/Y}$ statistic, which quantifies differences in the number of derived alleles between populations under the hypothesis that genetic variation acts additively, was examined to provide additional insights. There are no detectable differences with regards to this statistic in any of the non-African groups analysed relative to the Africans if missense mutations were assumed as a proxy for deleterious variants (Table 3.12) consistent with previous studies (Simons et al., 2014; Do et al., 2015).

One caveat for this interpretation is that synonymous sites are postulated to be neutral which might not hold for all variants in this class. Firstly, theoretical approaches attempting to infer the DFE based on the SFS have found that synonymous variants have higher average values of s than intergenic sites (Racimo and Schraiber, 2014). Secondly, case studies from model organisms as well as humans suggest that some synonymous variants can have considerable fitness costs and are causally associated with diseases (Brest et al., 2011) . Mechanisms that have been proposed relate to the stability of secondary mRNA structure and the regulation of translation (Knöppel et al., 2016). However, synonymous variants as a class should still be a reasonable proxy for neutrality in practice. The regression analyses in Table 3.10 also support this as synonymous variant counts can almost perfectly be predicted by "neutral" explanatory variables. This is probably because of the low N_e of our species for most of its recent history compared to many other taxa.

In conclusion, it was confirmed that the number of deleterious homozygotes is sensitive to recent population history as suggested by Simons et al. (2014) as explicitly incorporating post-OOA N_e based on MSMC inferences improved the amount of explained variance of multiple regression approaches for most deleterious variant classes (Table 3.10).

The relationship between another parameter of interest, the fraction of non-synonymous sites on a macro-group level, and recent N_e does not appear to be linear (Figure 3.2). In this context, it is tempting to speculate that the high value for this statistic in Andean Native Americans results from additional range expansions their ancestors underwent which could be tested with more empirical data and spatially explicit modelling (Peischl et al., 2013). The latter observation highlights a more general point. Genomic data from non-Africans with demographic histories that are very different from those experienced by well-studied references, especially if these include extreme bottlenecks and/or range expansions, are of great importance for examining theoretical predictions about the impact of population history on deleterious mutation load.

The analyses presented and discussed here confirm that if most deleterious variants act additively the differences in population histories between extant modern human groups are not extreme enough to cause robustly detectable differences in genetic load (Simons and Sella, 2016). However, if a considerable fraction of these loci is recessive, as recent research suggests, populations with low long-term N_e exhibit a moderate increase in genetic load. Most of this

excess derives from moderately, not extremely, deleterious variants. Geographical stratification is less pronounced for the latter (Appendix C.46) indicating that analogous results from Henn et al. (2016) obtained using the GERP score are robust if the CADD is applied.

A much-debated concept is this context is the "efficacy of [purifying] selection" (Fisher, 1930), which is difficult to separate from the role of drift for multiple reasons. Firstly, if the per-generation change in the frequency of deleterious alleles is considered then drift, selection and mutational forces always act jointly on this quantity. Secondly, population genetic theory and simulations suggest that selection and drift modulate each other through their past effects on the allele frequency spectrum (Gravel, 2016; Koch and Novembre, 2017). Therefore, in agreement with Lohmueller (2014b), here only their joint effect was considered, most directly through the $R_{X/Y}$ statistic. More practically, studies which have formally analysed the contribution of selection to genetic load note that for the relevant time frames it is often "overwhelmed" by genetic drift except for extremely deleterious mutations (Do et al., 2015; Gravel, 2016).

The impact of potential mutation load differences on the disease burden of an individual or a population is currently unclear. While generally the deleteriousness and pathogenicity of a variant should be correlated (Kircher et al., 2014; Shendure and Akey, 2015) this relationship is not straightforward. Firstly, there is experimental evidence suggesting that mutations under purifying selection over long-term periods can have no detectable adverse effects on the phenotype, especially for moderate to small selection coefficients (Miosge et al., 2015). Their effects on evolutionary fitness might be very subtle or only manifest under certain environmental conditions. Secondly, even if a variant is pathogenic in some individuals there is a wide range of factors modulating the relationship between genotype and phenotype known as "modifier genes" (see section 3.3.4). Besides some well-studied Mendelian disease examples that can have highly population-specific incidence rates there is some evidence that the burden of (rare) predicted deleterious variants can explain a fraction of intrapopulation variance in complex pathologies (see the Alzheimer's example in section 1.6.2).

3.3.3 Phenotype-specific effects of purifying selection

Here, the patterns of (deleterious) variation have mostly been discussed in the context of how they relate to drift and purifying selection as genome-wide forces without a phenotype-related biological context. The $R_{X/Y}$ was however also calculated for different groups of genes associated with variation in particular traits. The most pervasive signal is observed for genes related to pigmentation. The Africans exhibit a significant reduction of derived missense diversity in these genes relative to all non-Africans except Australo-Papuans (Table 3.12). This is consistent with earlier studies demonstrating strong functional constraint of these genes in Sub-Saharan Africans, e.g. *MC1R* where the allelic combination dominant in these groups ensures very effective melanisation (Harding et al., 2000). The main reason for this observation is thought to be selection for dark skin pigmentation at low latitudes (high UV radiation) due to a range of selection pressures such as folate loss resulting from photolysis (Jablonski and Chaplin, 2000) and skin cancer (Greaves, 2014). Recent efforts to generate large-scale genotype and skin pigmentation data from a wide range of Sub-Saharan African populations have hinted that the pattern of reduced diversity observed here is not necessarily generalisable to all Africans (Crawford et al., 2017). Furthermore, the same study found that many variants associated with dark pigmentation in Australo-Papuans are IBD with Africans. While this is not further explored here, it seems plausible that this contributes to them being not significantly differentiated from the West/Central African group according to $R_{X/Y}$.

The signal of reduced derived diversity in Africans relative to all non-Africans (reaching the significance threshold for ca. 50% of all comparisons) concerning genes related to the viral immune response is more difficult to interpret. The most likely and parsimonious explanation would be stronger purifying selection acting across this gene class in Africans. This seems plausible given that this mode of selection has been demonstrated for many immunity genes (Quintana-Murci and Clark, 2013) and that tropical environments are rich in viral pathogens. The theoretical alternative would be balancing selection, which has been demonstrated for some immunity genes, most prominently those located in the HLA region (DeGiorgio et al., 2014), causing an excess of derived missense polymorphisms in non-Africans. As balancing selection is thought to mostly act in longer evolutionary timeframes (Siewert and Voight, 2017) it seems likely that this signal would have been originally caused by factors affecting the common ancestral group from which all non-Africans descend. This could be tested by computing local TMRCAs across the whole genome for all individuals and checking whether the observed

excess of non-African diversity in these genes is primarily caused by regions with old TMRCAs which would indicate balancing selection.

3.3.4 LoF and Mendelian disease-related variation

Each individual in the Variant-Based Analysis Set carries on average 97 variants classified as HC LoF SNPs comprising 23 homozygotes (Table 3.9). Compared to other studies accessing LoF variation based on high coverage exome or WGS data this value is somewhat below the usually reported range of 100-150 LoF variants (Table 1.2). The main cause of this difference is likely that small frameshift indels were not considered here. This becomes apparent when different types of LoF variants are assessed separately, e.g. the SGDP (Mallick et al. (2016) reported an individual mean of 55.2 stop-gain variants which is similar to the value of 62.3 (Appendix C.19) obtained here with the VEP+LOFTEE plugin, if stop-gain variants inferred from IVA are considered the total of 56.1 (Table 3.5) is almost identical to the SGDP data. Other underlying factors for differences between studies are the choice of transcript sets and the effect-predicting software, where the concordance is particularly low for splice-site-disrupting mutations (see section 3.2.2). Furthermore, there is variation in the strictness of the downstream filtering after the initial annotation as LoF, e.g. the LOFTEE pipeline used here accounts for the position of LoF in the CDS and whether a splice-site affecting variant occurs in a canonical splicing context.

Genes containing at least one homozygous LoF variant (n = 352) were enriched for having one or more tandem duplicates and depleted for genomic elements contributing to the development of anatomical structures (Appendix C.27). This reflects theoretical expectations and confirms prior work (MacArthur et al., 2012) according to which many genes affected by homozygous LoFs in viable adults should be (partially) redundant and functionally non-essential.

The finding that genes coding for metal ion-binding proteins are overrepresented among genomic elements carrying homozygous LoF is, however, unexpected given the key role that the former play in many physiological processes (reviewed by Finney and O'Halloran, 2003). One possible confounder could be CDS length. However, this subset of genes is not significantly longer than all other genes carrying homozygous LoF ($\overline{L}_{homozygous\_lof\_metal}$ ~2038 bp vs. $\overline{L}_{homozygous\_lof\_rest}$ ~2022 bp, two-sample-t-test, t = -0.07814, p = 0.9378).

The largest functional subgroup among this category are zinc finger (ZNF) proteins, 20/107 of the corresponding genes code for molecules containing this motif. A study of WGS data from Danish individuals found that 4.1% of all LoF variants detected were located in ZNF genes (Besenbacher et al., 2015). This and the results presented in this chapter could partly be driven by the abundance of ZNFs: they are found in 3% of all human genes (Klug, 2010). Therefore, an effect of comparable size is more easily observed than for a very small group. Functionally, ZNF proteins are the most abundant type of eukaryotic transcription factors due their DNA-binding properties, even though they also interact with RNA and proteins (Gamsjaeger et al., 2007).

Work on yeast has shown that ZNF proteins encoded by genes which are paralogs within a species' genome often share a common core set of binding sites besides more specialised functions (Siggers et al., 2014). This has been interpreted as support for a "modular" model of transcription factor evolution where these specialised sites evolve while the core sites ensure redundancy for essential functions in gene transcription. There is currently no evidence for this mechanism in humans, however it has been shown that ZNF proteins have undergone accelerated evolution on different human and non-human primate lineages (Nowick et al., 2011). It is tempting to speculate that the excess of LoF variation observed in ZNF proteins is a by-product of redundancies allowing for variation that adaptive evolution can act on and is also buffering against purifying selection. This is furthermore supported by the observation that 11/20 of these ZNF proteins are part of (often large) groups of duplicated genes according to the Duplicated Genes Database. In conclusion, it seems possible that ZNF proteins exhibit redundancies that at least partially explain the overrepresentation of metal ion-binding proteins among genes carrying homozygous LoF.

These homozygote HC LoF genotypes were then filtered by frequency and their presence in published datasets. The resulting set consists of 29 rhLoF variants each located in a different gene (Appendix C.30). These genes have not previously been described as knockouts in healthy adults. Most of them are found in Central Siberians, Northeast Siberians and the ISEA group. This result is in line with predictions that besides populations practicing endogamy (Saleheen et al., 2017) isolated small N_e groups that are geographically distant from Africa should be the best source for detecting previously undescribed rhLoF variants due their elevated homozygosity (Kaiser et al., 2015).

A general caveat concerning the reported rhLoF variants is whether they truly lead to a functional knockout. The first confounders are problems related to the correct mapping of reads and the subsequent functional annotations. Assuming that the quality criteria and the high coverage of the data analysed minimise the first issue the impact of adjacent polymorphisms that could "rescue" the LoF has not been accounted for here. Saleheen et al. (2017) found that this confounder affects ca. 4% of all rhLoF sites. Secondly, experimental validation is necessary to prove the absence/strong reduction in the level of the relevant mRNA and the protein. Here, it was attempted to cross-check mRNA expression levels of potential heterozygotes for the 29 rhLoF SNPs in the GTEx dataset. However, unsurprisingly none of these variants was present in a heterozygous state in GTEx. Prior work on rare LoF variation in individuals for whom expression data were generated suggests that more than 2/3 of rare LoFs result in measurably lower mRNA levels compared to the reference allele (Rivas et al., 2015). Therefore, it seems plausible that at least a fraction of the rhLoFs reported here represent true LoF cases.

The mean number of 16.3 HGMD-DM sites per individual (Table 3.13) is lower than the count of variants of this kind observed in phase 3 of the 1000 Genomes Project, where on average 19 were observed (two-sample-t-test, $t = -12.316$, $p < 10^{-15}$, data retrieved from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/functional_annotation /filtered/functional_categories_summary_per_individual.20150208.txt). While this outcome is statistically significant, given the discrepancies in sample composition and sequencing technologies (Illumina vs Complete Genomics) the relatively small magnitude of the difference indicates that the general patterns are replicable across WGS datasets.

The total HGMD SNP count correlates very well with whole-genome heterozygosity, however there is a remarkable depletion of homozygous genotypes in Europeans and closely related populations. To contextualise this finding, it is important to recall that HGMD contains variants inferred from GWAS on highly polygenic traits. A systematic survey of all studies published in the GWAS Catalog found that as of 08/2016 81% of all participants were of European descent (Popejoy and Fullerton, 2016). Recently, Manrai et al. (2016) and Martin et al. (2017) demonstrated the consequences of this for inferring the heritable risk component of traits with monogenic and more complex genomic architectures respectively. Predictions of height, diabetes type II risk and other phenotypes for non-European populations using a polygenic score approach showed marked directional inconsistencies with anthropological and epidemiological data.

As disease-related variation is much better characterised in Europeans than in other global superpopulations, the excess HGMD homozygotes in non-Europeans reflects most likely false positives and any genuine biological differences are confounded by this effect. Besides the false positives, the HGMD totals are likely also lowered due to false negatives as they miss many rare and therefore often population-specific deleterious variants. Several factors contribute to this, firstly GWAS studies until very recently lacked the power to detect the effect of rare variants and secondly, as already described, many non-European populations remain underrepresented in medical genomics.

Many traits listed in HGMD are polygenic and/or of limited penetrance. Therefore, the individuals analysed in this chapter were scanned for a more restrictive panel of SNPs causative for recessive Mendelian childhood disorders (Chen et al., 2016). Almost all the 32 individuals exhibiting such variants carry them in a heterozygous state and the total carrier burden of Mendelian disease equals 0.086 alleles per individual which is much lower than previously reported (Bell et al., 2011). The divergence can mostly be explained by the filtering for clinically verified high penetrance variants by Chen et al. and the already discussed European study bias. If only Europeans are considered the carrier burden rises to 0.193 alleles per individual which is similar to the value of 0.17 reported for 4,313 exomes of European Americans under similarly strict filtering criteria (Tabor et al., 2014). In conclusion, based on HGMD homozygotes the relative disease burden in non-Europeans is likely overestimated while using a stricter Mendelian disease panel has the opposite effect.

3.3.5 A homozygous LoF mutation in the CFTR gene

The homozygous genotype found in the individual Murut13 has been recorded previously in a heterozygous state as rs77646904 in healthy individuals (frequency of 0.08% in ExAC East Asians). While HGMD describes the variant as disease-causing ClinVar (Landrum et al., 2016) currently lists it as of uncertain significance (Accession RCV000046339.3 as of 02/09/2017). According to the latter, the current evidence is conflicting. In support of its pathogenic status are studies where it was recorded in a homozygous state in some affected individuals without other known cystic fibrosis-related mutations being present, notably for a less severe "nonclassic" cystic fibrosis phenotype (Groman et al., 2002). Furthermore, a different missense mutation at the same position, a G to T transversion has been determined to be pathogenic,

which is also why this site was highlighted by Chen et al. (2016) . Against this it can be argued that on the protein level the G>A mutation causes a change from valine to isoleucine that is conservative and also found in other mammals for the CFTR protein at a homologous position.

Without further phenotypic data from the Murut individual in question and/or functional evidence regarding the impact of this mutation on the CFTR protein, two scenarios seem plausible: i) the mutation does not cause a cystic fibrosis phenotype or a mild atypical version or ii) it causes a full cystic fibrosis phenotype in other homozygote carriers, however the Murut individual either exhibits somatic mosaicism or other unknown secondary genetic or epigenetic modifiers that protect against the disease. The second possibility points to a more general problem in the interpretation of deleterious and putatively disease-causing mutations.

Leaving the absence of phenotypic data aside the predictions based on genetic data are necessarily incomplete as all variants were considered in isolation. Differences in the genetic background surrounding a causal variant can have a strong impact on how it manifests phenotypically. Mostly these background effects are thought to result from modifier genes, many of which are hypothesised to be part of interaction networks that evolved to coordinate gene activity in space and time (Riordan and Nadeau, 2017). Intriguingly, one prominent example for this gene class are expressivity modifiers of the severity of cystic fibrosis (Emond et al., 2012; Corvol et al., 2015).

3.3.6 Signals of positive selection

The primary signature of natural selection explored in this chapter is population differentiation. Through the DIND statistic, changes in nucleotide diversity surrounding the putative selected sites are also considered.

The main criticism of the empirical outlier approach applied here is that demographic history and heterogeneity in local genomic architectures can generate patterns mimicking those created by positive selection (Teshima et al., 2006). Simulations of neutral demographic history have been proposed as an approach to correct for this. The details vary depending on the statistic of interest, but the general concept is to quantify the fraction of simulation replicates that produce outcomes equal to or more extreme than those in the empirical data. False positive rates were not explicitly estimated here, however previous studies can provide helpful intuition.

Colonna et al. (2014) analysed a subset of empirical outliers with high ΔDAF values among 911 individuals from the 1000 Genomes Project. They simulated sequence data based on established demographic models and obtained high ΔDAF sites defined by the same criteria applied to the empirical data. The numbers of high ΔDAF sites were very sensitive to underlying assumptions about long-term intercontinental migration rates. As this parameter is difficult to estimate Colonna et al. concluded that currently such simulations are not a practical tool to estimate the false positive fraction of high ΔDAF sites. Instead they compared the genes these highly differentiated sites were located in with a list of previously identified putative selection targets. They found that 65% of all protein coding genes containing extremely high ΔDAF loci between continents and 30% of such genes that were highly differentiated within continents had been reported previously. The first figure can with some caution be taken as the approximate lower boundary of true positives for the high ΔDAF missense sites inferred here, as most of them cluster by continental groups (Figure 3.17). It is uncertain how relevant these estimates are for stop-gain variants given that their geographical differentiation is less clear, and the DAF differences are less marked.

A further demographic factor that can create strong allele frequency differences is population substructure. Excoffier et al. (2009) demonstrated by modelling that the F_{ST} variance under neutrality, which unsurprisingly is almost perfectly correlated with ΔDAF (Colonna et al., 2014), is considerably higher when strong substructure between the demes of populations is present. In addition to variance in between- and within-in population migration rates the already discussed effect of range expansions could also increase population differentiation through drift. Taken together, this also implies that using a uniform cut-off, i.e. the top 20 ΔDAF sites should lead to an enrichment of drifted groups.

Given the results from previous subchapters, the Northeast Siberians, Central Siberians and the ISEA group are good candidates for this and consistent with expectations, they are part of 44/65 top ΔDAF pairs (Tables 3.14-3.15).

The ΔDAF approach reproduced well-known examples of natural selection such as missense variants in *ABCC11*, *EDAR, MC1R, PRSS53, SLC24A5, SLC45A2* and *TLR1* (Table 3.14). Broad functional categories for the genes these variants are located in comprise pigmentation, immunity, ciliogenesis and motor activity. Whereas the former two can be more easily linked to selective pressures (UV radiation and pathogen exposure, Fan et al., 2016) the latter

categories are relevant to many phenotypes, e.g. the cilia as organelles are involved in a wide range of cellular transport, sensory and signalling processes (Ishikawa and Marshall, 2011). Therefore, it would be too speculative to propose underlying selective factors.

The question whether the nonsense variants highlighted by the ΔDAF represent true instances of positive selection cannot be answered conclusively based on the current state of variant annotation (Table 3.15). Given that many of the genes they lie in are part of larger gene families and that most of them do not lead to NMD it can be speculated that the phenotypic impact of the LoF events is likely very limited. This would point towards a scenario where most of these variants are selectively neutral and reached population differentiation through drift. In concordance with earlier systematic studies on the subject (Yngvadottir, 2008) it can therefore be stated that gene loss via stop-gain mutations does not appear to have been a major evolutionary force in the recent history of our species.

However, this does not wholly invalidate Olson's "*less is more hypothesis*" as splice-site altering mutations, small indels and large deletions were not evaluated. Furthermore, it seems possible that gene loss has a macro-evolutionary role in hominids as there many instances of lineage-specific gene loss, though its functional consequences are not yet well understood (Albalat and Cañestro, 2016).

In a final step, the outliers obtained from the ΔDAF, d_i and DIND approaches were systematically overlapped with the largest currently publicly available GWAS and eQTL datasets to support the interpretation of these signals. A few limitations of the latter methodologies and those of the overlap approach used here will be highlighted. Many of them apply to both GWAS and eQTL data as the underlying principles of QTL association mapping are similar. It should be noted that a considerable fraction of the sites resulting from the overlap is non-synonymous which might seem counterintuitive. However, in line with previous studies (Ye et al., 2013) these variants were kept as they still can be informative. Firstly, some non-synonymous mutations might represent genuine eQTLs, the simplest scenario would be a stop-gain mutation that lowers mRNA levels via NMD. However, there is little empirical evidence for this from the high ΔDAF nonsense mutations. The first of the two that are eQTLs for the gene they are located in counterintuitively leads to a higher expression (rs34358 in *ANKDD1B*) and while the directionality is consistent for the other (rs1790218 in *SLC22A10*) it is in LD with many other potentially causative loci. A further scenario is strong LD with a genic but non-

coding regulatory variant, i.e. in a promoter that is the true eQTL and could potentially also explain the signal of population differentiation. Finally, the non-synonymous variant could have a regulatory function in another transcription context where it falls in a non-coding region.

The focus of the selection scans on a small number of high frequency SNPs makes it inherently more likely that the overlapping GWAS results will be common high effect size variants from studies on traits that have a relatively simple architecture.

A good example for this is rs16891982 in *SLC45A2* (Appendix C.36), a major contributing factor to eye colour variation (Duffy, 2015), even though there can be interesting secondary effects on related traits, e.g. skin cancer risk in case of rs16891982 (see section 3.2.8). However, for most complex traits the proportion of variance explained by one single variant is relatively small (Visscher et al., 2017). In these cases, methods to detect polygenic selection applied to larger sets of variants that explain a significant fraction of heritability (if available) are likely more suited as demonstrated for height (Turchin et al., 2012; Mathieson et al., 2015). Furthermore, in absence of mechanistic evidence the exact variant highlighted is not necessarily the true selection target as the population differentiation signals as well as the relevant trait association are often spread across a region due to LD (Appendices C.42-C.43). While the signals reported in the GWAS Catalogue are usually lead SNPs, fine-mapping approaches suggest that, e.g. for a trait such as autoimmunity only as few as 5% of these variants may represent true causal SNPs (Farh et al., 2014). Furthermore, as already described (see section 3.3.3) for the GWAS databases currently the majority of individuals included in GTEx are of European descent, which is problematic as there likely is considerable interpopulation variation in gene expression patterns (Kelly et al., 2017). This also applies for age-dependent patterns, the GTEx dataset is biased towards older adult individuals with little data available on earlier life history stages (Ardlie and Guigó, 2017).

A problem specific to gene expression data is the role of trans-eQTLs. Evidence from model organisms suggests that they explain a substantial fraction of the genetic contribution to mRNA levels (Albert et al., 2017; Osada et al., 2017). Therefore, the 3,946 SNPs that act as trans- and long distance intra-chromosomal eQTLs identified from V6 of the GTEx datasets are very likely a substantial underestimate relative to the 152,869 cis-eQTLs identified for this stage of the project. However, it is difficult to quantify their role on a genome-wide level in humans as

the limited size of presently available datasets and the need to account for extreme multiple testing both contribute to low statistical power.

One of the motivations for combining highly differentiated putatively selected loci and cis-eQTLs was that changes in gene expression have been proposed to play a major role in adaptive evolution. Support for the importance of this mechanism comes from studies on model organisms such as yeast (Fraser et al., 2010) and also on specific phenotypes in humans, e.g. exposure to the pathogen *Mycobacterium leprae* (Manry et al., 2017). There have, however, been relatively few studies examining the systematic overlap of selection signals and eQTL data. The two most extensive analyses (Fraser, 2013; Ye et al., 2013) of this nature confirm the general pattern of an enrichment of cis-eQTLs among selection targets, however, both of these studies report a stronger effect than observed here.

3.3.7 A cis-eQTL upregulating the "speed gene" ACTN3 as potential selection target

Evidence for positive selection on rs11227639, an upregulating SNP impacting the *ACTN3* gene, in Northeast Siberians was examined based on a different metrics (ΔDAF, iHS, LD with the putatively epistatically interacting LoF rs1815739). For this signal to be confirmed in the future additional phenotypical evidence will be crucial. Individuals from Northeast Siberian groups carrying the upregulating derived allele at rs11227639 (and at least one functional allele at rs1815739) should exhibit greater muscularity and improved explosive power performance in a dosage-dependent manner mediated by elevated ACTN3 mRNA levels. Should this signal be confirmed the causative selective pressures for the greater muscularity could have been activity patterns and/or thermoregulation, which both in turn influence many other phenotypes.

The biological interpretation of the observed allele frequency patterns at rs11227639 in different populations is further complicated by past studies on the common LoF allele for *ACTN3* at rs1815739. Molecular changes in skeletal muscle cells of *Actn3* knockout mice, in particular increased cellular leakage and pumping of Ca^{2+} coupled with the elevated activity of mitochondrial oxidative enzymes, could have a possible thermogenic effect as they are similar to those of mice exposed to prolonged cold (Bruton et al., 2010). This has been interpreted by some authors (Head et al., 2015) as evidence supporting a scenario where the higher frequencies of the LoF allele at rs1815739 in Asians and Europeans relative to Africans indicate an adaptive benefit of this LoF phenotype in cold environments. The high frequencies of the functional

allele and putatively increased expression of *ACTN3* in Northeast and West Siberians, who live in very cold environments, as observed here (Appendix C.39) contradicts the role of the LoF allele at rs1815739 as a cold adaptation, at least in recent evolutionary times.

A contrasting interpretation of the high frequencies of the functional rs1815739 and the upregulating rs11227639 in some Siberian groups would be that these polymorphisms together lead to increased muscularity, as has been demonstrated for rs1815739 in isolation (Broos et al., 2016). This greater muscularity increases body volume. Following Bergmann's rule, which has been confirmed for modern humans in the northern hemisphere (Foster and Collard, 2013), this results in a reduction of the volume to surface ratio that in turn decreases heat loss. Future physiological work will be needed to examine whether and how these apparently contradictory effects at the cellular and organismal level affect the overall heat balance.

4. Analyses of rare variant sharing patterns

This chapter explores the distribution of rare variants in a global dataset of 447 whole genomes. Comprehensive surveys of rare variation from large-scale WGS studies have only become available in the last few years. They have revealed an excess of such variants below an MAF threshold of <1% in the respective samples, in which they constitute the majority of the SFS, compared to older empirical data and theoretical expectations. This implies accelerated recent growth for many populations (Gao and Keinan, 2016).

The accuracy of methods relying on local ancestry inference to detect IBD and population structure in general, is greatly improved by incorporating low frequency variants (O'Connor et al., 2015). Their availability also permits the development of novel approaches detecting continuous runs of rare variants indicative of shared ancestry (Mathieson and McVean, 2014; Fedorova et al., 2016). How and to what extent are rare variants shared between different worldwide populations in the EGDP dataset? Is there evidence for previously undescribed cryptic interpopulation relationships?

4.1 Material and Methods

4.1.1 Calculation of f_2 counts and related metrics

Dataset

The Diversity Set is a subset of the EGDP consisting of 447 individuals from 147 distinct local populations (Table 3.1, for the sampling strategy see Section 3.1.1). In contrast to the Variant-Based Analysis Set it additionally contains a heterogeneous collection of samples from Central Asia.

Computational Methodology

Unless stated explicitly otherwise, all data analysis steps described in this subchapter and sections 4.1.2 and 4.1.3 were run using R (R Core Team, 2017) and UNIX shell scripts custom-written for the respective purposes by the author. In the following, f_2 variants/doubletons are defined as variants for which only two individuals in the total dataset share the non-reference allele and for which those two individuals are heterozygous (Mathieson and McVean, 2014; The 1000 Genomes Project Consortium, 2012). This corresponds to a global allele frequency of ~0.22% in the Diversity Set.

All f_2 sites in the dataset were extracted using --maf and –max-maf flags in PLINK 1.9 (Chang et al., 2015a). The filtered dataset was then recoded as a VCF file which was transformed into a simple text file by custom-written bash and R scripts respectively. In this text file each line contained the IDs of the two individuals sharing an f_2 variant at a particular position in the genome. The f_2 counts for individual pairs were summed up using a modified version of a Perl script kindly provided by Ms Yuan Chen.

To adjust for the differences in overall genomic diversity between different individuals the raw f_2 counts were modified as follows. Let the two individuals of interest be i1 and i2 and designate the allele counts of all doubletons (shared with any other individual) recorded in these individuals as n(i1) and n(i2). The number of shared doubletons between i1 and i2 in absolute terms is n(i1∩i2). In order to convert this output to a relative measure of rare allele sharing d(i1∩i2), the following normalisation for differences in genomic diversity was applied:

$$d(i1 \cap i2) = n(i1 \cap i2) / \left(\frac{n(i1)+n(i2)}{2}\right) \quad (4.1)$$

To account for the influence of sample size a subset of samples designated as the "model world" was created. For this dataset all Africans were kept, however non-Africans were downsampled to eight groups of six individuals each. Some clusters of related populations, here Europeans and Siberians respectively, were summarised to a single group each. No Africans were removed as the original sampling strategy meant that African variation was underrepresented in the Diversity Set. To account for the substructure contained in the non-African clusters six individuals were drawn 20 times without replacement (see description of general strategy in section 3.1.3) and the f_2 statistics were calculated for each replicate, i.e. the total set of Africans and the respective replicate of non-Africans.

The relationships between individual pairwise similarity matrices of f_2 variant sharing, CP sharing (for details on CP analyses see section 3.1.1) and a geographic distance matrix were assessed using (partial) Mantel correlations (Mantel, 1967) as implemented in the vegan (Oksanen et al., 2017) R package. For each pair of individuals geographic distances in kilometres were inferred based on the great circle distance using the haversine formula as realised in the fields (Nychka et al., 2017) R package. To make the distance estimates between continents more reflective of human migration patterns the following waypoints were introduced: a) for all African/non-African pairs the Sinai (29.5N, 34E), b) for all American/non-American pairs the Beringia region (66.1N, 168.7W). The insertion of an additional corrective waypoint was necessary for a subset of individuals to avoid introducing unrealistic distance estimates. This applied to 20 Western Europeans whose sampling locations lie on longitude 11E or west of it and their distance to Native Americans. In these cases, the distance of the Western Europeans to the Beringian waypoint was inferred across the Atlantic and not across the Eurasian landmass as the great circle distance approach attempts to minimise the distance between coordinates. The corrective waypoint was Mount Narodnaya (60.0N, 65.1E) in the Urals. The total distance between two intercontinental points was then the sum of the great circle distances between the points and the waypoints connecting them plus the great circle distances between (potential) linking waypoints.

The significance of these matrix correlations was assessed using an approach where the respective dependent matrix was randomly permuted 1000 times. Negative values of the Mantel statistic occurred because the test was originally designed to assess the relationship between two dissimilarity matrices whereas the f_2/CP matrices describe similarity. As the permutation

approach only evaluates the significance of positive values of the Mantel statistic in these cases the dependent (f_2/FS) matrices were multiplied by -1.

For some Mantel tests the rare variant similarity matrices were natural log transformed. There is no general agreement on how this approach should be used if the untransformed dataset contains zeros, as is the case here, and what offset should be chosen (van den Berg et al., 2006; O'Hara and Kotze, 2010). One option is to fix the offset at one-half of the detection limit of the method (Zhang et al., 2015), i.e. here the smallest non-zero value for d(ix∩iy). This resulted in an offset of ~$6.87*10^{-6}$ for normalised f_2 sharing.

To further investigate the relationship between genetic similarity and geographical distance across space Mantel correlograms (Oden and Sokal, 1986) for the f_2/CP matrices and the great circle distance divided into 500-km bins (ranging from 0-500 km to 26,500-27,000 km) were generated using the mantel.correlog function in the vegan R package. For each of the distance classes the significance of the correlation between presence/absence (coded as 1/0) of an individual pair in this bin and its genetic similarity was assessed using a similar permutation approach as for the overall Mantel test. Due to the high number of bins a modified version of the Bonferroni correction following Holm (1979) was applied to adjust for multiple comparisons.

The goal of the next step in the analyses was to detect pairs of outlier individuals that share more f_2 variants than would theoretically be expected given their geographical distance. The plot of the raw data (Appendices D.1A-D.1B) is consistent with the hypothesis that there is a negative exponential relationship between geographical distance and normalised f_2 sharing. This can be related to historical concepts in population genetics such as the Malécot-Morton equation (Malécot, 1973; Morton, 1973a; b) expressing genetic similarity through common descent as a function of geographical distance under a one-dimensional isolation by distance model.

$$\theta_d = (1 - L) * a * e^{-bd} + L \tag{4.2}$$

In the above equation, d is the geographical distance between individuals/populations and θ_d a kinship coefficient. The parameter a can be interpreted as a measure of local kinship, b as the rate of exponential kinship decline and L as a scalar correction factor to eliminate negative kinship over large distances (the biological and mathematical validity of the latter has been debated, see Harpending, 1971).

The geographical distance data and normalised f_2 sharing were therefore fitted to a negative exponential model as given in Eq. (4.3) using a nonlinear least squares regression approach implemented in the R function nls (R Core Team, 2017). Note that the variable d and the parameters a and b are the same as in equation 4.2, whereas the scalar correction factor L has been summarised to a single constant c. More details on model fitting can be found in Appendix D.1C.

$$f_2 = a * e^{-bd} + c \tag{4.3}$$

4.1.2 Detection of rare variant clusters

The rare variant cluster (RVC) detection method proposed by Fedorova et al. (2016) applied here infers nearest-neighbour type IBD segments by exploiting the observation that rare variants shared between two individuals are often located in close proximity on the genome.

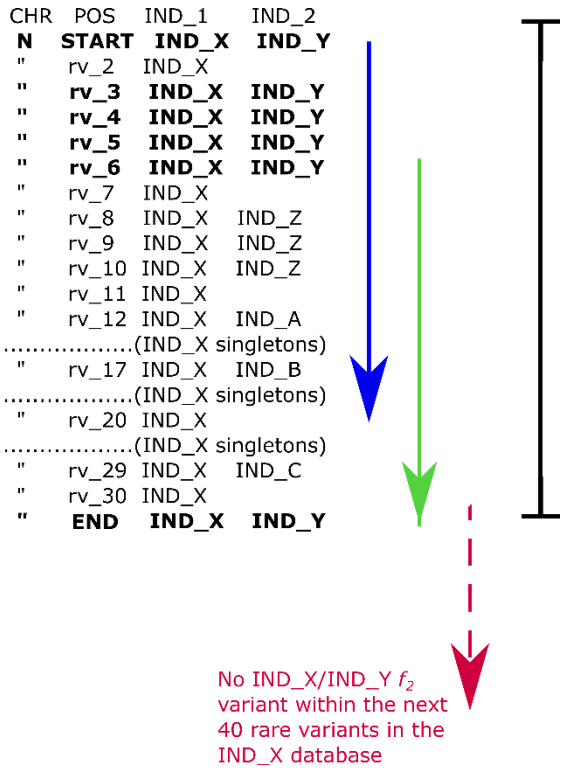


Figure 4.1: Example of RVC detection algorithm. 31 rows of an individual-specific vrGV (here singletons and doubletons) database are displayed. A core RVC is defined based on a scanning window (blue arrow) representing 20 consecutive rows of this database within which at least five f_2 variants have to be shared between individuals X and Y (bolded rows). The RVC is expanded if there is another uniquely shared vrGV between the two individuals of interest within 40 rows of the last vrGV of the core RVC (green arrow). This extension step is iterated as long as there is another uniquely shared vrGV within 40 rows of the previous one (red dotted arrow shows potential further extension which is not realised due to lack of shared doubletons). The length of this total segment was counted from the first vrGV of the initial core RVC to the last shared rare variant (black line). Abbreviations: IND – individual, rv – rare variant.

The frequency class of interest was denoted as very rare genetic variants (vrGVs) by the authors of the method and the frequency threshold was defined at a MAF of 0.2% in the respective dataset. For the Diversity Set, this approximately corresponds to singletons and doubletons. Firstly, individual-level databases of vrGVs were generated from a VCF file containing all individuals in the Diversity Set.

In the second step core regions of RVCs were defined in these databases according to the following rules:

i) A group of five or more vrGVs should be shared between two individuals and these variants should be adjacent to each other.

ii) There should be no more than 15 other non-shared vrGVs between these five shared rare variants detected in the individual which is eponymous for the database (this includes singletons).

These clusters were identified using a scanning window representing 20 consecutive rows of an individual-specific database of vrGVs and a stretching window representing 40 such rows respectively (for details see Figure 4.1). Fedorova et al. (2016) obtained the scanning and stretching window parameters by calibration on close relatives from the 1000 Genomes Project. They were adjusted in such a manner that the number and length of the resulting RVCs shared was consistent with the characteristics of IBD segments between close relatives predicted by population genetic theory.

The probability of random sharing of a run of five or more vrGVs with a frequency p of 0.0022 between two individuals in a window of n = 20 could be approximated as $Pr[binom(20,0.0022)\geq 5] \sim 7.74*10^{-10}$ (see also Hochreiter, 2013), i.e. it seems exceedingly unlikely to occur as IBS without any underlying IBD.

If the numbers and lengths of RVC chunks are calculated from individual-level vrGV databases as described above the resulting sharing matrices are non-symmetric, mainly because of variation in rare variant density between populations (see also 4.2.1). This additional information has been retained to describe the length distributions of RVCs and patterns (putatively) resulting from admixture events. However, for some general descriptions of the data the subtotals derived from the individual-level vrGV databases were averaged for each pair to create symmetric matrices. The individual-level vrGV and RVC databases were created using the Perl scripts provided by Fedorova et al. (2016).

The physical lengths of the inferred RVCs were converted to genetic lengths with base2genetic.jar (https://faculty.washington.edu/browning/beagle_utilities/ utilities.html) and a genetic map based on HapMap phase 2 data (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20110106_recombination_hotspots/). RVC statistics on the macro-group and population level were summarised in a manner similar to Beagle Refined IBD (see section 2.1.2).

The gnomAD database (Lek et al., 2016), which contains 15,496 genomes from a range of ancestries, was used as a reference to assess the continental frequencies of f_2 variants from the Diversity Set.

A plot of long-term N_e (0-30 kya, weighting as described in section 3.1.3) vs median RVC length $\tilde{L}$ (Appendix D.16A) is consistent with the hypothesis that there is a negative exponential relationship between these two quantities (a power law model would also be consistent with the data, this is briefly explored in Appendix D.18) This particular analysis was limited to a subset of 218 individuals from the Diversity Set for whom N_e could be inferred based on four individuals (Appendix C.2).

The exponential model was formulated as follows:

$$\tilde{L} = a * e^{-b*N_e} \tag{4.4}$$

It was fitted as described for Eq. (4.3) (see Appendix D.1C). Fisher's Z transformation (Fisher, 1915) as realised in the r.test function from the R package psych (Revelle, 2018) was applied to the Spearman's correlation coefficients for N_e over specific time intervals with $\tilde{L}$ to assess whether they differ significantly from each other. Vuong tests (Vuong, 1989) using the function vuongtest from the nonnest2 R package (Merkle et al., 2018) were conducted to compare model fits for N_e values representing the harmonic means for different time slices. Distributions of RVCs shared between different pairs of populations were compared using the k-sample Anderson-Darling test (Scholz and Stephens, 1987) implemented by the R package kSamples (Scholz and Zhu, 2018).

Furthermore, RVCs were assessed for consistency with the underlying complete genotype and haplotype data based on two main criteria. Firstly, it was investigated whether inside the boundaries of the RVCs there were any positions where the two individuals sharing the RVC exhibit incompatible homozygote genotypes. This was also the main criterion for the f_2 haplotype definition by Mathieson and McVean (2014) and IBD detection from unphased data

by Henn et al. (2012b). Secondly, potential inconsistencies with the haplotypes inferred using SHAPEIT2 (see section 3.1.1) were described. These RVC consistency analyses were run using a pipeline written by the author (described in detail in Appendix D.2).

4.1.3 f_3 and ALDER analyses

Pickrell's and Pritchard's (2012) implementation of the three population (f_3) test was used to investigate the detectability of a putative African admixture event in the Andean Calchaquíes. Four populations of interest (Calchaquíes, Siberian Eskimos, Yoruba and Wichi, $n_{total} = 22$) were considered. Following the recommendations of the authors of the method (Reich et al., 2009) the chosen approximate size for jackknife blocks to determine the significance of the f_3 statistic was ~5 cM which equals 708 windows each containing 15,005 SNPs. Population trios yielding a z-score smaller than -2 were considered significantly admixed. To assess the impact of two additional procedures for f_3 estimation the same dataset was analysed using ADMIXTOOLS (version 4.1) (Patterson et al., 2012) These modifications were: a) a normalisation for an abundance of SNPs shifted towards low frequencies, like in the case of WGS data analysed here, and b) a correction for inbreeding in the potentially admixed target population, analogous to what was proposed for the F_{ST} by Reich et al. (2009). ALDER v1.03, (Loh et al., 2013) was used to date the putative admixture event.

4.1.4 Demographic simulations using cosi2

Simulations were run to better understand rare variant sharing under a range of demographic scenarios. The first set of analyses (scenario A) consists of 20 independent replicates of a dataset designed to encapsulate the general features of the Diversity Set without any recent admixture between a Calchaquíes-like group and West Africans. The goal was to generate a null distribution for the RVC sharing statistics to which the empirical results can be compared. The second set of analyses (scenarios B-F) mostly assume that such a recent gene flow occurred, however admixture dates and fractions were set to vary between different runs.

The simulated data were generated using the coalescent simulation software cosi2 (Shlyakhter et al., 2014) in exact mode. The simulations encompassed five populations labelled Pop1 to Pop5 (Table 4.1) with group sizes corresponding to the sample distribution of the Diversity Set.

Table 4.1: Real groups whose demographic histories the simulated populations depicted in Figure 4.2 were designed to approximate and relevant sample sizes.

Simulated population	Real population represented	Number of sampled individuals
Pop1	Yoruba/West Africans	9
Pop2	Sandawe/East Africans	30
Pop3	Northwest Europeans/West Eurasians	209
Pop4	Han Chinese/East Eurasians	194
Pop5	Calchaquíes	5

The starting point for describing human population history was a simple demographic model for three well-studied reference groups (YRI, CEU, CHB) representing West Africans, West and East Eurasians respectively. It was estimated by Gravel et al. (2011) based on low-coverage data from phase 1 of the 1000 Genomes Project.

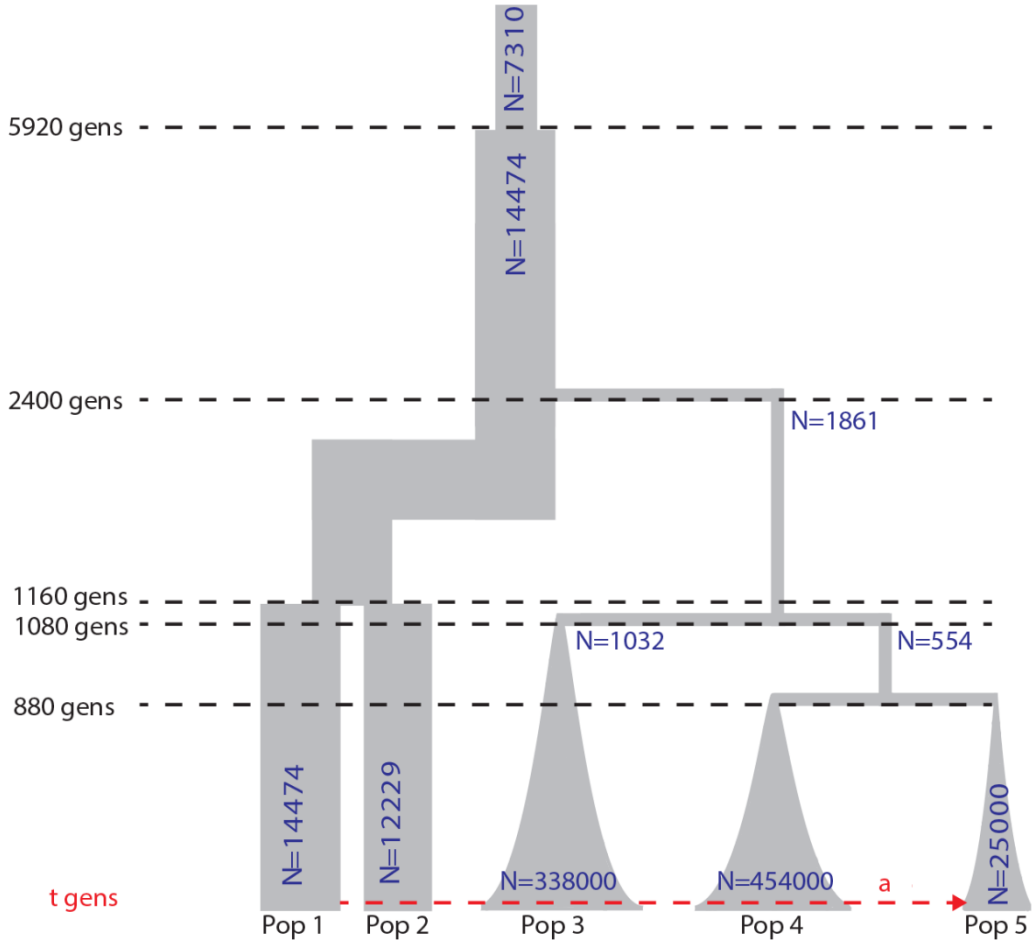


Figure 4.2: Five populations simulated to replicate general features of major continental groups from the Diversity Set. In some scenarios a recent admixture pulse of magnitude a from Pop1 (West-Africans) into Pop5 (Native Americans) that occurred t generations ago was added. Low-level intercontinental migration was included in the model; it is not depicted here for reasons of clarity.

Only two important modifications were made (Figure 4.2). The split time of Africans and non-Africans was increased to 2,400 generations ago based on the means of MSMC split times estimated for YRI vs CEU and YRI vs CHB respectively (generation time 30 years, Appendix D.3, see section 3.1.1). Again, using the MSMC results as reference, the split of East and West Eurasians was estimated as having occurred 1,080 generations ago. To mimic the populations in the Diversity Set two additional groups were introduced. The first (Pop2) is meant to represent a branch of East Africans. Its present-day N_e was the mean of the values obtained using MSMC for the Hadza and the Sandawe (Appendix C.3) and the split date from Pop1 was derived in a similar manner.

The values for the continuous low-level background migration between Pop2 and all non-African groups were assumed to be the same as for Pop1 and the non-Africans. This migration parameter between Pop1 and Pop2, i.e. within Africa was set to the same value proposed by Gravel et al. (2011) for Pop1 and Pop3. The second group (Pop5) was designed to mimic the Calchaquíes. It split from Pop4 880 generations ago, which is in the range of MSMC split dates calculated for the three Andean populations and the Han. The background migration rate of Pop5 with Pops1-3 was estimated as a tenth of the migration rate between the latter and the East-Asian like Pop4. Finally, between Pop4 and Pop5 this parameter was set to the same value as between Pop3 and Pop4.

The first of the added hypothetical scenarios considers a case where the intercontinental migration rate between Pop5 and Pops1/2 equals that between Pop4 and Pops1/2, i.e. it assumes higher background migration rates between Native Americans and Africans. The other four scenarios include a recent gene flow with admixture fraction of $a \in \{0.005,0.01\}$ from Pop1 into Pop5 $t \in \{15,18,20\}$ generations ago (combinations of $a = 0.01$ and $t = 15$, $a = 0.01$ and $t = 20$ were not run).

Each simulation run produced 22 chromosomes (lengths identical to autosomes in hg19, http://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/hg19.chrom.sizes). The mutation rate was assumed to be $\mu = 1.25*10^{-8}$ bp^{-1} per generation and the rate of gene conversion (relative to crossover recombination rate at the same locus) was kept at the default value of 0.45. The chromosome-specific recombination map for the cosi2 simulations was the one used for the MSMC analyses presented in Pagani et al. (2016) (see section 3.1.1) and derived from HapMap phase 2 recombination rates. The only significant change was that recombination rates

given as 0 in the original files were recoded as 10^{-15} as cosi2 cannot process recombination rates of 0.

After the simulations were run (an example parameter file can be found in Appendix D.4) the outputs were converted to VCF-like files using a customised pipeline consisting of bash and R scripts written by the author. To test whether the simulations can reproduce empirical patterns of global genetic diversity chromosome 22 from one randomly chosen replicate of scenario A was compared to chromosome 22 from the Diversity Set using the R/Bioconductor package SNPRelate (Zheng et al., 2012) implementation of PCA. First, both chromosomes were pruned for LD by recursively removing one SNP from each pair with an r^2 value greater than 0.2 within a 1000-SNP window. Then the SNPRelate implementation of PCA was run on the LD pruned sets of SNPs. Only markers with a MAF >1% were included in this analysis.

RVCs were extracted from the simulated sequence data and converted to genetic lengths for all scenarios as described for the real data in section 4.1.2.

4.2 Results

4.2.1 Global sharing patterns of f_2 variants

In the whole Diversity Set (n = 447) 4,206,999 f_2 variants were found, which translates to ~9,412 variants per haploid genome (Table 4.2). In this sample the African subgroup exhibits by far the highest number of f_2 variants with an average of ~44,400. The Oceanians are the next most diverse group with a per-individual count of ~16,800 f_2 variants, followed by the South Asian and ISEA macro-groups whereas most other Eurasians fall in a range of 5,000-6,000 doubletons per haploid genome copy. However, these counts depend both of the sample sizes of the populations, which vary considerably, and population history.

Table 4.2: Average counts of f_2 variants per haploid genome displayed a) for the unadjusted sample sizes, b) for 20 replicates of a subset of 87 individuals designated as the "model world". The latter was designed to reduce the relative overrepresentation of non-African populations and to control for the effect of sample size. The population-wise mean and SD across 20 replicates are given. Even though the set of Africans was fixed their sharing with non-Africans varies depending on the exact composition of the rest of the replicate. Abbreviations taken from Table 3.1.

	f_2 **counts Diversity Set**	f_2 **counts "model world"**
Afr	44386.1	47314.5+/-116.6
MiE	5768.8	13381.2+/-430.4

	f_2 counts Diversity Set	f_2 counts "model world"
WEu	5036.8	13720.9+/-223.7 (European supergroup)
EEu	4849.2	
Vol	5100.1	-
SoA	7841.9	11330.2+/-371.9
CeA	5613.6	-
WSi	4909.1	11681.8+/-392.0 (Siberian supergroup)
SSi	5696.6	
CSi	5149.0	
NSi	4793.8	
SeM	5858.5	10077.9+/-197.3
SeI	8626.4	11938.4+/-595.2
Ame	5151.0	11276.7+/-1255.7
Oce	16765.4	20191.1+/-2344.6
Whole dataset	9411.6	28354.6+/-453.0

When the overrepresentation of non-Africans is reduced and the effect of sample size among them is normalised for in the "model world" sampling scheme both the f_2 numbers for each haploid genome as well as the order of macro-groups change considerably. The increase in the average non-African f_2 counts (approximate doubling from 6,068.5 to 12,949.8) is a consequence of variants that occurred in three or more individuals before the reduction of the dataset now being counted as doubletons. This compensates for the loss of doubletons that became either singletons or were lost in downsampling. In terms of f_2 totals among non-African macro-groups. Oceanians remain the most diverse group. However, now Western Eurasians carry more doubletons than East Eurasians. Furthermore, Americans exhibit a high variance compared to other Eurasian groups, most likely because of the different f_2 sharing properties of admixed and unadmixed subgroups forming this cluster.

An additional analysis was run to rule out that the particularly f_2 low diversity in the East Asian/mainland SEA group is primarily the result of a relative overrepresentation of East Eurasians in the "model world". The latter would cause what would otherwise be doubletons to be recorded as more frequent variants. In a single dataset with the same sample size as the model world but less Eastern and more Western Eurasians (Appendix D.5) a merged group of East Asians and island/mainland SEA populations still only has an average of 11,124.4 f_2 variants per individual. Figure 4.3 (raw data in Appendix D.6) displays the proportion of doubletons a given population shares with any other group in the dataset. Overall, almost two thirds (65.4%) of f_2 variants are found in two individuals from the same macro-group. However, in continental Eurasia the majority of doubletons (~59.6%) represent sharing between two

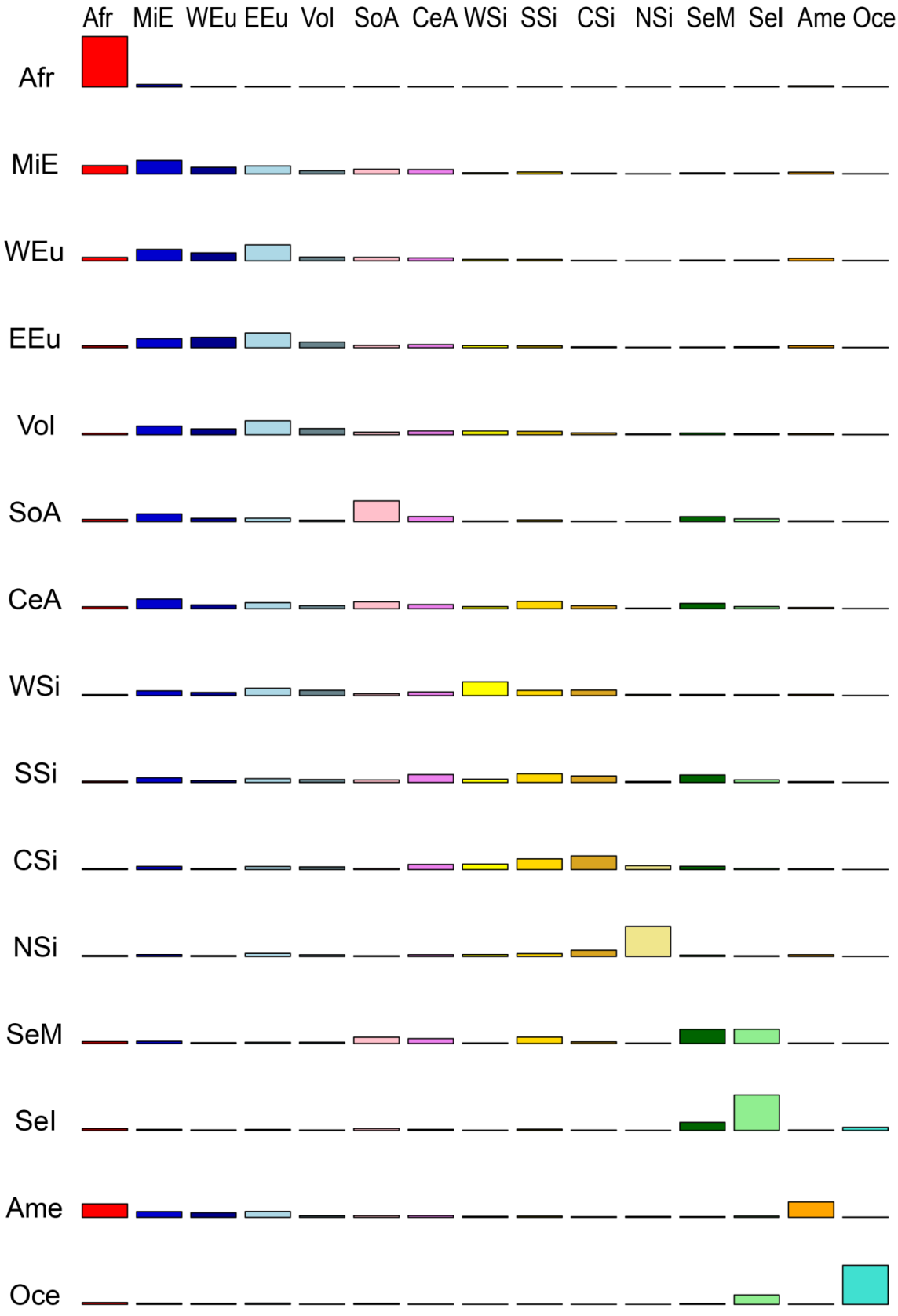


Figure 4.3: Average f_2 sharing between macro-groups. Each row represents how frequently f_2 variants are shared between individuals from the target macro-group (indicated on the left) and all other macro-groups (given on top) relative to the f_2 variant total in the target macro-group. Abbreviations taken from Table 3.1.

distinct macro-groups, meaning that only individuals from African, Oceanian, Island Southeast Asian and Northeast Siberian groups share the majority of doubletons with themselves. For example, Western Europeans share more of their f_2 variants with Eastern Europeans (31.2%) than with individuals from their own group (13.1%). This pattern of diverse inter-group ancestry is most visible in Central Asians, who share considerable amounts of f_2 variants with populations from the Middle East and the Caucasus (16.8%), Eastern/Northern Europe (10.6%), South Asia (11.9%), South Siberia (12.7%) and East Asia/MSEA (9.2%). As suggested by the variance of f_2 counts among the resampling replicates for the "model world" (Table 4.2) these macro-population labels contain considerable substructure which here is explored in form of individual-level doubleton coancestry matrices.

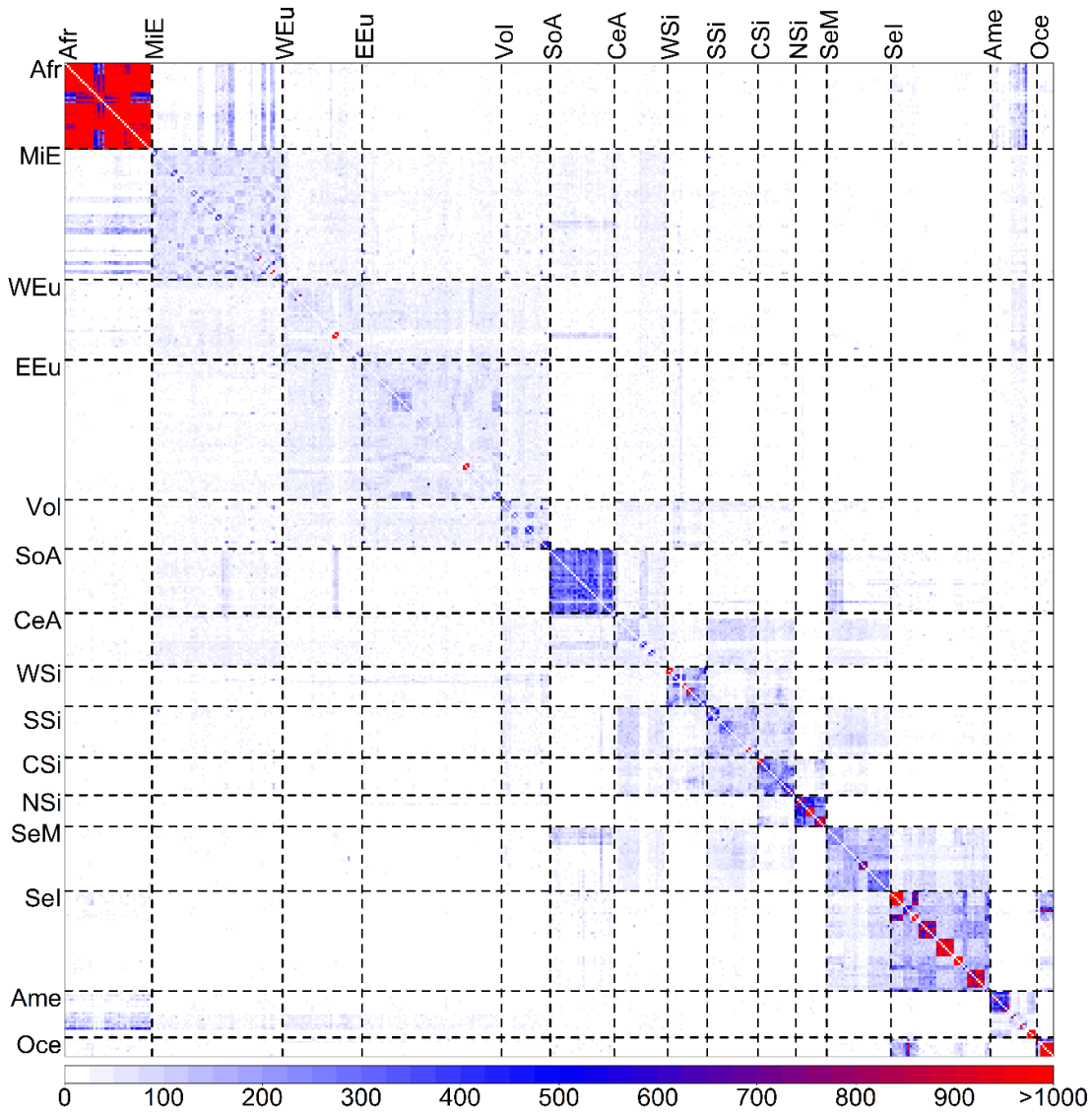


Figure 4.4: Raw sharing totals of f_2 variants between all individuals from 15 macro-groups belonging to the Diversity Set. Abbreviations taken from Table 3.1.

When the raw individual-level numbers (Appendix D.7A) are plotted in form of a heatmap (Figure 4.4) multiple clusters of high f_2 sharing (>1000 variants per pair) emerge. Consistent with the highest genetic diversity in Africa, African individuals, regardless of their population affinity, exhibit the highest f_2 sharing. Outside Africa, uniformly high doubleton sharing (~300 per average pair) is observed in the Indian subcontinent.

Most East Eurasian macro-groups appear to be more stratified than West Eurasians according to this metric, especially Northeast Siberians and ISEA groups. Among Americans the three Andean populations form very distinct clusters and the two Australians exhibit the highest raw sharing of any pair of individuals ($n_{f2\_Aus1 \cap Aus2}$ = 36,123). Finally, Papuans display strongly elevated sharing among each other while the Koinanbe and Kosipe subpopulations in turn share considerably more doubletons with members of their own group (not visible in Figure 4.4 due to scale).

The observed differences between African and non-African groups primarily reflect how strongly f_2 variant diversity is driven by overall genomic heterozygosity (even though this relationship is dependent on sample size and composition, see also section 4.3.1). Therefore, an f_2 statistic that normalises for differences in overall f_2 diversity was defined (Eq. 4.1).

On a general level, this diversity-normalised metric paints a similar picture to the macro-group and raw individual-level similarity statistics. The quantitatively strongest sharing can be observed within macro-groups, however there are a multitude of distinct out-sharing events away from the matrix diagonal (Figure 4.5, raw data in Appendix D.7B). The median normalised individual-level f_2 sharing within macro-groups is ~0.0070, ca. 12 times higher than the median for the normalised sharing between macro-groups (~$5.9*10^{-3}$). While this cut-off is necessarily subjective, doubleton sharing of groups from different macro-populations with a pairwise $d_{ix \cap iy}$ (Eq. 4.1) > 0.005 can therefore be potentially informative about recent gene flow. Here, the aim is not to provide an exhaustive descriptive analysis of all such links, but to highlight a few of interest.

The outcomes of a quantitative outlier approach of out-sharing given geographical distance will be reported at the end of this section. African populations mostly appear to be very distinct from non-Africans, however they exhibit elevated sharing with some populations from the Middle East (Saudi Arabians and Muslim Arabs from Israel) and the Americas (Mexicans and Puerto Ricans). Similarly, Oceanians show little f_2 sharing with other groups apart from clearly visible

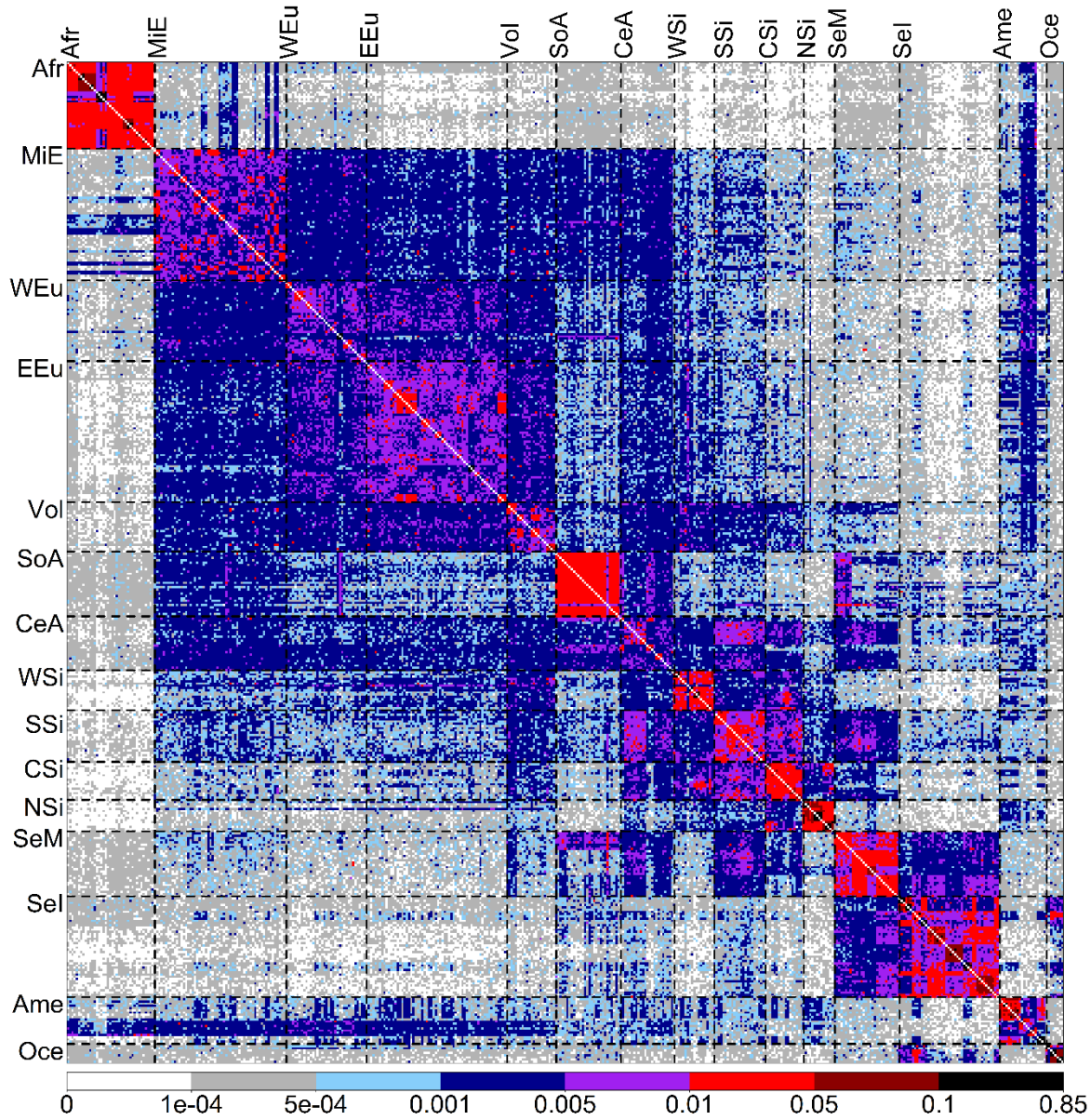


Figure 4.5: Sharing of f_2 variants between all individuals from 15 populations belonging to the Diversity Set. The metric is normalised to account for differences in overall genomic diversity. Abbreviations taken from Table 3.1.

interactions with various ISEA populations of Negrito as well as non-Negrito origins consistent with the analyses on SNP chip data in chapter 2. There is a high degree of inter-individual sharing across macro-populations within Eurasia (see also Figure 4.5).

The perhaps most striking example, which also demonstrates the amount of substructure in some macro-groups, occurs among Central Asians. A subgroup of Indo-Iranian speakers (Ishkashim, Rushan-Vanch, Shugnan, Tajiks and Yaghnobi) exhibits a three times higher average rare variant sharing with South Asians ($d_{CeA\_IndoIranian \cap SoA} = 5.4*10^{-3} \pm 1.8 *10^{-3}$, here

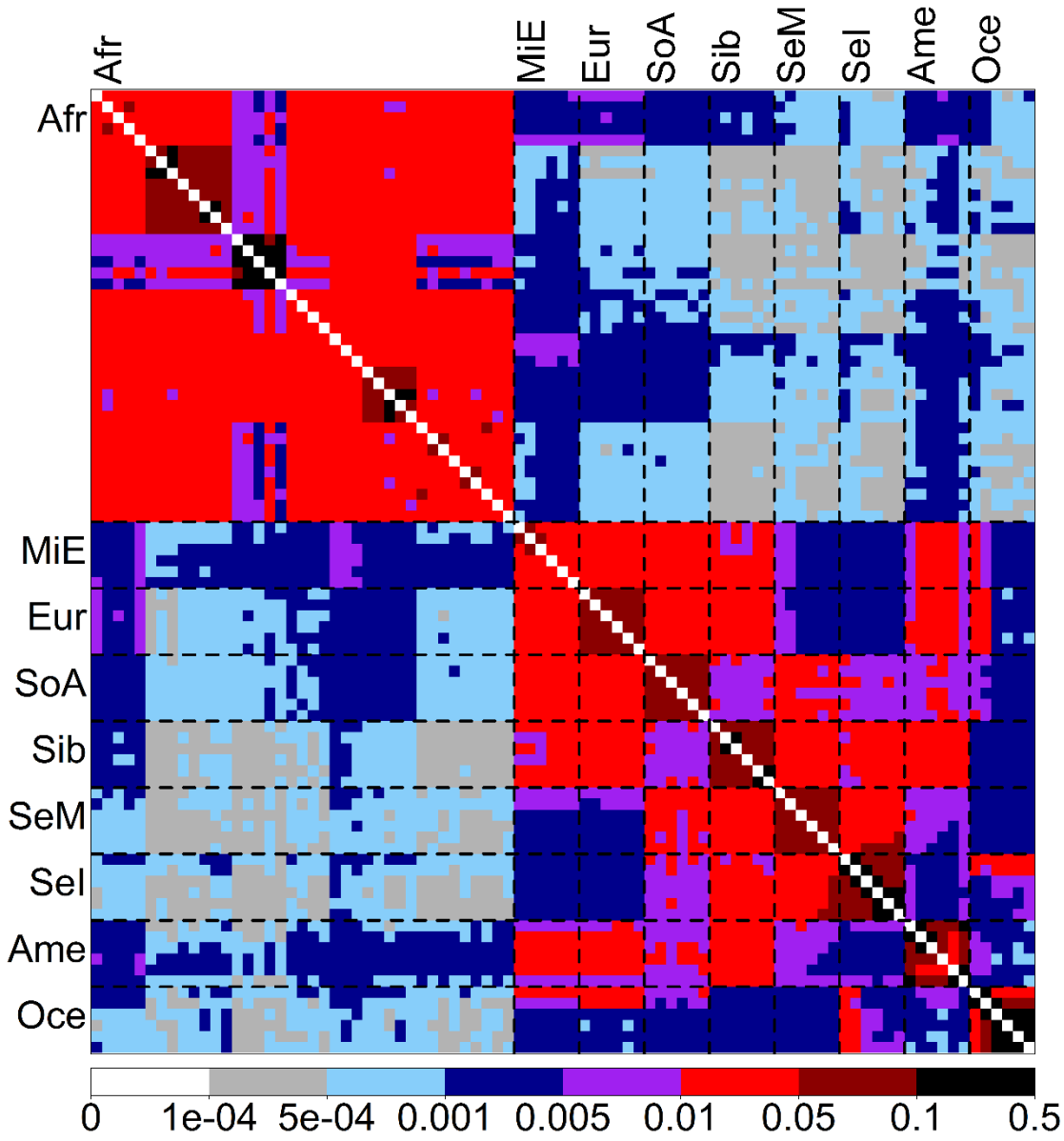


Figure 4.6: Mean sharing of f_2 variants derived from 20 replicates of a reduced subset of individuals (n=87) ("model world") belonging to the Diversity Set. The set of Africans was fixed and the non-Africans were subsampled 20 times. The metric is normalised to account for differences in overall genomic diversity. Abbreviations taken from Table 3.1., except: Eur – Europeans, Sib – Siberians.

and in the following "± x" denotes the standard deviation) than a subgroup of Turkic speakers (Kazakhs and Kyrgyz) ($d_{CeA\_Turkic \cap SoA} = 1.8*10^{-3} \pm 1.1 *10^{-3}$). For affinities to South Siberian/Mongolian populations the ratios are reversed ($d_{CeA\_Turkic \cap SSi} = 7.5*10^{-3} \pm 2.1 *10^{-3}$, $d_{CeA\_IndoIranian \cap SSi} = 2.0*10^{-3} \pm 1.0 *10^{-3}$). Three other Turkic-speaking populations (Turkmens, Uighurs and Uzbeks) fall somewhere between the two former clusters ($d_{CeA\_Turkic\_mixed \cap SoA} = 2.8*10^{-3} \pm 1.2 *10^{-3}$, $d_{CeA\_Turkic\_mixed \cap SSi} = 4.7*10^{-3} \pm 1.5 *10^{-3}$).

The major sharing patterns of f_2 variants remain consistent in the "model world" (Figure 4.6, raw data in Appendix D.8) compared to the full dataset. The clustering of all non-Africans as one group becomes more visible with the Oceanians on average being less similar to the Eurasians and Americans than the latter are to each other. The mean rare variant sharing in this Eurasian-American cluster is however still slightly below that of all Africans ($d_{African \cap African} = 0.025 \pm 0.025$, $d_{nonAfrican \cap nonAfrican} = 0.021 \pm 0.022$).

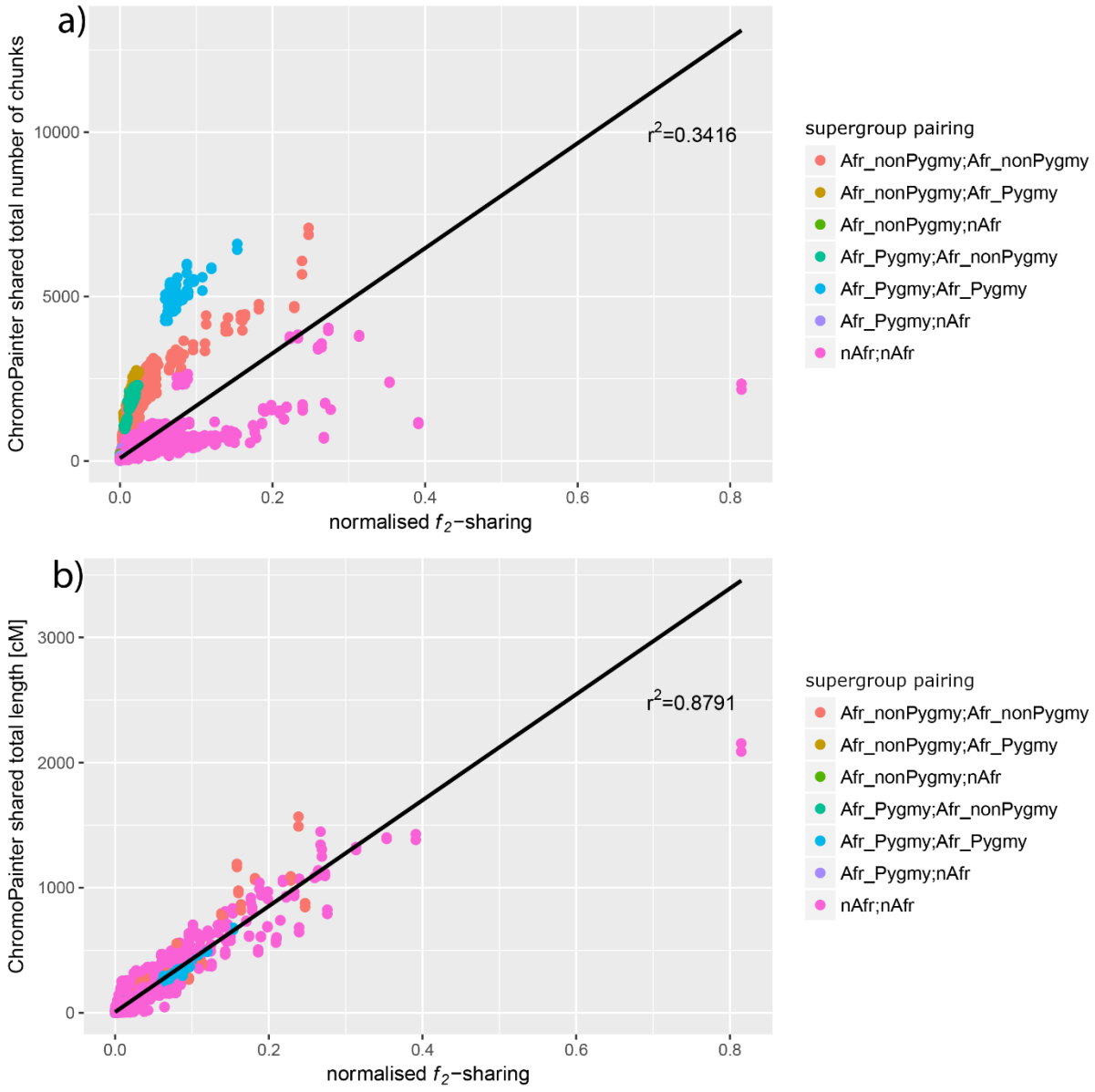


Figure 4.7: Normalised f_2 sharing between individuals from the Diversity Set plotted against a) the total number of shared chunks inferred by CP and b) the total length of all these chunks for any particular pair of individuals. The colours indicate which macro-groups the two individuals forming the respective pair belong to. The CP sharing is slightly asymmetric, i.e. there are cases where the chunks individual x receives from individual y are not fully identical to the ones received by y from x in turn. Therefore, each x value (f_2 sharing) from a pair is matched to two y values. Regression lines and coefficients of determination are displayed on the plot. Abbreviations: Afr – African, nAfr – non-African.

In this non-African group the divergence between West Eurasians (Europeans and Middle Easterners) and East Eurasians (East and Southeast Asians) is apparent while the (mixed) Siberian and South Asian groups exhibit affinities to both subclusters.

To assess the information content provided by the normalised f_2 metric more systematically it was here compared to the output of CP (raw data in Appendix D.9), a widely used method to infer recent shared ancestry on a haplotype level. Normalised f_2 sharing is moderately correlated to the number of haplotype segments two individuals have in common (r ~0.584). Compared to non-Africans, pairs of individuals from African populations share an excess of CP chunks relative to their f_2 sharing (Figure 4.7a). This effect is even more pronounced for Pygmies. The opposite can be observed for some non-African pairs, who fall below the regression line, where there is a high amount of f_2 sharing (>0.2) corresponding to >2000 shared CP segments for Oceanians and >1000 shared segments for some intrapopulation pairs of Philippine Negritos and Andeans. Using the total length of all shared CP chunks, however, results in a much stronger linear correlation (r ~0.938) (Figure 4.7b).

When normalised f_2 sharing is used as a linear predictor for total CP sharing, which a mean value of 15.69 cM across all pairs of the Diversity Set, the resulting mean absolute error is ~5.34 cM. Rank correlation tests yield somewhat lower correlation coefficients (Spearman's $\rho = 0.865$, Kendall's $\tau = 0.681$). The most notable outlier pair remaining consists of two Aboriginal Australians, who share the vast majority of their doubletons ($d_{Aus1 \cap Au2} = 0.815$) in contrast to being closest matches for 30% of the whole genome as indicated by CP. The likely underlying reasons for this excess f_2 sharing are that a) the two individuals were identified as second-degree relatives with the KING method (see section 3.1.1 and column "Comments" in Appendix C.1, they were originally retained despite being more closely related than the threshold to maximise geographical spread) and b) they are the only representatives for the populations of a whole continent. If the two Australians are removed, the correlations for both outcomes of CP and the normalised f_2 sharing improve further ($r_{chunkcount\_f2} = 0.613$, $r_{totallength\_f2} = 0.946$).

To more systematically analyse outlier pairs standardised residuals from the linear regression of CP and f_2 sharing were used. For a total of 1,741 combinations of two individuals the absolute value of the standardised residual is greater than 3 (Appendix D.10). There is a moderate excess of pairs with less CP sharing than would be predicted based on f_2 relative to pairs with more CP affinities than expected (989 vs 752). Both groups mostly consist of pairs from the same macro-

group with a substantial fraction belonging to the same local population (84.8%/31.8% for CP depletion and 88.9%/24.4% for CP excess respectively). The group with less than expected CP sharing given their shared doubletons contains five pairs of known relatives.

Another interesting observation is that among the CP-sharing depleted group there is a significant enrichment of pairs of individuals previously identified as exhibiting unusual metrics in the files provided by Complete Genomics (77/989 observed vs 2/989 expected based on the fraction of unusual metrics pairs among all possible pairs, $X^2 = 72.2$, $p < 10^{-15}$). Finally, the macro-groups which most pairs in this class belong to are Africans, South Asians and ISEA groups (primarily Philippine Negritos). On the other hand, in the excess CP group there are no pairs where both samples have unusual metrics and the respective dominant macro-groups are Native Americans, ISEA groups (most prominently the Kankanaey Igorot), all Siberians and one cross-population set of pairs (n = 27), North Americans and Siberian Eskimos.

To analyse spatial structure in rare variant sharing a Mantel test was applied. It investigates whether there is a linear relationship between the f_2 similarity matrix and a geographic (great circle) distance matrix. The resulting negative Mantel correlation coefficient ($r_{M\_f2\_D} = -0.229$, $p < 0.001$) is significant. However, its magnitude implies that given these assumptions geographical distance would only explain ca. 5.2% of the overall variance in the rare variant similarity matrix. This low explanatory power for a simple linear model is not unexpected given that the relationship between great circle distance and f_2 sharing appears to be primarily non-linear (Appendix D.1). The latter observation is supported by the considerable increase of the correlation when the normalised f_2 sharing matrix is log transformed ($r_{M\_logf2\_D} = -0.499$, $p < 0.001$). Furthermore, Spearman's rank correlation ($\rho = -0.721$, $p < 2.2*10^{-16}$) and Kendall's tau ($\tau = -0.531$, $p < 2.2*10^{-16}$) support a robust monotonic negative relationship between geographical distance and pairwise doubleton sharing.

To visualise the autocorrelation of rare variant sharing in relationship to the great circle distance between the pairs of individuals a Mantel correlogram was generated (Appendix D.11A). Individuals whose sampling locations are very close to each other (midpoint of first distance class: 250 km) tend to share more rare variants than all other pairs on average ($r_{M\_f2\_250kmbin} = 0.419$, $p < 0.001$). The Mantel correlation already decreases sharply for the second closest distance class and then progresses to decline gradually in an approximately linear manner. From 3000 km onwards, the Mantel correlation becomes statistically indistinguishable from zero and beyond 4000 km it turns negative, i.e. now individuals located at the respective distances are

less similar than the average across all pairs. The values reach a minimum at ~7000-7500 km ($r_{M\_f2\_7250kmbin}$ = -0.058, p = 0.015). This is followed by a stabilisation and a subsequent slight increase accompanied by fluctuations.

The Mantel test results for a matrix containing the total shared genetic length according to CP and the great circle distance matrix are broadly comparable to those obtained for rare variant sharing. However, the correlations of the raw CP sharing matrix ($r_{M\_CP\_D}$ = -0.293, p < 0.001) as well as that of its log transform ($r_{M\_logf2\_D}$ = -0.558, p < 0.001) with geographical distance are slightly higher. Furthermore, the Mantel correlogram of CP sharing exhibits a very similar shape to the one obtained with f_2 sharing as the response variable (Appendix D.11B). The only notable difference is that the distance-class-specific Mantel correlation coefficient for the total shared CP length remains significantly greater than zero until ~4000 km, i.e. over a somewhat greater distance than for the f_2 sharing.

Finally, the goal was to detect outliers that exhibit more sharing of rare variants than would be expected given their geographical distance. Prior theoretical work as well as examination of the plotted raw data suggests that a possible approximation would be a three-parameter exponential model as described in Eq. (4.3). The best fitting model (plotted in Appendix D.1) with all coefficients rounded to the third decimal to predict f_2 sharing from great circle distance d is:

$$f_2 = 0.058 * e^{-0.017d} + 0.002 \quad (4.5)$$

Outliers were defined based on the empirical top 1% of a normalised residual metric given in Eq. (4.4) for four broad distance bins: 1,000-4,999 km,>5,000-9999 km,>10,000-19,999 km, >20,000 km. In the following a few inter macro-group pairs exhibiting such excess sharing are highlighted (all top 1% outlier pairs given in Appendix D.12, the full dataset of predicted f_2 sharing and the normalised residuals in Appendix D.13).

The most frequently detected outliers in the first distance bin comprise among others Central Asians (Kazakhs, Kyrgyz) sharing with Siberians/Mongolians and the Vietnamese showing affinities to ISEA groups from Northern Borneo (Dusun, Murut). Furthermore, the Bajo and the Papuans as well as the Tamang and the Burmese share more rare variation than other similarly geographically distant population pairs.

In the following distance class the Japanese display elevated sharing with some Central Asians (Kazakhs, Kyrgyz). Some individuals from Middle Eastern groups (Saudi Arabians, Iranians) exhibit an excess of rare variant similarity with a broad range of Sub-Saharan African groups.

Across the same distance the Roma share a high number of doubletons with many South Asian populations. The excess sharing of a British individual (UK1) with two Chinese (Chn2, Chn3) likely represents a methodological artefact (see section 4.3.1).

The two distance classes beyond 10,000 km are dominated by sharing of known admixed groups from the Americas (Mexicans, Puerto Ricans) with a range of Old-World populations of Western European as well as Sub-Saharan African ancestry. An intriguing exception to this pattern is the finding that some individuals from the Andean Calchaquíes population, who in common-variant-based analyses (Pagani et al. 2016) did not show signs of admixture from outside the Americas, exhibit elevated sharing with the Yoruba as well as some West Eurasian groups (Druze, Lezgins, North West Europeans). The possibility that this unexpected sharing is the result of recent admixture is examined more rigorously in section 4.2.3.

4.2.2 Global sharing patterns of rare variant clusters

A total of 301,920 RVCs were detected based on the f_2 variants in the Diversity Set (all runs given in Appendix D.14). These span an extent of ~262 Gb in physical length and 2,359 M in genetic length reflecting approximately 20.4% and 16.0% of the autosomes for each individual on average, respectively. The corresponding median values for these two different measures of run length are ~308 kb and ~0.20 cM (Figure 4.8) while the median number of doubletons contained in each cluster is 8.

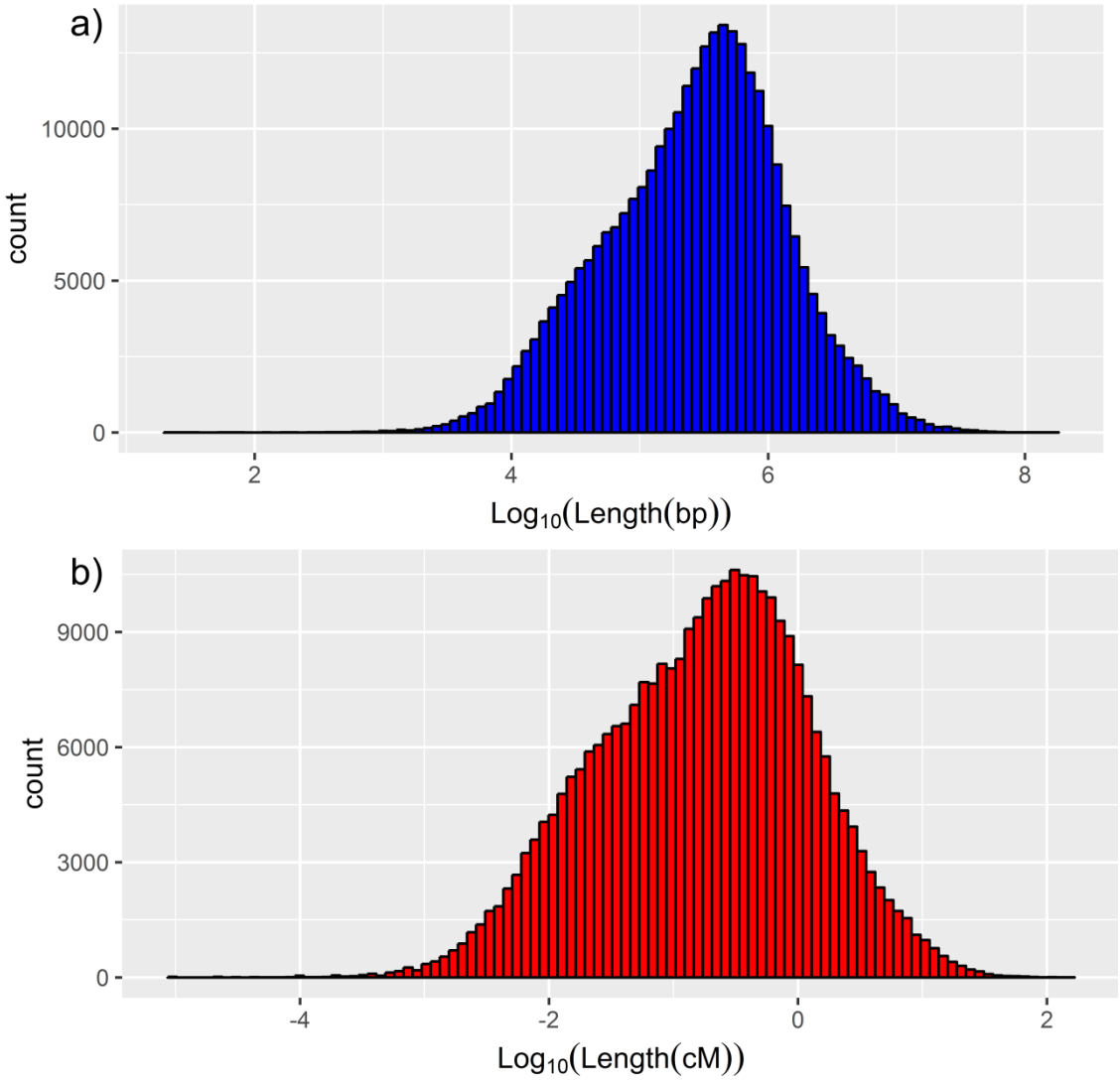


Figure 4.8: Histograms of RVC numbers as a function of physical and genetic length.

Of the RVCs 61,601 (20.4%) are shared within the same local population and 225,403 (74.7%) within the same macro-group. To understand the relationship between CP sharing and the rare-variant related metrics calculated in this chapter, a matrix displaying all possible correlations was generated (Table 4.3).

Table 4.3: Matrix displaying Pearson's correlation coefficient for different genomic sharing metrics between pairs of individuals from the Diversity Set. All correlations are significant ($p < 2.2*10^{-16}$). Abbreviations: norm - normalised.

	CP total length	f_2 sharing norm	RVCs shared	RVCs shared averaged	RVCs total length	RVCs total length averaged	Fraction of total f_2 in RVCs
CP chunk count	0.617	0.585	0.932	0.932	0.437	0.437	0.342
CP total length		0.938	0.604	0.604	0.859	0.859	0.364
f_2 sharing norm			0.62	0.62	0.91	0.911	0.362
RVCs shared				1	0.521	0.521	0.328
RVCs shared averaged					0.521	0.521	0.328
RVCs total length						0.999	0.206
RVCs total length averaged							0.206

The first important insight from these analyses is that the correlation between total CP sharing and the total length of all RVCs is worse than that between the former and a normalised f_2 sharing metric. Furthermore, under the assumption of linearity adding the genomic length covered by shared RVCs between two individuals to a simple model of f_2 sharing as the only predictor for CP does not lead to increased explanatory power. This is supported by both the coefficient of determination which remains at ~0.879, and the mean absolute error, which even slightly increases to 5.43 cM (the regression approach attempts to minimise the mean squared error, the latter decreases marginally from 127.97 cM^2 to 127.81 cM^2). Secondly, the number of RVCs shared between pairs of individuals correlates tightly with CP chunk counts. This implies that the different distributions of RVC lengths between pairs contain additional information about population history beyond what is captured by a genome-wide summary statistic describing normalised f_2 sharing. Finally, the fraction of shared f_2 variants located in RVCs is not strongly correlated with any other metric considered.

Table 4.4 gives an overview of the patterns of intra-macro-group RVC sharing. The number of pairwise shared RVCs is highest for Africans and Oceanians. Among Eurasians the Northeast

Table 4.4: Different metrics describing the properties of RVCs shared within macro-groups in the Diversity Set. Abbreviations taken from Table 3.1.

Macro-group	Number of RVCs per pair	Total length of RVCs per pair [cM]	Mean lengths of RVCs [cM]	Median length of RVCs [cM]
Afr	92.33	33.83	0.37	0.07
MiE	2.46	2.17	0.88	0.34
WEu	2.08	4.62	2.22	0.52
EEu	2.65	3.56	1.34	0.65
Vol	4.38	10.40	2.37	1.36
SoA	11.84	5.15	0.44	0.21
CeA	2.61	3.58	1.37	0.61
WSi	10.58	43.87	4.15	1.77
SSi	6.27	13.21	2.11	0.97
CSi	12.03	40.73	3.39	1.92
NSi	26.46	114.81	4.34	1.99
SeM	5.89	3.88	0.66	0.31
SeI	13.20	24.42	1.85	0.66
Ame	7.27	31.99	4.40	2.29
Oce	100.36	245.63	2.45	0.86

Siberians represent an outlier with a particularly strong excess of region-specific rare variant chunks while the within-macro-group RVC number is generally higher for East Eurasians and lower for West Eurasians. The cumulative pairwise length covered by RVCs as well as the mean/median lengths of RVCs shared within macro-groups present a less easily generalisable picture with notable differentiation between geographically adjacent macro-groups and even neighbouring local populations (Appendix D.15). The median length of RVCs is highest in Native Americans and Siberians (except South Siberians) and lowest in the African macro-group. The sharing properties of intra-macro-group RVCs like raw f_2 metrics are affected by sample size and composition as well as parameters from demographic history in a complex manner (see section 4.3.1).

To explore the latter relationship further the weighted harmonic mean of N_e inferred by MSMC over the last 30 kya was compared to the median intra-macro-group RVC length. The scatter plot of the raw data hints at a negative exponential relationship (see Eq. 4.5). The best-fitting model (Appendix D.16A) resulting from nonlinear least squares approximation is:

$$\tilde{L} = 4.345 * e^{-1.865*10^{-4}N_e} \quad (4.6)$$

From Appendix D.16A it is apparent that in the range between a long-term N_e of 5,000 and 10,000 the fit of an exponential model is worse than for the rest of the dataset. There are at least

two potential underlying causes: a) factors which are not captured by N_e influence median RVC length, b) the harmonic mean weighs smaller and therefore generally more ancient values of N_e too heavily for this type of analysis. The second statement would imply that there are differences in goodness of fit for different slices of this period. Therefore, the harmonic means of N_e for 5,000-year intervals were analysed separately. Spearman's rank correlation coefficients suggest that the relationship between N_e and median intrapopulation RVC length is strongest for the three most recent time intervals, i.e. 0-5 kya, 5-10 kya, 10-15 kya (ρ between -0.670 and -0.812, Appendix D.17A) whereas the correlations for earlier time intervals are significantly lower (Appendix D.17B).

To distinguish between the recent time intervals three separate exponential regression models were constructed (Appendix D.16B, Appendix D.16C). The Vuong test indicates that using N_e from the period 5-10 kya results in a significantly better fit than with N_e from the most recent time interval ($Z = 2.652$, $p = 0.004$), whereas the fits for 5-10 kya and 10-15 kya cannot be distinguished by this test statistic ($Z = 0.573$, $p = 0.283$).

Each pair of macro-groups shares at least one RVC if the latter are summed across all individuals in the respective macro-groups (Appendices D.19-D.22). The absence of recent gene flow and a sufficiently deep split between two macro-groups should cause a pattern where there are no or only very short RVCs present. Northeast Siberians and Africans should provide an example for such a scenario. These two macro-groups share 6.5 RVCs (this value is a non-integer due to the asymmetry described in section 4.1.2, Appendix D.23) with each other. While the majority of these are short, one Chukchi individual (Chuk8) has two long RVCs (2.81 cM, 6.42 cM) in common with the East-African Maasai. One plausible explanation for this observation would be that these runs represent more recently shared West Eurasian ancestry. Common variant-based analyses suggest that some sampled individuals from both populations have a West Eurasian-like ancestry component (Pagani et al., 2016). West Eurasians and Africans are also known to share more recent gene flow (Moorjani et al., 2011). The individual in question, Chuk8 exhibits an almost sixfold excess of CP sharing with Europeans relative to the other analysed Chukchi ($CP\_L_{Chuk8 \cap Eur}$ ~1,665 cM vs $CP\_L_{Chuk\_other \cap Eur}$ ~288 cM).

To further test this hypothesis, it is important to know in which other populations the rare variants in these two RVCs occur (Appendix D.24). The gnomAD database, a public repository of more than 15,000 genomes, provides a resource to address this question. Two of the five f_2 variants which make up the first (10:47,100,598-48,653,790) of these RVCs can be found in

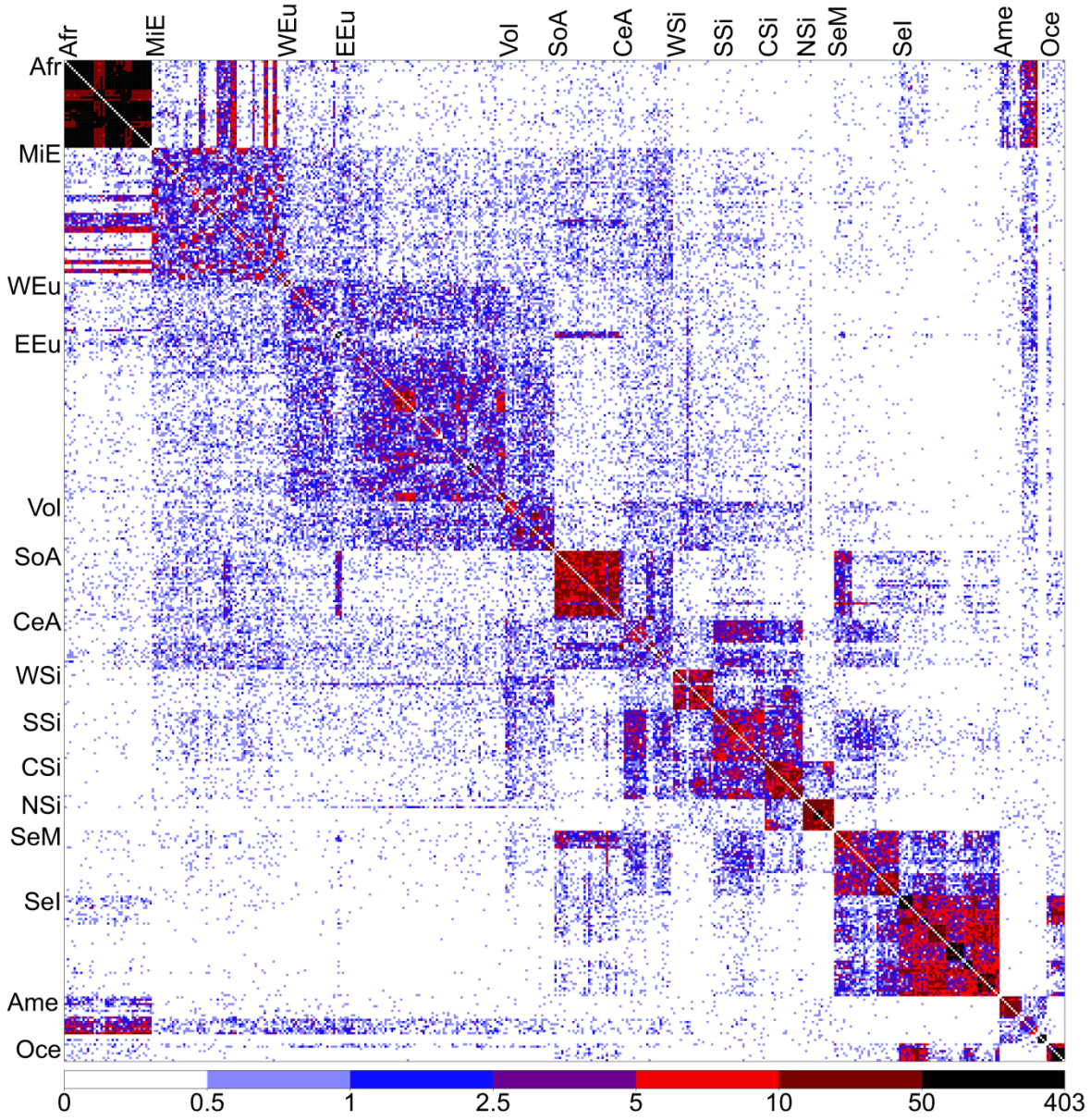


Figure 4.9: Pairwise number of shared RVCs. The average over the counts based on the two individual-level vrGV databases was taken. Abbreviations taken from Table 3.1.

gnomAD and these are most common in Europeans indicating a West Eurasian origin of this chunk (Appendix D.25). The second run (17:1,963,596-5,052,505) presents a more complex case where the first five f_2 variants, which are clustered in a short stretch with a genetic length of 0.03 cM, suggest an African origin whereas the last SNP reaches its highest frequency in Europeans.

A possible explanation for this pattern would be failure of the RVC method to distinguish two distinct runs. The lengths of the other runs (< 0.23 cM) are not inconsistent with ancient sharing predating the OOA event as runs with a length of 0.23 cM are still located within 2 SDs of the mean length of Neanderthal chunks detected in West Eurasians (Sankararaman et al., 2014),

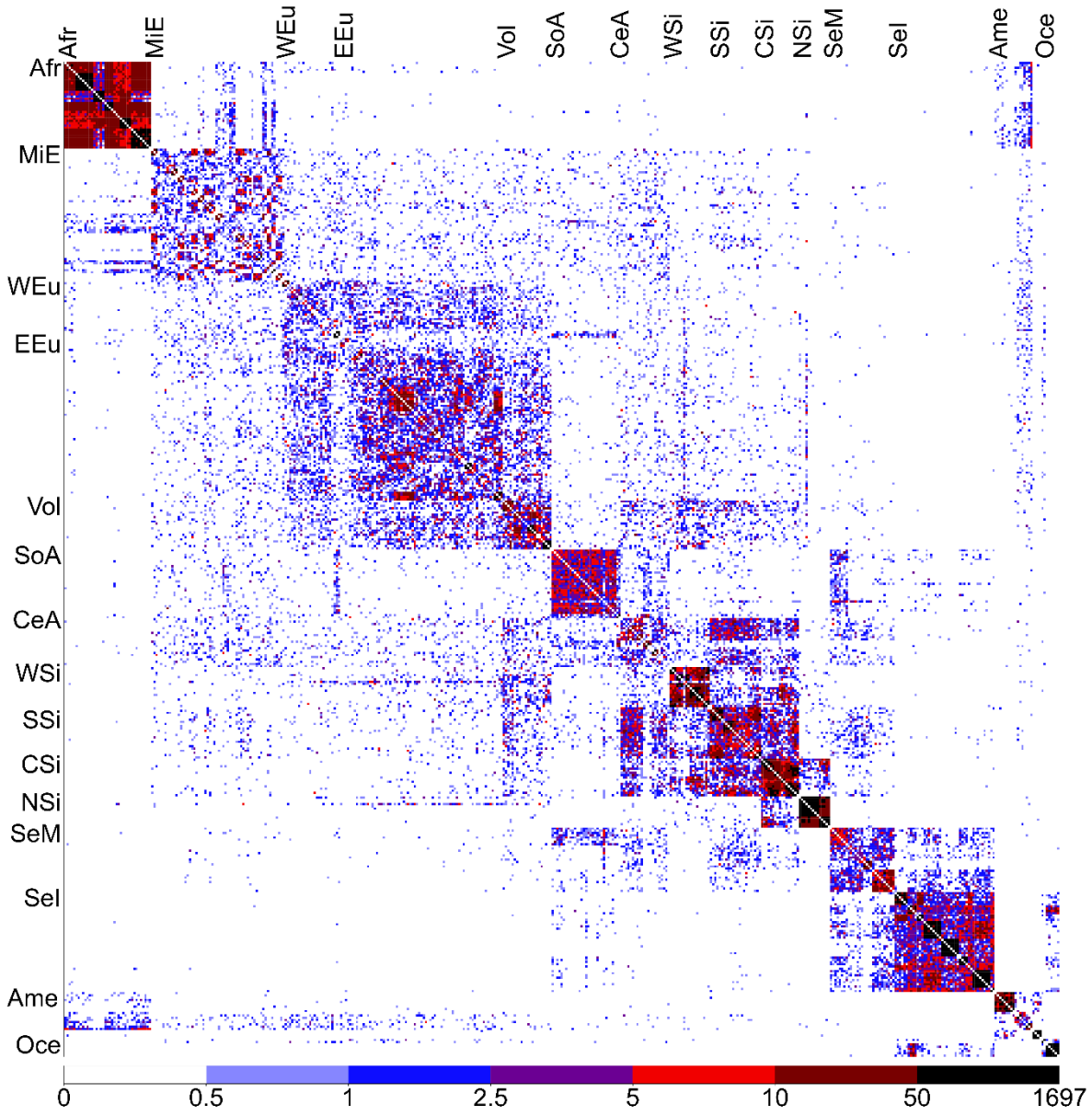


Figure 4.10: Pairwise total sharing across all RVCs in cM. The average over the RVCs based on the two individual-level vrGV databases was taken. Abbreviations taken from Table 3.1.

reflecting an event that occurred putatively 47-65 kya (Sankararaman et al., 2012). For the above interpretation of RVC lengths in Chuk8 it is relevant whether these can be directly understood as representing unbroken haplotypes, i.e. regions without internal recombination. The latter assumption will be examined in more detail in section 4.2.3.

As the total length of RVCs between two individuals (Figures 4.9-4.10, respective raw data in Appendices D.26-D.27) is tightly correlated with normalised f_2 sharing (Table 4.3), the pairwise interpopulation sharing of RVCs will not be described in detail here. However, it will be briefly highlighted that RVC sharing captures aspects of expected demographic history well, i.e. it

appears to be a meaningful measure of genetic similarity as population genetic theory would suggest.

Firstly, the average total shared RVC length (in a balanced sample) between groups should correlate to the intensity of recent gene flow between them. An instructive example is that African populations (excluding African Americans) exhibit considerable variation with regards to the total genetic length of RVCs shared with non-Africans (Appendix D.28A). All pairwise comparisons except for those inside a cluster consisting of Luhya, Sandawe and Yoruba are significant (two sample t-tests, $p < 0.05$, Appendix D.28B). In particular, the Hadza exhibit the least amount of sharing with non-African groups which is consistent with previous research indicating their genetic isolation as well as a recent bottleneck (Lachance et al., 2012). On the other hand, the Maasai share the highest fraction of their genome with non-Africans. This is expected due to their known substantial West Eurasian affinities, even though the exact modalities of the underlying admixture events are still debated (Wang et al., 2013).

Secondly, the timeframe over which this gene flow occurred should be reflected in the length distribution of RVCs. Generally, there should be an inverse relationship between shared RVC length and the last interpopulation contact. More precisely, for a given MRCA between two individuals equal to a separation of m meioses, IBD segment lengths should be exponentially distributed with a mean of $100*m^{-1}$ cM (Browning and Browning, 2012). However, the inference of the underlying age distributions of f_2 haplotypes and IBD segments from their lengths in general is a complex problem (Ralph and Coop, 2013; Mathieson and McVean, 2014). The relationship between admixture dates, which should be correlated to MRCAs, and RVC length will be explored using simulations in section 4.2.3.

Nevertheless, the general pattern can be demonstrated using three groups that exhibit elevated sharing with Africans (again excluding African-Americans). These are populations with known African admixture from the Americas ($\bar{n}_{RVC\_Afr\_admixed\_Ame} = 4.787$), Middle Easterners excluding Armenians ($\bar{n}_{RVC\_Afr\_MiE} = 4.269$) and Philippine Negritos ($\bar{n}_{RVC\_Afr\_Negrito} = 0.428$ which represents approximately 8-fold increase relative to non-Negrito ISEA groups). As theoretically expected, the length of RVCs shared between Africans and admixed Americans is of a comparable magnitude but still somewhat greater than that of RVCs shared by Africans and Middle Easterners ($\bar{L}_{RVC\_Afr\_admixed\_Ame} = 0.192$ cM, $\tilde{L}_{RVC\_Afr\_admixed\_Ame} = 0.057$ cM vs $\bar{L}_{RVC\_Afr\_MiE} = 0.153$ cM, $\tilde{L}_{RVC\_Afr\_MiE} = 0.045$ cM) (Figure 4.11). The RVCs with African

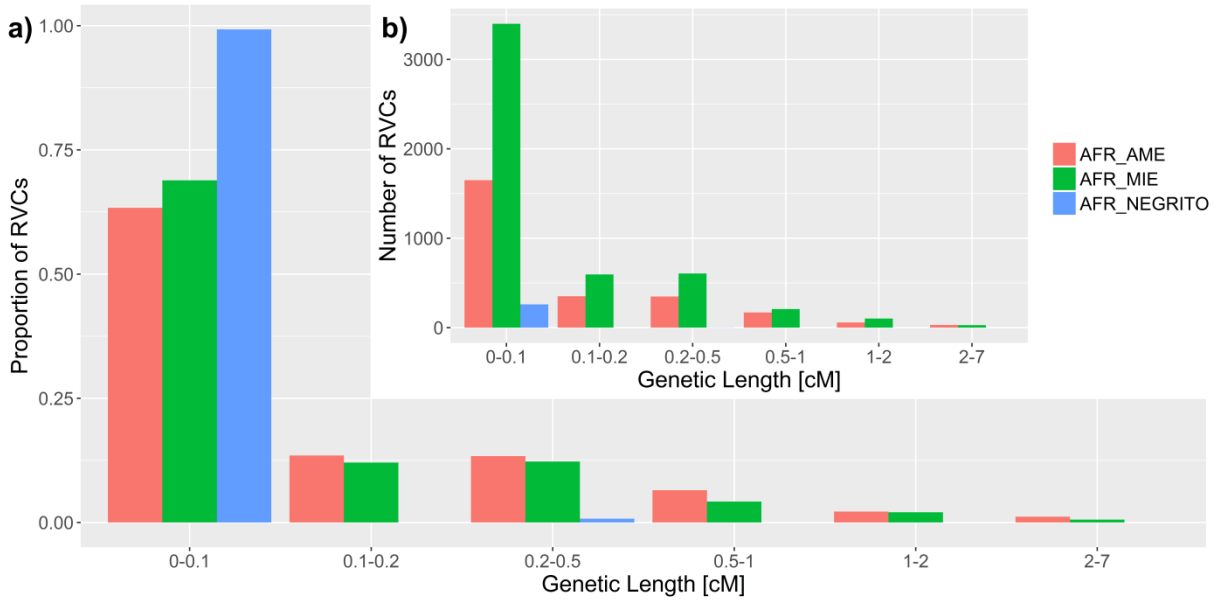


Figure 4.11: a) Relative frequency histogram of the length of RVCs shared between Africans (without African-Americans) (AFR) and i) admixed American populations (AME), iii) Middle Easterners (Arabs and Iranians) (MIE), iii) Philippine Negritos. b) Histogram of raw counts underlying a).

affinities in Negritos are by far the shortest ($\bar{L}_{RVC\_Afr\_Negrito}$ = 0.012 cM, $\tilde{L}_{RVC\_Afr\_Negrito}$ = 0.005 cM). The Anderson-Darling Criterion (ADC), a metric based on the average weighted distance between empirical cumulative distribution functions (Appendix D.29), further supports this and indicates that all three RVC length distributions are significantly different from each other ($ADC_{RVC\_Ame\_MiE}$ = 16.2, $p_{RVC\_Ame\_MiE}$ = $4.018*10^{-9}$; $ADC_{RVC\_Ame\_Negrito}$ = 227.6, $p_{RVC\_Ame\_Negrito}$ = $2.854*10^{-125}$; $ADC_{RVC\_MiE\_Negrito}$ = 217.6, $p_{RVC\_MiE\_Negrito}$ = $9.833*10^{-120}$).

4.2.3 The Andean Calchaquíes - a case of "cryptic" low level admixture?

The West African Yoruba represent one potential proxy for recent non-American gene flow into the Calchaquíes (see section 4.2.1, the possibility of European admixture will be briefly examined below). When the whole genome sequences from these samples were first published (Pagani et al., 2016) common-variant-based approaches did not suggest any admixture from non-American sources. To re-examine this result f_3 statistics were calculated to formally test for such an event. Potential African admixture into the Calchaquíes cannot be reliably inferred on a genome-wide level using this allele-frequency-based method as none of the obtained f_3 statistics are significantly negative (Table 4.5).

However, genomic segments originating from African admixture are likely to be unequally distributed across the genomes of the analysed Calchaquíes. This is plausible if the admixture event is relatively recent and is supported by differences in the density of Calchaquíes-African shared RVCs between different chromosomes (Appendix D.30). Therefore, f_3 statistics were calculated separately for each chromosome (Appendix D.31). The obtained statistics reach significance only for chromosome 13. Also, for some runs (Table 4.5) different methodologies to calculate the f_3 were applied. While normalising for an abundance of low frequency SNPs has little effect, correcting for inbreeding in the target population (Calchaquíes) results in a detectable shift of the f_3 towards more negative values. Lastly, while the Eskimos have the same position in a tree with the Yoruba and the Calchaquíes when they instead of the Wichi are used as a source population representative of Native-American ancestry all resulting f_3 statistics remain positive.

Table 4.5: f_3(C;A,B) statistics to investigate whether the Calchaquíes (C) can be modelled as a mixture of a Native-American-related source (A) and the Yoruba (B). Abbreviations: CAC – Calchaquíes, chr - chromosome, ESK – Eskimo, inbreed_correct – correction for inbreeding in admixed target population, lowfreq_norm - normalisation for abundance of low frequency SNPs, wg - whole genome, WIC- Wichi, YOR -Yoruba

Dataset	Software	Corrections	A	B	C	f_3(C;A,B)	Z
wg	threepop	none	ESK	YOR	CAC	0.008164	36.88
wg	threepop	none	WIC	YOR	CAC	0.000213	1.24
wg	ADMIXTOOLS	lowfreq_norm	ESK	YOR	CAC	0.059261	32.55
wg	ADMIXTOOLS	lowfreq_norm	WIC	YOR	CAC	0.001545	1.14
wg	ADMIXTOOLS	lowfreq_norm, inbreed_correct	ESK	YOR	CAC	0.055729	30.88
wg	ADMIXTOOLS	lowfreq_norm, inbreed_correct	WIC	YOR	CAC	-0.001623	-1.21
chr13	threepop	none	ESK	YOR	CAC	0.03784	5.80
chr13	threepop	none	WIC	YOR	CAC	-0.003460	-4.35
chr13	ADMIXTOOLS	lowfreq_norm	ESK	YOR	CAC	0.024878	4.81
chr13	ADMIXTOOLS	lowfreq_norm	WIC	YOR	CAC	-0.022794	-4.42
chr13	ADMIXTOOLS	lowfreq_norm, inbreed_correct	ESK	YOR	CAC	0.019336	3.86
chr13	ADMIXTOOLS	lowfreq_norm, inbreed_correct	WIC	YOR	CAC	-0.027835	-5.35

To test whether statistics exploiting other signals generated by admixture can provide more unequivocal support for recent African gene flow into the Calchaquíes the ALDER method, which relies on admixture LD, was applied. A significant ($p = 6.6*10^{-4}$) result is obtained when the Calchaquíes are modelled as a mixture of Eskimos and Yoruba with the admixture event dated to 13.4+/-3.9 generations ago (Figure 4.12A).

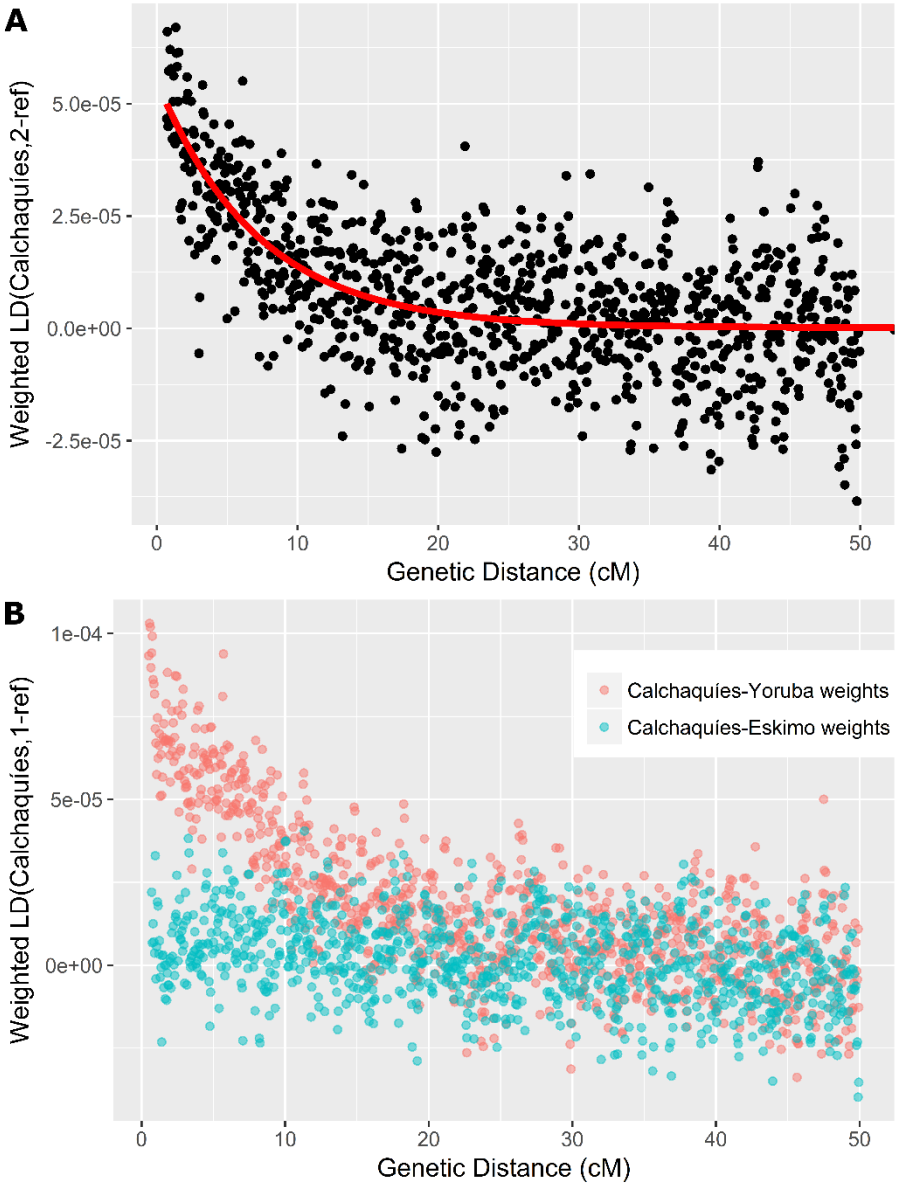


Figure 4.12: Weighted LD plots for Calchaquíes with Eskimo and Yoruba as reference populations including an exponential curve fit to the data (a) and (b) using Calchaquíes itself as reference in combination with the two different source populations.

The method can also be used with the admixed population itself as one of the references, then the weights for the LD curve reflect allele frequencies differences between one source and the target (Figure 4.12B). Accordingly, with the Yoruba as the other reference the putative African admixture in the Calchaquíes is inferred to have occurred 8.9+/-2.0 generations ago while the initial African mixture fraction is estimated at 2.2+/-0.5% (for the historical context of these admixture events see section 4.3.1).

These different admixture dates occur because the LD decay rates obtained for this scenario are inconsistent. This means that the exponential model which ALDER fits to the decrease in

admixture LD as a function of genetic distance infers significantly different exponential decay constants (which equal admixture dates) for all scenarios tested when ALDER is run for this population trio, i.e. a) Calchaquíes a mixture of Eskimo and Yoruba, b) Calchaquíes with Eskimo as single reference, c) Calchaquíes with Yoruba as single reference.

One way to assess the evidence for recent African gene flow into the Calchaquíes is to compare the characteristics of RVCs shared between these groups to empirical rare variant sharing patterns of non-Africans who are known to be African-admixed. Beforehand, it is however necessary to ascertain that the genomic runs of rare variants shared by Calchaquíes and Africans do not represent ancestry from a third source that mixed with both groups. Continental frequencies of 3,290 doubletons that form the 241 RVCs detected in Calchaquíes with African populations (Appendix D.32) were retrieved from gnomAD. Of the doubletons that could be matched unambiguously to the database 97.9% (3,191/3,261) reach their highest global frequencies in Africans. This together with the fact that the average African allele frequency at these sites is at least one hundred-fold higher than that observed in East Asians and Europeans, is consistent with an African origin of these mutations (Table 4.6).

Table 4.6: Mean allele frequencies of doubletons forming RVCs shared between Africans and Calchaquíes in continental groups from the gnomAD dataset. Sample sizes given in the table header. Run length intervals exclude their lower and include their upper boundary.

RVC physical length	Number of RVCs	Africans (includes African Americans) (n = 4,368)	East Asians (n = 811)	Europeans (Non-Finnish) (n = 7,509)
0-200 kb	143	0.025	$2.27*10^{-5}$	$1.14*10^{-4}$
200-1,000 kb	88	0.024	$4.39*10^{-5}$	$1.40*10^{-4}$
>1,000 kb	10	0.019	$6.08*10^{-5}$	$1.85*10^{-4}$

Only three of these genomic runs are suspect as many sites in them show a pattern where either a) Europeans are polymorphic and no Africans exhibit the variant or b) European and African allele frequencies are comparable (Appendix D.33). The longest of these runs (3:157,703,340-159,568,153 between Cachi5 and ASW_3) likely indicates recent European admixture whereas the shorter two, which have lengths of ~0.023 cM and ~0.165 cM, could also be interpreted as originating from residual pre-OOA sharing. Regardless, only a very small fraction (maximum of 1.39 cM/59.55 cM) of all RVCs between Calchaquíes and Africans is consistent with being the result of introgression from a non-African source.

When the vrGV databases of the Calchaquíes individuals are used to define their sharing with other populations, the average Calchaquí-African pair has a total of ~1.23 RVCs in common

which span a mean cumulative length of ~0.305 cM. Both metrics are significantly elevated relative to the geographically adjacent Wichi in whom African-shared runs are almost totally absent ($\bar{n}_{RVC\_Wic\_Afr} = 0.115$, $\bar{L}_{RVC\_cumulative\_Wic\_Afr} = 0.016$ cM; two-sample t-tests: $t_{RVC} = 11.477$, $p < 10^{-15}$, $t_{L\_RVC\_cumulative} = 6.727$, $p = 1.729*10^{-10}$).

However, when compared to known African-admixed groups from the Americas ($\bar{n}_{RVC\_Afr\_admixed\_Ame} = 5.105$, $\bar{L}_{RVC\_cumulative\_Afr\_admixed\_Ame} = 1.182$ cM; two-sample t-tests: $t_{RVC} = -12.174$, $p < 10^{-15}$, $t_{L\_RVC\_cumulative} = -7.646$, $p = 1.541*10^{-13}$) and the Middle East ($\bar{n}_{RVC\_Afr\_MiE} = 4.243$, $\bar{L}_{RVC\_cumulative\_Afr\_MiE} = 0.728$ cM; two-sample t-tests: $t_{RVC} = -19.103$, $p < 10^{-15}$, $t_{L\_RVC\_cumulative} = -7.174$, $p = 2.269*10^{-12}$) the signal in the Calchaquíes is approximately 3-4 times weaker.

Analysing the RVC length distributions (Figure 4.13) with the ADC supports broadly similar dates for the African gene flow into the Calchaquíes compared to Mexicans and Puerto Ricans as they cannot be distinguished by this metric ($ADC_{RVC\_Cac\_Ame\_admixed} = 1.402$, $p_{RVC\_Cac\_Ame\_admixed} = 0.2014$).

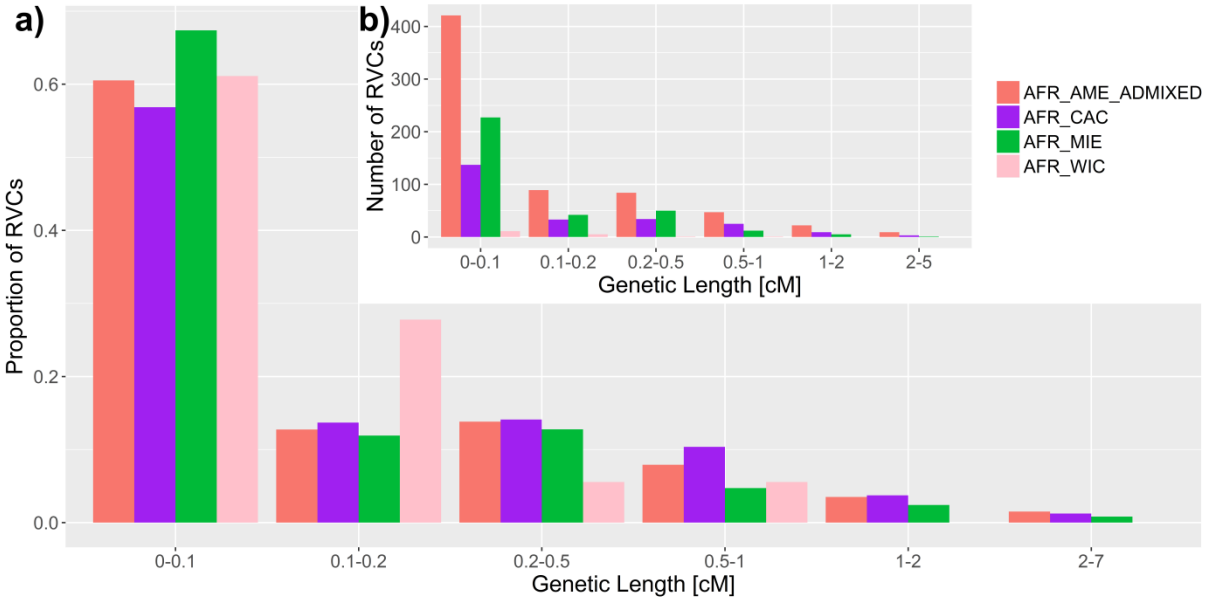


Figure 4.13: a) Relative frequency histogram of the length of RVCs shared between Africans (AFR) and i) known admixed American populations (AME_ADMIXED), ii) Calchaquíes (CAC), iii) Middle Easterners (Arabs and Iranians) (MIE), iv) Wichi (WIC), b) Histogram of raw counts analogous to a), however Mexicans and Druze were chosen as representatives for their larger clusters to make absolute histogram interpretable. RVCs were defined based on non-African vrGV databases.

Consistent with this, RVCs shared between Calchaquíes and Africans are shifted towards greater lengths relative to those the latter and Middle Easterners ($ADC_{RVC\_Cac\_MiE} = 7.521$, $p_{RVC\_Cac\_MiE} = 1.88*10^{-4}$) have in common. Finally, the ADC does not indicate a significant

shift in the lengths of African-Wichi vs African-Calchaquíes runs ($ADC_{RVC\_Cac\_MiE}$ = 0.583, $p_{RVC\_Cac\_MiE}$ = 0.6646).

One factor contributing to this outcome is likely that only 18 runs fall into the African-Wichi category, which reduces the statistical power of the Anderson–Darling test. Low frequency European admixture could contribute to the observed excess f_2 affinities between the Calchaquíes and some West Eurasian groups. The 99 RVCs shared between Calchaquíes and Europeans from the Diversity Set were again compared to the gnomAD dataset (Appendix D.34). Out of 22 such RVCs, which encompass a total length of 7.11 cM, two exhibit a pattern where the doubletons constituting them have their highest recorded frequencies in Africans. For the other 20 RVCs the mean European frequency of the f_2 variants is ca. 3.7-fold higher than in Africans. The latter runs are mostly short enough that they could reflect ancient background sharing; however, five runs have a length greater than 0.5 cM. These could represent potential traces of European ancestry. While this possibility is not further investigated here it underlines that any small-scale European contribution does not appear to be a plausible confounder of the African admixture signal in the Calchaquíes.

The African admixture fraction can be estimated from nearest neighbour sharing segments as the average cumulative CP length a Calchaquí individual shares with all Africans minus the same average quantity for a Wichi genome. This results in an estimated average length of ~57 cM corresponding to 0.81% of the Calchaquí genomes.

To test whether a signal of rare variant sharing of that magnitude can result from background sharing rather than a recent admixture, plausible demographic scenarios (Figure 4.2) were run in simulations. PCA suggests that the expected continental genetic clustering of extant human populations can be reproduced well from the baseline scenario (Figure 4.2) simulated here (Appendix D.35). Additional simulated demographic histories explore the nature of a hypothetical recent admixture event from Pop1 (Yoruba-like) into Pop5 (Calchaquíes-like). None of the 20 replicates of a null model without recent admixture from Pop1 into Pop5 (scenario A) yielded a total RVC count or cumulative RVC tract length exceeding the amount of empirically observed sharing between the Calchaquíes and all Africans (Figure 4.14, relevant RVCs from simulations listed in Appendix D.36). This implies a simulation-based p-value < 0.05.

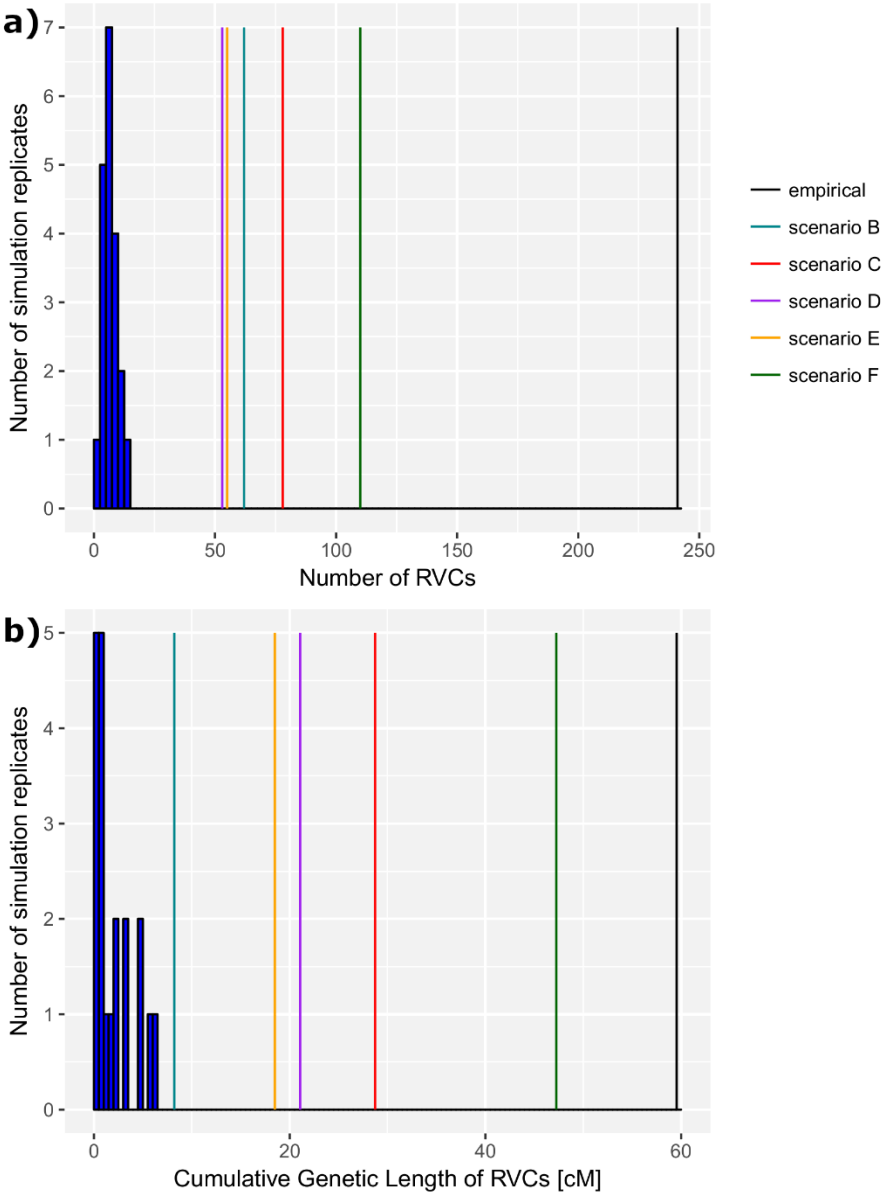


Figure 4.14: Histograms of different measures of abundance of rare variant sharing between Pop5 (Calchaquíes-like) and Pops1/2 (African-like) in twenty simulation replicates without recent excess gene flow. The metrics used are a) the total number of RVCs and b) the cumulative genetic length of RVCs. Vertical lines indicate outcomes from empirical data and from scenarios for which only one set of simulated genomes was created. The latter differ from the null-model/scenario A as follows: scenario B- 10 times elevated background migration rate between Pop5 and Pops1/2, scenario C- a single admixture pulse from Pop1 into Pop5 with an admixture fraction of 0.5% occurred 15 generations ago, scenario D- identical to scenario C with admixture event 18 generations ago, scenario E- identical to scenario C with admixture event 20 generations ago, scenario F: identical to scenario D with admixture fraction of 1%.

This estimate is likely too conservative, as depending on the abundance metric used, the observed sharing signal in the real data ($L_{RVC\_cumulative\_Cac\_Afr}$ = 59.546 cM, n_{RVC} = 241) is 10-19 times greater than the highest value obtained from any simulation replicate ($L_{RVC\_cumulative\_scenarioA\_Pop5\_Pops1/2}$ = 6.192 cM, $n_{RVC\_scenarioA\_Pop5\_Pops1/2}$ = 13).

Scenario B that includes a higher background migration rate between Pops1/2 and Pop5 produces an increased signal with a cumulative length of 8.201 cM spread across 62 runs. Interestingly, the outputs from scenarios C-E assuming a very low admixture fraction of 0.5% have a lower total of RVCs even though the genetic length these cover is 2-3 times greater than for scenario B. Within the very low-level admixture scenarios the RVC signal is stronger the more recent the simulated admixture date is. The final scenario F with an admixture fraction of 1% yields RVC output totals that are closest to those observed in the real Diversity Set, though still somewhat lower ($n_{RVC\_\ scenarioF\_Pop5\_Pops1/2} = 110$, $L_{RVC\_cumulative\_scenarioF\_Pop5\_Pops1/2} = 47.252$ cM).

Another aspect of RVC sharing for which empirical and simulated datasets can be compared is how the genomic runs the Calchaquíes have in common with Africans are apportioned between populations. Across all scenarios that include recent admixture from Pop1 into Pop5 86.2% (255/296) of all RVCs and 97.0% of the cumulative length of the runs Pop5 shares with African-like groups are accounted for by Pop1.

In contrast, only a third of all RVCs and less than half of the total cumulative RVC length shared between Calchaquíes and Africans are contributed by the Yoruba, though the overall sharing profile matches that of the Yoruba themselves very well (Table 4.7). In a larger geographical context more than 80% of the cumulative total for both abundance metrics derives from pairs which involve either groups whose predominant ancestry has been linked a cluster corresponding to the present-day distribution of the Niger-Congo language family (Tishkoff et al., 2009) or Pygmies from the region who are known to have received gene flow from the former (Patin et al., 2014).

Table 4.7: Percentages of total RVC sharing a) between Calchaquíes and all Africans, b) between Yoruba and all Africans that is accounted for by each African population from the Diversity Set.

Population	N	African-shared RVCs in the Calchaquíes		Yoruba	
		Cumulative RVC genetic length	Number of RVCs	Cumulative RVC genetic length	Number of RVCs
African-Americans	5	14.27%	15.35%	18.20%	16.55%
Hadza	5	0.83%	3.32%	1.69%	3.41%
Luhya	4	8.7%	13.69%	11.26%	12.26%
Maasai	3	2.98%	3.73%	3.78%	5.25%
Pygmies (Western Central Africa)	8	19.37%	19.92%	13.96%	19.74%
Sandawe	5	9.01%	9.96%	4.36%	6.8%
Yoruba	9	44.85%	34.02%	46.76%	35.99%

The RVC length distributions resulting from most of the simulated scenarios and the empirical Calchaquíes-African sharing (Figure 4.15) cannot be distinguished using the Anderson Darling test (Appendix D.37A). Even for cases of known African admixture (Mexicans, Puerto-Ricans), the length distributions of runs they share with Africans are not significantly different from the two simulated non-admixture scenarios (Appendix D.37B). The only detectable pattern is that the RVCs shared between Pop1 and Pop5 under scenario F are significantly shifted towards longer runs relative to the empirical data and simulations without gene flow.

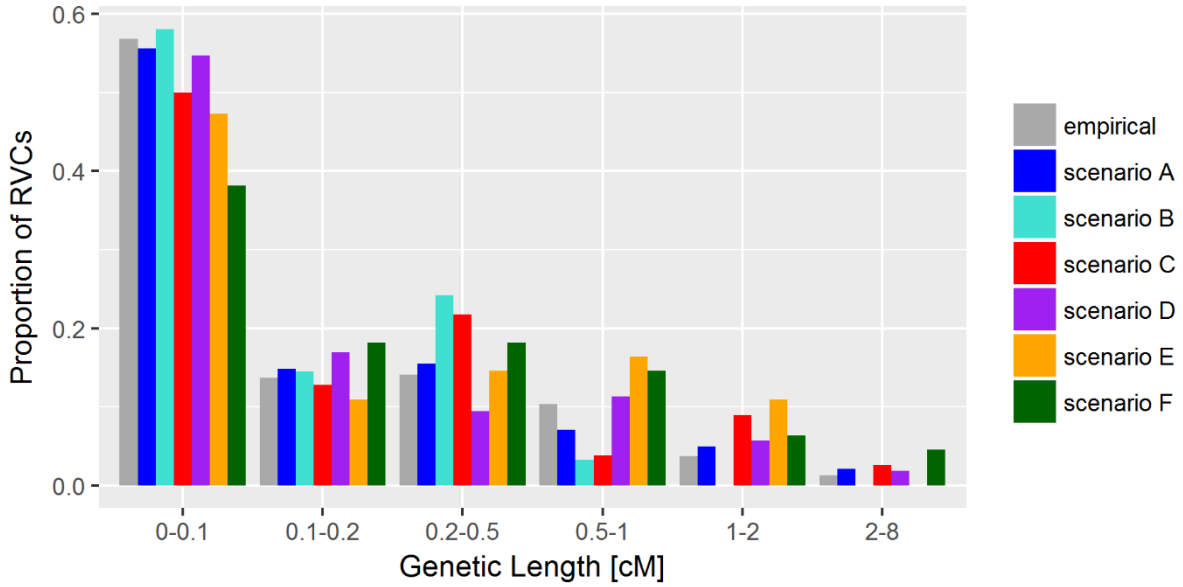


Figure 4.15: Relative frequency histogram of the length of RVCs shared between all Africans from the Diversity Set and the Calchaquíes (empirical) compared to sharing between Pop5 (Calchaquíes-like) and Pops1/2 (African-like) from six different simulated scenarios. The distribution for scenario A is based on 20 replicates and on a single replicate for all other scenarios. RVCs were defined based on Calchaquíes/Pop5 databases.

4.2.4 Consistency of RVCs with computationally phased haplotypes

Lastly, it will be examined how consistent the detected RVCs shared between Calchaquíes and Africans are with the complete genotypes and computationally phased haplotypes under the assumption that RVCs represent genomic segments uninterrupted by recombination events. Around three quarters of them (180/241) do not contain any sites where the two individuals sharing the run exhibit inconsistent homozygote genotypes. The remaining "broken" runs were fixed by splitting them at the homozygous inconsistent sites and retaining the resulting sub-runs if they contained five or more shared doubletons (see Appendix D.2 for scripts).

POSITION	STATUS	YRI_12.HAP1	YRI_12.HAP2	Cachi4.HAP1	Cachi4.HAP2
183,288,831	DOUBLETON	1	0	1	0
183,289,075	DOUBLETON	1	0	1	0
183,290,050	VARIANT	1	1	0	1
183,290,145	VARIANT	0	1	0	0
183,290,194	VARIANT	1	1	0	1
183,290,647	VARIANT	1	1	0	1
183,293,267	VARIANT	1	1	0	1
.....					
183,295,777	DOUBLETON	1	0	0	1
.....					
183,298,745	DOUBLETON	1	0	0	1
183,298,818	DOUBLETON	1	0	0	1
.....					
183,304,394	DOUBLETON	1	0	1	0

Figure 4.16: Subsection of the RVC 1:183,212,668-184,089,833 spanning a total of 16 kb. Only some segregating sites from this subregion are displayed; dots represent variants which are not recorded here. Green background shading indicates that for this position the inferred haplotypes YRI_12.HAP1 and Cachi4.HAP1 are consistent with the assumption that this RVC represents a continuous long shared haplotype. Red shading indicates sites which are inconsistent with it and polymorphic in other individuals, which are not displayed here, whereas yellow shading represents inconsistent doubletons. Under the assumption that YRI_12.HAP1 represents a correctly phased haplotype there would have to be eight switch errors in Cachi4.HAP1 alone. If to the contrary the hypothesis is that the computational phasing of this region is correct it would require four short independent haplotypes carrying at least one shared doubleton with YRI_12 within 16 kb in the genome of Cachi4.

These 65 fixed runs (Appendix D.30C) together with original RVCs without inconsistent homozygotes span a cumulative genetic length of 36.94 cM. A less unambiguous criterion is that all heterozygous positions within the RVC boundaries are correctly phased. Switch error rate, which is formally defined as the proportion of heterozygous sites whose phase 1s wrongly deduced relative to the previous heterozygote (Stephens and Donnelly, 2003) was not estimated. This is because it cannot be determined which of the two computationally inferred haplotypes that share the first doubleton of an RVC is closest to the unknown true haplotype.

If RVC consistency is just assessed by a simple count of (non-singleton) heterozygous sites which differ between the two inferred haplotypes underlying the RVC 91.4% (224/245) of all RVCs contain at least one such site. However, in turn 84.3% (189/224) of these RVCs contain at least one doubleton shared by the two respective individuals which is inferred to switch haplotypes within an individual relative to the previous uniquely shared doubleton. The latter could be suggestive of phasing errors.

An illustrative example for this is the RVC 1:183,212,668-184,089,833 (Figure 4.16): if the statistical phasing using SHAPEIT2 is assumed to be entirely correct there would have to be four independent very short haplotypes each carrying one doubleton uniquely shared between

the same two individuals within 16 kb. If, on the other hand, the RVC is interpreted as a continuous shared haplotype eight switch errors would have to occur in this short genomic segment. While these probabilities are not formally assessed here, a qualitative argument can be made that the latter is biologically more plausible, especially given that statistical phasing is known to have difficulties with correctly inferring rare haplotypes (see section 4.3.1).

4.3 Discussion

4.3.1 General insights

As with any method it is important to know whether the obtained rare variant patterns are broadly consistent with theoretical expectations and whether errors are nonrandomly distributed. Human populations are known to exhibit a considerable degree of structure in common variant allele frequency distributions. If the rare variants mainly reflect recent history there should be a strong excess of individual-level f_2 sharing within macro-groups vs out-sharing, which is exactly what is observed (Figures 4.3-4.5).

Furthermore, inferences about interpopulation relationships from rare variants rely on the assumption that if two individuals share f_2 at a site they are likely to share a recent common ancestor at that locus. This is plausible because the average per-site mutation rate across the human genome per generation is on the order of 10^{-8} (see section 1.2.1) and therefore the probability of the same mutation occurring independently at the same site multiple times throughout recent human evolutionary history is very low. In the large ExAC dataset (Lek et al., 2016) such events, also known as recurrent mutations, have been reported at high mutability sites, e.g. CpG transitions. This means that doubleton sharing does not necessarily reflect IBD but can occur also due to homoplasy. In the ExAC dataset mutability at doubletons is correlated with the genetic distance between the individuals sharing the f_2 variant. While there is strong evidence that in the Diversity Set the window-specific *de novo* mutation rate is the primary determining factor for rare variant density, there is no detectable excess of between-population doubletons in high mutability windows (Appendix D.38). This implies that for datasets of this size (two orders of magnitude smaller than ExAC) doubletons do with very high probability indicate IBD.

Wall et al. (2014) estimated a lower boundary for genotype errors of rare variants (MAF < 1% in Western Europeans) from Complete Genomics data at 6.3%, mostly due to false positives. While this is non-negligible, as long as it can be assumed that these errors are randomly distributed across the genome, it seems unlikely that two individuals would share a high number of doubletons due to genotype errors. Only one such case could be confidently inferred from the dataset; it was originally detected from excess-sharing relative to what can be expected based on geographical distance (Appendix D.10). In this case a British individual (UK1) shares a total of 352 f_2 variants with two Chinese (Chn2 and Chn3) (Appendix D.39). The nature of this finding as a methodological artefact is supported by multiple lines of evidence. Firstly, UK1 has almost no doubletons in common with the other five Chinese in dataset. Furthermore, ca. 20% of the doubletons UK1 shares with Chn2/Chn3 are clustered within one read length (35 bp) indicating potential alignment errors. Most importantly, the three samples in question were the only ones from the whole Diversity Set for which version 2.0.3.x of the Complete Genomics pipeline was used for mapping and variant calling, suggesting a batch effect. This trio of individuals was flagged but not removed, as their aberrant pattern seems limited to their sharing with each other and does not measurably influence how they relate to other samples in the dataset.

The excess of doubletons in African relative to non-African genomes (Table 4.2) suggests that the f_2 totals still mostly reflect the consequences of OOA, as observed for synonymous variants when all frequencies were considered (Table 3.5). This is due to the sampling strategy for the Diversity Set. While many populations are covered, only a few individuals were sampled from each and no more than 63 from any macro-group. A considerable amount of private variation is not captured and some of the variants designated as f_2 here are likely at intermediate frequencies in subgroups.

Interpreting the order of non-Africans genomes by f_2 numbers in terms of underlying processes is not straightforward due to the complexities hinted at above. For instance, the Middle Eastern group is more diverse than other West Eurasians which most plausibly reflects higher N_e before very recent times and/or admixture with the more diverse Africans. On the other hand, elevated doubleton counts in Oceanians vs other non-Africans most likely result from lack of gene flow between them leading to very distinct population-specific branches in the former. The different underlying forces become clearer if f_2 sharing is considered (Figures 4.4-4.5) but very dissimilar population histories can produce effects of the same directionality in terms of total f_2 counts.

Therefore, the finding that the East Asian/mainland SEA group is the least diverse according to the sample-size normalised f_2 totals only indicates that these f_2 sites are at intermediate frequency on a population-level (~16.7% for downsampled groups if population-specific) and in consequence mainly reflect older demographic history. Some models have suggested that the post-OOA bottleneck was more severe for East Asians than for Europeans (Gravel et al., 2011), which could have caused the observed pattern, even though MSMC data (Figure 3.21) seemingly contradict this.

In a comparative perspective, these results at a first glance appear very different from those compiled in the final stage of the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2015). There, a total of 8,769,658 f_2 variants in 2,504 individuals was reported, which equals ~3,502 variants per haploid genome. However, when sample sizes are comparable most of the 2.7-fold increase in f_2 variant counts in the Diversity Set disappears (see Appendix D.40 for analyses based on 1000 Genomes genotypes, and Appendix D.41 for raw f_2 sharing matrices). Even if sample sizes are comparable, for each macro-group except East Asians the Diversity Set still exhibits a moderate excess of f_2 variants relative to the 1000 Genomes data. This could conceivably reflect either that the macro-groups in the Diversity Set contain more diverse subpopulations or potentially the higher average coverage in the Diversity Set vs the final 1000 Genome data (for the latter estimated at ~7.8×). Finally, it could also indicate platform-specific differences between the Illumina and Complete Genomics technologies.

When the non-normalised datasets are compared, the distinction in diversity between Africans and non-Africans is much less clear in the 1000 Genomes data. The latter is perhaps not surprising as the samples from single populations are much larger and therefore the f_2 variants often have an MAF of ~1% in the specific subpopulations. For the higher and relatively balanced sample sizes in the 1000 Genomes data East Asians/SEA groups and South Asians exhibit higher f_2 totals than West Eurasians, which is also observable in the full Diversity Set (Table 4.2).

An interesting property of the normalised f_2 counts as well as the total length of RVC sharing between two individuals is that these metrics strongly correlate with the total genetic length of pairwise shared CP chunks (Figure 4.7, Table 4.3). This is not an unexpected outcome as the f_2 methodology has conceptual similarities to the most widely used IBD detection approaches, the key idea being the identification of low frequency haplotypes. When these are shared between

individuals they are most likely derived from a particular MRCA (Browning and Browning, 2012).

However, IBD sharing as detected by these approaches does not in all cases measure the same quantity as f_2 sharing which can be illustrated by considering the underlying genealogies. The amount of f_2 sharing between two individuals *i* and *j* can be interpreted as proportional to the probability that the first coalescence of both the lineages of *i* and *j* is with each other. On the other hand, the amount of IBD sharing between *i* and *j* is directly related to the probability that these two lineages have coalesced by a specified time. This time depth is approximately determined by the minimum candidate segment length that can be inferred reliably by the IBD detection method. Therefore, f_2 sharing can highlight older coalescent events than those detectable by IBD. However, f_2 by definition misses IBD chunks without doubletons, i.e. IBD segments where three or more individuals share a common ancestor within the proposed timeframe.

Intriguingly, Mathieson (2013), who proposed this genealogical interpretation of IBD and f_2, also remarked on the close relationship between f_2 and the haplotype copying model from Li and Stephens (2003); the latter underlies the CP method (see section 1.3.3). Mathieson (2013) also compared a log-transformed and z-scored f_2 metric with similarly treated IBD sharing; the resulting correlation was moderate (r = 0.58). To ensure that the high correlation obtained in this thesis is robust to transformations (i.e. not driven by just a few outliers, even though the high rank correlation coefficients suggest that this is not the case) f_2 and CP sharing were analysed as proposed by Mathieson. While the correlation coefficient is somewhat reduced (r = 0.83) (Appendix D.42), it is still much higher than for Mathieson's f_2 metric and IBD. This can be partly attributed to CP sharing being more closely related to f_2 than other IBD detection approaches as described above. Furthermore, using the mean of the total f_2 diversities of the individuals forming the pair for normalisation (Eq. 4.1) yields a metric that is less biased by individual-level f_2 diversity (when Mathieson's metric is calculated from the Diversity Set the correlation with CP decreases to r = 0.71).

The pairs with large standardised residuals from the regression between CP and f_2 (Appendix D.7) demonstrate two major patterns. Firstly, even if f_2 sharing is normalised by baseline diversity, its ability to detect potential pairwise coalescences is directly related to the density of f_2 variants. Therefore, groups with low N_e might share nearest-neighbour segments identifiable for CP, e.g. through particular combinations of more frequent alleles which are not picked up

by f_2 due to lack of shared doubletons. Similar outcomes should occur for populations without closely related groups in the dataset as CP assigns every chromosome segment to a best match even if those matches are remote. Conversely, as a linear predictor variable f_2 sharing will overestimate CP sharing for groups with higher nucleotide diversity as the true underlying CP chunks will have higher SNP density and therefore carry proportionally more f_2 variants. Secondly, excess f_2 sharing relative to CP affinities could indicate underlying non-random sequencing or variant calling errors as a measurable fraction of these pairs consists of individuals flagged as having unusual metrics by Complete Genomics.

Overall, normalised f_2 sharing represents a computationally efficient and simple heuristic method to describe recent shared ancestry as demonstrated by its very strong correlation with the results of the powerful state-of-the-art method CP. Like other non-parametric approaches, it requires balanced sampling and should correlate best with CP if the overall nucleotide diversities of included individuals are comparable.

Many of the statements made above about normalised f_2 counts also apply to the RVC approach. Again, the 1000 Genomes Project phase 3 dataset and an f_2 haplotype metric according to Mathieson and McVean (2014) are used as a reference. The median physical run length of the RVCs appears very similar to the value of ~336 kb obtained for f_2 haplotypes from the 1000 Genomes data. However, the latter have a median genetic length of ~0.41 cM, about twice that of the RVCs (medians obtained from raw data kindly provided by Dr Iain Mathieson). This difference is unlikely to be explained by variation in the assumed recombination rates, as both maps used are derived from HapMap phase 2 data. The most plausible reason is that the criterion which defines the core region of an RVC is more restrictive than for the f_2 haplotype metric, therefore the regions covered by RVCs should mostly be a subset of f_2 haplotypes.

Regarding the potential and broader applications of the RVC method two aspects deserve closer attention. First, it is important to reconsider how RVCs are defined and second, what inferences can be made based on how they are shared. One important difference between the RVC and the f_2 haplotype metrics is how the endpoints of the shared haplotype segments are determined. Recall that the definition of RVCs relies on individual-level rare variant databases and that an RVC is extended beyond the core region of five shared doubletons if there is another doubleton within 40 rows (i.e. variants). At least one empirical example suggests that multiple independent rare variant segments can thereby be conflated into one (Appendix D.25).

One method to account for this is to apply additional stringency criteria, i.e. assume a break of RVC segment continuity in case of homozygote inconsistency, as was done here for Calchaquíes-African RVCs. An alternative criterion was suggested by Hochreiter (2013), whose HapFABIA method has conceptual similarities to the RVC approach. A HapFABIA IBD segment ends when the next shared SNP between the individuals in question has a distance to the previous shared SNP that has a very low probability of occurrence assuming an exponential distribution with the median distance between the other SNPs constituting the IBD segment as parameter.

A more comprehensive approach would be to use simulations to generate ground-truth IBD segments derived from a software which produces ancestral recombination graphs (e.g. the MaCS program by Chen et al., 2008). A potential framework to compare RVCs inferred on these simulated datasets to ground-truth IBD was provided by Chiang et al. (2016). They demonstrated that at least for high coverage SNP data state-of-the-art IBD detection algorithms conflate independently inherited segments and therefore overestimate the length of a considerable fraction (~22%) of shorter (1-2 cM) fragments. Ideally, such simulations would also include diverse demographic histories as these influence the abundance of rare variants available for inference and in the vrGV databases affect which physical/genetic distance one row on average represents. This would allow testing whether a uniform threshold for the stretching window across very diverse populations may be a problematic assumption. Simultaneously, there should be cases where the length of true IBD segments is underestimated when they extend beyond the last shared doubleton.

Simulations could also help to contextualise the finding that many heterozygote inconsistencies occur between the RVCs and the computational phasing using SHAPEIT2 (Appendix D.30, Figure 4.16). The two most plausible non-exclusive explanations are that either many statistically phased haplotypes carrying doubletons exhibit an excess of phasing errors or that many RVC segments do not represent regions of contiguous IBD.

There is some circumstantial evidence that at least for the core regions of the RVCs the former factor is likely more relevant. Firstly, switch error rates of 8.7% have been reported for doubletons phased using SHAPEIT2 based on simulations, even though the simulated dataset (n = 262) was somewhat smaller than the empirical sample used here (Delaneau et al., 2013a). Secondly, it should be considered that SHAPEIT2 models the haplotypes in each individual as an imperfect mosaic of the haplotypes estimated from the unphased genotypes for all other

individuals. Intuitively, the RVCs that contain many uniquely shared doubletons between two individuals should indicate long genealogical branches which would imply that for these specific regions the haplotypes of other sampled individuals might not represent suitable proxies. Interestingly, the authors of SHAPEIT2 themselves recently published a method (SHAPEITR) incorporating this idea. They modified the SHAPEIT2 algorithm to give greater weight to rare variant sharing patterns and demonstrated that this improves phasing accuracy (Sharp et al., 2016).

At least theoretically, the biggest advantage of the RVC compared to many other IBD detection algorithms should be its lower false positive rates for short IBD segments (Durand et al., 2014; Ralph and Coop, 2013). However, the RVC method should also suffer from generally reduced power as IBD segments lacking a cluster of uniquely shared doubletons will be missed. Another relevant aspect of the RVC/f_2 approaches are their fast running times and that they should scale well with increasing sample sizes. The RVCs based on empirical data were generated from the chromosome-specific VCF files (n_{sites} = 42,971,058) using the respective Perl scripts (section 4.1.2) in ~242 minutes on one core of a server with an Intel(R) Xeon(R) CPU E5-2650 0 @ 2.00GHz and 24 GB RAM. Finally, for very large samples when a substantial fraction of doubletons will reflect recurrent mutations the RVC should be more robust than the standard f_2 haplotype approach due to its minimum requirement of five adjacent doubletons.

IBD segments can be used to infer demographic history, to estimate relatedness, either as a general kinship coefficient or a specific pedigree (Sun et al., 2016) and to perform IBD mapping comparing cases vs controls in medical genetics. Here, RVC segments generated from a range of simulated scenarios describing a hypothesised admixture event of interest were compared to empirical observations (for a more in-depth discussion see section 4.3.2). A potential follow-up would be to use the maximum likelihood age estimation method described by Mathieson and McVean (2014) for f_2 haplotypes on the detected RVCs. This method falls into the broader context of recently developed approaches which use the information contained in IBD segment distributions to infer a range of quantities describing demographic history. The latter include recent N_e (Browning and Browning, 2015; Palamara et al., 2012), the number of MRCAs between individuals from different populations in a particular time period (Ralph and Coop, 2013) and spatial parameters such as dispersal rates and population densities (Ringbauer et al., 2017).

In practice, one could envision that for future analyses, the f_2/RVC methodology will be combined with an approach that detects longer IBD segments reliably, i.e. Refined IBD. Combining well-performing haplotype-based IBD detection methods yields only modest improvements in accuracy as measured by relatedness inference relative to known pedigrees (Ramstetter et al., 2017). However, the underlying genealogical interpretations as well as empirical evidence (Lawson and Falush, 2012) suggest that the two types of algorithms emphasise different historical signals, i.e. additional information could be gained by combining them.

The challenge would lie in designing the actual framework for such a composite approach. A starting point could be to use both the distributions of long IBD as well as shorter RVC segments to compute (approximate) maximum likelihood estimates for demographic parameters of interest and then evaluate based on simulations or "ground-truth" from independent methods whether there is a significant improvement in accuracy compared to the methods that only rely on the distributions of long IBD segments. A further advantage of the RVC approach for parameter inference from demographic models is that it detects IBS segments that are very likely to be IBD but does not require many assumptions about demographic history. The latter is the case for many other IBD detection approaches (Spence et al., 2018). Applying the RVC method should at least partly address the underlying problem of circularity when demographic parameters are then inferred from the IBD segment distribution.

Assuming an exponential regression model, the median lengths of intra-macro-group RVCs are best explained by the harmonic means of N_e values from 5-10 kya relative to N_e over other time periods. For this time interval, a power law model of the following form outperforms an exponential model using the Vuong test statistic as criterion (Appendix D.18A):

$$\tilde{L} = \frac{86.385}{{N_e}^{0.445}} \tag{4.7}$$

Eq. (4.7) is of interest as a more complex expression that has been derived to describe the expected lengths of IBD segments shared within a population as a function of population size N is also a power law distribution (Palamara et al., 2012). The mean of this distribution of (uniformly sampled) IBD segment lengths is 1/(4*N) (Wakeley and Wilton, 2016). The differences between the latter expression and Eq. (4.7) are due to multiple reasons. Firstly, the mean of a power law distribution is self-evidently different from its median and secondly the RVCs represent a very specific subset of all IBD segments within each population. Finally, the

theoretical derivation is based on a Wright-Fisher population whereas real population histories are much more dynamic and complex. Taking this into account it can still be stated that N_e for different macro-groups between 5-10 kya has a relationship to median RVC segment length that is of a similar nature to that of Wright-Fisher populations of different sizes and their respective mean IBD segment lengths.

A potential criticism of this analysis is its redundancy, as N_e inferred from MSMC is obtained from rates of first pairwise coalescences between sequences along the genome whereas the RVC method focusses on a subset of first pairwise coalescences. However, this type of analysis can still be informative about the approximate time frame of the processes generating the RVCs and it should provide a lower bound of the ages of the doubletons the RVCs carry.

The observed decline of rare variant sharing with geographical distance (Appendix D.1, Appendix D.11) is not an unexpected result and consistent with previous analyses of allele frequency (Ramachandran et al., 2005) as well as haplotype sharing (Martin et al., 2018). It occurs in an approximately exponential manner consistent with predictions about genetic similarity in space made as early as the 1940s, e.g. by Wright and Malécot (see section 4.1.1). While it would go beyond the scope of this thesis to discuss the vast amount of scholarship on this subject since these early investigations, it should be noted that Eq. (4.2) by Malécot and its extension to two dimensions have been criticised because the theoretical foundations they were derived from contained inconsistent assumptions (Felsenstein, 1975). Furthermore, there are patterns observed in biological data, e.g. long-range correlations of allele frequencies which cannot be explained by "classical" theory (Sokal and Thomson, 1998). However, as described above f_2 sharing correlates very well with haplotype-based IBD detection methods. The latter should mostly be robust for the confounding effect of ancestral structure (Barton et al., 2013; Ringbauer et al., 2017). Therefore, the rare variant sharing patterns by geographical distance observed here are best interpreted as primarily resulting from recent isolation-by-distance.

A few more aspects of the Mantel correlograms (Appendix D.11) are worth highlighting briefly. Firstly, the slightly higher values of the Mantel statistic for CP vs f_2 sharing likely reflect that the former is less affected by differences in rare variant densities as mentioned above. Secondly, for both sharing matrices the Mantel correlations becomes non-significant at ca. 3,000-4,000 km, this distance indicates a quantity defined as "patch size" by (Sokal and Wartenberg, 1983). While in theory it can be used to gain insights about migration rates and other aspects of population dynamics this is not further pursued here. Finally, the slight increase in the Mantel

statistic in the distance bins >7,500 km is most plausibly interpreted as reflecting the impact of Post-Columbian long-distance gene flow between the Old World and the Americas.

4.3.2 Detecting low level admixtures using rare variant approaches

Overall, the analyses presented here provide strong support for a previously undetected low-level African admixture in the Andean Calchaquíes. This conclusion rests on rare allele sharing, relative to both known African-admixed groups and sequence data simulated under realistic scenarios, as well as a signal of admixture LD. The magnitude of this gene flow is estimated at 0.8-2.2% and it is dated to 8.9-13.6 generations ago. In a supra-regional context this finding is plausible, given that post-Columbian South American population history has been characterised by widespread admixture of Native Americans with Europeans and Africans. The upper boundary of this signal is comparable to the mean value of 2.5% African ancestry reported for Peruvians from Lima (Martin et al., 2017). The date obtained using ALDER also is consistent with the range of 9-14 generations which has been reported as the (starting) date of African/European admixture into Native Americans in continental South America (Adhikari et al., 2016b).

The Calchaquí samples were collected in the town of Cachi, located in the Salta Province of Northwest Argentina (Eichstaedt et al., 2015). Generally, urban Argentinians exhibit high European ancestry proportions compared to the populations of most neighbouring Latin American countries. A study using multiple genetic systems detected autosomal averages of 78.6% European, 17.3% Native American and 4.2% West African ancestry with a relative overrepresentation of European Y-chromosomal and Native American mtDNA lineages (Corach et al., 2010). Recent work has shown that for the Northwest of Argentina these patterns are almost exactly reversed with a genomic make-up of only 13.7% autosomal European ancestry and >80% Native American ancestry (Muzzio et al., 2018). These populations cluster closely genetically with indigenous Central Andeans such as the Quechua and Aymara (Chacon-Duque et al., 2018). The latter is consistent with historical evidence that before European contact Northwest Argentina was more densely populated than other regions of present-day Argentina (Martínez Sarasola, 2005) and that it was less affected by the massive influx of European immigrants into the country from 1870-1950 (Motti et al., 2012).

There are two previous studies that have analysed genetic data from the Calchaquíes. Pagani et al. (2016) presented the five Calchaquí WGS which are the focus of this chapter. These in turn

were a subset of a larger group of 24 individuals. 730k SNP data from these was published by Eichstaedt et al. (2015). Both studies applied the allele-frequency-based ADMIXUTRE method to analyse population structure and were unable to detect the African admixture into the Calchaquíes inferred here from rare variant sharing patterns.

Very low-level European admixture into the five Calchaquíes analysed here cannot be ruled out, even though the total cumulative genetic length (7.11 cM) covered by the respective RVCs is less than an eighth of that covered by Calchaquíes-African RVCs (59.55 cM). These numbers can however not be compared directly, as Europeans and Native Americans are not as deeply diverged and Europeans exhibit a lower rare variant density than Africans which decreases the power of the RVC approach. An interesting contrast between the genotyping SNP assay and WGS studies is that the former estimated a mean 8.6% fraction of European ancestry in the Calchaquíes whereas the latter, in closer agreement with the RVC method, did not detect such a signal. At least two factors can explain this apparent contradiction. Firstly, the SNP data from the subset of the five Calchaquíes for whom WGS was later generated had an average European ancestry proportion of 5.8%, i.e. somewhat below the mean of all samples and all of the three sampled males carried the Y chromosome haplogroup Q1a, which is most prevalent in Native Americans (Dr Christina Eichstaedt, personal communication, 28/10/2018). Secondly, it has been demonstrated that even for a moderately high number of markers (10k) the unsupervised standard ADMIXTURE algorithm can erroneously introduce patterns resembling low-level admixture (Lange et al. 2011). This bias can be reduced by increasing the number of analysed SNPs further (>100k). Such a tenfold difference corresponds approximately to the sizes of the sets of markers the respective ADMIXTURE analyses by Eichstaedt et al. (2015) vs. Pagani et al. (2016) were run on.

The low African ancestry proportion in the Calchaquíes described here most likely reflects forced migration due to the transatlantic slave trade. African slaves were first introduced to the region in the late 16th century and a substantial inflow continued until 1812. The latter is estimated at ca. 70,000 slave arrivals for the whole Viceroyalty of the Río de la Plata (mostly comprising modern Argentina) from 1777-1812 alone (Borucki, 2011). Correspondingly, historical records suggest that during the late 18th century as much as one third of the population of Buenos Aires was of African origin (Andrews, 1980). According to the current Argentine census, however, < 0.5% of the country's population identifies as Afro-Argentines (http://www.indec.gov.ar/definitivos_bajarArchivoNacionales.asp?idc=58&arch=x&c=2010).

The ongoing debate about the causes of the subsequent decline in their numbers can only be hinted at here. Most likely it reflects a combination of demographic pressures against individuals with a large African ancestry proportion as they were overrepresented among the victims of epidemics and wars, the already mentioned large-scale European immigration and the categorisation of individuals of partial African ancestry as "white" by the authorities (Cottrol, 2007).

So perhaps the African ancestry detected here in the Calchaquíes can be interpreted as a population genetic remnant of this earlier period of Argentinian history. It can be speculated that the contributing ancestors were slaves or freed Africans who had been brought to the interior of the country to work in agriculture or mining. It is also interesting to note that compared to the communities of the other Andean populations studied (Colla, Wichi) the town of Cachi, where also almost all the grandparents of the Calchaquíes analysed here originated from, is geographically somewhat more accessible from Salta, the provincial capital (Dr Christina Eichstaedt, personal communication, 16/06/2016).

Beyond the anthropological and historical context there are several methodological aspects worth highlighting. They concern the sensitivity of different approaches for detecting admixture, the reliability of the estimates of admixture fractions and date and finally what can be learned about the power of rare variant approaches from simulations.

The possibility of admixture events between two sources that do not result in a significantly negative f_3 statistic with the respective target population was already considered by the authors of the method (Patterson et al., 2012). Based on the equations underlying the f_3 introduced by Reich et al. (2009) there are several aspects of population history to consider.

The value of the f_3 statistic for a trio of populations C (target population for admixture test) A and B for a single polymorphism is defined as follows:

$$E[f_3(C;A,B)] = E[(C-A)(C-B)] = (c'-a')(c'-b') \quad (4.8)$$

Note that a', b' and c' are the population alleles frequencies at the respective marker in A, B and C. In their paper Reich et al. (2009) also demonstrated that it is valid to calculate expected values for the f_3 by tracing drift paths through a tree and decomposing an admixture tree relating A,B and C. Figure 4.17 illustrates that in such a tree under the hypothesis of admixture there are two possible paths each from C to A and C to B (two are "short" and two are "long", i.e. go

through the root of the tree). Note that here A' and B' are defined as the (unknown) true admixing populations who are related to A and B.

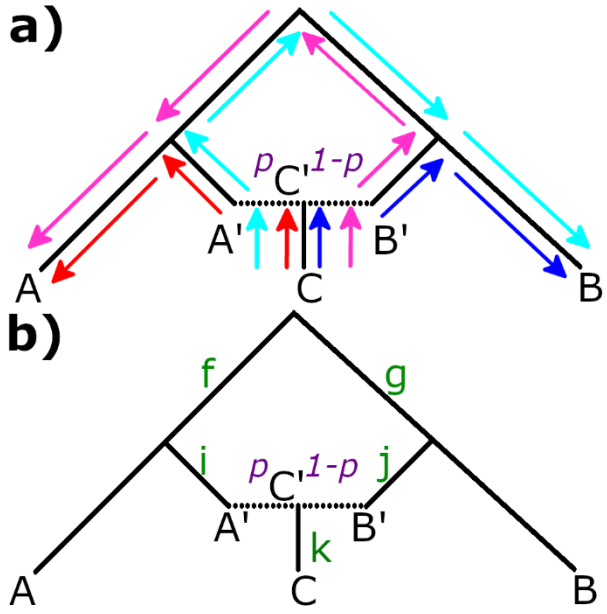


Figure 4.17: Visual interpretation of the f_3 statistic between a putatively admixed group C and source proxies A and B. A' and B' represent the actual groups who admixed to form C' and who might be unknown. The arrows in a) indicate the four possible ways C can be connected to either A or B. The shorter paths are highlighted in red and blue, whereas pink and turquoise are used for the longer paths. P and 1-p denote the putative admixture fractions A' and B' contributed to C'. The green letters in b) label the segments which are shared by both the paths from C to A as well as the paths from C to B and therefore contribute to the expected f_3 statistic (see Eq.4.9) (Modified from Reich et al., 2009).

To explore the product of the shared drift between C and A vs C and B visually in a tree four possible combinations of two drift paths need to be considered. They are: CA(short) + CB(short); CA(short) + CB(long); CA(long) + CB(long); CA(long) + CB(short). These subtrees can be added up linearly and are weighted by the admixture fraction p contributed by A' if a path goes through A' and/or (1-p), the complementary fraction contributed by B', if a path goes through B'. This results in the following expression (for the exact tree segments represented by f-k see Figure 4.17):

$$E[f_3(\mathrm{C};\mathrm{A},\mathrm{B})] = \mathrm{p}(1-\mathrm{p})\mathrm{k} + \mathrm{p}^2(\mathrm{k}+\mathrm{i}) + \mathrm{p}(1-\mathrm{p})(\mathrm{k}-\mathrm{f}-\mathrm{g}) + (1-\mathrm{p})^2(\mathrm{k}+\mathrm{j}) \qquad (4.9)$$

This can be simplified to:

$$E[f_3(\mathrm{C};\mathrm{A},\mathrm{B})] = \mathrm{k} + \mathrm{p}^2\mathrm{i} - \mathrm{p}(1-\mathrm{p})(\mathrm{f}+\mathrm{g}) + (1-\mathrm{p})^2\mathrm{j} \qquad (4.10)$$

From Eq. (4.10) it can be seen that $f_3(\mathrm{C};\mathrm{A},\mathrm{B})$ is likely to be highly negative if a) p is close to 0.5, i.e. both admixing groups A' and B' contributed comparable amounts of ancestry, b) population-specific drift k from C' to C, i.e. since the population was formed, has been small,

c) the common ancestors of A/A' and B/B' respectively were deeply diverged, i.e. f and g are large, d) A' and B' were not drifted far from their common ancestors with A and B, i.e. i and j are small.

Therefore, likely the primary reason that the African admixture in the Calchaquíes is not detectable with the f_3 on a genome-wide level is its small magnitude. In addition, the effect of the application of the inbreeding correction (Table 4.5), which gives a greater weight to homozygous sites for contributing negatively to the f_3 statistic, suggests that drift in the Calchaquíes, even in the very brief period since the admixture, has reduced the likelihood of the signal being identified.

It is interesting to note that when the f_3 test was run with the Siberian Eskimos as surrogate population for the Native American ancestors of the present-day Calchaquíes condition d) becomes relevant as the population split between these two groups has been estimated at ~14 kya (Appendix D.4). The strong drift the ancestors of the Andeans have experienced since this divergence results in a considerably more positive value for f_3(Calchaquíes;Yoruba,Eskimo) than f_3(Calchaquíes;Yoruba,Wichi). Peter (2016) who explored the power of admixture f_3 statistics in the context of population genetic theory and simulations also derived conditions a)-c), he however did not consider d) as it involves "internal" populations which cannot be observed empirically.

The example of the substitution of the Wichi by the Eskimos and the subsequent change in the f_3 statistic however demonstrates that for certain cases the choice of good surrogate populations is crucial even if the underlying topologies of the target and source populations are identical, as for f_3(Calchaquíes;Yoruba,Eskimo/Wichi). Taken in isolation, the significant result of the f_3 statistic obtained for chromosome 13 should be treated with some caution. The significance of the z-score depends on the accuracy of the jackknife standard error, which for this chromosome only reflects 20 independent outcomes (i.e. 5 cM blocks). Although in the literature (Moorjani et al., 2011) standard errors for population genetic statistics have been calculated from a comparable number of blocks.

The inconsistent exponential decay rates observed for the ALDER method when the Calchaquíes are fit as a mixture of Siberian Eskimos and Yoruba raise the question whether these results should be treated with low confidence. However, the authors of ALDER demonstrated that when one group, as Native Americans here, contributes the vast majority (≥ 90%) of ancestry to a two-way mixture the amplitude of the one-reference weighted LD curve

between the target population and the reference representing this majority ancestry becomes very small (Figure 4.12B) or the signal vanishes completely. Therefore, the estimate of the exponential decay rate from fitting a one-pulse admixture model to the weighted LD Calchaquíes-Eskimo curve becomes very imprecise and the inconsistency with the Calchaquíes-Yoruba decay rate occurs.

The signal is further weakened by the complex demographic history of the Siberian Eskimos. Recent aDNA studies suggest that they are best modelled as deriving one third of their ancestry from a Native-American-like group and two thirds from a source represented by a 9.8 kya old individual from Northeast Siberia ("Ancient Palaeosiberians"), the latter being an outgroup to all Native Americans (Sikora et al., 2018).

Colla and Wichi could not be fit as a references for ALDER (data not shown). They are also affected by the fact that the Native American contribution to the Calchaquí genome is much larger than the African. Additionally, the z-scores obtained for the Calchaquíes-Wichi/Calchaquíes-Colla weighted LD curves estimated from jackknifing over chromosomes do not reach significance. This perhaps reflects effects resulting from strong locally specific bottlenecks these Andean populations experienced after they diverged which overshadow the already weak real signal. The bottleneck-related explanation gains further plausibility as the much more distantly related Han Chinese could be fit as a reference (data not shown).

Regardless of these issues, a theoretical framework (Loh et al., 2013; Pickrell et al., 2012) exists which also allows reliable parameter inference from one-reference weighted LD curves. As already mentioned, the admixture date from the Calchaquíes-Yoruba one-reference curve is consistent with expectations based on historical records and genetic studies of other admixed groups from the Americas. The ALDER estimated admixture fraction of 2.2% African ancestry in the Calchaquíes represents a lower bound. However, it seems unlikely that the true African contribution is much higher, as in that case the f_3 signal would have been less ambivalent and the other approaches used here suggest a lower mixture proportion. Furthermore, the rare variant sharing profile of the African-like chunks in the Calchaquíes matches that of the Yoruba well (Table 4.7) suggesting that the latter represent a reasonably good proxy for the true admixing source.

A final caveat concerning the ALDER results is that it has been argued that with small sample sizes ALDER inferences become noisy (Ayub et al., 2015). Simulations suggest that with larger sample sizes both the allele frequencies as well as the LD estimates underlying the weighted

LD inference method should become more accurate (Loh et al., 2013). There is, however, little evidence that low sample sizes produce spurious signals, instead the uncertainty manifests in form of larger standard errors, as observed for the Calchaquíes and the respective reference populations here.

The approximate estimate of the African admixture fraction in the Calchaquíes (0.81%) using the output of CP should likewise be interpreted with caution. It could represent an underestimate of the true African proportion in the Calchaquíes as only haplotypes found in sampled Africans are considered as donors. Therefore, some haplotypes the Calchaquíes "receive" from other non-Africans which could represent shared African gene flow into both groups are excluded. On the other hand, the Wichi, whose African CP sharing is used a proxy for a non-admixture scenario, appear to have an even smaller long-term N_e than the Calchaquíes (Appendix C.3). Therefore, their amount of old short CP haplotypes shared with any other population would be reduced as has been demonstrated for other severely bottlenecked groups (van Dorp et al., 2015). A potential follow-up analysis would be to use more statistically rigorous approaches that model the CP sharing/"painting" profile of a particular group as a mixture of sharing profiles from other groups and use either non-linear least squares (Leslie et al., 2015) or Bayesian (Chacon-Duque et al., 2018) methodologies to infer the respective mixture coefficients.

When the total number/genetic lengths and RVC length distributions are combined the African-Calchaquíes shared tracks fall in-between populations with known larger fractions of African admixture and groups that are thought not to have received such gene flow (Figure 4.13). One weakness of this approach is that RVC length distributions for groups which received African gene flow and those that did not were indistinguishable in certain cases, e.g. Calchaquíes vs Wichi. To some extent larger sample sizes could improve this situation as the Anderson-Darling test should become more powerful with more RVCs. However, for very low admixture proportions the randomness of the underlying true IBD sharing reflected by the RVCs would make the detection of the signal difficult in any case.

The observed African-Calchaquíes RVCs are clearly much more abundant in terms of numbers as well as cumulative genetic length compared to the RVCs from all 20 simulation runs between groups with a broadly similar demography but lacking a recent admixture pulse. Therefore, such an extreme outcome in terms of rare variant sharing is very unlikely to occur without recent excess gene flow.

Between scenarios including a very small admixture pulse of 0.5% at different recent time points there is a decline of total RVC sharing as a function of the admixture date (Figure 4.14). While this is in line with theoretical expectations of the breakdown of IBD segments due to recombination over time, there is not enough data to infer the exact shape of this decay curve. It is worth highlighting again that an admixture pulse of 0.5% which lies further back than 15 generations starts to become indistinguishable from a scenario with relatively modest (African-East Asian) background migration rates under the given sampling scheme (Figure 4.14). While the cumulative genetic lengths of the RVCs still appear to be differentiated between admixture and non-admixture scenarios, this summary statistic also seems to be more prone to statistical noise than the raw RVC numbers due to the overestimation of the length of putatively underlying true IBD segments (see section 4.2.3).

When background migration rates are high enough, they will introduce more interpopulation RVCs which might either reflect continuous IBD segments due to very distant relatedness or conflation of smaller very old segments that are picked up as one RVC. It can be speculated that for introgression from a group that is less deeply diverged from Native Americans and/or has a lower rare variant diversity than Sub-Saharan Africans the detectability threshold would lie at higher admixture fractions.

Scenario F with an admixture fraction of 1% dated to 18 generations ago seems to be a good fit to the empirical cumulative genetic length of RVCs. However, the total number of RVCs and the raw numbers of doubletons forming these clusters, which is ca. threefold higher for the empirical data (3,290) relative to this specific scenario (1,092), suggest that simulations with either more recent admixture dates or higher admixture fractions would provide a better fit to the observed patterns.

Despite the simulated admixture date already being older than the most plausible range for the real date, the RVC length distribution obtained from scenario F is shifted towards longer runs compared to the empirical distributions for Africans vs each of the admixed American groups respectively (Figure 4.15, Appendix D.37). Several factors likely contribute to this difference. Firstly, the simulations most likely overestimate the recent relatedness of the Yoruba to the individuals who introgressed into the ancestors of the Calchaquíes. Secondly, the simplification of African demographic history for the simulations could reduce the amount of very old lineages represented by very short RVCs; underestimating the Yoruba's long-term N_e would have a similar effect. Finally, this pattern could be interpreted as representing an excess of short RVCs

in the empirical data, a phenomenon also observed by Harris and Nielsen (2013) for IBS segments. Besides complex demography, these could also reflect the non-uniformity of mutation rates across the genome (Appendix D.38) not incorporated into simulations here.

There are several aspects which were not considered explicitly or simplified for the simulations. One such factor was hypothetical selection against African alleles after admixture. There is some limited evidence that variants representing adaptations to moderate hypoxia (as a consequence of intermediate altitude) have potentially been selected for in the Calchaquíes (Eichstaedt et al., 2015). A large-scale study on Peruvian WGS found support for asymmetric migration from the Andes towards coastal regions and the Amazon basin (Harris et al., 2018). This could suggest a disadvantage to migrants from the latter two into the Andean region who would have lacked genetic adaptations to high altitude but could also reflect settlement policies of the Inca and Spanish empires. However, even in much larger and well-studied recently admixed populations such as African-Americans the extent of ancestry-specific selection is debated (Bhatia et al., 2014; Jin et al., 2012). Therefore, it is unclear whether this phenomenon has measurably impacted genome-wide ancestry proportions in the Calchaquíes.

Furthermore, all simulated genomes are effectively (i.e. in terms of SNP-containing regions) longer than real genomes as a substantial fraction of the reference genome (GRCh37/hg19) is inaccessible. This portion consists of ~200 Mbp representing centromeres, telomeres and the short arms of the acrocentric chromosomes as well as a further ~30 Mbp of interstitial gaps, the latter mostly regions that could not be cloned or assembled with confidence (Genovese et al., 2013). For the genetic lengths of the RVCs this does not impact the comparisons of empirical to simulated data as the relevant genetic map is also limited by the accessible regions. Therefore, simulated RVCs that fall completely into the inaccessible portion are assigned a genetic length of 0 cM. However, assuming uniformity of mutation rates and considering that each simulated genome effectively has a length of 2881 Mbp (see section 4.1.4), the simulations would yield an 8.6% higher total number of doubletons compared to a real genome that has undergone the same demographic history.

The baseline demographic history simulated using cosi2 contains several simplifications which could partly explain differences in RVC sharing patterns compared to the empirical data (see also Appendix D.35C). Firstly, Native American population history is more complex than modelled here. The most important large-scale event is gene flow from another East Eurasian group labelled ancient North Eurasians after the Native American ancestors diverged from East

Asians (Raghavan et al., 2014). The downstream impact of this is that the recent population growth of the Calchaquíes could have been overstated as the MSMC estimates of present-day N_e are known to be inflated for admixed populations (Li and Durbin, 2011). It is uncertain though for how long after the admixture event this would measurably bias MSMC inferences. If this was the case this overestimation of recent growth would lead to a higher population mutation rate which in turn would increase the rare variant density in simulated data which can impact the detectability of RVCs.

Indeed, the number of singletons and doubletons private to the Calchaquíes ($\overline{n}$ = 32,234) relative to similar variants in Pop5 ($\overline{n}$ = 49,290) is considerably lower. This effect is also partly due to the other Native Americans and Northeast Siberians not being explicitly modelled in the simulations. Therefore, doubletons of the Calchaquíes with these groups are recorded as variants private to Pop5. However, the practical impact of these differences in rare variant density is likely only modest given that the average individual-level vrGV database for an African contains 5-6 times more rare variants than for a Calchaquí. Even under these extreme circumstances, African-Calchaquí RVCs detected with the Native American group as reference mostly overlap with those detected using Africans as reference and the former cover only a 1.4 times greater genomic length than the latter (Appendix D.30).

The perhaps most important aspect is the considerable uncertainty surrounding long-term background migration rates between (South) American Indians and other continental populations. Raghavan et al. (2015) estimated the background migration rate between the Amazonian Karitiana and the Han at $9.5*10^{-5}$ migrants per generation using *diCal2.0* (Steinrücken et al., 2015), a composite likelihood approach applied to haplotype data. This is about three times higher than the rate assumed here ($3.11*10^{-5}$) between Pop4 and Pop5, but of a similar order of magnitude.

While it seems intuitive that the pre-Columbian continental migration rates between Native Americans and Africans/West Eurasians were lower than the analogous rates for East Asians there is no consensus about the magnitude of this difference. Some authors (Browning et al., 2018) have circumvented this issue by approximating Native Americans in simulations by Gravel et al.'s (2011) parameter estimates for East Asians. However, the cosi2 simulations conducted here demonstrate that even assuming East-Asian-like background migration rates with Africans a signal of rare variant sharing as strong as that between Calchaquíes and Sub-Saharan Africans cannot be generated without a recent admixture event (Figure 4.14).

5. Conclusions and future directions

This thesis consists of three projects connected by an overarching aim to improve our understanding of human genetic diversity in understudied regions. It addresses a broad range of questions from population-specific patterns of functional and deleterious mutations to the application of rare variant approaches to detect cryptic relatedness in population history.

The first subproject described population structure in 196 individuals from nine Southeast Asian populations genotyped using an Illumina OmniExpress 730k SNP array and placed them in a global as well as regional context. Allele frequency-based methods suggested that the Kankanaey from the Northern Philippines could represent either the best extant genetic proxy or a closely related sister group to the original Austronesian settlers. Their homozygosity metrics were comparable to other ISEA groups suggesting that their ancestry profile is unlikely to result solely from genetic drift. Minor contributions from South Asian sources to the ancestral make-up of the Malay, Bajo and lowland Filipinos, dating to the last two millennia, were detected. This represented a contrast to the documented strong cultural and linguistic impact from the Indian subcontinent on the region, indicating that these connections likely did not involve large-scale population movements.

Since the publication of the paper that chapter 2 is based on the amount of genomic data from ISEA and Oceania has greatly increased. One potential follow-up could be a comparison of the populations studied here to these new samples. This would greatly increase marker density and especially the power of haplotype-based approaches relative to using the Pan-Asia panel as reference. One interesting aspect would be re-evaluating the status of the Kankanaey. In a recent study of the Eastern Polynesian Leeward Society Islanders (Hudjashov et al., 2018) allele frequency-based tests were consistent with the Kankanaey being the best extant match to the early Austronesian colonists. However, GLOBETROTTER analyses by the same authors indicated that lowland Filipinos (though recently admixed from other sources) are more representative of the haplotypic variation in those ancestors. Another task would be fine-mapping the precise Indian sources which introgressed into various ISEA populations. The range of ALDER dates (Table 2.3) and subsequent analyses of other Bajo groups by Kusuma et al. (2017) suggest that the “Indian signal” reflects multiple admixture events.

The two following chapters presented analyses on a global dataset of 483 high coverage whole genomes. Chapter 3 investigated the distribution of functional and deleterious variants in a subset of 382 genomes. It was demonstrated that when the latter dataset was annotated using the two different software packages Ingenuity Variant Analysis and Ensembl's Variant Effect Predictor the overall rate of matching exonic annotations was 74.9%. The two main reasons for this outcome were differences in the underlying transcript sets, with RefSeq generally being more conservative than Ensembl, and that IVA did not detect most splice site variants. However, the major patterns reported in the following were consistent between both methodologies.

Across all populations there was a shift of missense and nonsense variants towards lower frequencies relative to synonymous variants consistent with the predicted effect of purifying selection. An interesting novel result was that Andean Native Americans exhibited the highest ratio of non-synonymous to synonymous mutations of all populations. Almost all DAF spectra for weakly deleterious variants (missense and CADD score >20) were significantly differentiated between macro-groups whereas the amount of geographic structure was reduced for strongly deleterious mutations (LoF and CADD score >30) suggesting more uniformly acting purifying selection on the latter. In agreement with this, long-term N_e obtained using MSMC in addition to distance from Africa improved the explanatory power of a simple linear model for predicting the count of weakly but not strongly deleterious variants.

The $R_{X/Y}$-statistic, which assesses the per-genome accumulation of non-synonymous mutations between different macro-groups, provided support for purifying selection on pigmentation genes in West/Central Africans. Per-individual counts of putative LoF as well as potentially disease-causing variants from the HGMD database were consistent with other large-scale sequencing efforts such as the 1000 Genomes and Simons Genome Diversity Project. This is important given the differences in sample composition, sequencing technologies and annotation methods. Furthermore, 29 previously unreported homozygous rare LoF were detected, mostly in isolated Siberian and ISEA groups with low N_e.

Applying the ΔDAF statistic to missense variants to detect sites under positive selection confirmed many known signals and suggested potential new targets, though uncertainties regarding false positive rates remain. The current evidence concerning nonsense variants highly differentiated between populations was not sufficient to reject a scenario where the majority of these are selectively neutral. Lastly, integrating selection scans with functional databases

highlighted rs11227639 in Northeast Siberians. This SNP upregulates the expression of *ACTN3* which encodes the structural muscle protein actinin-3. Evidence from rs1815739, a well-studied LoF mutation affecting this gene, suggests that ACNT3 mRNA and protein levels impact muscularity and explosive power performance in a dosage-dependent manner. Future work should focus on these Northeast Siberian populations to investigate whether individuals carrying one or two copies of the upregulating allele at rs11227639 exhibit a) more lean muscle mass and b) elevated grip strength.

Additional follow-ups to chapter 3 should again benefit from the progress of the field during last few years. Firstly, it could be tested whether the convergence of annotations increases with newer versions of the transcript sets, though this requires lifting over genomic coordinates which would result in minor information loss. Secondly, the efficacy of purifying selection could be quantified using model-based approaches. One example would be the recently developed Fit∂a∂I (Kim et al., 2017) which approximates the DFE of new non-synonymous mutations from fitted model parameters. From these estimates the fixation probability of new deleterious mutations over neutral mutations can in turn be derived (Lopez et al., 2018). However, to be robust these approaches would require larger samples from homogeneous populations.

In chapter 4 global sharing patterns of rare variants were analysed in a subset of 447 genomes from the EGDP designed to maximise coverage of ethnic and linguistic diversity. The results obtained for per-individual rare variant totals and interpopulation sharing patterns were consistent with the vast majority of f_2 variants being of recent origin and reflecting sharing due to common descent. The f_2 as well as the related rare variant cluster (RVC) methodologies appear to be fast and powerful heuristic approaches to describe recent shared ancestry. Normalised f_2 sharing correlated very strongly (r = 0.938) with total ChromoPainter sharing, a state-of-the-art method describing pairwise haplotype sharing. This normalised f_2 metric generally declined exponentially with geographical distance. Furthermore, it was demonstrated that the lengths of shared RVCs contain information about population history beyond what a single genome-wide summary statistic can provide.

Strong evidence was presented that excess rare variant sharing of the Andean Calchaquíes with West Africans represents a "cryptic" post-Columbian admixture event which was not detectable on a genome-wide level using allele frequency-based methods. Realistic demographic scenarios were not able to produce rare variant sharing patterns as extreme as those observed between

these two groups without incorporating such a recent admixture. The simulations also suggested that admixtures as low as 1% between deeply diverged groups can be confidently discriminated from non-admixture scenarios using the RVC methodology if the gene flow is relatively recent (≤20 generations ago). Finally, if RVCs were interpreted as continuous haplotypes inconsistencies with statistical phasing obtained using SHAPEIT2 occurred, either reflecting phasing errors or the conflation of short adjacent haplotypes marked by doubletons into longer RVCs.

Follow-ups to the rare variant analyses were already outlined in chapter 4, the most important being more extensive simulations including ground-truth IBD segments to better characterise the statistical power of the methodology. A complementary approach to the cosi2 simulations conducted here would be the rarecoal (Flegontov et al., 2017; Schiffels et al., 2016) framework. This method maximises the likelihood of the observed joint distribution of rare alleles to estimate demographic model parameters. However, like Fit∂a∂I the ideal sample sizes for local populations to exploit the power of this method need to be 10-100 times larger than those comprising the EGDP.

A second future avenue would be to explicitly connect rare variant analyses to functional annotations to explore how f_2 sharing patterns differ across classes of deleterious mutations. Another possibility, which might appear counterintuitive given that on average they are of recent mutational origin, is that long SNP-dense RVCs could be informative about archaic introgression. This idea is motivated by a method conceptually related to the detection of RVCs, the S* statistic (Plagnol and Wall, 2006), which highlights population-specific SNPs that are in (almost) complete LD, that been shown to detect regions enriched for archaic introgression (Browning et al., 2018; Vernot and Akey, 2014).

While the work presented in this thesis was conducted the field of human evolutionary genetics has continued to develop at a rapid pace. Projecting into the future is inherently challenging, however a few trends as well topics which in the author's opinion require further research will be briefly touched upon here.

A trend which is certain to continue is the increase in data volume. In the very near future in principle hundreds to thousands of high-coverage whole genomes from a specific local population could be generated relatively cost-efficiently. Approaches relying on shared genetic drift/allele frequency patterns to reconstruct demography reach saturation at relatively modest (n = 20 from a local population) sample sizes and are unlikely to offer any radically new

conclusions. However, methods exploiting different features of the data, e.g. haplotype information or the rare site frequency spectrum should become more powerful, especially for recent population history and analyses involving deleterious variants which are shifted towards lower frequencies. An example for the potential of haplotype-based approaches to provide new insights into fine-scale population history from large samples would be a recently described signal of wide-spread Eastern Mediterranean/North Africans ancestry in Latin Americans (Chacon-Duque et al., 2018). Some methods widely used in the field do not scale well with the vast increases in sample size which requires innovative new approaches. An example is the recently published tsinfer method (Kelleher et al., 2018). It infers the so-called tree sequences, which can be thought of as a summary encoding of the genealogies of a sample and can process ca. four orders of magnitude more sequences than previous methods while generating genealogies with comparable information content.

Functional and mechanistic inferences from genomic data should greatly benefit from increased sample sizes, e.g. there is increasing evidence for population-specific mutation rates in particular genomic contexts (e.g. Mathieson and Reich, 2017). This in turn leads to new questions concerning the biological drivers of these processes and their relevance for the inference of population history. Another idea which should be pursued further is a "human knockout project". A crucial component for it is the availability of large exome datasets, preferably from populations enriched for homozygous LoF variants resulting in gene loss. Pioneering studies by the ExAC (Lek et al., 2016) and PROMIS (Saleheen et al., 2017) suggest a great potential to improve our understanding of gene function, especially in combination with data from medical records and targeted experimental follow-ups.

There are other ongoing concerns which have started being addressed. Non-Europeans and Sub-Saharan Africans in particular, are still underrepresented in genomic studies. More data from these groups will not only be crucial for understanding population-specific components of genetic disease risk but also to make more reliable inferences about the deep history of our species.

The relative increase in available aDNA data has even outstripped the growth of the field at large. There is a need for methods which better incorporate the uncertainty associated with genotypes inferred from aDNA when analysing ancient samples. To contextualise these results, it will be crucial to intensify collaborations with archaeologists, linguists, historians and palaeontologists.

Overall, the coming years still hold great promise for a richer, more fine-grained understanding of our species' history as well as new perspectives into the relationship between genomic variation and human health and disease.

6. Bibliography

Acabado S. 2017. The Archaeology of Pericolonialism: Responses of the "Unconquered" to Spanish Conquest and Colonialism in Ifugao, Philippines. Int J Hist Archaeol 21:1–26.

Adhikari K, Fontanil T, Cal S, Mendoza-Revilla J, Fuentes-Guajardo M, Chacón-Duque J-C, Al-Saadi F, Johansson JA, Quinto-Sanchez M, Acuña-Alonzo V, Jaramillo C, Arias W, Barquera Lozano R, Macín Pérez G, Gómez-Valdés J, Villamil-Ramírez H, Hunemeier T, Ramallo V, Silva de Cerqueira CC, Hurtado M, Villegas V, Granja V, Gallo C, Poletti G, Schuler-Faccini L, Salzano FM, Bortolini M-C, Canizales-Quinteros S, Rothhammer F, Bedoya G, Gonzalez-José R, Headon D, López-Otín C, Tobin DJ, Balding D, Ruiz-Linares A. 2016a. A genome-wide association scan in admixed Latin Americans identifies loci influencing facial and scalp hair features. Nat Commun 7:10815.

Adhikari K, Mendoza-Revilla J, Chacón-Duque JC, Fuentes-Guajardo M, Ruiz-Linares A. 2016b. Admixture in Latin America. Curr Opin Genet Dev 41:106–114.

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. Nat Methods 7:248–249.

Agrawal AF, Whitlock MC. 2011. Inferences About the Distribution of Dominance Drawn From Yeast Gene Knockout Data. Genetics 187:553–566.

Aguet F, Ardlie KG, Cummings BB, Gelfand ET, Getz G, Hadley K, Handsaker RE, Huang KH, Kashin S, Karczewski KJ, Lek M, Li X, MacArthur DG, Nedzel JL, Nguyen DT, Noble MS, Segrè AV, Trowbridge CA, Tukiainen T, Abell NS, Balliu B, Barshir R, Basha O, Battle A, Bogu GK, Brown A, Brown CD, Castel SE, Chen LS, Chiang C, Conrad DF, Cox NJ, Damani FN, Davis JR, Delaneau O, Dermitzakis ET, Engelhardt BE, Eskin E, Ferreira PG, Frésard L, Gamazon ER, Garrido-Martín D, Gewirtz ADH, Gliner G, Gloudemans MJ, Guigo R, Hall IM, Han B, He Y, Hormozdiari F, Howald C, Kyung Im H, Jo B, Yong Kang E, Kim Y, Kim-Hellmuth S, Lappalainen T, Li G, Li X, Liu B, Mangul S, McCarthy MI, McDowell IC, Mohammadi P, Monlong J, Montgomery SB, Muñoz-Aguirre M, Ndungu AW, Nicolae DL, Nobel AB, Oliva M, Ongen H, Palowitch JJ, Panousis N, Papasaikas P, Park Y, Parsana P, Payne AJ, Peterson CB, Quan J, Reverter F, Sabatti C, Saha A, Sammeth M, Scott AJ, Shabalin AA, Sodaei R, Stephens M, Stranger BE, Strober BJ, Sul JH, Tsang EK, Urbut S, van de Bunt M, Wang G, Wen X, Wright FA, Xi HS, et al. 2017. Genetic effects on gene expression across human tissues. Nature 550:204–213.

Akey JM. 2002. Interrogating a High-Density SNP Map for Signatures of Natural Selection. Genome Res 12:1805–1814.

Akey JM. 2009. Constructing genomic maps of positive selection in humans: Where do we go from here? Genome Res 19:711–722.

Akey JM, Ruhe AL, Akey DT, Wong AK, Connelly CF, Madeoy J, Nicholas TJ, Neff MW. 2010. Tracking footprints of artificial selection in the dog genome. Proc Natl Acad Sci 107:1160–1165.

Albalat R, Cañestro C. 2016. Evolution by gene loss. Nat Rev Genet 17:379–391.

Albert FW, Bloom JS, Siegel J, Day L, Kruglyak L. 2017. Genetics of trans-regulatory variation in gene expression.

Alexander DH, Lange K. 2011. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. BMC Bioinformatics 12:246.

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Res 19:1655–1664.

Al-Khudhair A, Qiu S, Wyse M, Chowdhury S, Cheng X, Bekbolsynov D, Saha-Mandal A, Dutta R, Fedorova L, Fedorov A. 2015. Inference of Distant Genetic Relations in Humans Using "1000 Genomes." Genome Biol Evol 7:481–492.

Allen JL. 2014. Kankanaey: a role and reference grammar analysis. Dallas, Texas: SIL International Publications.

Allentoft ME, Sikora M, Sjögren K-G, Rasmussen S, Rasmussen M, Stenderup J, Damgaard PB, Schroeder H, Ahlström T, Vinner L, Malaspinas A-S, Margaryan A, Higham T, Chivall D, Lynnerup N, Harvig L, Baron J, Casa PD, Dąbrowski P, Duffy PR, Ebel AV, Epimakhov A, Frei K, Furmanek M, Gralak T, Gromov A, Gronkiewicz S, Grupe G, Hajdu T, Jarysz R, Khartanovich V, Khokhlov A, Kiss V, Kolář J, Kriiska A, Lasak I, Longhi C, McGlynn G, Merkevicius A, Merkyte I, Metspalu M, Mkrtchyan R, Moiseyev V, Paja L, Pálfi G, Pokutta D, Pospieszny Ł, Price TD, Saag L, Sablin M, Shishlina N, Smrčka V, Soenov VI, Szeverényi V, Tóth G, Trifanova SV, Varul L, Vicze M, Yepiskoposyan L, Zhitenev V, Orlando L, Sicheritz-Pontén T, Brunak S, Nielsen R, Kristiansen K, Willerslev E. 2015. Population genomics of Bronze Age Eurasia. Nature 522:167–172.

Alqallaf AK, Alkoot FM, Aldabbous MS. 2013. Discovering the Genetics of Autism. In: Fitzgerald M, editor. Recent Advances in Autism Spectrum Disorders - Volume I. InTech. Available from: http://www.intechopen.com/books/recent-advances-in-autism-spectrum-disorders-volume-i/discovering-the-genetics-of-autism

Amorim CEG, Acuña-Alonzo V, Salzano FM, Bortolini MC, Hünemeier T. 2015. Differing Evolutionary Histories of the ACTN3*R577X Polymorphism among the Major Human Geographic Groups. PLOS ONE 10:e0115449.

Amos W. 2013. Variation in Heterozygosity Predicts Variation in Human Substitution Rates between Populations, Individuals and Genomic Regions. PLoS ONE 8:e63048.

Amos W. 2016. The quantity of Neanderthal DNA in modern humans: a reanalysis relaxing the assumption of constant mutation rate. Available from: http://biorxiv.org/lookup/doi/10.1101/065359

Amos W, Flint J, Xu X. 2008. Heterozygosity increases microsatellite mutation rate, linking it to demographic history. BMC Genet 9:72.

Andrews GR. 1980. The Afro-Argentines of Buenos Aires, 1800-1900. Madison, Wisconsin: University of Wisconsin Press.

Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N. 1999. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat Genet 23:147.

Antonarakis SE. 2019. Carrier screening for recessive disorders. Nat Rev Genet 20:549–561.

Ardika W, Bellwood PS. 1991. Sembiran: the beginnings of Indian contact with Bali. Antiquity 65:221–232.

Ardika W, Bellwood PS, Sutaba IM, Yuliati KC. 1997. Sembiran and the first Indian contacts with Bali: an update. Antiquity 71:193–95.

Ardlie KG, Guigó R. 2017. Data resources for human functional genomics. Curr Opin Syst Biol 1:75–79.

Ayadi A, Birling M-C, Bottomley J, Bussell J, Fuchs H, Fray M, Gailus-Durner V, Greenaway S, Houghton R, Karp N, Leblanc S, Lengger C, Maier H, Mallon A-M, Marschall S, Melvin D, Morgan H, Pavlovic G, Ryder E, Skarnes WC, Selloum M, Ramirez-Solis R, Sorg T, Teboul L, Vasseur L, Walling A, Weaver T, Wells S, White JK, Bradley A, Adams DJ, Steel KP, Hrabě de Angelis M, Brown SD, Herault Y. 2012. Mouse large-scale phenotyping initiatives: overview of the European Mouse Disease Clinic (EUMODIC) and of the Wellcome Trust Sanger Institute Mouse Genetics Project. Mamm Genome 23:600–610.

Ayub Q, Mezzavilla M, Pagani L, Haber M, Mohyuddin A, Khaliq S, Mehdi SQ, Tyler-Smith C. 2015. The Kalash Genetic Isolate: Ancient Divergence, Drift, and Selection. Am J Hum Genet 96:775–783.

Bae CJ, Douka K, Petraglia MD. 2017. On the origin of modern humans: Asian perspectives. Science 358:eaai9067.

Bakdash JZ, Marusich LR. 2017. Repeated Measures Correlation. Front Psychol [Internet] 8. Available from: http://journal.frontiersin.org/article/10.3389/fpsyg.2017.00456/full

Balasubramanian S, Habegger L, Frankish A, MacArthur DG, Harte R, Tyler-Smith C, Harrow J, Gerstein M. 2011. Gene inactivation and its implications for annotation in the era of personal genomics. Genes Dev 25:1–10.

Barreiro LB, Ben-Ali M, Quach H, Laval G, Patin E, Pickrell JK, Bouchier C, Tichit M, Neyrolles O, Gicquel B, Kidd JR, Kidd KK, Alcaïs A, Ragimbeau J, Pellegrini S, Abel L, Casanova J-L, Quintana-Murci L. 2009. Evolutionary Dynamics of Human Toll-Like Receptors and Their Different Contributions to Host Defense. PLoS Genet 5:e1000562.

Barton NH, Etheridge AM, Véber A. 2013. Modelling evolution in a spatial continuum. J Stat Mech Theory Exp 2013:P01002.

Basu Mallick C, Iliescu FM, Möls M, Hill S, Tamang R, Chaubey G, Goto R, Ho SYW, Gallego Romero I, Crivellaro F, Hudjashov G, Rai N, Metspalu M, Mascie-Taylor CGN, Pitchappan R, Singh L, Mirazon-Lahr M, Thangaraj K, Villems R, Kivisild T. 2013. The Light Skin Allele of SLC24A5 in South Asians and Europeans Shares Identity by Descent. PLoS Genet 9:e1003912.

Beaumont MA, Zhang W, Balding DJ. 2002. Approximate Bayesian Computation in Population Genetics. Genetics 162:2025–2035.

Beck T, Hastings RK, Gollapudi S, Free RC, Brookes AJ. 2014. GWAS Central: a comprehensive resource for the comparison and interrogation of genome-wide association studies. Eur J Hum Genet 22:949–952.

Bell CJ, Dinwiddie DL, Miller NA, Hateley SL, Ganusova EE, Mudge J, Langley RJ, Zhang L, Lee CC, Schilkey FD, Sheth V, Woodward JE, Peckham HE, Schroth GP, Kim RW, Kingsmore SF. 2011. Carrier Testing for Severe Childhood Recessive Diseases by Next-Generation Sequencing. Sci Transl Med 3:65ra4-65ra4.

Bellenguez C, Charbonnier C, Grenier-Boley B, Quenez O, Le Guennec K, Nicolas G, Chauhan G, Wallon D, Rousseau S, Richard AC, Boland A, Bourque G, Munter HM, Olaso R, Meyer V, Rollin-Sillaire A, Pasquier F, Letenneur L, Redon R, Dartigues J-F, Tzourio C, Frebourg T, Lathrop M, Deleuze J-F, Hannequin D, Genin E, Amouyel P, Debette S, Lambert J-C, Campion D. 2017. Contribution to Alzheimer's disease risk of rare variants in TREM2, SORL1 and ABCA7 in 1,779 cases and 1,273 controls. Neurobiol Aging [Internet]. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0197458017302324

Bellina B, Silapanth P, Chaisuwan B, Thongcharoenchaikit C, Bernard V, Borell B, Bouvet P. 2014. The Development of Coastal Polities in the Upper Thai-Malay Peninsula in the Late First Millennium BCE. In: Revire N, Murphy SA, editors. Before Siam: Essays in Art and Archaeology. Bangkok: River Books. p 69–89.

Bellwood P. 2006. Austronesian Prehistory in Southeast Asia: Homeland, Expansion and Transformation. In: The Austronesians: Historical and Comparative Perspectives. Comparative Austronesian Series. ANU Press. p 103–118. Available from: http://www.jstor.org/stable/j.ctt2jbjx1.8

Bellwood P, Oxenham M. 2008. The Expansions of Farming Societies and the Role of the Neolithic Demographic Transition. In: Bocquet-Appel J-P, Bar-Yosef O, editors. The Neolithic Demographic Transition and its Consequences. Dordrecht: Springer Netherlands. p 13–34. Available from: http://dx.doi.org/10.1007/978-1-4020-8539-0_2

Bellwood PS. 2005. First Farmers: the origins of agricultural societies. Malden, MA: Blackwell Pub.

Bellwood PS. 2007. Prehistory of the Indo-Malaysian Archipelago. Canberra: ANU E Press. Available from: http://epress.anu.edu.au/?p=80041

Benjamini Y, Speed TP. 2012. Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res 40:e72–e72.

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ, Keira Cheetham R, Cox AJ, Ellis DJ, Flatbush MR, Gormley NA, Humphray SJ, Irving LJ, Karbelashvili MS, Kirk SM, Li H, Liu X, Maisinger KS, Murray LJ, Obradovic B, Ost T, Parkinson ML, Pratt MR, Rasolonjatovo IMJ, Reed MT, Rigatti R, Rodighiero C, Ross MT, Sabot A, Sankar SV, Scally A, Schroth GP, Smith ME, Smith VP, Spiridou A, Torrance PE, Tzonev SS, Vermaas EH, Walter K, Wu X, Zhang L, Alam MD, Anastasi C, Aniebo IC, Bailey DMD, Bancarz IR,

Banerjee S, Barbour SG, Baybayan PA, Benoit VA, Benson KF, Bevis C, Black PJ, Boodhun A, Brennan JS, Bridgham JA, Brown RC, Brown AA, Buermann DH, Bundu AA, Burrows JC, Carter NP, Castillo N, Chiara E. Catenazzi M, Chang S, Neil Cooley R, Crake NR, Dada OO, Diakoumakos KD, Dominguez-Fernandez B, Earnshaw DJ, Egbujor UC, Elmore DW, Etchin SS, Ewan MR, Fedurco M, Fraser LJ, Fuentes Fajardo KV, Scott Furey W, George D, Gietzen KJ, Goddard CP, Golda GS, Granieri PA, Green DE, Gustafson DL, Hansen NF, Harnish K, Haudenschild CD, Heyer NI, Hims MM, et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456:53–59.

van den Berg RA, Hoefsloot HC, Westerhuis JA, Smilde AK, van der Werf MJ. 2006. Centering, scaling, and transformations: improving the biological information content of metabolomics data. BMC Genomics 7:142.

Bergström A. 2018. Genomic insights into the human population history of Australia and New Guinea. Available from: https://www.repository.cam.ac.uk/handle/1810/273775

Bergström A, Oppenheimer SJ, Mentzer AJ, Auckland K, Robson K, Attenborough R, Alpers MP, Koki G, Pomat W, Siba P, Xue Y, Sandhu MS, Tyler-Smith C. 2017. A Neolithic expansion, but strong genetic structure, in the independent history of New Guinea. Science 357:1160–1163.

Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Rhodes M, Reich DE, Hirschhorn JN. 2004. Genetic Signatures of Strong Recent Positive Selection at the Lactase Gene. The American Journal of Human Genetics 74:1111–1120.

Besenbacher S, Liu S, Izarzugaza JMG, Grove J, Belling K, Bork-Jensen J, Huang S, Als TD, Li S, Yadav R, Rubio-García A, Lescai F, Demontis D, Rao J, Ye W, Mailund T, Friborg RM, Pedersen CNS, Xu R, Sun J, Liu H, Wang O, Cheng X, Flores D, Rydza E, Rapacki K, Damm Sørensen J, Chmura P, Westergaard D, Dworzynski P, Sørensen TIA, Lund O, Hansen T, Xu X, Li N, Bolund L, Pedersen O, Eiberg H, Krogh A, Børglum AD, Brunak S, Kristiansen K, Schierup MH, Wang J, Gupta R, Villesen P, Rasmussen S. 2015. Novel variation and de novo mutation rates in population-wide de novo assembled Danish trios. Nat Commun 6:5969.

Bhatia G, Tandon A, Patterson N, Aldrich MC, Ambrosone CB, Amos C, Bandera EV, Berndt SI, Bernstein L, Blot WJ, Bock CH, Caporaso N, Casey G, Deming SL, Diver WR, Gapstur SM, Gillanders EM, Harris CC, Henderson BE, Ingles SA, Isaacs W, De Jager PL, John EM, Kittles RA, Larkin E, McNeill LH, Millikan RC, Murphy A, Neslund-Dudas C, Nyante S, Press MF, Rodriguez-Gil JL, Rybicki BA, Schwartz AG, Signorello LB, Spitz M, Strom SS, Tucker MA, Wiencke JK, Witte JS, Wu X, Yamamura Y, Zanetti KA, Zheng W, Ziegler RG, Chanock SJ, Haiman CA, Reich D, Price AL. 2014. Genome-wide Scan of 29,141 African Americans Finds No Evidence of Directional Selection since Admixture. Am J Hum Genet 95:437–444.

Blench R. 2011. Was there an Austroasiatic Presence in Island Southeast Asia prior to the Austronesian Expansion? Bull Indo-Pac Prehistory Assoc [Internet] 30. Available from: http://journals.lib.washington.edu/index.php/JIPA/article/view/10637

Bobo D, Lipatov M, Rodriguez-Flores JL, Auton A, Henn BM. 2016. False Negatives Are a Significant Feature of Next Generation Sequencing Callsets.

Borucki A. 2011. The Slave Trade to the Río de la Plata, 1777–1812: Trans-Imperial Networks and Atlantic Warfare. Colon Lat Am Rev 20:81–107.

Bowcock AM, Ruiz-Linares A, Tomfohrde J, Minch E, Kidd JR, Cavalli-Sforza LL. 1994. High resolution of human evolutionary trees with polymorphic microsatellites. Nature 368:455–457.

Boyd JL, Skove SL, Rouanet JP, Pilaz L-J, Bepler T, Gordân R, Wray GA, Silver DL. 2015. Human-Chimpanzee Differences in a FZD8 Enhancer Alter Cell-Cycle Dynamics in the Developing Neocortex. Curr Biol 25:772–779.

Brandt DYC, Aguiar VRC, Bitarello BD, Nunes K, Goudet J, Meyer D. 2015. Mapping Bias Overestimates Reference Allele Frequencies at the *HLA* Genes in the 1000 Genomes Project Phase I Data. G3amp58 GenesGenomesGenetics 5:931–941.

Brandvain Y, Wright SI. 2016. The Limits of Natural Selection in a Nonequilibrium World. Trends Genet 32:201–210.

Brest P, Lapaquette P, Souidi M, Lebrigand K, Cesaro A, Vouret-Craviari V, Mari B, Barbry P, Mosnier J-F, Hébuterne X, Harel-Bellan A, Mograbi B, Darfeuille-Michaud A, Hofman P. 2011. A synonymous variant in IRGM alters a binding site for miR-196 and causes deregulation of IRGM-dependent xenophagy in Crohn's disease. Nat Genet 43:242–245.

Bronkhorst J. 2011. The Spread of Sanskrit in Southeast Asia. In: Manguin P-Y, Mani A, Wade G, editors. Early interactions between South and Southeast Asia: reflections on cross-cultural exchange. Nalanda-Sriwijaya series. Singapore : New Delhi: Institute of Southeast Asian Studies ; Manohar India. p 263–275.

Broos S, Malisoux L, Theisen D, van Thienen R, Ramaekers M, Jamart C, Deldicque L, Thomis MA, Francaux M. 2016. Evidence for ACTN3 as a Speed Gene in Isolated Human Muscle Fibers. PLOS ONE 11:e0150594.

Browning BL, Browning SR. 2013. Improving the Accuracy and Efficiency of Identity-by-Descent Detection in Population Data. Genetics 194:459–471.

Browning SR, Browning BL. 2011. Haplotype phasing: existing methods and new developments. Nat Rev Genet 12:703–714.

Browning SR, Browning BL. 2012. Identity by Descent Between Distant Relatives: Detection and Applications. Annu Rev Genet 46:617–633.

Browning SR, Browning BL. 2015. Accurate Non-parametric Estimation of Recent Effective Population Size from Segments of Identity by Descent. Am J Hum Genet 97:404–418.

Browning SR, Browning BL, Daviglus ML, Durazo-Arvizu RA, Schneiderman N, Kaplan RC, Laurie CC. 2018a. Ancestry-specific recent effective population size in the Americas. PLOS Genet 14:e1007385.

Browning SR, Browning BL, Zhou Y, Tucci S, Akey JM. 2018b. Analysis of Human Sequence Data Reveals Two Pulses of Archaic Denisovan Admixture. Cell 173:53-61.e9.

Bruton JD, Aydin J, Yamada T, Shabalina IG, Ivarsson N, Zhang S-J, Wada M, Tavi P, Nedergaard J, Katz A, Westerblad H. 2010. Increased fatigue resistance linked to Ca^{2+} -stimulated mitochondrial biogenesis in muscle fibres of cold-acclimated mice: Increased fatigue resistance in cold-acclimated skeletal muscle fibres. J Physiol 588:4275–4288.

Bryois J, Buil A, Evans DM, Kemp JP, Montgomery SB, Conrad DF, Ho KM, Ring S, Hurles M, Deloukas P, Davey Smith G, Dermitzakis ET. 2014. Cis and Trans Effects of Human Genomic Variants on Gene Expression. PLoS Genet 10:e1004461.

Bulbeck F. 2008. An integrated perspective on the Austronesian Diaspora: the switch from cereal agriculture to maritime foraging in the colonisation of Island Southeast Asia. Aust Archaeol 67:31–52.

Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Todd Hubisz M, Glanowski S, Tanenbaum DM, White TJ, Sninsky JJ, Hernandez RD, Civello D, Adams MD, Cargill M, Clark AG. 2005. Natural selection on protein-coding genes in the human genome. Nature 437:1153–1157.

Calo A. 2014. Ancient trade between India and Indonesia. Science 345:1255–1255.

Calo A, Prasetyo B, Bellwood P, Lankton JW, Gratuze B, Pryce TO, Reinecke A, Leusch V, Schenk H, Wood R, Bawono RA, Gede IDK, Yuliati NLKC, Fenner J, Reepmeyer C, Castillo C, Carter AK. 2015. Sembiran and Pacung on the north coast of Bali: a strategic crossroads for early trans-Asiatic exchange. Antiquity 89:378–396.

Campbell MC, Tishkoff SA. 2008. African Genetic Diversity: Implications for Human Demographic History, Modern Human Origins, and Complex Disease Mapping. Annu Rev Genomics Hum Genet 9:403–433.

Cann HM, de Toma C, Cazes L, Legrand M-F, Morel V, Piouffre L, Bodmer J, Bodmer WF, Bonne-Tamir B, Cambon-Thomsen A, Chen Z, Chu J, Carcassi C, Contu L, Du R, Excoffier L, Ferrara GB, Friedlaender JS, Groot H, Gurwitz D, Jenkins T, Herrera RJ, Huang X, Kidd J, Kidd KK, Langaney A, Lin AA, Mehdi SQ, Parham P, Piazza A, Pistillo MP, Qian Y, Shu Q, Xu J, Zhu S, Weber JL, Greely HT, Feldman MW, Thomas G, Dausset J, Cavalli-Sforza LL. 2002. A human genome diversity cell line panel. Science 296:261–262.

Cann RL, Stoneking M, Wilson AC. 1987. Mitochondrial DNA and human evolution. Nature 325:31–36.

Capra JA, Hubisz MJ, Kostka D, Pollard KS, Siepel A. 2013. A Model-Based Analysis of GC-Biased Gene Conversion in the Human and Chimpanzee Genomes. PLoS Genet 9:e1003684.

Carmi S, Hui KY, Kochav E, Liu X, Xue J, Grady F, Guha S, Upadhyay K, Ben-Avraham D, Mukherjee S, Bowen BM, Thomas T, Vijai J, Cruts M, Froyen G, Lambrechts D, Plaisance S, Van Broeckhoven C, Van Damme P, Van Marck H, Barzilai N, Darvasi A, Offit K, Bressman S, Ozelius LJ, Peter I, Cho JH, Ostrer H, Atzmon G, Clark LN, Lencz T, Pe'er I. 2014. Sequencing an Ashkenazi reference panel supports population-targeted personal genomics and illuminates Jewish and European origins. Nat Commun 5:4835.

Castillo C. 2013. The Archaeobotany of Khao Sam Kaeo and Phu Khao Thong: The Agriculture of Late Prehistoric Southern Thailand.

Cavalli-Sforza LL. 1973. Analytic review: some current problems of human population genetics. Am J Hum Genet 25:82–104.

Chacon-Duque JC, Adhikari K, Fuentes-Guajardo M, Mendoza-Revilla J, Acuna-Alonzo V, Barquera Lozano R, Quinto-Sanchez M, Gomez-Valdes J, Everardo Martinez P, Villamil-Ramirez H, Hunemeier T, Ramallo V, Silva de Cerqueira CC, Hurtado M, Villegas V, Granja V, Villena M, Vasquez R, Llop E, Sandoval JR, Salazar-Granara AA, Parolin M-L, Sandoval K, Penaloza-Espinosa RI, Rangel-Villalobos H, Winkler C, Klitz W, Bravi C, Molina J, Corach D, Barrantes R, Gomes V, Resende C, Gusmao L, Amorim A, Xue Y, Dugoujon J-M, Moral P, Gonzalez-Jose R, Schuler-Faccini L, Salzano FM, Bortolini M-C, Canizales-Quinteros S, Poletti G, Gallo C, Bedoya G, Rothhammer F, Balding D, Hellenthal G, Ruiz-Linares A. 2018. Latin Americans show wide-spread Converso ancestry and the imprint of local Native ancestry on physical appearance. Available from: http://biorxiv.org/lookup/doi/10.1101/252155

Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. 2015a. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience [Internet] 4. Available from: https://academic.oup.com/gigascience/article-lookup/doi/10.1186/s13742-015-0047-8

Chang C-S, Liu H-L, Moncada X, Seelenfreund A, Seelenfreund D, Chung K-F. 2015b. A holistic picture of Austronesian migrations revealed by phylogeography of Pacific paper mulberry. Proc Natl Acad Sci 112:13537–13542.

Charlesworth B. 2009. Fundamental concepts in genetics: Effective population size and patterns of molecular evolution and variation. Nat Rev Genet 10:195–205.

Chaubey G, Endicott P. 2013. The Andaman Islanders in a regional genetic context: reexamining the evidence for an early peopling of the archipelago from South Asia. Hum Biol 85:153–172.

Chaubey G, Metspalu M, Choi Y, Magi R, Romero IG, Soares P, van Oven M, Behar DM, Rootsi S, Hudjashov G, Mallick CB, Karmin M, Nelis M, Parik J, Reddy AG, Metspalu E, van Driem G, Xue Y, Tyler-Smith C, Thangaraj K, Singh L, Remm M, Richards MB, Lahr MM, Kayser M, Villems R, Kivisild T. 2011. Population Genetic Structure in Indian Austroasiatic Speakers: The Role of Landscape Barriers and Sex-Specific Admixture. Mol Biol Evol 28:1013–1024.

Chen GK, Marjoram P, Wall JD. 2008. Fast and flexible simulation of DNA sequence data. Genome Res 19:136–142.

Chen H, Hey J, Chen K. 2015. Inferring Very Recent Population Growth Rate from Population-Scale Sequencing Data: Using a Large-Sample Coalescent Estimator. Mol Biol Evol 32:2996–3011.

Chen R, Shi L, Hakenberg J, Naughton B, Sklar P, Zhang J, Zhou H, Tian L, Prakash O, Lemire M, Sleiman P, Cheng W, Chen W, Shah H, Shen Y, Fromer M, Omberg L, Deardorff MA, Zackai E, Bobe JR, Levin E, Hudson TJ, Groop L, Wang J, Hakonarson H, Wojcicki A, Diaz GA, Edelmann L, Schadt EE, Friend SH. 2016. Analysis of 589,306 genomes identifies individuals resilient to severe Mendelian childhood diseases. Nat Biotechnol 34:531–538.

Chiang CWK, Ralph P, Novembre J. 2016. Conflation of Short Identity-by-Descent Segments Bias Their Inferred Length Distribution. G3amp58 GenesGenomesGenetics 6:1287–1296.

Christiansen FB. 2008. Theories of Population Variation in Genes and Genomes. Princeton: Princeton University Press.

Clark AG, Nielsen R, Signorovitch J, Matise TC, Glanowski S, Heil J, Winn-Deen ES, Holden AL, Lai E. 2003a. Linkage Disequilibrium and Inference of Ancestral Recombination in 538 Single-Nucleotide Polymorphism Clusters across the Human Genome. Am J Hum Genet 73:285–300.

Clark JD, Beyene Y, WoldeGabriel G, Hart WK, Renne PR, Gilbert H, Defleur A, Suwa G, Katoh S, Ludwig KR, Boisserie J-R, Asfaw B, White TD. 2003b. Stratigraphic, chronological and behavioural contexts of Pleistocene Homo sapiens from Middle Awash, Ethiopia. Nature 423:747–752.

Clarkson C, Smith M, Marwick B, Fullagar R, Wallis LA, Faulkner P, Manne T, Hayes E, Roberts RG, Jacobs Z, Carah X, Lowe KM, Matthews J, Florin SA. 2015. The archaeology, chronology and stratigraphy of Madjedbebe (Malakunanja II): A site in northern Australia with early occupation. J Hum Evol 83:46–64.

Clemente FJ, Cardona A, Inchley CE, Peter BM, Jacobs G, Pagani L, Lawson DJ, Antão T, Vicente M, Mitt M, DeGiorgio M, Faltyskova Z, Xue Y, Ayub Q, Szpak M, Mägi R, Eriksson A, Manica A, Raghavan M, Rasmussen M, Rasmussen S, Willerslev E, Vidal-Puig A, Tyler-Smith C, Villems R, Nielsen R, Metspalu M, Malyarchuk B, Derenko M, Kivisild T. 2014. A Selective Sweep on a Deleterious Mutation in CPT1A in Arctic Populations. Am J Hum Genet 95:584–589.

Cohen J, Pertsemlidis A, Kotowski IK, Graham R, Garcia CK, Hobbs HH. 2005. Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9. Nat Genet 37:161–165.

Colonna V, Ayub Q, Chen Y, Pagani L, Luisi P, Pybus M, Garrison E, Xue Y, Tyler-Smith C, The 1000 Genomes Project Consortium. 2014. Human genomic regions with exceptionally high levels of population differentiation identified from 911 whole-genome sequences. Genome Biol 15:R88.

Coffey AJ, Kokocinski F, Calafato MS, Scott CE, Palta P, Drury E, Joyce CJ, LeProust EM, Harrow J, Hunt S, Lehesjoki A-E, Turner DJ, Hubbard TJ, Palotie A. 2011. The GENCODE exome: sequencing the complete human exome. Eur J Hum Genet 19:827–831.

Coop G, Pickrell JK, Novembre J, Kudaravalli S, Li J, Absher D, Myers RM, Cavalli-Sforza LL, Feldman MW, Pritchard JK. 2009. The Role of Geography in Human Adaptation. PLoS Genet 5:e1000500.

Coop G, Witonsky D, Di Rienzo A, Pritchard JK. 2010. Using Environmental Correlations to Identify Loci Underlying Local Adaptation. Genetics 185:1411–1423.

Cooper DN, Krawczak M. 1990. The mutational spectrum of single base-pair substitutions causing human genetic disease: patterns and predictions. Hum Genet 85:55–74.

Cooper DN, Krawczak M, Polychronakos C, Tyler-Smith C, Kehrer-Sawatzki H. 2013. Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease. Hum Genet 132:1077–1130.

Cooper GM. 2005. Distribution and intensity of constraint in mammalian genomic sequence. Genome Res 15:901–913.

Corach D, Lao O, Bobillo C, Van Der Gaag K, Zuniga S, Vermeulen M, Van Duijn K, Goedbloed M, Vallone PM, Parson W, De Knijff P, Kayser M. 2010. Inferring Continental Ancestry of Argentineans from Autosomal, Y-Chromosomal and Mitochondrial DNA: Genetic Ancestry in Extant Argentineans. Ann Hum Genet 74:65–76.

Corvol H, Blackman SM, Boëlle P-Y, Gallins PJ, Pace RG, Stonebraker JR, Accurso FJ, Clement A, Collaco JM, Dang H, Dang AT, Franca A, Gong J, Guillot L, Keenan K, Li W, Lin F, Patrone MV, Raraigh KS, Sun L, Zhou Y-H, O'Neal WK, Sontag MK, Levy H, Durie PR, Rommens JM, Drumm ML, Wright FA, Strug LJ, Cutting GR, Knowles MR. 2015. Genome-wide association meta-analysis identifies five modifier loci of lung disease severity in cystic fibrosis. Nat Commun 6:8382.

Cottrol RJ. 2007. Beyond Invisibility: Afro-Argentines in Their Nation's Culture and Memory. Lat Am Res Rev 42:139–156.

Coventry A, Bull-Otterson LM, Liu X, Clark AG, Maxwell TJ, Crosby J, Hixson JE, Rea TJ, Muzny DM, Lewis LR, Wheeler DA, Sabo A, Lusk C, Weiss KG, Akbar H, Cree A, Hawes AC, Newsham I, Varghese RT, Villasana D, Gross S, Joshi V, Santibanez J, Morgan M, Chang K, Iv WH, Templeton AR, Boerwinkle E, Gibbs R, Sing CF. 2010. Deep resequencing reveals excess rare recent variants consistent with explosive population growth. Nat Commun 1:131.

Crawford NG, Kelly DE, Hansen MEB, Beltrame MH, Fan S, Bowman SL, Jewett E, Ranciaro A, Thompson S, Lo Y, Pfeifer SP, Jensen JD, Campbell MC, Beggs W, Hormozdiari F, Mpoloka SW, Mokone GG, Nyambo T, Wolde Meskel D, Belay G, Haut J, Rothschild H, Zon L, Zhou Y, Kovacs MA, Xu M, Zhang T, Bishop K, Sinclair J, Rivas C, Elliot E, Choi J, Li SA, Hicks B, Burgess S, Abnet C, Watkins-Chow DE, Oceana E, Song YS, Eskin E, Brown KM, Marks MS, Loftus SK, Pavan WJ, Yeager M, Chanock S, Tishkoff S. 2017. Loci associated with skin pigmentation identified in African populations. Science:eaan8433.

Crisci JL, Poh Y-P, Mahajan S, Jensen JD. 2013. The impact of equilibrium assumptions on tests of selection. Front Genet [Internet] 4. Available from: http://journal.frontiersin.org/article/10.3389/fgene.2013.00235/abstract

Crow JF. 1958. Some Possibilities for Measuring Selection Intensities in Man. Hum Biol 30:1–13.

Crow JF, Kimura M. 1970. An introduction to population genetics theory. New York: Harper and Row.

Dal GM, Ergüner B, Sağıroğlu MS, Yüksel B, Onat OE, Alkan C, Özçelik T. 2014. Early postzygotic mutations contribute to de novo variation in a healthy monozygotic twin pair. J Med Genet 51:455–459.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R, 1000 Genomes Project Analysis Group. 2011. The variant call format and VCFtools. Bioinformatics 27:2156–2158.

Davidson I. 2010. The Colonization of Australia and Its Adjacent Islands and the Evolution of Modern Cognition. Curr Anthropol 51:S177–S189.

Dean L. 2005. Chapter 10, The Kidd blood group. In: Blood Groups and Red Cell Antigens [Internet]. Bethesda (MD): National Center for Biotechnology Information (US). Available from: https://www.ncbi.nlm.nih.gov/books/NBK2272/

DeGiorgio M, Lohmueller KE, Nielsen R. 2014. A Model-Based Approach for Identifying Signatures of Ancient Balancing Selection in Genetic Data. PLoS Genet 10:e1004561.

Delaneau O, Howie B, Cox AJ, Zagury J-F, Marchini J. 2013a. Haplotype Estimation Using Sequencing Reads. Am J Hum Genet 93:687–696.

Delaneau O, Zagury J-F, Marchini J. 2013b. Improved whole-chromosome phasing for disease and population genetic studies. Nat Methods 10:5–6.

Delfin F, Min-Shan Ko A, Li M, Gunnarsdóttir ED, Tabbada KA, Salvador JM, Calacal GC, Sagum MS, Datar FA, Padilla SG, De Ungria MCA, Stoneking M. 2014. Complete mtDNA genomes of Filipino ethnolinguistic groups: a melting pot of recent and ancient lineages in the Asia-Pacific region. Eur J Hum Genet 22:228–237.

Demeter F, Shackelford LL, Bacon A-M, Duringer P, Westaway K, Sayavongkhamdy T, Braga J, Sichanthongtip P, Khamdalavong P, Ponche J-L, Wang H, Lundstrom C, Patole-Edoumba E, Karpoff A-M. 2012. Anatomically modern human in Southeast Asia (Laos) by 46 ka. Proc Natl Acad Sci 109:14375–14380.

Denham T. 2011. Early Agriculture and Plant Domestication in New Guinea and Island Southeast Asia. Curr Anthropol 52:S379–S395.

Deshpande O, Batzoglou S, Feldman MW, Luca Cavalli-Sforza L. 2009. A serial founder effect model for human settlement out of Africa. Proc R Soc B Biol Sci 276:291–300.

Diamond J. 2002. Evolution, consequences and future of plant and animal domestication. Nature 418:700–707.

Do R, Balick D, Li H, Adzhubei I, Sunyaev S, Reich D. 2015. No evidence that selection has been less effective at removing deleterious mutations in Europeans than in Africans. Nat Genet 47:126–131.

van Dorp L, Balding D, Myers S, Pagani L, Tyler-Smith C, Bekele E, Tarekegn A, Thomas MG, Bradman N, Hellenthal G. 2015. Evidence for a Common Origin of Blacksmiths and Cultivators in the Ethiopian Ari within the Last 4500 Years: Lessons for Clustering-Based Inference. PLOS Genet 11:e1005397.

Drake JW, Charlesworth B, Charlesworth D, Crow JF. 1998. Rates of Spontaneous Mutation. Genetics 148:1667–1686.

Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, Kermani BG, Carnevali P, Nazarenko I, Nilsen GB, Yeung G, Dahl F, Fernandez A, Staker B, Pant KP, Baccash J, Borcherding AP, Brownley A, Cedeno R, Chen L, Chernikoff D, Cheung A, Chirita R, Curson B, Ebert JC, Hacker CR, Hartlage R, Hauser B, Huang S, Jiang Y, Karpinchyk V, Koenig M, Kong C, Landers T, Le C, Liu J, McBride CE, Morenzoni M, Morey RE, Mutch K, Perazich H, Perry K, Peters BA, Peterson J, Pethiyagoda CL, Pothuraju K, Richter C, Rosenbaum AM, Roy S, Shafto J, Sharanhovich U, Shannon KW, Sheppy CG, Sun M, Thakuria JV, Tran A, Vu

D, Zaranek AW, Wu X, Drmanac S, Oliphant AR, Banyai WC, Martin B, Ballinger DG, Church GM, Reid CA. 2010. Human Genome Sequencing Using Unchained Base Reads on Self-Assembling DNA Nanoarrays. Science 327:78–81.

Duffy D. 2015. Genetics of Eye Colour. In: John Wiley & Sons Ltd, editor. eLS. Chichester, UK: John Wiley & Sons, Ltd. p 1–9. Available from: http://doi.wiley.com/10.1002/9780470015902.a0024646

Durand EY, Eriksson N, McLean CY. 2014. Reducing Pervasive False-Positive Identical-by-Descent Segments Detected by Large-Scale Pedigree Analysis. Mol Biol Evol 31:2212–2222.

Durand JD. 1960. The population statistics of China, A.D. 2–1953. Popul Stud 13:209–256.

Duret L, Galtier N. 2009. Biased Gene Conversion and the Evolution of Mammalian Genomic Landscapes. Annu Rev Genomics Hum Genet 10:285–311.

Eichstaedt CA, Antão T, Cardona A, Pagani L, Kivisild T, Mormina M. 2015. Genetic and phenotypic differentiation of an Andean intermediate altitude population. Physiol Rep 3:e12376.

Eilbeck K, Lewis SE, Mungall CJ, Yandell M, Stein L, Durbin R, Ashburner M. 2005. The Sequence Ontology: a tool for the unification of genome annotations. Genome Biol 6:R44.

Emond MJ, Louie T, Emerson J, Zhao W, Mathias RA, Knowles MR, Wright FA, Rieder MJ, Tabor HK, Nickerson DA, Barnes KC, Gibson RL, Bamshad MJ. 2012. Exome sequencing of extreme phenotypes identifies DCTN4 as a modifier of chronic Pseudomonas aeruginosa infection in cystic fibrosis. Nat Genet 44:886–889.

Erickson RP, Mitchison NA. 2014. The low frequency of recessive disease: insights from ENU mutagenesis, severity of disease phenotype, GWAS associations, and demography: an analytical review. J Appl Genet 55:319–327.

Eriksson A, Betti L, Friend AD, Lycett SJ, Singarayer JS, von Cramon-Taubadel N, Valdes PJ, Balloux F, Manica A. 2012. Late Pleistocene climate change and the global expansion of anatomically modern humans. Proc Natl Acad Sci 109:16089–16094.

Eriksson A, Manica A. 2012. Effect of ancient population structure on the degree of polymorphism shared between modern human populations and ancient hominins. Proc Natl Acad Sci 109:13956–13960.

Excoffier L, Hofer T, Foll M. 2009. Detecting loci under selection in a hierarchically structured population. Heredity 103:285–298.

Eynon N, Hanson ED, Lucia A, Houweling PJ, Garton F, North KN, Bishop DJ. 2013. Genes for Elite Power and Sprint Performance: ACTN3 Leads the Way. Sports Med 43:803–817.

Fagny M, Patin E, Enard D, Barreiro LB, Quintana-Murci L, Laval G. 2014. Exploring the Occurrence of Classic Selective Sweeps in Humans Using Whole-Genome Sequencing Data Sets. Mol Biol Evol 31:1850–1868.

Fan S, Hansen MEB, Lo Y, Tishkoff SA. 2016. Going global by adapting local: A review of recent human adaptation. Science 354:54–59.

Farh KK-H, Marson A, Zhu J, Kleinewietfeld M, Housley WJ, Beik S, Shoresh N, Whitton H, Ryan RJH, Shishkin AA, Hatan M, Carrasco-Alfonso MJ, Mayer D, Luckey CJ, Patsopoulos NA, De Jager PL, Kuchroo VK, Epstein CB, Daly MJ, Hafler DA, Bernstein BE. 2014. Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature 518:337–343.

Fedorova L, Qiu S, Dutta R, Fedorov A. 2016. Atlas of Cryptic Genetic Relatedness Among 1000 Human Genomes. Genome Biol Evol 8:777–790.

Felsenstein J. 1975. A Pain in the Torus: Some Difficulties with Models of Isolation by Distance. Am Nat 109:359–368.

Fenner JN. 2005. Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. Am J Phys Anthropol 128:415–423.

Fernandes V, Alshamali F, Alves M, Costa MD, Pereira JB, Silva NM, Cherni L, Harich N, Cerny V, Soares P, Richards MB, Pereira L. 2012. The Arabian Cradle: Mitochondrial Relicts of the First Steps along the Southern Route out of Africa. Am J Hum Genet 90:347–355.

Field Y, Boyle EA, Telis N, Gao Z, Gaulton KJ, Golan D, Yengo L, Rocheleau G, Froguel P, McCarthy MI, Pritchard JK. 2016. Detection of human adaptation during the past 2000 years. Science 354:760–764.

Finney LA, O'Halloran TV. 2003. Transition Metal Speciation in the Cell: Insights from the Chemistry of Metal Ion Receptors. Science 300:931–936.

Firnberg E, Labonte JW, Gray JJ, Ostermeier M. 2014. A Comprehensive, High-Resolution Map of a Gene's Fitness Landscape. Mol Biol Evol 31:1581–1592.

Fisher RA. 1915. Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population. Biometrika 10:507.

Fisher RA. 1930. The genetical theory of natural selection. Oxford University Press.

Flegontov P, Altinisik NE, Changmai P, Rohland N, Mallick S, Bolnick DA, Candilio F, Flegontova O, Jeong C, Harper TK, Keating D, Kennett DJ, Kim AM, Lamnidis TC, Olalde I, Raff J, Sattler RA, Skoglund P, Vajda EJ, Vasilyev S, Veselovskaya E, Hayes MG, O'Rourke DH, Pinhasi R, Krause J, Reich D, Schiffels S. 2017. Paleo-Eskimo genetic legacy across North America. bioRxiv:203018.

Fort J. 2015. Demic and cultural diffusion propagated the Neolithic transition across different regions of Europe. J R Soc Interface 12:20150166–20150166.

Foster F, Collard M. 2013. A Reassessment of Bergmann's Rule in Modern Humans. PLoS ONE 8:e72269.

Franchini LF, Pollard KS. 2015. Can a few non-coding mutations make a human brain? BioEssays 37:1054–1061.

Fraser HB. 2013. Gene expression drives local adaptation in humans. Genome Res 23:1089–1096.

Fraser HB, Moses AM, Schadt EE. 2010. Evidence for widespread adaptive evolution of gene expression in budding yeast. Proc Natl Acad Sci 107:2977–2982.

Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, Pasternak S, Wheeler DA, Willis TD, Yu F, Yang H, Zeng C, Gao Y, Hu H, Hu W, Li C, Lin W, Liu S, Pan H, Tang X, Wang J, Wang W, Yu J, Zhang B, Zhang Q, Zhao H, Zhao H, Zhou J, Gabriel SB, Barry R, Blumenstiel B, Camargo A, Defelice M, Faggart M, Goyette M, Gupta S, Moore J, Nguyen H, Onofrio RC, Parkin M, Roy J, Stahl E, Winchester E, Ziaugra L, Altshuler D, Shen Y, Yao Z, Huang W, Chu X, He Y, Jin L, Liu Y, Shen Y, Sun W, Wang H, Wang Y, Wang Y, Xiong X, Xu L, Waye MMY, Tsui SKW, Xue H, Wong JT-F, Galver LM, Fan J-B, Gunderson K, Murray SS, Oliphant AR, Chee MS, Montpetit A, Chagnon F, Ferretti V, Leboeuf M, Olivier J-F, Phillips MS, Roumy S, Sallée C, Verner A, Hudson TJ, Kwok P-Y, Cai D, Koboldt DC, Miller RD, Pawlikowska L, Taillon-Miller P, Xiao M, Tsui L-C, Mak W, Qiang Song Y, Tam PKH, Nakamura Y, Kawaguchi T, Kitamoto T, Morizono T, et al. 2007. A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851–861.

Friedlander SM, Herrmann AL, Lowry DP, Mepham ER, Lek M, North KN, Organ CL. 2013. ACTN3 Allele Frequency in Humans Covaries with Global Latitudinal Gradient. PLoS ONE 8:e52282.

Fry JD. 2004. On the rate and linearity of viability declines in Drosophila mutation-accumulation experiments: genomic mutation rates and synergistic epistasis revisited. Genetics 166:797–806.

Fu Q, Hajdinjak M, Moldovan OT, Constantin S, Mallick S, Skoglund P, Patterson N, Rohland N, Lazaridis I, Nickel B, Viola B, Prüfer K, Meyer M, Kelso J, Reich D, Pääbo S. 2015. An early modern human from Romania with a recent Neanderthal ancestor. Nature 524:216–219.

Fu Q, Li H, Moorjani P, Jay F, Slepchenko SM, Bondarev AA, Johnson PLF, Aximu-Petri A, Prüfer K, de Filippo C, Meyer M, Zwyns N, Salazar-García DC, Kuzmin YV, Keates SG, Kosintsev PA, Razhev DI, Richards MP, Peristov NV, Lachmann M, Douka K, Higham TFG, Slatkin M, Hublin J-J, Reich D, Kelso J, Viola TB, Pääbo S. 2014a. Genome sequence of a 45,000-year-old modern human from western Siberia. Nature 514:445–449.

Fu Q, Meyer M, Gao X, Stenzel U, Burbano HA, Kelso J, Paabo S. 2013. DNA analysis of an early modern human from Tianyuan Cave, China. Proc Natl Acad Sci 110:2223–2227.

Fu W, Akey JM. 2013. Selection and Adaptation in the Human Genome. Annu Rev Genomics Hum Genet 14:467–489.

Fu W, Gittelman RM, Bamshad MJ, Akey JM. 2014b. Characteristics of Neutral and Deleterious Protein-Coding Variation among Individuals and Populations. Am J Hum Genet 95:421–436.

Fumagalli M, Pozzoli U, Cagliani R, Comi GP, Bresolin N, Clerici M, Sironi M. 2010. The landscape of human genes involved in the immune response to parasitic worms. BMC Evol Biol 10:264.

Fumagalli M, Sironi M, Pozzoli U, Ferrer-Admettla A, Pattini L, Nielsen R. 2011. Signatures of Environmental Genetic Adaptation Pinpoint Pathogens as the Main Selective Pressure through Human Evolution. PLoS Genet 7:e1002355.

Gamsjaeger R, Liew C, Loughlin F, Crossley M, Mackay J. 2007. Sticky fingers: zinc-fingers as protein-recognition motifs. Trends Biochem Sci 32:63–70.

Gao F, Keinan A. 2016. Explosive genetic evidence for explosive human population growth. Curr Opin Genet Dev 41:130–139.

Gao Z, Waggoner D, Stephens M, Ober C, Przeworski M. 2015. An Estimate of the Average Number of Recessive Lethal Mutations Carried by Humans. Genetics 199:1243–1254.

Genome of the Netherlands Consortium. 2014. Whole-genome sequence variation, population structure and demographic history of the Dutch population. Nat Genet 46:818–825.

Genovese G, Handsaker RE, Li H, Kenny EE, McCarroll SA. 2013. Mapping the Human Reference Genome's Missing Sequence by Three-Way Admixture in Latino Genomes. Am J Hum Genet 93:411–421.

Glémin S, Arndt PF, Messer PW, Petrov D, Galtier N, Duret L. 2015. Quantification of GC-biased gene conversion in the human genome. Genome Res 25:1215–1228.

Gonda J. 1973. Sanskrit in Indonesia. New Dehli: International Academy of Indian Culture.

Goodwin S, McPherson JD, McCombie WR. 2016. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet 17:333–351.

Grasberger H, Refetoff S. 2006. Identification of the Maturation Factor for Dual Oxidase: *EVOLUTION OF AN EUKARYOTIC OPERON EQUIVALENT*. J Biol Chem 281:18269–18272.

Gravel S. 2016. When Is Selection Effective? Genetics 203:451–462.

Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, Clark AG, Yu F, Gibbs RA, The 1000 Genomes Project, Bustamante CD, Altshuler DL, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, Collins FS, De La Vega FM, Donnelly P, Egholm M, Flicek P, Gabriel SB, Gibbs RA, Knoppers BM, Lander ES, Lehrach H, Mardis ER, McVean GA, Nickerson DA, Peltonen L, Schafer AJ, Sherry ST, Wang J, Wilson RK, Gibbs RA, Deiros D, Metzker M, Muzny D, Reid J, Wheeler D, Wang J, Li J, Jian M, Li G, Li R, Liang H, Tian G, Wang B, Wang J, Wang W, Yang H, Zhang X, Zheng H, Lander ES, Altshuler DL, Ambrogio L, Bloom T, Cibulskis K, Fennell TJ, Gabriel SB, Jaffe DB, Shefler E, Sougnez CL, Bentley DR, Gormley N, Humphray S, Kingsbury Z, Koko-Gonzales P, Stone J, McKernan KJ, Costa GL, Ichikawa JK, Lee CC, Sudbrak R, Lehrach H, Borodina TA, Dahl A, Davydov AN, Marquardt P, Mertes F, Nietfeld W, Rosenstiel P, Schreiber S, Soldatov AV, Timmermann B, Tolzmann M, Egholm M, Affourtit J, Ashworth D, Attiya S, Bachorski M, Buglione E, Burke A, Caprio A, Celone C, Clark S, Conners D, Desany B, et al. 2011. Demographic history and rare allele sharing among human populations. Proc Natl Acad Sci 108:11983–11988.

Gray RD, Drummond AJ, Greenhill SJ. 2009. Language Phylogenies Reveal Expansion Pulses and Pauses in Pacific Settlement. Science 323:479–483.

Greaves M. 2014. Was skin cancer a selective force for black pigmentation in early hominin evolution? Proc R Soc B Biol Sci 281:20132955–20132955.

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MHY, Hansen NF, Durand EY, Malaspinas AS, Jensen JD, Marques-Bonet T, Alkan C, Prufer K, Meyer M, Burbano HA, Good JM, Schultz R, Aximu-Petri A, Butthof A, Hober B, Hoffner B, Siegemund M, Weihmann A, Nusbaum C, Lander ES, Russ C, Novod N, Affourtit J, Egholm M, Verna C, Rudan P, Brajkovic D, Kucan Z, Gusic I, Doronichev VB, Golovanova LV, Lalueza-Fox C, de la Rasilla M, Fortea J, Rosas A, Schmitz RW, Johnson PLF, Eichler EE, Falush D, Birney E, Mullikin JC, Slatkin M, Nielsen R, Kelso J, Lachmann M, Reich D, Paabo S. 2010. A Draft Sequence of the Neandertal Genome. Science 328:710–722.

Groman JD, Meyer ME, Wilmott RW, Zeitlin PL, Cutting GR. 2002. Variant cystic fibrosis phenotypes in the absence of CFTR mutations. N Engl J Med 347:401–407.

Gronau I, Hubisz MJ, Gulko B, Danko CG, Siepel A. 2011. Bayesian inference of ancient human demography from individual genome sequences. Nat Genet 43:1031–1034.

Grossman SR, Andersen KG, Shlyakhter I, Tabrizi S, Winnicki S, Yen A, Park DJ, Griesemer D, Karlsson EK, Wong SH, Cabili M, Adegbola RA, Bamezai RNK, Hill AVS, Vannberg FO, Rinn JL, Lander ES, Schaffner SF, Sabeti PC. 2013. Identifying Recent Adaptations in Large-Scale Genomic Data. Cell 152:703–713.

Grossman SR, Shylakhter I, Karlsson EK, Byrne EH, Morales S, Frieden G, Hostetter E, Angelino E, Garber M, Zuk O, Lander ES, Schaffner SF, Sabeti PC. 2010. A Composite of Multiple Signals Distinguishes Causal Variants in Regions of Positive Selection. Science 327:883–886.

Groucutt HS, Petraglia MD, Bailey G, Scerri EML, Parton A, Clark-Balzan L, Jennings RP, Lewis L, Blinkhorn J, Drake NA, Breeze PS, Inglis RH, Devès MH, Meredith-Williams M, Boivin N, Thomas MG, Scally A. 2015. Rethinking the dispersal of Homo sapiens out of Africa. Evol Anthropol 24:149–164.

Grün R, Stringer C, McDermott F, Nathan R, Porat N, Robertson S, Taylor L, Mortimer G, Eggins S, McCulloch M. 2005. U-series and ESR analyses of bones and teeth relating to the human burials from Skhul. J Hum Evol 49:316–334.

GTEx Consortium, Ardlie KG, Deluca DS, Segre AV, Sullivan TJ, Young TR, Gelfand ET, Trowbridge CA, Maller JB, Tukiainen T, Lek M, Ward LD, Kheradpour P, Iriarte B, Meng Y, Palmer CD, Esko T, Winckler W, Hirschhorn JN, Kellis M, MacArthur DG, Getz G, Shabalin AA, Li G, Zhou Y-H, Nobel AB, Rusyn I, Wright FA, Lappalainen T, Ferreira PG, Ongen H, Rivas MA, Battle A, Mostafavi S, Monlong J, Sammeth M, Mele M, Reverter F, Goldmann JM, Koller D, Guigo R, McCarthy MI, Dermitzakis ET, Gamazon ER, Im HK, Konkashbaev A, Nicolae DL, Cox NJ, Flutre T, Wen X, Stephens M, Pritchard JK, Tu Z, Zhang B, Huang T, Long Q, Lin L, Yang J, Zhu J, Liu J, Brown A, Mestichelli B, Tidwell D, Lo E, Salvatore M, Shad S, Thomas JA, Lonsdale JT, Moser MT, Gillard BM, Karasik E, Ramsey K, Choi C, Foster BA, Syron J, Fleming J, Magazine H, Hasz R, Walters GD, Bridge JP, Miklos M, Sullivan S, Barker LK, Traino HM, Mosavel M, Siminoff LA, Valley DR, Rohrer DC, Jewell SD, Branton PA, Sobin LH, Barcus M, Qi L, McLean J, Hariharan P, Um KS, Wu S, Tabor D,

et al. 2015. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science 348:648–660.

Gudbjartsson DF, Helgason H, Gudjonsson SA, Zink F, Oddson A, Gylfason A, Besenbacher S, Magnusson G, Halldorsson BV, Hjartarson E, Sigurdsson GT, Stacey SN, Frigge ML, Holm H, Saemundsdottir J, Helgadottir HT, Johannsdottir H, Sigfusson G, Thorgeirsson G, Sverrisson JT, Gretarsdottir S, Walters GB, Rafnar T, Thjodleifsson B, Bjornsson ES, Olafsson S, Thorarinsdottir H, Steingrimsdottir T, Gudmundsdottir TS, Theodors A, Jonasson JG, Sigurdsson A, Bjornsdottir G, Jonsson JJ, Thorarensen O, Ludvigsson P, Gudbjartsson H, Eyjolfsson GI, Sigurdardottir O, Olafsson I, Arnar DO, Magnusson OT, Kong A, Masson G, Thorsteinsdottir U, Helgason A, Sulem P, Stefansson K. 2015. Large-scale whole-genome sequencing of the Icelandic population. Nat Genet 47:435–444.

Gurdasani D, Carstensen T, Tekola-Ayele F, Pagani L, Tachmazidou I, Hatzikotoulas K, Karthikeyan S, Iles L, Pollard MO, Choudhury A, Ritchie GRS, Xue Y, Asimit J, Nsubuga RN, Young EH, Pomilla C, Kivinen K, Rockett K, Kamali A, Doumatey AP, Asiki G, Seeley J, Sisay-Joof F, Jallow M, Tollman S, Mekonnen E, Ekong R, Oljira T, Bradman N, Bojang K, Ramsay M, Adeyemo A, Bekele E, Motala A, Norris SA, Pirie F, Kaleebu P, Kwiatkowski D, Tyler-Smith C, Rotimi C, Zeggini E, Sandhu MS. 2015. The African Genome Variation Project shapes medical genetics in Africa. Nature 517:327–332.

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. 2009. Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. PLoS Genet 5:e1000695.

Guy J. 2011. Tamil merchants and the Hindu-Buddhist Diaspora in early Southeast Asia. In: Manguin P-Y, Mani A, Wade G, editors. Early interactions between South and Southeast Asia: reflections on cross-cultural exchange. Singapore : New Delhi: Institute of Southeast Asian Studies ; Manohar India. p 243–262.

Haak W, Lazaridis I, Patterson N, Rohland N, Mallick S, Llamas B, Brandt G, Nordenfelt S, Harney E, Stewardson K, Fu Q, Mittnik A, Bánffy E, Economou C, Francken M, Friederich S, Pena RG, Hallgren F, Khartanovich V, Khokhlov A, Kunst M, Kuznetsov P, Meller H, Mochalov O, Moiseyev V, Nicklisch N, Pichler SL, Risch R, Rojo Guerra MA, Roth C, Szécsényi-Nagy A, Wahl J, Meyer M, Krause J, Brown D, Anthony D, Cooper A, Alt KW, Reich D. 2015. Massive migration from the steppe was a source for Indo-European languages in Europe. Nature 522:207–211.

Haber M, Mezzavilla M, Xue Y, Tyler-Smith C. 2016. Ancient DNA and the rewriting of human history: be sparing with Occam's razor. Genome Biol [Internet] 17. Available from: http://genomebiology.com/2016/17/1/1

Hajdinjak M, Fu Q, Hübner A, Petr M, Mafessoni F, Grote S, Skoglund P, Narasimham V, Rougier H, Crevecoeur I, Semal P, Soressi M, Talamo S, Hublin J-J, Gušić I, Kućan Ž, Rudan P, Golovanova LV, Doronichev VB, Posth C, Krause J, Korlević P, Nagel S, Nickel B, Slatkin M, Patterson N, Reich D, Prüfer K, Meyer M, Pääbo S, Kelso J. 2018. Reconstructing the genetic history of late Neanderthals. Nature 555:652–656.

Haldane JBS. 1937. The Effect of Variation on Fitness. Am Nat 71:337–349.

Haldane JBS. 1949. The rate of mutation of human genes. Hereditas 35:267–273.

Hamblin MT, Thompson EE, Di Rienzo A. 2002. Complex Signatures of Natural Selection at the Duffy Blood Group Locus. Am J Hum Genet 70:369–383

Harding RM, Healy E, Ray AJ, Ellis NS, Flanagan N, Todd C, Dixon C, Sajantila A, Jackson IJ, Birch-Machin MA, Rees JL. 2000. Evidence for Variable Selective Pressures at MC1R. Am J Hum Genet 66:1351–1361.

Harding RM, McVean G. 2004. A structured ancestral population for the evolution of modern humans. Curr Opin Genet Dev 14:667–674.

Harpending H. 1971. Inference in population structure studies. Am J Hum Genet 23:536–538.

Harris DN, Song W, Shetty AC, Levano KS, Cáceres O, Padilla C, Borda V, Tarazona D, Trujillo O, Sanchez C, Kessler MD, Galarza M, Capristano S, Montejo H, Flores-Villanueva PO, Tarazona-Santos E, O'Connor TD, Guio H. 2018. Evolutionary genomic dynamics of Peruvians before, during, and after the Inca Empire. Proc Natl Acad Sci 115:E6526–E6535.

Harris K. 2015. Evidence for recent, population-specific evolution of the human mutation rate. Proc Natl Acad Sci 112:3439–3444.

Harris K, Nielsen R. 2013. Inferring Demographic History from a Spectrum of Shared Haplotype Lengths. PLoS Genet 9:e1003521.

Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, Barnes I, Bignell A, Boychenko V, Hunt T, Kay M, Mukherjee G, Rajan J, Despacio-Reyes G, Saunders G, Steward C, Harte R, Lin M, Howald C, Tanzer A, Derrien T, Chrast J, Walters N, Balasubramanian S, Pei B, Tress M, Rodriguez JM, Ezkurdia I, van Baren J, Brent M, Haussler D, Kellis M, Valencia A, Reymond A, Gerstein M, Guigo R, Hubbard TJ. 2012. GENCODE: The reference human genome annotation for The ENCODE Project. Genome Res 22:1760–1774.

Head SI, Chan S, Houweling PJ, Quinlan KGR, Murphy R, Wagner S, Friedrich O, North KN. 2015. Altered Ca2+ Kinetics Associated with α-Actinin-3 Deficiency May Explain Positive Selection for ACTN3 Null Allele in Human Evolution. PLOS Genet 11:e1004862.

Hellenthal G, Busby GBJ, Band G, Wilson JF, Capelli C, Falush D, Myers S. 2014. A Genetic Atlas of Human Admixture History. Science 343:747–751.

Henn BM, Botigué LR, Bustamante CD, Clark AG, Gravel S. 2015. Estimating the mutation load in human genomes. Nat Rev Genet 16:333–343.

Henn BM, Botigué LR, Peischl S, Dupanloup I, Lipatov M, Maples BK, Martin AR, Musharoff S, Cann H, Snyder MP, Excoffier L, Kidd JM, Bustamante CD. 2016. Distance from sub-Saharan Africa predicts mutational load in diverse human genomes. Proc Natl Acad Sci 113:E440–E449.

Henn BM, Cavalli-Sforza LL, Feldman MW. 2012a. The great human expansion. Proc Natl Acad Sci U S A 109:17758–17764.

Henn BM, Hon L, Macpherson JM, Eriksson N, Saxonov S, Pe'er I, Mountain JL. 2012b. Cryptic Distant Relatives Are Common in Both Isolated and Cosmopolitan Genetic Samples. PLoS ONE 7:e34267.

Hernandez RD, Kelley JL, Elyashiv E, Melton SC, Auton A, McVean G, Sella G, Przeworski M. 2011. Classic Selective Sweeps Were Rare in Recent Human Evolution. Science 331:920–924.

Hochreiter S. 2013. HapFABIA: Identification of very short segments of identity by descent characterized by rare variants in large sequencing data. Nucleic Acids Res 41:e202–e202.

Hofreiter M, Paijmans JLA, Goodchild H, Speller CF, Barlow A, Fortes GG, Thomas JA, Ludwig A, Collins MJ. 2015. The future of ancient DNA: Technical advances and conceptual shifts: Prospects & Overviews. BioEssays 37:284–293.

Holm S. 1979. A Simple Sequentially Rejective Multiple Test Procedure. Scand J Stat 6:65–70.

Hoogervorst T. 2015. Detecting pre-modern lexical influence from South India in Maritime Southeast Asia. Archipel 89:63–93.

Horton J, Cohen J, Hobbs H. 2007. Molecular biology of PCSK9: its role in LDL metabolism. Trends Biochem Sci 32:71–77.

Horton R, Wilming L, Rand V, Lovering RC, Bruford EA, Khodiyar VK, Lush MJ, Povey S, Talbot CC, Wright MW, Wain HM, Trowsdale J, Ziegler A, Beck S. 2004. Gene map of the extended human MHC. Nat Rev Genet 5:889–899.

Housley RA, Gamble CS, Street M, Pettitt P. 1997. Radiocarbon evidence for the Lateglacial Human Recolonisation of Northern Europe. Proc Prehist Soc 63:25–54.

Hovers E. 2009. The lithic assemblages of Qafzeh Cave. Oxford ; New York: Oxford University Press.

Hsieh P, Veeramah KR, Lachance J, Tishkoff SA, Wall JD, Hammer MF, Gutenkunst RN. 2016. Whole-genome sequence analyses of Western Central African Pygmy hunter-gatherers reveal a complex demographic history and identify candidate genes under positive natural selection. Genome Res 26:279–290.

Hu J, Ng PC. 2012. Predicting the effects of frameshifting indels. Genome Biol 13:R9.

Hu J, Ng PC. 2013. SIFT Indel: Predictions for the Functional Effects of Amino Acid Insertions/Deletions in Proteins. PLoS ONE 8:e77940.

Huber CD, Durvasula A, Hancock AM, Lohmueller KE. 2017. Gene expression drives the evolution of dominance. bioRxiv [Internet]. Available from: http://biorxiv.org/content/early/2017/08/31/182865.abstract

Hublin J-J, Ben-Ncer A, Bailey SE, Freidline SE, Neubauer S, Skinner MM, Bergmann I, Le Cabec A, Benazzi S, Harvati K, Gunz P. 2017. New fossils from Jebel Irhoud, Morocco and the pan-African origin of Homo sapiens. Nature 546:289–292.

Hudjashov G, Endicott P, Post H, Nagle N, Ho SYW, Lawson DJ, Reidla M, Karmin M, Rootsi S, Metspalu E, Saag L, Villems R, Cox MP, Mitchell RJ, Garcia-Bertrand RL, Metspalu M, Herrera RJ. 2018. Investigating the origins of eastern Polynesians using genome-wide data from the Leeward Society Isles. Sci Rep [Internet] 8. Available from: http://www.nature.com/articles/s41598-018-20026-8

Hudson RR, Kreitman M, Aguadé M. 1987. A test of neutral molecular evolution based on nucleotide data. Genetics 116:153–159.

Huerta-Sánchez E, Jin X, Asan, Bianba Z, Peter BM, Vinckenbosch N, Liang Y, Yi X, He M, Somel M, Ni P, Wang B, Ou X, Huasang, Luosang J, Cuo ZXP, Li K, Gao G, Yin Y, Wang W, Zhang X, Xu X, Yang H, Li Y, Wang J, Wang J, Nielsen R. 2014. Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. Nature 512:194–197.

Hug N, Longman D, Cáceres JF. 2016. Mechanism and regulation of the nonsense-mediated decay pathway. Nucleic Acids Res 44:1483–1495.

Hughes AL, Packer B, Welch R, Bergen AW, Chanock SJ, Yeager M. 2003. Widespread purifying selection at polymorphic sites in human protein-coding loci. Proc Natl Acad Sci 100:15754–15757.

HUGO Pan-Asian SNP Consortium, Abdulla MA, Ahmed I, Assawamakin A, Bhak J, Brahmachari SK, Calacal GC, Chaurasia A, Chen C-H, Chen J, Chen Y-T, Chu J, Cutiongco-de la Paz EMC, De Ungria MCA, Delfin FC, Edo J, Fuchareon S, Ghang H, Gojobori T, Han J, Ho S-F, Hoh BP, Huang W, Inoko H, Jha P, Jinam TA, Jin L, Jung J, Kangwanpong D, Kampuansai J, Kennedy GC, Khurana P, Kim H-L, Kim K, Kim S, Kim W-Y, Kimm K, Kimura R, Koike T, Kulawonganunchai S, Kumar V, Lai PS, Lee J-Y, Lee S, Liu ET, Majumder PP, Mandapati KK, Marzuki S, Mitchell W, Mukerji M, Naritomi K, Ngamphiw C, Niikawa N, Nishida N, Oh B, Oh S, Ohashi J, Oka A, Ong R, Padilla CD, Palittapongarnpim P, Perdigon HB, Phipps ME, Png E, Sakaki Y, Salvador JM, Sandraling Y, Scaria V, Seielstad M, Sidek MR, Sinha A, Srikummool M, Sudoyo H, Sugano S, Suryadi H, Suzuki Y, Tabbada KA, Tan A, Tokunaga K, Tongsima S, Villamor LP, Wang E, Wang Y, Wang H, Wu J-Y, Xiao H, Xu S, Yang JO, Shugart YY, Yoo H-S, Yuan W, Zhao G, Zilfalil BA, Indian Genome Variation Consortium. 2009. Mapping human genetic diversity in Asia. Science 326:1541–1545.

Hurst LD. 2002. The Ka/Ks ratio: diagnosing the form of sequence evolution. Trends Genet 18:486–487.

Inchley CE, Larbey CDA, Shwan NAA, Pagani L, Saag L, Antão T, Jacobs G, Hudjashov G, Metspalu E, Mitt M, Eichstaedt CA, Malyarchuk B, Derenko M, Wee J, Abdullah S, Ricaut F-X, Mormina M, Mägi R, Villems R, Metspalu M, Jones MK, Armour JAL, Kivisild T. 2016. Selective sweep on human amylase genes postdates the split with Neanderthals. Sci Rep 6:37198.

Ingman M, Kaessmann H, Paabo S, Gyllensten U. 2000. Mitochondrial genome variation and the origin of modern humans. Nature 408:708–713.

International HapMap Consortium, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, Pasternak S, Wheeler DA, Willis TD, Yu F, Yang H, Zeng C, Gao Y, Hu H, Hu W, Li C, Lin W, Liu S, Pan H, Tang X, Wang J, Wang W, Yu J, Zhang B, Zhang Q, Zhao H, Zhao H, Zhou J, Gabriel SB, Barry R, Blumenstiel

B, Camargo A, Defelice M, Faggart M, Goyette M, Gupta S, Moore J, Nguyen H, Onofrio RC, Parkin M, Roy J, Stahl E, Winchester E, Ziaugra L, Altshuler D, Shen Y, Yao Z, Huang W, Chu X, He Y, Jin L, Liu Y, Shen Y, Sun W, Wang H, Wang Y, Wang Y, Xiong X, Xu L, Waye MMY, Tsui SKW, Xue H, Wong JT-F, Galver LM, Fan J-B, Gunderson K, Murray SS, Oliphant AR, Chee MS, Montpetit A, Chagnon F, Ferretti V, Leboeuf M, Olivier J-F, Phillips MS, Roumy S, Sallée C, Verner A, Hudson TJ, Kwok P-Y, Cai D, Koboldt DC, Miller RD, Pawlikowska L, Taillon-Miller P, Xiao M, Tsui L-C, Mak W, Song YQ, Tam PKH, Nakamura Y, Kawaguchi T, Kitamoto T, et al. 2007. A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851–861.

International HapMap3 Consortium, Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu F, Peltonen L, Dermitzakis E, Bonnen PE, Altshuler DM, Gibbs RA, de Bakker PIW, Deloukas P, Gabriel SB, Gwilliam R, Hunt S, Inouye M, Jia X, Palotie A, Parkin M, Whittaker P, Yu F, Chang K, Hawes A, Lewis LR, Ren Y, Wheeler D, Gibbs RA, Marie Muzny D, Barnes C, Darvishi K, Hurles M, Korn JM, Kristiansson K, Lee C, McCarroll SA, Nemesh J, Dermitzakis E, Keinan A, Montgomery SB, Pollack S, Price AL, Soranzo N, Bonnen PE, Gibbs RA, Gonzaga-Jauregui C, Keinan A, Price AL, Yu F, Anttila V, Brodeur W, Daly MJ, Leslie S, McVean G, Moutsianas L, Nguyen H, Schaffner SF, Zhang Q, Ghori MJR, McGinnis R, McLaren W, Pollack S, Price AL, Schaffner SF, Takeuchi F, Grossman SR, Shlyakhter I, Hostetter EB, Sabeti PC, Adebamowo CA, Foster MW, Gordon DR, Licinio J, Cristina Manca M, Marshall PA, Matsuda I, Ngare D, Ota Wang V, Reddy D, Rotimi CN, Royal CD, Sharp RR, Zeng C, Brooks LD, McEwen JE. 2010. Integrating common and rare genetic variation in diverse human populations. Nature 467:52–58.

International Human Genome Sequencing Consortium, Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng J-F, Olsen A, Lucas S, et al. 2001. Initial sequencing and analysis of the human genome. Nature 409:860–921.

Irshaid NM, Henry SM, Olsson ML. 2000. Genomic characterization of the Kidd blood group gene:different molecular basis of the Jk(a-b-) phenotypein Polynesians and Finns. Transfusion (Paris) 40:69–74.

Ishikawa H, Marshall WF. 2011. Ciliogenesis: building the cell's antenna. Nat Rev Mol Cell Biol 12:222–234.

Itan Y, Shang L, Boisson B, Ciancanelli MJ, Markle JG, Martinez-Barricarte R, Scott E, Shah I, Stenson PD, Gleeson J, Cooper DN, Quintana-Murci L, Zhang S-Y, Abel L, Casanova J-L.

2016. The mutation significance cutoff: gene-level thresholds for variant predictions. Nat Methods 13:109–110.

Jablonski NG, Chaplin G. 2000. The evolution of human skin coloration. J Hum Evol 39:57–106.

Jacobs GS, Sluckin TJ, Kivisild T. 2016. Refining the Use of Linkage Disequilibrium as a Robust Signature of Selective Sweeps. Genetics 203:1807–1825.

Jesaitis A. 2017. The State of Variant Annotation in 2017. Available from: http://andrewjesaitis.com/2017/03/the-state-of-variant-annotation-in-2017/

Jin W, Xu S, Wang H, Yu Y, Shen Y, Wu B, Jin L. 2012. Genome-wide detection of natural selection in African Americans pre- and post-admixture. Genome Res 22:519–527.

Jobling MA, Hollox E, Kivisild T, Tyler-Smith C. 2013. Human Evolutionary Genetics. 2nd ed. New York: Garland Science.

Johnson MJ, Wallace DC, Ferris SD, Rattazzi MC, Cavalli-Sforza LL. 1983. Radiation of human mitochondria DNA types analyzed by restriction endonuclease cleavage patterns. J Mol Evol 19:255–271.

Kaback MM, Desnick RJ. 2011. Hexosaminidase A Deficiency. In: Pagon RA, Adam MP, Ardinger HH, Wallace SE, Amemiya A, Bean LJ, Bird TD, Ledbetter N, Mefford HC, Smith RJ, Stephens K, editors. GeneReviews(®). Seattle (WA): University of Washington, Seattle. Available from: http://www.ncbi.nlm.nih.gov/books/NBK1218/

Kaiser VB, Svinti V, Prendergast JG, Chau Y-Y, Campbell A, Patarcic I, Barroso I, Joshi PK, Hastie ND, Miljkovic A, Taylor MS, Generation Scotland, UK10K, Enroth S, Memari Y, Kolb-Kokocinski A, Wright AF, Gyllensten U, Durbin R, Rudan I, Campbell H, Polašek O, Johansson Å, Sauer S, Porteous DJ, Fraser RM, Drake C, Vitart V, Hayward C, Semple CA, Wilson JF. 2015. Homozygous loss-of-function variants in European cosmopolitan and isolate populations. Hum Mol Genet 24:5464–5474.

Kamberov YG, Wang S, Tan J, Gerbault P, Wark A, Tan L, Yang Y, Li S, Tang K, Chen H, Powell A, Itan Y, Fuller D, Lohmueller J, Mao J, Schachar A, Paymer M, Hostetter E, Byrne E, Burnett M, McMahon AP, Thomas MG, Lieberman DE, Jin L, Tabin CJ, Morgan BA, Sabeti PC. 2013. Modeling Recent Human Evolution in Mice by Expression of a Selected EDAR Variant. Cell 152:691–702.

Karafet TM, Hallmark B, Cox MP, Sudoyo H, Downey S, Lansing JS, Hammer MF. 2010. Major East-West Division Underlies Y Chromosome Stratification across Indonesia. Mol Biol Evol 27:1833–1844.

Karafet TM, Lansing JS, Redd AJ, Reznikova S, Watkins JC, Surata SPK, Arthawiguna WA, Mayer L, Bamshad M, Jorde LB, Hammer MF. 2005. Balinese Y-chromosome perspective on the peopling of Indonesia: genetic contributions from pre-neolithic hunter-gatherers, Austronesian farmers, and Indian traders. Hum Biol 77:93–114.

Karam R, Lou C-H, Kroeger H, Huang L, Lin JH, Wilkinson MF. 2015. The unfolded protein response is shaped by the NMD pathway. EMBO Rep 16:599–609.

Karczewski KJ. 2016. LOFTEE (Loss-Of-Function Transcript Effect Estimator).

Karczewski KJ, Weisburd B, Thomas B, Solomonson M, Ruderfer DM, Kavanagh D, Hamamsy T, Lek M, Samocha KE, Cummings BB, Birnbaum D, The Exome Aggregation Consortium, Daly MJ, MacArthur DG. 2017. The ExAC browser: displaying reference data information from over 60 000 exomes. Nucleic Acids Res 45:D840–D845.

Keinan A, Clark AG. 2012. Recent Explosive Human Population Growth Has Resulted in an Excess of Rare Genetic Variants. Science 336:740–743.

Kelleher J, Wong Y, Albers P, Wohns AW, McVean G. 2018. Inferring the ancestry of everyone. Available from: http://biorxiv.org/lookup/doi/10.1101/458067

Kelly DE, Hansen MEB, Tishkoff SA. 2017. Global variation in gene expression and the value of diverse sampling. Curr Opin Syst Biol 1:102–108.

Kim BY, Huber CD, Lohmueller KE. 2017. Inference of the Distribution of Selection Coefficients for New Nonsynonymous Mutations Using Large Samples. Genetics 206:345–361.

Kimura M. 1968. Evolutionary Rate at the Molecular Level. Nature 217:624–626.

Kimura M. 1994. Population genetics, molecular evolution, and the neutral theory: selected papers. Chicago: University of Chicago Press.

Kimura M, Maruyama T, Crow JF. 1963. The mutation load in small populations. Genetics 48:1303–1312.

Kimura M, Ohta T. 1973. The age of a neutral mutant persisting in a finite population. Genetics 75:199–212.

King M, Wilson A. 1975. Evolution at two levels in humans and chimpanzees. Science 188:107–116.

Kingman JFC. 1982. The coalescent. Stoch Process Their Appl 13:235–248.

Kirch PV. 2000. On the road of the winds : an archaeological history of the Pacific islands before European contact. Berkeley, Calif.: University of California Press.

Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. 2014. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 46:310–315.

Kloss-Brandstätter A, Pacher D, Schönherr S, Weissensteiner H, Binna R, Specht G, Kronenberg F. 2011. HaploGrep: a fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups. Hum Mutat 32:25–32.

Klug A. 2010. The Discovery of Zinc Fingers and Their Applications in Gene Regulation and Genome Manipulation. Annu Rev Biochem 79:213–231.

Knöppel A, Näsvall J, Andersson DI. 2016. Compensating the Fitness Costs of Synonymous Mutations. Mol Biol Evol 33:1461–1477.

Ko AM-S, Chen C-Y, Fu Q, Delfin F, Li M, Chiu H-L, Stoneking M, Ko Y-C. 2014. Early Austronesians: Into and Out Of Taiwan. Am J Hum Genet 94:426–436.

Koch E, Novembre J. 2017. A Temporal Perspective on the Interplay of Demography and Selection on Deleterious Variation in Humans. G3amp58 GenesGenomesGenetics 7:1027–1037.

Kohnen N. 1986. "Natural" childbirth among the Kankanaly-Igorot. Bull N Y Acad Med 62:768–777.

Kong A, Thorleifsson G, Gudbjartsson DF, Masson G, Sigurdsson A, Jonasdottir A, Walters GB, Jonasdottir A, Gylfason A, Kristinsson KT, Gudjonsson SA, Frigge ML, Helgason A, Thorsteinsdottir U, Stefansson K. 2010. Fine-scale recombination rate differences between sexes, populations and individuals. Nature 467:1099–1103.

Kuhlwilm M, Gronau I, Hubisz MJ, de Filippo C, Prado-Martinez J, Kircher M, Fu Q, Burbano HA, Lalueza-Fox C, de la Rasilla M, Rosas A, Rudan P, Brajkovic D, Kucan Ž, Gušic I, Marques-Bonet T, Andrés AM, Viola B, Pääbo S, Meyer M, Siepel A, Castellano S. 2016. Ancient gene flow from early modern humans into Eastern Neanderthals. Nature 530:429–433.

Kumar P, Henikoff S, Ng PC. 2009. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc 4:1073–1081.

Kumar S, Stecher G, Tamura K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol Biol Evol 33:1870–1874.

Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, Ziller MJ, Amin V, Whitaker JW, Schultz MD, Ward LD, Sarkar A, Quon G, Sandstrom RS, Eaton ML, Wu Y-C, Pfenning A, Wang X, ClaussnitzerYaping Liu M, Coarfa C, Alan Harris R, Shoresh N, Epstein CB, Gjoneska E, Leung D, Xie W, David Hawkins R, Lister R, Hong C, Gascard P, Mungall AJ, Moore R, Chuah E, Tam A, Canfield TK, Scott Hansen R, Kaul R, Sabo PJ, Bansal MS, Carles A, Dixon JR, Farh K-H, Feizi S, Karlic R, Kim A-R, Kulkarni A, Li D, Lowdon R, Elliott G, Mercer TR, Neph SJ, Onuchic V, Polak P, Rajagopal N, Ray P, Sallari RC, Siebenthall KT, Sinnott-Armstrong NA, Stevens M, Thurman RE, Wu J, Zhang B, Zhou X, Abdennur N, Adli M, Akerman M, Barrera L, Antosiewicz-Bourget J, Ballinger T, Barnes MJ, Bates D, Bell RJA, Bennett DA, Bianco K, Bock C, Boyle P, Brinchmann J, Caballero-Campo P, Camahort R, Carrasco-Alfonso MJ, Charnecki T, Chen H, Chen Z, Cheng JB, Cho S, Chu A, Chung W-Y, Cowan C, Athena Deng Q, Deshpande V, Diegel M, Ding B, Durham T, Echipare L, Edsall L, et al. 2015. Integrative analysis of 111 reference human epigenomes. Nature 518:317–330.

Kusuma P, Brucato N, Cox MP, Letellier T, Manan A, Nuraini C, Grangé P, Sudoyo H, Ricaut F-X. 2017. The last sea nomads of the Indonesian archipelago: genomic origins and dispersal. Eur J Hum Genet 25:1004–1010.

Kusuma P, Cox MP, Brucato N, Sudoyo H, Letellier T, Ricaut F-X. 2015. Western Eurasian genetic influences in the Indonesian archipelago. Quat Int [Internet]. Available from: http://linkinghub.elsevier.com/retrieve/pii/S1040618215006485

Kwiatkowski DP. 2005. How Malaria Has Affected the Human Genome and What Human Genetics Can Teach Us about Malaria. Am J Hum Genet 77:171–192.

Lachance J, Tishkoff SA. 2013. SNP ascertainment bias in population genetic analyses: Why it is important, and how to correct it: Prospects & Overviews. BioEssays 35:780–786.

Lachance J, Vernot B, Elbers CC, Ferwerda B, Froment A, Bodo J-M, Lema G, Fu W, Nyambo TB, Rebbeck TR, Zhang K, Akey JM, Tishkoff SA. 2012. Evolutionary History and Adaptation from High-Coverage Whole-Genome Sequences of Diverse African Hunter-Gatherers. Cell 150:457–469.

LaFramboise T. 2009. Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. Nucleic Acids Res 37:4181–4193.

Lahr MM, Foley R. 1994. Multiple dispersals and modern human origins. Evol Anthropol Issues News Rev 3:48–60.

Lam HYK, Clark MJ, Chen R, Chen R, Natsoulis G, O'Huallachain M, Dewey FE, Habegger L, Ashley EA, Gerstein MB, Butte AJ, Ji HP, Snyder M. 2011. Performance comparison of whole-genome sequencing platforms. Nat Biotechnol 30:78–82.

Lam TH, Shen M, Tay MZ, Ren EC. 2017. Unique Allelic eQTL Clusters in Human MHC Haplotypes. G3amp58 GenesGenomesGenetics 7:2595–2604.

Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Hoover J, Jang W, Katz K, Ovetsky M, Riley G, Sethi A, Tully R, Villamarin-Salomon R, Rubinstein W, Maglott DR. 2016. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res 44:D862–D868.

Larson G, Cucchi T, Fujita M, Matisoo-Smith E, Robins J, Anderson A, Rolett B, Spriggs M, Dolman G, Kim T-H, Thuy NTD, Randi E, Doherty M, Due RA, Bollt R, Djubiantono T, Griffin B, Intoh M, Keane E, Kirch P, Li K-T, Morwood M, Pedrina LM, Piper PJ, Rabett RJ, Shooter P, Van den Bergh G, West E, Wickler S, Yuan J, Cooper A, Dobney K. 2007. Phylogeny and ancient DNA of Sus provides insights into neolithic expansion in Island Southeast Asia and Oceania. Proc Natl Acad Sci 104:4834–4839.

Lawler A. 2014. Sailing Sinbad's seas. Science 344:1440–1445.

Lawson DJ, Falush D. 2012. Population Identification Using Genetic Data. Annu Rev Genomics Hum Genet 13:337–361.

Lawson DJ, Hellenthal G, Myers S, Falush D. 2012. Inference of Population Structure using Dense Haplotype Data. PLoS Genet 8:e1002453.

Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, Kirsanow K, Sudmant PH, Schraiber JG, Castellano S, Lipson M, Berger B, Economou C, Bollongino R, Fu Q, Bos KI, Nordenfelt S, Li H, de Filippo C, Prüfer K, Sawyer S, Posth C, Haak W, Hallgren F, Fornander E, Rohland N, Delsate D, Francken M, Guinet J-M, Wahl J, Ayodo G, Babiker HA, Bailliet G, Balanovska E, Balanovsky O, Barrantes R, Bedoya G, Ben-Ami H, Bene J, Berrada F, Bravi CM, Brisighelli F, Busby GBJ, Cali F, Churnosov M, Cole DEC, Corach D, Damba L, van Driem G, Dryomov S, Dugoujon J-M, Fedorova SA, Gallego Romero I, Gubina M, Hammer M, Henn BM, Hervig

T, Hodoglugil U, Jha AR, Karachanak-Yankova S, Khusainova R, Khusnutdinova E, Kittles R, Kivisild T, Klitz W, Kučinskas V, Kushniarevich A, Laredj L, Litvinov S, Loukidis T, Mahley RW, Melegh B, Metspalu E, Molina J, Mountain J, Näkkäläjärvi K, Nesheva D, Nyambo T, Osipova L, Parik J, Platonov F, Posukh O, Romano V, Rothhammer F, Rudan I, Ruizbakiev R, Sahakyan H, Sajantila A, Salas A, Starikovskaya EB, Tarekegn A, Toncheva D, Turdikulova S, Uktveryte I, Utevska O, Vasquez R, Villena M, Voevoda M, Winkler CA, et al. 2014. Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature 513:409–413.

Leberg PL. 2002. Estimating allelic richness: effects of sample size and bottlenecks. Mol Ecol 11:2445–2449.

Lee FXZ, Houweling PJ, North KN, Quinlan KGR. 2016. How does α-actinin-3 deficiency alter muscle function? Mechanistic insights into ACTN3, the 'gene for speed.' Biochim Biophys Acta BBA - Mol Cell Res 1863:686–693.

Lee Y, Park S, Lee JS, Kim SY, Cho J, Yoo Y, Lee S, Yoo T, Lee M, Seo J, Lee J, Kneissl J, Lee J, Jeon H, Jeon EY, Hong SE, Kim E, Kim H, Kim WJ, Kim JS, Ko JM, Cho A, Lim BC, Kim WS, Choi M, Chae J-H. 2019. Genomic profiling of 553 uncharacterized neurodevelopment patients reveals a high proportion of recessive pathogenic variant carriers in an outbred population. bioRxiv:675868.

Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, Tukiainen T, Birnbaum DP, Kosmicki JA, Duncan LE, Estrada K, Zhao F, Zou J, Pierce-Hoffman E, Berghout J, Cooper DN, Deflaux N, DePristo M, Do R, Flannick J, Fromer M, Gauthier L, Goldstein J, Gupta N, Howrigan D, Kiezun A, Kurki MI, Moonshine AL, Natarajan P, Orozco L, Peloso GM, Poplin R, Rivas MA, Ruano-Rubio V, Rose SA, Ruderfer DM, Shakir K, Stenson PD, Stevens C, Thomas BP, Tiao G, Tusie-Luna MT, Weisburd B, Won H-H, Yu D, Altshuler DM, Ardissino D, Boehnke M, Danesh J, Donnelly S, Elosua R, Florez JC, Gabriel SB, Getz G, Glatt SJ, Hultman CM, Kathiresan S, Laakso M, McCarroll S, McCarthy MI, McGovern D, McPherson R, Neale BM, Palotie A, Purcell SM, Saleheen D, Scharf JM, Sklar P, Sullivan PF, Tuomilehto J, Tsuang MT, Watkins HC, Wilson JG, Daly MJ, MacArthur DG, Exome Aggregation Consortium. 2016. Analysis of protein-coding genetic variation in 60,706 humans. Nature 536:285–291.

Leslie S, Winney B, Hellenthal G, Davison D, Boumertit A, Day T, Hutnik K, Royrvik EC, Cunliffe B, Wellcome Trust Case Control Consortium 2, International Multiple Sclerosis Genetics Consortium, Lawson DJ, Falush D, Freeman C, Pirinen M, Myers S, Robinson M, Donnelly P, Bodmer W. 2015. The fine-scale genetic structure of the British population. Nature 519:309–314.

Lewis MP, Simons GF, Fennig CD eds. 2015. Ethnologue: Languages of the World, Eighteenth edition. Dallas, Texas: SIL International. Available from: http://www.ethnologue.com

Li H, Durbin R. 2011. Inference of human population history from individual whole-genome sequences. Nature 475:493–496.

Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL, Myers RM. 2008. Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation. Science 319:1100–1104.

Li N, Stephens M. 2003. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics 165:2213–2233.

Li W-H. 1997. Molecular evolution. Sunderland, Mass: Sinauer Associates.

Li WH, Sadler LA. 1991. Low nucleotide diversity in man. Genetics 129:513–523.

Liang M, Nielsen R. 2014. The Lengths of Admixture Tracts. Genetics 197:953–967.

Lim ET, Würtz P, Havulinna AS, Palta P, Tukiainen T, Rehnström K, Esko T, Mägi R, Inouye M, Lappalainen T, Chan Y, Salem RM, Lek M, Flannick J, Sim X, Manning A, Ladenvall C, Bumpstead S, Hämäläinen E, Aalto K, Maksimow M, Salmi M, Blankenberg S, Ardissino D, Shah S, Horne B, McPherson R, Hovingh GK, Reilly MP, Watkins H, Goel A, Farrall M, Girelli D, Reiner AP, Stitziel NO, Kathiresan S, Gabriel S, Barrett JC, Lehtimäki T, Laakso M, Groop L, Kaprio J, Perola M, McCarthy MI, Boehnke M, Altshuler DM, Lindgren CM, Hirschhorn JN, Metspalu A, Freimer NB, Zeller T, Jalkanen S, Koskinen S, Raitakari O, Durbin R, MacArthur DG, Salomaa V, Ripatti S, Daly MJ, Palotie A, for the Sequencing Initiative Suomi (SISu) Project. 2014. Distribution and Medical Impact of Loss-of-Function Variants in the Finnish Founder Population. PLoS Genet 10:e1004494.

Lipson M, Cheronet O, Mallick S, Rohland N, Oxenham M, Pietrusewsky M, Pryce TO, Willis A, Matsumura H, Buckley H, Domett K, Nguyen GH, Trinh HH, Kyaw AA, Win TT, Pradier B, Broomandkhoshbacht N, Candilio F, Changmai P, Fernandes D, Ferry M, Gamarra B, Harney E, Kampuansai J, Kutanan W, Michel M, Novak M, Oppenheimer J, Sirak K, Stewardson K, Zhang Z, Flegontov P, Pinhasi R, Reich D. 2018. Ancient genomes document multiple waves of migration in Southeast Asian prehistory. Science 361:92–95.

Lipson M, Loh P-R, Patterson N, Moorjani P, Ko Y-C, Stoneking M, Berger B, Reich D. 2014. Reconstructing Austronesian population history in Island Southeast Asia. Nat Commun 5:4689.

Lipson M, Loh P-R, Sankararaman S, Patterson N, Berger B, Reich D. 2015. Calibrating the Human Mutation Rate via Ancestral Recombination Density in Diploid Genomes. PLOS Genet 11:e1005550.

Liu W, Martinón-Torres M, Cai Y, Xing S, Tong H, Pei S, Sier MJ, Wu X, Edwards RL, Cheng H, Li Y, Yang X, de Castro JMB, Wu X. 2015. The earliest unequivocally modern humans in southern China. Nature 526:696–699.

Liu X, Fu Y-X. 2015. Exploring population size changes using SNP frequency spectra. Nat Genet 47:555–559.

Loh P-R, Lipson M, Patterson N, Moorjani P, Pickrell JK, Reich D, Berger B. 2013. Inferring admixture histories of human populations using linkage disequilibrium. Genetics 193:1233–1254.

Lohmueller KE. 2014a. The distribution of deleterious genetic variation in human populations. Curr Opin Genet Dev 29:139–146.

Lohmueller KE. 2014b. The Impact of Population Demography and Selection on the Genetic Architecture of Complex Traits. PLoS Genet 10:e1004379.

Lohmueller KE, Indap AR, Schmidt S, Boyko AR, Hernandez RD, Hubisz MJ, Sninsky JJ, White TJ, Sunyaev SR, Nielsen R, Clark AG, Bustamante CD. 2008. Proportionally more deleterious genetic variation in European than in African populations. Nature 451:994–997.

Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, Hasz R, Walters G, Garcia F, Young N, Foster B, Moser M, Karasik E, Gillard B, Ramsey K, Sullivan S, Bridge J, Magazine H, Syron J, Fleming J, Siminoff L, Traino H, Mosavel M, Barker L, Jewell S, Rohrer D, Maxim D, Filkins D, Harbach P, Cortadillo E, Berghuis B, Turner L, Hudson E, Feenstra K, Sobin L, Robb J, Branton P, Korzeniewski G, Shive C, Tabor D, Qi L, Groch K, Nampally S, Buia S, Zimmerman A, Smith A, Burges R, Robinson K, Valentino K, Bradbury D, Cosentino M, Diaz-Mayoral N, Kennedy M, Engel T, Williams P, Erickson K, Ardlie K, Winckler W, Getz G, DeLuca D, MacArthur D, Kellis M, Thomson A, Young T, Gelfand E, Donovan M, Meng Y, Grant G, Mash D, Marcus Y, Basile M, Liu J, Zhu J, Tu Z, Cox NJ, Nicolae DL, Gamazon ER, Im HK, Konkashbaev A, Pritchard J, Stevens M, Flutre T, Wen X, Dermitzakis ET, Lappalainen T, Guigo R, Monlong J, Sammeth M, Koller D, Battle A, Mostafavi S, McCarthy M, Rivas M, Maller J, Rusyn I, Nobel A, Wright F, Shabalin A, et al. 2013. The Genotype-Tissue Expression (GTEx) project. Nat Genet 45:580–585.

Lopez M, Kousathanas A, Quach H, Harmant C, Mouguiama-Daouda P, Hombert J-M, Froment A, Perry GH, Barreiro LB, Verdu P, Patin E, Quintana-Murci L. 2018. The demographic history and mutational load of African hunter-gatherers and farmers. Nat Ecol Evol 2:721–730.

Lopez S, van Dorp L, Hellenthal G. 2016. Human Dispersal Out of Africa: A Lasting Debate. Evol Bioinforma:57.

Lykke-Andersen S, Jensen TH. 2015. Nonsense-mediated mRNA decay: an intricate machinery that shapes transcriptomes. Nat Rev Mol Cell Biol 16:665–677.

Mabbett IW. 1977. The “Indianization” of Southeast Asia: Reflections on the Historical Sources. J Southeast Asian Stud 8:143–161.

MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, Jostins L, Habegger L, Pickrell JK, Montgomery SB, Albers CA, Zhang ZD, Conrad DF, Lunter G, Zheng H, Ayub Q, DePristo MA, Banks E, Hu M, Handsaker RE, Rosenfeld JA, Fromer M, Jin M, Mu XJ, Khurana E, Ye K, Kay M, Saunders GI, Suner M-M, Hunt T, Barnes IHA, Amid C, Carvalho-Silva DR, Bignell AH, Snow C, Yngvadottir B, Bumpstead S, Cooper DN, Xue Y, Romero IG, 1000 Genomes Project Consortium, Wang J, Li Y, Gibbs RA, McCarroll SA, Dermitzakis ET, Pritchard JK, Barrett JC, Harrow J, Hurles ME, Gerstein MB, Tyler-Smith C. 2012. A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes. Science 335:823–828.

MacArthur DG, Seto JT, Chan S, Quinlan KGR, Raftery JM, Turner N, Nicholson MD, Kee AJ, Hardeman EC, Gunning PW, Cooney GJ, Head SI, Yang N, North KN. 2008. An Actn3 knockout mouse provides mechanistic insights into the association between -actinin-3 deficiency and human athletic performance. Hum Mol Genet 17:1076–1086.

MacArthur DG, Seto JT, Raftery JM, Quinlan KG, Huttley GA, Hook JW, Lemckert FA, Kee AJ, Edwards MR, Berman Y, Hardeman EC, Gunning PW, Easteal S, Yang N, North KN. 2007.

Loss of ACTN3 gene function alters mouse muscle metabolism and shows evidence of positive selection in humans. Nat Genet 39:1261–1265.

MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, Junkins H, McMahon A, Milano A, Morales J, Pendlington ZM, Welter D, Burdett T, Hindorff L, Flicek P, Cunningham F, Parkinson H. 2017. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res 45:D896–D901.

Malaspinas A-S, Westaway MC, Muller C, Sousa VC, Lao O, Alves I, Bergström A, Athanasiadis G, Cheng JY, Crawford JE, Heupink TH, Macholdt E, Peischl S, Rasmussen S, Schiffels S, Subramanian S, Wright JL, Albrechtsen A, Barbieri C, Dupanloup I, Eriksson A, Margaryan A, Moltke I, Pugach I, Korneliussen TS, Levkivskyi IP, Moreno-Mayar JV, Ni S, Racimo F, Sikora M, Xue Y, Aghakhanian FA, Brucato N, Brunak S, Campos PF, Clark W, Ellingvåg S, Fourmile G, Gerbault P, Injie D, Koki G, Leavesley M, Logan B, Lynch A, Matisoo-Smith EA, McAllister PJ, Mentzer AJ, Metspalu M, Migliano AB, Murgha L, Phipps ME, Pomat W, Reynolds D, Ricaut F-X, Siba P, Thomas MG, Wales T, Wall CM, Oppenheimer SJ, Tyler-Smith C, Durbin R, Dortch J, Manica A, Schierup MH, Foley RA, Lahr MM, Bowern C, Wall JD, Mailund T, Stoneking M, Nielsen R, Sandhu MS, Excoffier L, Lambert DM, Willerslev E. 2016. A genomic history of Aboriginal Australia. Nature [Internet]. Available from: http://www.nature.com/doifinder/10.1038/nature18299

Malécot G. 1973. Isolation by distance. In: Morton NE, editor. Genetic Structure of Populations. Honolulu: University of Hawaii Press. p 72–75.

Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, Zhao M, Chennagiri N, Nordenfelt S, Tandon A, Skoglund P, Lazaridis I, Sankararaman S, Fu Q, Rohland N, Renaud G, Erlich Y, Willems T, Gallo C, Spence JP, Song YS, Poletti G, Balloux F, van Driem G, de Knijff P, Romero IG, Jha AR, Behar DM, Bravi CM, Capelli C, Hervig T, Moreno-Estrada A, Posukh OL, Balanovska E, Balanovsky O, Karachanak-Yankova S, Sahakyan H, Toncheva D, Yepiskoposyan L, Tyler-Smith C, Xue Y, Abdullah MS, Ruiz-Linares A, Beall CM, Di Rienzo A, Jeong C, Starikovskaya EB, Metspalu E, Parik J, Villems R, Henn BM, Hodoglugil U, Mahley R, Sajantila A, Stamatoyannopoulos G, Wee JTS, Khusainova R, Khusnutdinova E, Litvinov S, Ayodo G, Comas D, Hammer MF, Kivisild T, Klitz W, Winkler CA, Labuda D, Bamshad M, Jorde LB, Tishkoff SA, Watkins WS, Metspalu M, Dryomov S, Sukernik R, Singh L, Thangaraj K, Pääbo S, Kelso J, Patterson N, Reich D. 2016. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature [Internet]. Available from: http://www.nature.com/doifinder/10.1038/nature18964

Manguin P-Y, Mani A, Wade G eds. 2011. Early interactions between South and Southeast Asia: reflections on cross-cultural exchange. Singapore : New Delhi: Institute of Southeast Asian Studies ; Manohar India.

Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen W-M. 2010. Robust relationship inference in genome-wide association studies. Bioinformatics 26:2867–2873.

Manna F, Martin G, Lenormand T. 2011. Fitness Landscapes: An Alternative Theory for the Dominance of Mutation. Genetics 189:923–937.

Manrai AK, Funke BH, Rehm HL, Olesen MS, Maron BA, Szolovits P, Margulies DM, Loscalzo J, Kohane IS. 2016. Genetic Misdiagnoses and the Potential for Health Disparities. N Engl J Med 375:655–665.

Manry J, Nédélec Y, Fava VM, Cobat A, Orlova M, Thuc NV, Thai VH, Laval G, Barreiro LB, Schurr E. 2017. Deciphering the genetic control of gene expression following Mycobacterium leprae antigen stimulation. PLOS Genet 13:e1006952.

Mantel N. 1967. The detection of disease clustering and a generalized regression approach. Cancer Res 27:209–220.

Maquat LE. 2004. Nonsense-Mediated mRNA Decay: Splicing, Translation And mRNP Dynamics. Nat Rev Mol Cell Biol 5:89–99.

Margolis DJ, Gupta J, Apter AJ, Ganguly T, Hoffstad O, Papadopoulos M, Rebbeck TR, Mitra N. 2014. Filaggrin-2 variation is associated with more persistent atopic dermatitis in African American subjects. J Allergy Clin Immunol 133:784–789.

Martin AR, Gignoux CR, Walters RK, Wojcik GL, Neale BM, Gravel S, Daly MJ, Bustamante CD, Kenny EE. 2017. Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. Am J Hum Genet 100:635–649.

Martin AR, Karczewski KJ, Kerminen S, Kurki MI, Sarin A-P, Artomov M, Eriksson JG, Esko T, Genovese G, Havulinna AS, Kaprio J, Konradi A, Korányi L, Kostareva A, Männikkö M, Metspalu A, Perola M, Prasad RB, Raitakari O, Rotar O, Salomaa V, Groop L, Palotie A, Neale BM, Ripatti S, Pirinen M, Daly MJ. 2018. Haplotype Sharing Provides Insights into Fine-Scale Population History and Disease in Finland. Am J Hum Genet 102:760–775.

Martínez Sarasola C. 2005. Nuestros paisanos los indios: vida, historia y destino de las comunidades indígenas en la Argentina. Buenos Aires: Emecé. Available from: http://books.google.com/books?id=vyp7AAAAMAAJ

Mather CA, Mooney SD, Salipante SJ, Scroggins S, Wu D, Pritchard CC, Shirts BH. 2016. CADD score has limited clinical validity for the identification of pathogenic variants in noncoding regions in a hereditary cancer panel. Genet Med 18:1269–1275.

Mathieson I. 2013. Genes in space: selection, association and variation in spatially structured populations. PhD dissertation. University of Oxford

Mathieson I, Lazaridis I, Rohland N, Mallick S, Patterson N, Roodenberg SA, Harney E, Stewardson K, Fernandes D, Novak M, Sirak K, Gamba C, Jones ER, Llamas B, Dryomov S, Pickrell J, Arsuaga JL, de Castro JMB, Carbonell E, Gerritsen F, Khokhlov A, Kuznetsov P, Lozano M, Meller H, Mochalov O, Moiseyev V, Guerra MAR, Roodenberg J, Vergès JM, Krause J, Cooper A, Alt KW, Brown D, Anthony D, Lalueza-Fox C, Haak W, Pinhasi R, Reich D. 2015. Genome-wide patterns of selection in 230 ancient Eurasians. Nature 528:499–503.

Mathieson I, McVean G. 2014. Demography and the Age of Rare Variants. PLoS Genet 10:e1004528.

Mathieson I, Reich D. 2017. Differences in the rare variant spectrum among human populations. PLOS Genet 13:e1006581.

Matisoo-Smith EA. 2015. Tracking Austronesian expansion into the Pacific via the paper mulberry plant. Proc Natl Acad Sci 112:13432–13433.

McCarthy DJ, Humburg P, Kanapin A, Rivas MA, Gaulton K, asds, Cazier J-B, Donnelly P. 2014. Choice of transcripts and software has a large effect on variant annotation. Genome Med 6:26.

McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in Drosophila. Nature 351:652–654.

McDougall I, Brown FH, Fleagle JG. 2005. Stratigraphic placement and age of modern humans from Kibish, Ethiopia. Nature 433:733–736.

McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, Flicek P, Cunningham F. 2016. The Ensembl Variant Effect Predictor. Genome Biol [Internet] 17. Available from: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0974-4

McManus KF, Taravella AM, Henn BM, Bustamante CD, Sikora M, Cornejo OE. 2017. Population genetic analysis of the DARC locus (Duffy) reveals adaptation from standing variation associated with malaria resistance in humans. PLOS Genet 13:e1006560.

McVean GAT, Cardin NJ. 2005. Approximating the coalescent with recombination. Philos Trans R Soc Lond B Biol Sci 360:1387–1393.

Melé M, Javed A, Pybus M, Zalloua P, Haber M, Comas D, Netea MG, Balanovsky O, Balanovska E, Jin L, Yang Y, Pitchappan RM, Arunkumar G, Parida L, Calafell F, Bertranpetit J, the Genographic Consortium. 2012. Recombination Gives a New Insight in the Effective Population Size and the History of the Old World Human Populations. Mol Biol Evol 29:25–30.

Merkle E, You D, Schneider L, Seongho B. 2018. nonnest2: Tests of Non-Nested Models. Available from: https://CRAN.R-project.org/package=nonnest2

Metspalu M, Romero IG, Yunusbayev B, Chaubey G, Mallick CB, Hudjashov G, Nelis M, Mägi R, Metspalu E, Remm M, Pitchappan R, Singh L, Thangaraj K, Villems R, Kivisild T. 2011. Shared and Unique Components of Human Population Structure and Genome-Wide Signals of Positive Selection in South Asia. Am J Hum Genet 89:731–744.

Meyer M, Kircher M, Gansauge M-T, Li H, Racimo F, Mallick S, Schraiber JG, Jay F, Prufer K, de Filippo C, Sudmant PH, Alkan C, Fu Q, Do R, Rohland N, Tandon A, Siebauer M, Green RE, Bryc K, Briggs AW, Stenzel U, Dabney J, Shendure J, Kitzman J, Hammer MF, Shunkov MV, Derevianko AP, Patterson N, Andres AM, Eichler EE, Slatkin M, Reich D, Kelso J, Paabo S. 2012. A High-Coverage Genome Sequence from an Archaic Denisovan Individual. Science 338:222–226.

Mezzavilla M, Ghirotto S. 2015. Neon: An R Package to Estimate Human Effective Population Size and Divergence Time from Patterns of Linkage Disequilibrium between SNPS. J Comput Sci Syst Biol [Internet] 8. Available from: http://www.omicsonline.org/open-access/neon-an-r-package-to-estimate-human-effective-population-size-and-divergence-jcsb.1000168.php?aid=36380

Mi H, Huang X, Muruganujan A, Tang H, Mills C, Kang D, Thomas PD. 2017. PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. Nucleic Acids Res 45:D183–D189.

Migliano AB, Romero IG, Metspalu M, Leavesley M, Pagani L, Antao T, Huang D-W, Sherman BT, Siddle K, Scholes C, Hudjashov G, Kaitokai E, Babalu A, Belatti M, Cagan A, Hopkinshaw B, Shaw C, Nelis M, Metspalu E, Mägi R, Lempicki RA, Villems R, Lahr MM, Kivisild T. 2013. Evolution of the Pygmy Phenotype: Evidence of Positive Selection from Genome-wide Scans in African, Asian, and Melanesian Pygmies. Hum Biol 85:251.

Miosge LA, Field MA, Sontani Y, Cho V, Johnson S, Palkova A, Balakishnan B, Liang R, Zhang Y, Lyon S, Beutler B, Whittle B, Bertram EM, Enders A, Goodnow CC, Andrews TD. 2015. Comparison of predicted and actual consequences of missense mutations. Proc Natl Acad Sci 112:E5189–E5198.

Mirazón Lahr M. 2016. The shaping of human diversity: filters, boundaries and transitions. Philos Trans R Soc B Biol Sci 371:20150241.

Montgomery SB, Goode DL, Kvikstad E, Albers CA, Zhang ZD, Mu XJ, Ananda G, Howie B, Karczewski KJ, Smith KS, Anaya V, Richardson R, Davis J, The 1000 Genomes Project Consortium, MacArthur DG, Sidow A, Duret L, Gerstein M, Makova KD, Marchini J, McVean G, Lunter G. 2013. The origin, evolution, and functional impact of short insertion-deletion variants identified in 179 human genomes. Genome Res 23:749–761.

Moorjani P, Patterson N, Hirschhorn JN, Keinan A, Hao L, Atzmon G, Burns E, Ostrer H, Price AL, Reich D. 2011. The History of African Gene Flow into Southern Europeans, Levantines, and Jews. PLoS Genet 7:e1001373.

Moran CN, Yang N, Bailey MES, Tsiokanos A, Jamurtas A, MacArthur DG, North K, Pitsiladis YP, Wilson RH. 2007. Association analysis of the ACTN3 R577X polymorphism and complex quantitative body composition and performance phenotypes in adolescent Greeks. Eur J Hum Genet 15:88–93.

Moreno-Estrada A, Gravel S, Zakharia F, McCauley JL, Byrnes JK, Gignoux CR, Ortiz-Tello PA, Martínez RJ, Hedges DJ, Morris RW, Eng C, Sandoval K, Acevedo-Acevedo S, Norman PJ, Layrisse Z, Parham P, Martínez-Cruzado JC, Burchard EG, Cuccaro ML, Martin ER, Bustamante CD. 2013. Reconstructing the Population Genetic History of the Caribbean. PLoS Genet 9:e1003925.

Morton NE. 1973a. Kinship and population structure. In: Morton NE, editor. Genetic Structure of Populations. Honolulu: University of Hawaii Press. p 66–69.

Morton NE. 1973b. Isolation by distance. In: Morton NE, editor. Genetic Structure of Populations. Honolulu: University of Hawaii Press. p 76–77.

Morton NE, Crow JF, Muller HJ. 1956. An estimate of the mutational damage in man from data on consanguineous marriages. Proc Natl Acad Sci U S A 42:855–863.

Motti J, Muzzio M, Ramallo V, Kladniew BR, Alfaro E, Dipierri J, Bailliet G, Bravi C. 2012. Origen y distribución espacial de linajes maternos nativos en el noroeste y centro oeste argentinos/ Origin and spatial distribution of native maternal lineages in Northwest and Center

West of Argentina. Rev Argent Antropol Biológica [Internet] 15. Available from: https://revistas.unlp.edu.ar/raab/article/view/607

Muzzio M, Motti JMB, Paz Sepulveda PB, Yee M, Cooke T, Santos MR, Ramallo V, Alfaro EL, Dipierri JE, Bailliet G, Bravi CM, Bustamante CD, Kenny EE. 2018. Population structure in Argentina. PLOS ONE 13:e0196325.

Myers S, Spencer CCA, Auton A, Bottolo L, Freeman C, Donnelly P, McVean G. 2006. The distribution and causes of meiotic recombination in the human genome. Biochemical Society Transactions 34:526–530.

Nachman MW, Crowell SL. 2000. Estimate of the mutation rate per nucleotide in humans. Genetics 156:297–304.

Nagy E, Maquat LE. 1998. A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance. Trends Biochem Sci 23:198–199.

Narasimhan VM, Hunt KA, Mason D, Baker CL, Karczewski KJ, Barnes MR, Barnett AH, Bates C, Bellary S, Bockett NA, Giorda K, Griffiths CJ, Hemingway H, Jia Z, Kelly MA, Khawaja HA, Lek M, McCarthy S, McEachan R, ODonnell-Luria A, Paigen K, Parisinos CA, Sheridan E, Southgate L, Tee L, Thomas M, Xue Y, Schnall-Levin M, Petkov PM, Tyler-Smith C, Maher ER, Trembath RC, MacArthur DG, Wright J, Durbin R, van Heel DA. 2016. Health and population effects of rare gene knockouts in adult humans with related parents. Science 352:474–477.

Navarese EP, Kolodziejczak M, Schulze V, Gurbel PA, Tantry U, Lin Y, Brockmeyer M, Kandzari DE, Kubica JM, D'Agostino RB, Kubica J, Volpe M, Agewall S, Kereiakes DJ, Kelm M. 2015. Effects of Proprotein Convertase Subtilisin/Kexin Type 9 Antibodies in Adults With Hypercholesterolemia: A Systematic Review and Meta-analysis. Ann Intern Med 163:40.

Nei M, Suzuki Y, Nozawa M. 2010. The Neutral Theory of Molecular Evolution in the Genomic Era. Annu Rev Genomics Hum Genet 11:265–289.

Nelson MR, Wegmann D, Ehm MG, Kessner D, St. Jean P, Verzilli C, Shen J, Tang Z, Bacanu S-A, Fraser D, Warren L, Aponte J, Zawistowski M, Liu X, Zhang H, Zhang Y, Li J, Li Y, Li L, Woollard P, Topp S, Hall MD, Nangle K, Wang J, Abecasis G, Cardon LR, Zollner S, Whittaker JC, Chissoe SL, Novembre J, Mooser V. 2012. An Abundance of Rare Functional Variants in 202 Drug Target Genes Sequenced in 14,002 People. Science 337:100–104.

Ng PC, Henikoff S. 2001. Predicting deleterious amino acid substitutions. Genome Res 11:863–874.

Nielsen R, Bustamante C, Clark AG, Glanowski S, Sackton TB, Hubisz MJ, Fledel-Alon A, Tanenbaum DM, Civello D, White TJ, J. Sninsky J, Adams MD, Cargill M. 2005. A Scan for Positively Selected Genes in the Genomes of Humans and Chimpanzees. PLoS Biol 3:e170.

Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG. 2007. Recent and ongoing selection in the human genome. Nat Rev Genet 8:857–868.

Nigst PR, Haesaerts P, Damblon F, Frank-Fellner C, Mallol C, Viola B, Götzinger M, Niven L, Trnka G, Hublin J-J. 2014. Early modern human settlement of Europe north of the Alps

occurred 43,500 years ago in a cold steppe-type environment. Proc Natl Acad Sci 111:14394–14399.

NIST. 2013. NIST/SEMATECH e-Handbook of Statistical Methods. Available from: http://www.itl.nist.gov/div898/handbook/

Norton HL, Kittles RA, Parra E, McKeigue P, Mao X, Cheng K, Canfield VA, Bradley DG, McEvoy B, Shriver MD. 2006. Genetic Evidence for the Convergent Evolution of Light Skin in Europeans and East Asians. Mol Biol Evol 24:710–722.

Novembre J, Stephens M. 2008. Interpreting principal component analyses of spatial population genetic variation. Nat Genet 40:646–649.

Nowick K, Fields C, Gernat T, Caetano-Anolles D, Kholina N, Stubbs L. 2011. Gain, Loss and Divergence in Primate Zinc-Finger Genes: A Rich Resource for Evolution of Gene Regulatory Differences between Species. PLoS ONE 6:e21553.

Nychka D, Furrer R, Paige J, Sain S. 2017. fields: Tools for spatial data. Boulder, CO, USA: University Corporation for Atmospheric Research. Available from: www.image.ucar.edu/nychka/Fields

O'Connor TD, Fu W, Mychaleckyj JC, Logsdon B, Auer P, Carlson CS, Leal SM, Smith JD, Rieder MJ, Bamshad MJ, Nickerson DA, Akey JM. 2015. Rare Variation Facilitates Inferences of Fine-Scale Population Structure in Humans. Mol Biol Evol 32:653–660.

Oden NL, Sokal RR. 1986. Directional Autocorrelation: An Extension of Spatial Correlograms to Two Dimensions. Syst Zool 35:608.

O'Hara RB, Kotze DJ. 2010. Do not log-transform count data: Do not log-transform count data. Methods Ecol Evol 1:118–122.

Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Stevens MHH, Szoecs E, Wagner H. 2017. vegan: Community Ecology Package. Available from: https://CRAN.R-project.org/package=vegan

O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, Tatusova T, DiCuccio M, Kitts P, Murphy TD, Pruitt KD. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 44:D733–D745.

Oliveira TYK, Harris EE, Meyer D, Jue CK, Silva WA. 2012. Molecular evolution of a malaria resistance gene (DARC) in primates. Immunogenetics 64:497–505.

Olson MV. 1999. When Less Is More: Gene Loss as an Engine of Evolutionary Change. Am J Hum Genet 64:18–23.

Orlando L, Gilbert MTP, Willerslev E. 2015. Reconstructing ancient genomes and epigenomes. Nat Rev Genet 16:395–408.

Osada N, Miyagi R, Takahashi A. 2017. *Cis* - and *Trans* -regulatory Effects on Gene Expression in a Natural Population of *Drosophila melanogaster*. Genetics 206:2139–2148.

Ouedraogo M, Bettembourg C, Bretaudeau A, Sallou O, Diot C, Demeure O, Lecerf F. 2012. The Duplicated Genes Database: Identification and Functional Annotation of Co-Localised Duplicated Genes across Genomes. PLoS ONE 7:e50653.

Ouwens KG, Jansen R, Tolhuis B, Slagboom PE, Penninx BWJH, Boomsma DI. 2018. A characterization of postzygotic mutations identified in monozygotic twins. Hum Mutat 39:1393–1401.

van Oven M, Kayser M. 2009. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum Mutat 30:E386–E394.

Pääbo S. 2015. The diverse origins of the human gene pool. Nat Rev Genet 16:313–314.

Pagani L, Lawson DJ, Jagoda E, Mörseburg A, Eriksson A, Mitt M, Clemente F, Hudjashov G, DeGiorgio M, Saag L, Wall JD, Cardona A, Mägi R, Sayres MAW, Kaewert S, Inchley C, Scheib CL, Järve M, Karmin M, Jacobs GS, Antao T, Iliescu FM, Kushniarevich A, Ayub Q, Tyler-Smith C, Xue Y, Yunusbayev B, Tambets K, Mallick CB, Saag L, Pocheshkhova E, Andriadze G, Muller C, Westaway MC, Lambert DM, Zoraqi G, Turdikulova S, Dalimova D, Sabitov Z, Sultana GNN, Lachance J, Tishkoff S, Momynaliev K, Isakova J, Damba LD, Gubina M, Nymadawa P, Evseeva I, Atramentova L, Utevska O, Ricaut F-X, Brucato N, Sudoyo H, Letellier T, Cox MP, Barashkov NA, Škaro V, Mulahasanovic´ L, Primorac D, Sahakyan H, Mormina M, Eichstaedt CA, Lichman DV, Abdullah S, Chaubey G, Wee JTS, Mihailov E, Karunas A, Litvinov S, Khusainova R, Ekomasova N, Akhmetova V, Khidiyatova I, Marjanović D, Yepiskoposyan L, Behar DM, Balanovska E, Metspalu A, Derenko M, Malyarchuk B, Voevoda M, Fedorova SA, Osipova LP, Lahr MM, Gerbault P, Leavesley M, Migliano AB, Petraglia M, Balanovsky O, Khusnutdinova EK, Metspalu E, Thomas MG, Manica A, Nielsen R, Villems R, Willerslev E, Kivisild T, Metspalu M. 2016a. Genomic analyses inform on migration events during the peopling of Eurasia. Nature [Internet]. Available from: http://www.nature.com/doifinder/10.1038/nature19792

Pagani L, Lawson DJ, Jagoda E, Mörseburg A, Eriksson A, Mitt M, Clemente F, Hudjashov G, DeGiorgio M, Saag L, Wall JD, Cardona A, Mägi R, Sayres MAW, Kaewert S, Inchley C, Scheib CL, Järve M, Karmin M, Jacobs GS, Antao T, Iliescu FM, Kushniarevich A, Ayub Q, Tyler-Smith C, Xue Y, Yunusbayev B, Tambets K, Mallick CB, Saag L, Pocheshkhova E, Andriadze G, Muller C, Westaway MC, Lambert DM, Zoraqi G, Turdikulova S, Dalimova D, Sabitov Z, Sultana GNN, Lachance J, Tishkoff S, Momynaliev K, Isakova J, Damba LD, Gubina M, Nymadawa P, Evseeva I, Atramentova L, Utevska O, Ricaut F-X, Brucato N, Sudoyo H, Letellier T, Cox MP, Barashkov NA, Škaro V, Mulahasanovic´ L, Primorac D, Sahakyan H, Mormina M, Eichstaedt CA, Lichman DV, Abdullah S, Chaubey G, Wee JTS, Mihailov E, Karunas A, Litvinov S, Khusainova R, Ekomasova N, Akhmetova V, Khidiyatova I, Marjanović D, Yepiskoposyan L, Behar DM, Balanovska E, Metspalu A, Derenko M, Malyarchuk B, Voevoda M, Fedorova SA, Osipova LP, Lahr MM, Gerbault P, Leavesley M, Migliano AB, Petraglia M, Balanovsky O, Khusnutdinova EK, Metspalu E, Thomas MG,

Manica A, Nielsen R, Villems R, Willerslev E, Kivisild T, Metspalu M. 2016b. Genomic analyses inform on migration events during the peopling of Eurasia. Nature 538:238–242.

Pagani L, Schiffels S, Gurdasani D, Danecek P, Scally A, Chen Y, Xue Y, Haber M, Ekong R, Oljira T, Mekonnen E, Luiselli D, Bradman N, Bekele E, Zalloua P, Durbin R, Kivisild T, Tyler-Smith C. 2015. Tracing the route of modern humans out of Africa by using 225 human genome sequences from Ethiopians and Egyptians. Am J Hum Genet 96:986–991.

Palamara PF, Lencz T, Darvasi A, Pe'er I. 2012. Length Distributions of Identity by Descent Reveal Fine-Scale Demographic History. Am J Hum Genet 91:809–822.

Palo JU, Ulmanen I, Lukka M, Ellonen P, Sajantila A. 2009. Genetic markers and population history: Finland revisited. Eur J Hum Genet 17:1336–1346.

Parkin DM, Mesher D, Sasieni P. 2011. 13. Cancers attributable to solar (ultraviolet) radiation exposure in the UK in 2010. Br J Cancer 105:S66–S69.

Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, Genschoreck T, Webster T, Reich D. 2012. Ancient Admixture in Human History. Genetics 192:1065–1093.

Pavlidis P, Jensen JD, Stephan W, Stamatakis A. 2012. A Critical Assessment of Storytelling: Gene Ontology Categories and the Importance of Validating Genomic Scans. Mol Biol Evol 29:3237–3248.

Pawley A, Ross M. 1993. Austronesian Historical Linguistics and Culture History. Annu Rev Anthropol 22:425–459.

Peischl S, Dupanloup I, Kirkpatrick M, Excoffier L. 2013. On the accumulation of deleterious mutations during range expansions. Mol Ecol 22:5972–5982.

Peischl S, Excoffier L. 2015. Expansion load: recessive mutations and the role of standing genetic variation. Mol Ecol 24:2084–2094.

Pemberton TJ, Absher D, Feldman MW, Myers RM, Rosenberg NA, Li JZ. 2012. Genomic Patterns of Homozygosity in Worldwide Human Populations. Am J Hum Genet 91:275–292.

Pengelly RJ, Tapper W, Gibson J, Knut M, Tearle R, Collins A, Ennis S. 2015. Whole genome sequences are required to fully resolve the linkage disequilibrium structure of human populations. BMC Genomics [Internet] 16. Available from: http://www.biomedcentral.com/1471-2164/16/666

Peter BM. 2016. Admixture, Population Structure, and F-Statistics. Genetics 202:1485–1501.

Petit N, Barbadilla A. 2009. Selection efficiency and effective population size in *Drosophila* species. J Evol Biol 22:515–526.

Pickrell JK, Coop G, Novembre J, Kudaravalli S, Li JZ, Absher D, Srinivasan BS, Barsh GS, Myers RM, Feldman MW, Pritchard JK. 2009a. Signals of recent positive selection in a worldwide sample of human populations. Genome Res 19:826–837.

Pickrell JK, Patterson N, Barbieri C, Berthold F, Gerlach L, Güldemann T, Kure B, Mpoloka SW, Nakagawa H, Naumann C, Lipson M, Loh P-R, Lachance J, Mountain J, Bustamante CD, Berger B, Tishkoff SA, Henn BM, Stoneking M, Reich D, Pakendorf B. 2012. The genetic prehistory of southern Africa. Nat Commun 3:1143.

Pickrell JK, Pritchard JK. 2012. Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. PLoS Genet 8:e1002967.

Pickrell JK, Reich D. 2014. Toward a new history and geography of human genes informed by ancient DNA. Trends Genet TIG 30:377–389.

Pierron D, Razafindrazaka H, Pagani L, Ricaut F-X, Antao T, Capredon M, Sambo C, Radimilahy C, Rakotoarisoa J-A, Blench RM, Letellier T, Kivisild T. 2014. Genome-wide evidence of Austronesian-Bantu admixture and cultural reversion in a hunter-gatherer group of Madagascar. Proc Natl Acad Sci 111:936–941.

Pinhasi R, Stock J eds. 2011. Human bioarchaeology of the transition to agriculture. Chichester ; Hoboken, NJ: Wiley-Blackwell.

Plagnol V, Wall JD. 2006. Possible Ancestral Structure in Human Populations. PLoS Genet 2:e105.

Pneuman A, Budimlija ZM, Caragine T, Prinz M, Wurmbach E. 2012. Verification of eye and skin color predictors in various populations. Leg Med Tokyo Jpn 14:78–83.

Pool JE, Nielsen R. 2008. Inference of Historical Changes in Migration Rate From the Lengths of Migrant Tracts. Genetics 181:711–719.

Poolman EM, Galvani AP. 2007. Evaluating candidate agents of selective pressure for cystic fibrosis. J R Soc Interface 4:91–98.

Popejoy AB, Fullerton SM. 2016. Genomics is failing on diversity. Nature 538:161–164.

Posth C, Nägele K, Colleran H, Valentin F, Bedford S, Kami KW, Shing R, Buckley H, Kinaston R, Walworth M, Clark GR, Reepmeyer C, Flexner J, Maric T, Moser J, Gresky J, Kiko L, Robson KJ, Auckland K, Oppenheimer SJ, Hill AVS, Mentzer AJ, Zech J, Petchey F, Roberts P, Jeong C, Gray RD, Krause J, Powell A. 2018. Language continuity despite population replacement in Remote Oceania. Nat Ecol Evol 2:731–740.

Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B, Veeramah KR, Woerner AE, O'Connor TD, Santpere G, Cagan A, Theunert C, Casals F, Laayouni H, Munch K, Hobolth A, Halager AE, Malig M, Hernandez-Rodriguez J, Hernando-Herraez I, Prüfer K, Pybus M, Johnstone L, Lachmann M, Alkan C, Twigg D, Petit N, Baker C, Hormozdiari F, Fernandez-Callejo M, Dabad M, Wilson ML, Stevison L, Camprubí C, Carvalho T, Ruiz-Herrera A, Vives L, Mele M, Abello T, Kondova I, Bontrop RE, Pusey A, Lankester F, Kiyang JA, Bergl RA, Lonsdorf E, Myers S, Ventura M, Gagneux P, Comas D, Siegismund H, Blanc J, Agueda-Calpena L, Gut M, Fulton L, Tishkoff SA, Mullikin JC, Wilson RK, Gut IG, Gonder MK, Ryder OA, Hahn BH, Navarro A, Akey JM, Bertranpetit J, Reich D, Mailund T, Schierup MH, Hvilsom C, Andrés AM, Wall JD, Bustamante CD, Hammer MF, Eichler EE, Marques-Bonet T. 2013. Great ape genetic diversity and population history. Nature 499:471–475.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38:904–909.

Pritchard JK, Di Rienzo A. 2010. Adaptation- Not by sweep alone. Nat Rev Genet 11:665–667.

Pritchard JK, Pickrell JK, Coop G. 2010. The Genetics of Human Adaptation: Hard Sweeps, Soft Sweeps, and Polygenic Adaptation. Curr Biol CB 20:R208–R215.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. Genetics 155:945–959.

Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, Heinze A, Renaud G, Sudmant PH, de Filippo C, Li H, Mallick S, Dannemann M, Fu Q, Kircher M, Kuhlwilm M, Lachmann M, Meyer M, Ongyerth M, Siebauer M, Theunert C, Tandon A, Moorjani P, Pickrell J, Mullikin JC, Vohr SH, Green RE, Hellmann I, Johnson PLF, Blanche H, Cann H, Kitzman JO, Shendure J, Eichler EE, Lein ES, Bakken TE, Golovanova LV, Doronichev VB, Shunkov MV, Derevianko AP, Viola B, Slatkin M, Reich D, Kelso J, Pääbo S. 2013. The complete genome sequence of a Neanderthal from the Altai Mountains. Nature 505:43–49.

Prugnolle F, Manica A, Balloux F. 2005. Geography predicts neutral genetic diversity of human populations. Curr Biol 15:R159–R160.

Pugach I, Delfin F, Gunnarsdóttir E, Kayser M, Stoneking M. 2013. Genome-wide data substantiate Holocene gene flow from India to Australia. Proc Natl Acad Sci:201211927.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC. 2007. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am J Hum Genet 81:559–575.

Quintana-Murci L, Clark AG. 2013. Population genetic tools for dissecting innate immunity in humans. Nat Rev Immunol 13:280–293.

R Core Team. 2017. R: A language and environment for statistical computing. Available from: https://www.R-project.org/

Racimo F, Sankararaman S, Nielsen R, Huerta-Sánchez E. 2015. Evidence for archaic adaptive introgression in humans. Nat Rev Genet 16:359–371.

Racimo F, Schraiber JG. 2014. Approximation to the Distribution of Fitness Effects across Functional Categories in Human Segregating Polymorphisms. PLoS Genet 10:e1004697.

Raghavan M, Skoglund P, Graf KE, Metspalu M, Albrechtsen A, Moltke I, Rasmussen S, Stafford Jr TW, Orlando L, Metspalu E, Karmin M, Tambets K, Rootsi S, Mägi R, Campos PF, Balanovska E, Balanovsky O, Khusnutdinova E, Litvinov S, Osipova LP, Fedorova SA, Voevoda MI, DeGiorgio M, Sicheritz-Ponten T, Brunak S, Demeshchenko S, Kivisild T, Villems R, Nielsen R, Jakobsson M, Willerslev E. 2013. Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. Nature 505:87–91.

Raghavan M, Skoglund P, Graf KE, Metspalu M, Albrechtsen A, Moltke I, Rasmussen S, Stafford TW, Orlando L, Metspalu E, Karmin M, Tambets K, Rootsi S, Mägi R, Campos PF,

Balanovska E, Balanovsky O, Khusnutdinova E, Litvinov S, Osipova LP, Fedorova SA, Voevoda MI, DeGiorgio M, Sicheritz-Ponten T, Brunak S, Demeshchenko S, Kivisild T, Villems R, Nielsen R, Jakobsson M, Willerslev E. 2014. Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. Nature 505:87–91.

Raghavan M, Steinrucken M, Harris K, Schiffels S, Rasmussen S, DeGiorgio M, Albrechtsen A, Valdiosera C, Avila-Arcos MC, Malaspinas A-S, Eriksson A, Moltke I, Metspalu M, Homburger JR, Wall J, Cornejo OE, Moreno-Mayar JV, Korneliussen TS, Pierre T, Rasmussen M, Campos PF, Damgaard P d. B, Allentoft ME, Lindo J, Metspalu E, Rodriguez-Varela R, Mansilla J, Henrickson C, Seguin-Orlando A, Malmstrom H, Stafford T, Shringarpure SS, Moreno-Estrada A, Karmin M, Tambets K, Bergstrom A, Xue Y, Warmuth V, Friend AD, Singarayer J, Valdes P, Balloux F, Leboreiro I, Vera JL, Rangel-Villalobos H, Pettener D, Luiselli D, Davis LG, Heyer E, Zollikofer CPE, Ponce de Leon MS, Smith CI, Grimes V, Pike K-A, Deal M, Fuller BT, Arriaza B, Standen V, Luz MF, Ricaut F, Guidon N, Osipova L, Voevoda MI, Posukh OL, Balanovsky O, Lavryashina M, Bogunov Y, Khusnutdinova E, Gubina M, Balanovska E, Fedorova S, Litvinov S, Malyarchuk B, Derenko M, Mosher MJ, Archer D, Cybulski J, Petzelt B, Mitchell J, Worl R, Norman PJ, Parham P, Kemp BM, Kivisild T, Tyler-Smith C, Sandhu MS, Crawford M, Villems R, Smith DG, Waters MR, Goebel T, Johnson JR, Malhi RS, Jakobsson M, Meltzer DJ, Manica A, Durbin R, Bustamante CD, et al. 2015. Genomic evidence for the Pleistocene and recent population history of Native Americans. Science 349:aab3884–aab3884.

Raj SM, Pagani L, Gallego Romero I, Kivisild T, Amos W. 2013. A general linear model-based approach for inferring selection to climate. BMC Genet 14:87.

Ralph P, Coop G. 2013. The Geography of Recent Genetic Ancestry across Europe. PLoS Biol 11:e1001555.

Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL. 2005. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proc Natl Acad Sci 102:15942–15947.

Ramstetter MD, Dyer TD, Lehman DM, Curran JE, Duggirala R, Blangero J, Mezey JG, Williams AL. 2017. Benchmarking Relatedness Inference Methods with Genome-Wide Data from Thousands of Relatives. Genetics 207:75–82.

Rappaport N, Nativ N, Stelzer G, Twik M, Guan-Golan Y, Iny Stein T, Bahir I, Belinky F, Morrey CP, Safran M, Lancet D. 2013. MalaCards: an integrated compendium for diseases and their annotation. Database 2013:bat018–bat018.

Rasmussen M, Guo X, Wang Y, Lohmueller KE, Rasmussen S, Albrechtsen A, Skotte L, Lindgreen S, Metspalu M, Jombart T, Kivisild T, Zhai W, Eriksson A, Manica A, Orlando L, De La Vega FM, Tridico S, Metspalu E, Nielsen K, Avila-Arcos MC, Moreno-Mayar JV, Muller C, Dortch J, Gilbert MTP, Lund O, Wesolowska A, Karmin M, Weinert LA, Wang B, Li J, Tai S, Xiao F, Hanihara T, van Driem G, Jha AR, Ricaut F-X, de Knijff P, Migliano AB, Gallego Romero I, Kristiansen K, Lambert DM, Brunak S, Forster P, Brinkmann B, Nehlich O, Bunce M, Richards M, Gupta R, Bustamante CD, Krogh A, Foley RA, Lahr MM, Balloux F, Sicheritz-Ponten T, Villems R, Nielsen R, Wang J, Willerslev E. 2011. An Aboriginal Australian Genome Reveals Separate Human Dispersals into Asia. Science 334:94–98.

Razeto-Barry P, Diaz J, Vasquez RA. 2012. The Nearly Neutral and Selection Theories of Molecular Evolution Under the Fisher Geometrical Framework: Substitution Rate, Population Size, and Complexity. Genetics 191:523–534.

Reich D, Green RE, Kircher M, Krause J, Patterson N, Durand EY, Viola B, Briggs AW, Stenzel U, Johnson PLF, Maricic T, Good JM, Marques-Bonet T, Alkan C, Fu Q, Mallick S, Li H, Meyer M, Eichler EE, Stoneking M, Richards M, Talamo S, Shunkov MV, Derevianko AP, Hublin J-J, Kelso J, Slatkin M, Pääbo S. 2010. Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature 468:1053–1060.

Reich D, Patterson N, Kircher M, Delfin F, Nandineni MR, Pugach I, Ko AM-S, Ko Y-C, Jinam TA, Phipps ME, Saitou N, Wollstein A, Kayser M, Pääbo S, Stoneking M. 2011. Denisova Admixture and the First Modern Human Dispersals into Southeast Asia and Oceania. Am J Hum Genet 89:516–528.

Reich D, Thangaraj K, Patterson N, Price AL, Singh L. 2009. Reconstructing Indian population history. Nature 461:489–494.

Reppell M, Boehnke M, Zollner S. 2014. The Impact of Accelerating Faster than Exponential Population Growth on Genetic Variation. Genetics 196:819–828.

Revelle W. 2018. psych: Procedures for Psychological, Psychometric, and Personality Research. Evanston, Illinois: Northwestern University. Available from: https://CRAN.R-project.org/package=psych

Richter D, Grün R, Joannes-Boyau R, Steele TE, Amani F, Rué M, Fernandes P, Raynal J-P, Geraads D, Ben-Ncer A, Hublin J-J, McPherron SP. 2017. The age of the hominin fossils from Jebel Irhoud, Morocco, and the origins of the Middle Stone Age. Nature 546:293–296.

Rieber N, Zapatka M, Lasitschka B, Jones D, Northcott P, Hutter B, Jäger N, Kool M, Taylor M, Lichter P, Pfister S, Wolf S, Brors B, Eils R. 2013. Coverage Bias and Sensitivity of Variant Calling for Four Whole-genome Sequencing Technologies. PLoS ONE 8:e66621.

Ringbauer H, Coop G, Barton NH. 2017. Inferring Recent Demography from Isolation by Distance of Long Shared Sequence Blocks. Genetics 205:1335–1351.

Riordan JD, Nadeau JH. 2017. From Peas to Disease: Modifier Genes, Network Resilience, and the Genetics of Health. Am J Hum Genet 101:177–191.

Rivas MA, Pirinen M, Conrad DF, Lek M, Tsang EK, Karczewski KJ, Maller JB, Kukurba KR, DeLuca DS, Fromer M, Ferreira PG, Smith KS, Zhang R, Zhao F, Banks E, Poplin R, Ruderfer DM, Purcell SM, Tukiainen T, Minikel EV, Stenson PD, Cooper DN, Huang KH, Sullivan TJ, Nedzel J, The GTEx Consortium, The Geuvadis Consortium, Bustamante CD, Li JB, Daly MJ, Guigo R, Donnelly P, Ardlie K, Sammeth M, Dermitzakis ET, McCarthy MI, Montgomery SB, Lappalainen T, MacArthur DG, Segre AV, Young TR, Gelfand ET, Trowbridge CA, Ward LD, Kheradpour P, Iriarte B, Meng Y, Palmer CD, Esko T, Winckler W, Hirschhorn J, Kellis M, Getz G, Shablin AA, Li G, Zhou Y-H, Nobel AB, Rusyn I, Wright FA, Battle A, Mostafavi S, Mele M, Reverter F, Goldmann J, Koller D, Gamazon ER, Im HK, Konkashbaev A, Nicolae DL, Cox NJ, Flutre T, Wen X, Stephens M, Pritchard JK, Tu Z, Zhang B, Huang T, Long Q, Lin L, Yang J, Zhu J, Liu J, Brown A, Mestichelli B, Tidwell D, Lo E, Salvatore M, Shad S, Thomas JA, Lonsdale JT, Choi RC, Karasik E, Ramsey K, Moser MT, Foster BA, Gillard BM,

Syron J, Fleming J, et al. 2015. Effect of predicted protein-truncating genetic variants on the human transcriptome. Science 348:666–669.

Roach JC, Glusman G, Smit AFA, Huff CD, Hubley R, Shannon PT, Rowen L, Pant KP, Goodman N, Bamshad M, Shendure J, Drmanac R, Jorde LB, Hood L, Galas DJ. 2010. Analysis of Genetic Inheritance in a Family Quartet by Whole-Genome Sequencing. Science 328:636–639.

Romero-Hidalgo S, Ochoa-Leyva A, Garcíarrubio A, Acuña-Alonzo V, Antúnez-Argüelles E, Balcazar-Quintero M, Barquera-Lozano R, Carnevale A, Cornejo-Granados F, Fernández-López JC, García-Herrera R, García-Ortíz H, Granados-Silvestre Á, Granados J, Guerrero-Romero F, Hernández-Lemus E, León-Mimila P, Macín-Pérez G, Martínez-Hernández A, Menjivar M, Morett E, Orozco L, Ortíz-López G, Pérez-Villatoro F, Rivera-Morales J, Riveros-McKay F, Villalobos-Comparán M, Villamil-Ramírez H, Villarreal-Molina T, Canizales-Quinteros S, Soberón X. 2017. Demographic history and biologically relevant genetic variation of Native Mexicans inferred from whole-genome sequencing. Nat Commun [Internet] 8. Available from: http://www.nature.com/articles/s41467-017-01194-z

Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH, McCarroll SA, Gaudet R, Schaffner SF, Lander ES, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, Pasternak S, Wheeler DA, Willis TD, Yu F, Yang H, Zeng C, Gao Y, Hu H, Hu W, Li C, Lin W, Liu S, Pan H, Tang X, Wang J, Wang W, Yu J, Zhang B, Zhang Q, Zhao H, Zhao H, Zhou J, Gabriel SB, Barry R, Blumenstiel B, Camargo A, Defelice M, Faggart M, Goyette M, Gupta S, Moore J, Nguyen H, Onofrio RC, Parkin M, Roy J, Stahl E, Winchester E, Ziaugra L, Altshuler D, Shen Y, Yao Z, Huang W, Chu X, He Y, Jin L, Liu Y, Shen Y, Sun W, Wang H, Wang Y, Wang Y, Xiong X, Xu L, Waye MMY, Tsui SKW, Xue H, Wong JT-F, Galver LM, Fan J-B, Gunderson K, Murray SS, Oliphant AR, Chee MS, Montpetit A, Chagnon F, Ferretti V, Leboeuf M, Olivier J-F, Phillips MS, Roumy S, Sallée C, Verner A, Hudson TJ, Kwok P-Y, Cai D, Koboldt DC, et al. 2007. Genome-wide detection and characterization of positive selection in human populations. Nature 449:913–918.

Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, Mullikin JC, Mortimore BJ, Willey DL, Hunt SE, Cole CG, Coggill PC, Rice CM, Ning Z, Rogers J, Bentley DR, Kwok P-Y, Mardis ER, Yeh RT, Schultz B, Cook L, Davenport R, Dante M, Fulton L, Hillier L, Waterston RH, McPherson JD, Gilman B, Schaffner S, Van Etten WJ, Reich D, Higgins J, Daly MJ, Blumenstiel B, Baldwin J, Stange-Thomann N, Zody MC, Linton L, Lander ES, Altshuler D, The International SNP Map Working Group, Cold Spring Harbor Laboratories:, National Center for Biotechnology Information:, The Sanger Centre:, Washington University in St. Louis:, Whitehead/MIT Center for Genome Research: 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409:928–933.

Saleheen D, Natarajan P, Armean IM, Zhao W, Rasheed A, Khetarpal SA, Won H-H, Karczewski KJ, O'Donnell-Luria AH, Samocha KE, Weisburd B, Gupta N, Zaidi M, Samuel M, Imran A, Abbas S, Majeed F, Ishaq M, Akhtar S, Trindade K, Mucksavage M, Qamar N, Zaman KS, Yaqoob Z, Saghir T, Rizvi SNH, Memon A, Hayyat Mallick N, Ishaq M, Rasheed SZ, Memon F-R, Mahmood K, Ahmed N, Do R, Krauss RM, MacArthur DG, Gabriel S, Lander ES, Daly MJ, Frossard P, Danesh J, Rader DJ, Kathiresan S. 2017. Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity. Nature 544:235–239.

Sankararaman S, Mallick S, Dannemann M, Prüfer K, Kelso J, Pääbo S, Patterson N, Reich D. 2014. The genomic landscape of Neanderthal ancestry in present-day humans. Nature 507:354–357.

Sankararaman S, Mallick S, Patterson N, Reich D. 2016. The Combined Landscape of Denisovan and Neanderthal Ancestry in Present-Day Humans. Curr Biol 26:1241–1247.

Sankararaman S, Patterson N, Li H, Pääbo S, Reich D. 2012. The Date of Interbreeding between Neandertals and Modern Humans. PLoS Genet 8:e1002947.

Sarkisyan KS, Bolotin DA, Meer MV, Usmanova DR, Mishin AS, Sharonov GV, Ivankov DN, Bozhanova NG, Baranov MS, Soylemez O, Bogatyreva NS, Vlasov PK, Egorov ES, Logacheva MD, Kondrashov AS, Chudakov DM, Putintseva EV, Mamedov IZ, Tawfik DS, Lukyanov KA, Kondrashov FA. 2016. Local fitness landscape of the green fluorescent protein. Nature 533:397–401.

Sawyer S, Renaud G, Viola B, Hublin J-J, Gansauge M-T, Shunkov MV, Derevianko AP, Prüfer K, Kelso J, Pääbo S. 2015. Nuclear and mitochondrial DNA sequences from two Denisovan individuals. Proc Natl Acad Sci:201519905.

Sawyer SA, Hartl DL. 1992. Population genetics of polymorphism and divergence. Genetics 132:1161–1176.

Scally A. 2016. The mutation rate in human evolution and demographic inference. Curr Opin Genet Dev 41:36–43.

Scally A, Durbin R. 2012. Revising the human mutation rate: implications for understanding human evolution. Nat Rev Genet 13:745–753.

Scheinfeldt LB, Tishkoff SA. 2013. Recent human adaptation: genomic approaches, interpretation and insights. Nat Rev Genet 14:692–702.

Schiffels S, Haak W, Paajanen P, Llamas B, Popescu E, Loe L, Clarke R, Lyons A, Mortimer R, Sayer D, Tyler-Smith C, Cooper A, Durbin R. 2016. Iron Age and Anglo-Saxon genomes from East England reveal British migration history. Nature Communications [Internet] 7. Available from: http://www.nature.com/articles/ncomms10408

Schiffels S, Durbin R. 2014. Inferring human population size and separation history from multiple genome sequences. Nat Genet 46:919–925.

Schlebusch CM, Malmström H, Günther T, Sjödin P, Coutinho A, Edlund H, Munters AR, Vicente M, Steyn M, Soodyall H, Lombard M, Jakobsson M. 2017. Southern African ancient genomes estimate modern human divergence to 350,000 to 260,000 years ago. Science 358:652–655.

Scholz F, Zhu A. 2018. kSamples: K-sample Rank Tests and their Combinations. R Package version. Available from: https://CRAN.R-project.org/package=kSamples

Scholz FW, Stephens MA. 1987. K-Sample Anderson-Darling Tests. J Am Stat Assoc 82:918–924.

Schraiber JG, Akey JM. 2015. Methods and models for unravelling human evolutionary history. Nat Rev Genet 16:727–740.

Seguin-Orlando A, Korneliussen TS, Sikora M, Malaspinas A-S, Manica A, Moltke I, Albrechtsen A, Ko A, Margaryan A, Moiseyev V, Goebel T, Westaway M, Lambert D, Khartanovich V, Wall JD, Nigst PR, Foley RA, Lahr MM, Nielsen R, Orlando L, Willerslev E. 2014. Genomic structure in Europeans dating back at least 36,200 years. Science 346:1113–1118.

Seo J-S, Rhie A, Kim J, Lee S, Sohn M-H, Kim C-U, Hastie A, Cao H, Yun J-Y, Kim J, Kuk J, Park GH, Kim J, Ryu H, Kim J, Roh M, Baek J, Hunkapiller MW, Korlach J, Shin J-Y, Kim C. 2016. De novo assembly and phasing of a Korean human genome. Nature [Internet]. Available from: http://www.nature.com/doifinder/10.1038/nature20098

Shamseldin HE, Tulbah M, Kurdi W, Nemer M, Alsahan N, Al Mardawi E, Khalifa O, Hashem A, Kurdi A, Babay Z, Bubshait DK, Ibrahim N, Abdulwahab F, Rahbeeni Z, Hashem M, Alkuraya FS. 2015. Identification of embryonic lethal genes in humans by autozygosity mapping and exome sequencing in consanguineous families. Genome Biol [Internet] 16. Available from: http://genomebiology.com/2015/16/1/116

Sharp K, Kretzschmar W, Delaneau O, Marchini J. 2016. Phasing for medical sequencing using rare variants and large haplotype reference panels. Bioinformatics 32:1974–1980.

Shen Y-Y, Liang L, Li G-S, Murphy RW, Zhang Y-P. 2012. Parallel Evolution of Auditory Genes for Echolocation in Bats and Toothed Whales. PLoS Genet 8:e1002788.

Shendure J, Akey JM. 2015. The origins, determinants, and consequences of human mutations. Science 349:1478–1483.

Shlyakhter I, Sabeti PC, Schaffner SF. 2014. Cosi2: an efficient simulator of exact and approximate coalescent with selection. Bioinformatics 30:3427–3429.

Siepel A, Pollard KS, Haussler D. 2006. New Methods for Detecting Lineage-Specific Selection. In: Apostolico A, Guerra C, Istrail S, Pevzner PA, Waterman M, editors. Research in Computational Molecular Biology. Vol. 3909. Berlin, Heidelberg: Springer Berlin Heidelberg. p 190–205. Available from: http://link.springer.com/10.1007/11732990_17

Siewert KM, Voight BF. 2017. Detecting Long-Term Balancing Selection Using Allele Frequency Correlation. Mol Biol Evol 34:2996–3005.

Siggers T, Reddy J, Barron B, Bulyk ML. 2014. Diversification of Transcription Factor Paralogs via Noncanonical Modularity in C2H2 Zinc Finger DNA Binding. Mol Cell 55:640–648.

Sikora M, Pitulko V, Sousa V, Allentoft ME, Vinner L, Rasmussen S, Margaryan A, de Barros Damgaard P, de la Fuente Castro C, Renaud G, Yang M, Fu Q, Dupanloup I, Giampoudakis K, Bravo Nogues D, Rahbek C, Kroonen G, Peyrot M, McColl H, Vasilyev S, Veselovskaya E, Gerasimova M, Pavlova E, Chasnyk V, Nikolskiy P, Grebenyuk P, Fedorchenko A, Lebedintsev A, Malyarchuk B, Meldgaard M, Martiniano R, Arppe L, Palo J, Sundell T, Mannermaa K, Putkonen M, Alexandersen V, Primeau C, Mahli R, Sjögren K-G, Kristiansen K, Wessman A, Sajantila A, Mirazon Lahr M, Durbin R, Nielsen R, Meltzer D, Excoffier L,

Willerslev E. 2018. The population history of northeastern Siberia since the Pleistocene. Available from: http://biorxiv.org/lookup/doi/10.1101/448829

Simons YB, Sella G. 2016. The impact of recent population history on the deleterious mutation load in humans and close evolutionary relatives. Curr Opin Genet Dev 41:150–158.

Simons YB, Turchin MC, Pritchard JK, Sella G. 2014. The deleterious mutation load is insensitive to recent population history. Nat Genet 46:220–224.

Skoglund P, Posth C, Sirak K, Spriggs M, Valentin F, Bedford S, Clark GR, Reepmeyer C, Petchey F, Fernandes D, Fu Q, Harney E, Lipson M, Mallick S, Novak M, Rohland N, Stewardson K, Abdullah S, Cox MP, Friedlaender FR, Friedlaender JS, Kivisild T, Koki G, Kusuma P, Merriwether DA, Ricaut F-X, Wee JTS, Patterson N, Krause J, Pinhasi R, Reich D. 2016. Genomic insights into the peopling of the Southwest Pacific. Nature 538:510–513.

Slatkin M, Racimo F. 2016. Ancient DNA and human history. Proc Natl Acad Sci 113:6380–6387.

Smith BD. 1998. The emergence of agriculture. Paperback ed. New York: Scientific American Library.

Smith JM, Haigh J. 1974. The hitch-hiking effect of a favourable gene. Genet Res 23:23–35.

Soares P, Rito T, Trejaut J, Mormina M, Hill C, Tinkler-Hundal E, Braid M, Clarke DJ, Loo J-H, Thomson N, Denham T, Donohue M, Macaulay V, Lin M, Oppenheimer S, Richards MB. 2011. Ancient Voyaging and Polynesian Origins. Am J Hum Genet 88:239–247.

Sokal RR, Thomson BA. 1998. Spatial genetic structure of human populations in Japan. Hum Biol 70:1–22.

Sokal RR, Wartenberg DE. 1983. A Test of Spatial Autocorrelation Analysis Using an Isolation-by-Distance Model. Genetics 105:219–237.

Sousa V, Peischl S, Excoffier L. 2014. Impact of range expansions on current human genomic diversity. Curr Opin Genet Dev 29:22–30.

Spence JP, Steinrücken M, Terhorst J, Song YS. 2018. Inference of population history using coalescent HMMs: review and outlook. Curr Opin Genet Dev 53:70–76.

Spriggs M. 2011. Archaeology and the Austronesian expansion: where are we now? Antiquity 85:510–528.

Steinrücken M, Kamm JA, Song YS. 2015. Inference of complex population histories using whole-genome sequences from multiple populations. Available from: http://biorxiv.org/lookup/doi/10.1101/026591

Stenson PD, Mort M, Ball EV, Evans K, Hayden M, Heywood S, Hussain M, Phillips AD, Cooper DN. 2017. The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. Hum Genet 136:665–677.

Stringer C, Andrews P. 1988. Genetic and fossil evidence for the origin of modern humans. Science 239:1263–1268.

Sulem P, Helgason H, Oddson A, Stefansson H, Gudjonsson SA, Zink F, Hjartarson E, Sigurdsson GT, Jonasdottir A, Jonasdottir A, Sigurdsson A, Magnusson OT, Kong A, Helgason A, Holm H, Thorsteinsdottir U, Masson G, Gudbjartsson DF, Stefansson K. 2015. Identification of a large set of rare complete human knockouts. Nat Genet 47:448–452.

Sun M, Jobling MA, Taliun D, Pramstaller PP, Egeland T, Sheehan NA. 2016. On the use of dense SNP marker data for the identification of distant relative pairs. Theor Popul Biol 107:14–25.

Sutton JEG. 1977. The African aqualithic. Antiquity 51:25–34.

Tabor HK, Auer PL, Jamal SM, Chong JX, Yu J-H, Gordon AS, Graubert TA, O'Donnell CJ, Rich SS, Nickerson DA, Bamshad MJ. 2014. Pathogenic Variants for Mendelian and Complex Traits in Exomes of 6,517 European and African Americans: Implications for the Return of Incidental Results. Am J Hum Genet 95:183–193.

Taipale M, Kaminen N, Nopola-Hemmi J, Haltia T, Myllyluoma B, Lyytinen H, Muller K, Kaaranen M, Lindsberg PJ, Hannula-Jouppi K, Kere J. 2003. A candidate gene for developmental dyslexia encodes a nuclear tetratricopeptide repeat domain protein dynamically regulated in brain. Proc Natl Acad Sci 100:11553–11558.

Tajima F. 1989. Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. Genetics 123:585–595.

Takahata N. 1993. Allelic genealogy and human evolution. Mol Biol Evol 10:2–22.

Takahata N, Nei M. 1990. Allelic genealogy under overdominant and frequency-dependent selection and polymorphism of major histocompatibility complex loci. Genetics 124:967–978.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. 2013. MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. Mol Biol Evol 30:2725–2729.

Tang H, Peng J, Wang P, Risch NJ. 2005. Estimation of individual admixture: Analytical and study design considerations. Genet Epidemiol 28:289–301.

Tardif S, Brady HA, Breazeale KR, Bi M, Thompson LD, Bruemmer JE, Bailey LB, Hardy DM. 2010. Zonadhesin D3-polypeptides vary among species but are similar in Equus species capable of interbreeding. Biol Reprod 82:413–421.

Templeton AR. 2007. GENETICS AND RECENT HUMAN EVOLUTION. Evolution 61:1507–1519.

Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, McGee S, Do R, Liu X, Jun G, Kang HM, Jordan D, Leal SM, Gabriel S, Rieder MJ, Abecasis G, Altshuler D, Nickerson DA, Boerwinkle E, Sunyaev S, Bustamante CD, Bamshad MJ, Akey JM, Broad GO, Seattle GO, on behalf of the NHLBI Exome Sequencing Project. 2012. Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes. Science 337:64–69.

Teshima KM, Coop G, Przeworski M. 2006. How reliable are empirical genomic scans for selective sweeps? Genome Res 16:702–712.

Teslovich TM, Musunuru K, Smith AV, Edmondson AC, Stylianou IM, Koseki M, Pirruccello JP, Ripatti S, Chasman DI, Willer CJ, Johansen CT, Fouchier SW, Isaacs A, Peloso GM, Barbalic M, Ricketts SL, Bis JC, Aulchenko YS, Thorleifsson G, Feitosa MF, Chambers J, Orho-Melander M, Melander O, Johnson T, Li X, Guo X, Li M, Shin Cho Y, Jin Go M, Jin Kim Y, Lee J-Y, Park T, Kim K, Sim X, Twee-Hee Ong R, Croteau-Chonka DC, Lange LA, Smith JD, Song K, Hua Zhao J, Yuan X, Luan J, Lamina C, Ziegler A, Zhang W, Zee RYL, Wright AF, Witteman JCM, Wilson JF, Willemsen G, Wichmann H-E, Whitfield JB, Waterworth DM, Wareham NJ, Waeber G, Vollenweider P, Voight BF, Vitart V, Uitterlinden AG, Uda M, Tuomilehto J, Thompson JR, Tanaka T, Surakka I, Stringham HM, Spector TD, Soranzo N, Smit JH, Sinisalo J, Silander K, Sijbrands EJG, Scuteri A, Scott J, Schlessinger D, Sanna S, Salomaa V, Saharinen J, Sabatti C, Ruokonen A, Rudan I, Rose LM, Roberts R, Rieder M, Psaty BM, Pramstaller PP, Pichler I, Perola M, Penninx BWJH, Pedersen NL, Pattaro C, Parker AN, Pare G, Oostra BA, O'Donnell CJ, Nieminen MS, Nickerson DA, Montgomery GW, Meitinger T, et al. 2010. Biological, clinical and population relevance of 95 loci for blood lipids. Nature 466:707–713.

The 1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. Nature 467:1061–1073.

The 1000 Genomes Project Consortium. 2012. An integrated map of genetic variation from 1,092 human genomes. Nature 491:56–65.

The 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. Nature 526:68–74.

The Gene Ontology Consortium. 2015. Gene Ontology Consortium: going forward. Nucleic Acids Res 43:D1049–D1056.

Thomson R, Pritchard JK, Shen P, Oefner PJ, Feldman MW. 2000. Recent common ancestry of human Y chromosomes: evidence from DNA sequence data. Proc Natl Acad Sci U S A 97:7360–7365.

Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, Silverman JS, Powell K, Mortensen HM, Hirbo JB, Osman M, Ibrahim M, Omar SA, Lema G, Nyambo TB, Ghori J, Bumpstead S, Pritchard JK, Wray GA, Deloukas P. 2007. Convergent adaptation of human lactase persistence in Africa and Europe. Nat Genet 39:31–40.

Tran C, Gagnon F, Wigg KG, Feng Y, Gomez L, Cate-Carter TD, Kerr EN, Field LL, Kaplan BJ, Lovett MW, Barr CL. 2013. A family-based association analysis and meta-analysis of the reading disabilities candidate gene *DYX1C1*. Am J Med Genet B Neuropsychiatr Genet 162:146–156.

Trejaut JA, Kivisild T, Loo JH, Lee CL, He CL, Hsu CJ, Li ZY, Lin M. 2005. Traces of Archaic Mitochondrial Lineages Persist in Austronesian-Speaking Formosan Populations. PLoS Biol 3:e247.

Trejaut JA, Poloni ES, Yen J-C, Lai Y-H, Loo J-H, Lee C-L, He C-L, Lin M. 2014. Taiwan Y-chromosomal DNA variation and its relationship with Island Southeast Asia. BMC Genet 15:77.

Tsompana M, Buck MJ. 2014. Chromatin accessibility: a window into the genome. Epigenetics Chromatin 7:33.

Turchin MC, Chiang CW, Palmer CD, Sankararaman S, Reich D, Hirschhorn JN. 2012. Evidence of widespread selection on standing variation in Europe at height-associated SNPs. Nat Genet 44:1015–1019.

VanLiere JM, Rosenberg NA. 2008. Mathematical properties of the measure of linkage disequilibrium. Theor Popul Biol 74:130–137.

van der Velde KJ, Kuiper J, Thompson BA, Plazzer J-P, van Valkenhoef G, de Haan M, Jongbloed JDH, Wijmenga C, de Koning TJ, Abbott KM, Sinke R, Spurdle AB, Macrae F, Genuardi M, Sijmons RH, Swertz MA, InSiGHT Group. 2015. Evaluation of CADD Scores in Curated Mismatch Repair Gene Variants Yields a Model for Clinical Validation and Prioritization. Hum Mutat 36:712–719.

Vernot B, Akey JM. 2014. Resurrecting Surviving Neandertal Lineages from Modern Human Genomes. Science 343:1017–1021.

Vignal A, Milan D, SanCristobal M, Eggen A. 2002. A review on SNP and other types of molecular markers and their use in animal genetics. Genet Sel Evol 34:275–305.

Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, Yang J. 2017. 10 Years of GWAS Discovery: Biology, Function, and Translation. Am J Hum Genet 101:5–22.

Vitti JJ, Grossman SR, Sabeti PC. 2013. Detecting natural selection in genomic data. Annu Rev Genet 47:97–120.

Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A Map of Recent Positive Selection in the Human Genome. PLoS Biol 4:e72.

Vuong QH. 1989. Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. Econometrica 57:307.

Vy HMT, Kim Y. 2015. A Composite-Likelihood Method for Detecting Incomplete Selective Sweep from Population Genomic Data. Genetics 200:633–649.

Wagner A. 1998. The fate of duplicated genes: loss or new function? BioEssays 20:785–788.

Wakeley J, Wilton PR. 2016. Coalescent and Models of Identity by Descent. In: Kliman R, editor. Encyclopedia of Evolutionary Biology. Elsevier. p 287–292. Available from: https://linkinghub.elsevier.com/retrieve/pii/B9780128000496000330

Wall JD, Tang LF, Zerbe B, Kvale MN, Kwok P-Y, Schaefer C, Risch N. 2014. Estimating genotype error rates from high-coverage next-generation sequence data. Genome Res 24:1734–1739.

Wang S, Lachance J, Tishkoff SA, Hey J, Xing J. 2013. Apparent Variation in Neanderthal Admixture among African Populations is Consistent with Gene Flow from Non-African Populations. Genome Biol Evol 5:2075–2081.

Weidenreich F. 1947. Facts and Speculations concerning The Origin of Homo Sapiens. Am Anthropol 49:187–203.

Weischenfeldt J, Damgaard I, Bryder D, Theilgaard-Monch K, Thoren LA, Nielsen FC, Jacobsen SEW, Nerlov C, Porse BT. 2008. NMD is essential for hematopoietic stem and progenitor cells and for eliminating by-products of programmed DNA rearrangements. Genes Dev 22:1381–1396.

Weiss KM, Lambert BW. 2014. What Type of Person Are You? Old-Fashioned Thinking Even in Modern Science. Cold Spring Harb Perspect Biol 6:a021238–a021238.

Wellcome Trust Case Control Consortium, Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P, Duncanson A, Kwiatkowski DP, McCarthy MI, Ouwehand WH, Samani NJ, Todd JA, Donnelly P, Barrett JC, Burton PR, Davison D, Donnelly P, Easton D, Evans D, Leung H-T, Marchini JL, Morris AP, Spencer CCA, Tobin MD, Cardon LR, Clayton DG, Attwood AP, Boorman JP, Cant B, Everson U, Hussey JM, Jolley JD, Knight AS, Koch K, Meech E, Nutland S, Prowse CV, Stevens HE, Taylor NC, Walters GR, Walker NM, Watkins NA, Winzer T, Todd JA, Ouwehand WH, Jones RW, McArdle WL, Ring SM, Strachan DP, Pembrey M, Breen G, St Clair D, Caesar S, Gordon-Smith K, Jones L, Fraser C, Green EK, Grozeva D, Hamshere ML, Holmans PA, Jones IR, Kirov G, Moskvina V, Nikolov I, O'Donovan MC, Owen MJ, Craddock N, Collier DA, Elkin A, Farmer A, Williamson R, McGuffin P, Young AH, Ferrier IN, Ball SG, Balmforth AJ, Barrett JH, Bishop DT, Iles MM, Maqbool A, Yuldasheva N, Hall AS, Braund PS, Burton PR, Dixon RJ, Mangino M, Stevens S, Tobin MD, Thompson JR, Samani NJ, Bredin F, Tremelling M, Parkes M, Drummond H, Lees CW, Nimmo ER, Satsangi J, Fisher SA, et al. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447:661–678.

Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). Available from: www.genome.gov/sequencingcostsdata

Wigginton JE, Cutler DJ, Abecasis GR. 2005. A Note on Exact Tests of Hardy-Weinberg Equilibrium. Am J Hum Genet 76:887–893.

Williamson SH, Hernandez R, Fledel-Alon A, Zhu L, Nielsen R, Bustamante CD. 2005. Simultaneous inference of selection and population growth from patterns of variation in the human genome. Proc Natl Acad Sci 102:7882–7887.

Wiuf C, Hein J. 1999. Recombination as a point process along sequences. Theor Popul Biol 55:248–259.

Wollstein A, Lao O, Becker C, Brauer S, Trent RJ, Nürnberg P, Stoneking M, Kayser M. 2010. Demographic history of Oceania inferred from genome-wide data. Curr Biol CB 20:1983–1992.

Wollstein A, Stephan W. 2015. Inferring positive selection in humans from genomic data. Investig Genet 6:5.

Wolpoff MH, Wu X, Thome AG. 1984. Homo sapiens origins: a general theory of hominid evolution involving the fossil evidence from east Asia. In: Smith FH, Spencer F, editors. The Origins of modern humans: a world survey of the fossil evidence. New York: A.R. Liss. p 411–483.

Wong WSW, Solomon BD, Bodian DL, Kothiyal P, Eley G, Huddleston KC, Baker R, Thach DC, Iyer RK, Vockley JG, Niederhuber JE. 2016. New observations on maternal age effect on germline de novo mutations. Nat Commun 7:10486.

Wright S. 1931. Evolution in Mendelian Populations. Genetics 16:97–159.

Wright S. 1943. Isolation by Distance. Genetics 28:114–138.

Wright S. 1951. The genetical structure of populations. Ann Eugen 15:323–354.

Wright S. 1965. The Interpretation of Population Structure by F-Statistics with Special Regard to Systems of Mating. Evolution 19:395.

Wu G-S, Luo H-R, Dong C, Mastronardi C, Licinio J, Wong M-L. 2011. Sequence polymorphisms of MC1R gene and their association with depression and antidepressant response: Psychiatr Genet 21:14–18.

Wu P-Y, Phan JH, Wang MD. 2013. Assessing the impact of human genome annotation choice on RNA-seq expression estimates. BMC Bioinformatics 14:S8.

Xiang-Yu J, Yang Z, Tang K, Li H. 2016. Revisiting the false positive rate in detecting recent positive selection. Quant Biol 4:207–216.

Xu S, Pugach I, Stoneking M, Kayser M, Jin L. 2012. Genetic dating indicates that the Asian–Papuan admixture through Eastern Indonesia corresponds to the Austronesian expansion. Proc Natl Acad Sci:201118892.

Xue Y, Daly A, Yngvadottir B, Liu M, Coop G, Kim Y, Sabeti P, Chen Y, Stalker J, Huckle E, Burton J, Leonard S, Rogers J, Tyler-Smith C. 2006. Spread of an Inactive Form of Caspase-12 in Humans Is Due to Recent Positive Selection. Am J Hum Genet 78:659–670.

Xue Y, Zhang X, Huang N, Daly A, Gillson CJ, MacArthur DG, Yngvadottir B, Nica AC, Woodwark C, Chen Y, Conrad DF, Ayub Q, Mehdi SQ, Li P, Tyler-Smith C. 2009. Population Differentiation as an Indicator of Recent Positive Selection in Humans: An Empirical Evaluation. Genetics 183:1065–1077.

Yang MA, Malaspinas A-S, Durand EY, Slatkin M. 2012. Ancient Structure in Africa Unlikely to Explain Neanderthal and Non-African Genetic Similarity. Mol Biol Evol 29:2987–2995.

Ye K, Lu J, Raj SM, Gu Z. 2013. Human Expression QTLs Are Enriched in Signals of Environmental Adaptation. Genome Biol Evol 5:1689–1701.

Yeretssian G, Doiron K, Shao W, Leavitt BR, Hayden MR, Nicholson DW, Saleh M. 2009. Gender differences in expression of the human caspase-12 long variant determines susceptibility to Listeria monocytogenes infection. Proc Natl Acad Sci 106:9016–9020.

Yi X, Liang Y, Huerta-Sanchez E, Jin X, Cuo ZXP, Pool JE, Xu X, Jiang H, Vinckenbosch N, Korneliussen TS, Zheng H, Liu T, He W, Li K, Luo R, Nie X, Wu H, Zhao M, Cao H, Zou J, Shan Y, Li S, Yang Q, Asan, Ni P, Tian G, Xu J, Liu X, Jiang T, Wu R, Zhou G, Tang M, Qin J, Wang T, Feng S, Li G, Huasang, Luosang J, Wang W, Chen F, Wang Y, Zheng X, Li Z, Bianba Z, Yang G, Wang X, Tang S, Gao G, Chen Y, Luo Z, Gusang L, Cao Z, Zhang Q, Ouyang W, Ren X, Liang H, Zheng H, Huang Y, Li J, Bolund L, Kristiansen K, Li Y, Zhang Y, Zhang X, Li R, Li S, Yang H, Nielsen R, Wang J, Wang J. 2010. Sequencing of 50 Human Exomes Reveals Adaptation to High Altitude. Science 329:75–78.

Yngvadottir B. 2008. Evolution by Gene Loss? PhD dissertation. University of Cambridge

Yngvadottir B, Xue Y, Searle S, Hunt S, Delgado M, Morrison J, Whittaker P, Deloukas P, Tyler-Smith C. 2009. A Genome-wide Survey of the Prevalence and Evolutionary Forces Acting on Human Nonsense SNPs. Am J Hum Genet 84:224–234.

Zeder MA. 2008. Domestication and early agriculture in the Mediterranean Basin: Origins, diffusion, and impact. Proc Natl Acad Sci 105:11597–11604.

Zhai W, Nielsen R, Slatkin M. 2009. An Investigation of the Statistical Power of Neutrality Tests Based on Comparative and Population Genetic Data. Mol Biol Evol 26:273–283.

Zhang QS, Browning BL, Browning SR. 2015. Genome-wide haplotypic testing in a Finnish cohort identifies a novel association with low-density lipoprotein cholesterol. Eur J Hum Genet 23:672–677.

Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS. 2012. A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics 28:3326–3328.

Zook J, McDaniel J, Parikh H, Heaton H, Irvine SA, Trigg L, Truty R, McLean CY, De La Vega FM, Xiao C, Sherry S, Salit M. 2018. Reproducible integration of multiple sequencing datasets to form high-confidence SNP, indel, and reference calls for five human genome reference materials. Available from: http://biorxiv.org/lookup/doi/10.1101/281006

Zook JM, Chapman B, Wang J, Mittelman D, Hofmann O, Hide W, Salit M. 2014. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. Nat Biotechnol 32:246–251.

Zuk O, Schaffner SF, Samocha K, Do R, Hechter E, Kathiresan S, Daly MJ, Neale BM, Sunyaev SR, Lander ES. 2014. Searching for missing heritability: Designing rare variant association studies. Proc Natl Acad Sci 111:E455–E464.

Appendix A

Appendix A.1 A selection of available variant annotation algorithms

Machine learning approaches, score SNPs (all methods) and short indels (CADD and DANN)		
Name	**Principle and short description**	**Source**
CADD	A training dataset was constructed from observed high (>95%) frequency derived sites, which are depleted in deleterious variation, and simulated variants. A linear support vector machine (SVM) was then trained to discriminate between observed and simulated variants based on a set of 63 features (including the results from other variant annotation algorithms). The resulting SVM model can be used to assign a combined annotation score to every SNP or (small scale) indel in the human genome. The scores are then scaled in a phred-like manner, the higher the score the more dissimilar the respective variant is from the high DAF observed data and therefore more likely to be deleterious.	(Kircher et al., 2014)
DANN	The authors used the same training data and annotation features as for CADD. However, the algorithm was a neural network type learning approach which also incorporates nonlinear relationships between the annotation features. The resulting score is interpreted similarly to CADD.	(Quang et al., 2015)
FATHMM-MKL	This approach is conceptually similar to CADD, with some differences in the training data and annotations. The training data consisted of heritable germline mutations from HGMD as pathogenic and SNPs with a MAF > 1% in the 1000 Genomes dataset. The authors applied a linear kernel-based classifier and came up with two different models to categorise coding and non-coding variation with features mainly consisting of phylogenetic conservation scores and data from the ENCODE project, i.e. regulatory information. The former reaches optimal prediction accuracy for 10 groups of features (n_{total} = 1181), the latter for 4 groups of features (n_{total} = 663). Scores resulting from these models can be assigned to each SNP in the genome on a 0-1 scale with >0.5 being an indication of inferred deleteriousness.	(Shihab et al., 2013, 2015)

Evolutionary conservation approaches, score small- and large-scale variation (inferences for the latter are more challenging)		
Name	**Principle and short description**	**Source**
GERP	GERP identifies elements that have been conserved across multiple alignments consisting of different species. This is achieved by quantifying substitution deficits compared to what would be expected under a neutral scenario. The underlying assumption is that novel variation in these conserved regions is most likely to be deleterious as they are functionally constrained.	(Cooper et al., 2005)
PhyloP	PhyloP is conceptually very similar to GERP. A phylogenetic HMM is used to assign conservation scores to regions of alignments based on different evolutionary models. The inferences are assigned a p-value based on a comparison with the number of substitutions expected under a neutral 0-model, i.e. given a particular underlying tree it is assessed how likely the resulting alignments and a particular prediction for a region are under neutrality.	(Siepel et al., 2006)
Protein sequence and structure approaches, PolyPhen-2 scores missense SNPs, PROVEAN and SIFT also assess multiple substitution and small in-frame indels affecting the amino acid sequence, MutationTaster2 assesses all SNPs and short indels in genic regions (i.e. also synonymous and intronic variation)		
Name	**Principle and short description**	**Source**
PolyPhen-2	Two training datasets were constructed based on different sets of disease-related variants from the UniProt database with putatively neutral intra- and interspecies polymorphisms affecting protein structure as a control. The impact of a missense mutation on protein structure is assessed based on a functional comparison to the "wild-type" (i.e. mostly reference) human allele and the evolutionary conservation of the protein based on a multiple alignment with other mammalian species. Eleven features describing these properties were used as input for a machine learning algorithm (Naïve Bayes classifier). The probability that a particular mutation is damaging is calculated and the true and false positive rates are estimated; then qualitative labels are applied based on the scores.	(Adzhubei et al., 2010)

PROVEAN	PROVEAN compares a target sequence with homologous sequences from other species and clusters them based on similarity. The clusters most related to the human sequence are used as a supporting sequence set to compute an average delta alignment score. It describes the increase in dissimilarity in the alignment of the mutant allele relative to the non-human sequences compared to the human wildtype sequence. The more negative this alignment score the more likely the respective change is to be deleterious. A threshold of the delta score to separate putatively deleterious and non-deleterious variation was derived based on the application of PROVEAN to known deleterious vs neutral variation derived from UniProt/SwissProt so that their separation was maximised.	(Choi et al., 2012)
SIFT	Similarly to PROVEAN, SIFT compares a target protein containing the site of interest to a set of homologous mammalian sequences and builds an alignment. Then, based on this alignment the probabilities for all 20 amino acids potentially appearing at that particular site are calculated. For this the type of amino acid change is also considered, i.e. if in the original alignment only various amino acids with hydrophobic side chains occur, then it is assumed that this position can only contains this kind of amino acid. This probability is normalised by comparison to the most frequently observed amino acid and if it is below a certain threshold (which usually means that the position is highly conserved) then the observed change is thought to be deleterious. Later versions of SIFT which are not described in detail here use sequence homology data as features based on which the deleteriousness of indels in protein-coding regions can be assessed using machine learning algorithms (Hu and Ng, 2012, 2013).	(Kumar et al., 2009; Ng and Henikoff, 2001)
MutationTaster2	The general framework of MutationTester2 is similar to PolyPhen-2. Known disease mutations were extracted from the HGMD database and variants for which at least 20 homozygote individuals were found in the 1000 Genomes data were taken as control. The features used as input for the machine learning approach came from a broad range of databases. This input was processed by a Naïve Bayesian classifier resulting in three different classification models for changes a) leading to amino acid substitutions, b) involving more than one amino acid or which were c) non-coding or synonymous. It should be noted that the	(Schwarz et al., 2014)

	approach is designed to pick up rare large effect variants and that variants which are found >4 times in a homozygous state in the 1000 Genomes or HapMap datasets are automatically regarded as neutral.	

References

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. Nat Methods 7:248–249.

Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. 2012. Predicting the Functional Effect of Amino Acid Substitutions and Indels. PLoS ONE 7:e46688.

Cooper GM, Stone EA, Asimenos G, NISC Comparative Sequencing Program, Green ED, Batzoglou S, Sidow A. 2005. Distribution and intensity of constraint in mammalian genomic sequence. Genome Res 15:901–913.

Hu J, Ng PC. 2012. Predicting the effects of frameshifting indels. Genome Biol 13:R9.

Hu J, Ng PC. 2013. SIFT Indel: Predictions for the Functional Effects of Amino Acid Insertions/Deletions in Proteins. PLoS ONE 8:e77940.

Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. 2014. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 46:310–315.

Kumar P, Henikoff S, Ng PC. 2009. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc 4:1073–1081.

Ng PC, Henikoff S. 2001. Predicting deleterious amino acid substitutions. Genome Res 11:863–874.

Quang D, Chen Y, Xie X. 2015. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics 31:761–763.

Schwarz JM, Cooper DN, Schuelke M, Seelow D. 2014. MutationTaster2: mutation prediction for the deep-sequencing age. Nat Methods 11:361–362.

Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GLA, Edwards KJ, Day INM, Gaunt TR. 2013. Predicting the Functional, Molecular, and Phenotypic Consequences of Amino Acid Substitutions using Hidden Markov Models. Hum Mutat 34:57–65.

Shihab HA, Rogers MF, Gough J, Mort M, Cooper DN, Day INM, Gaunt TR, Campbell C. 2015. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. Bioinformatics 31:1536–1543.

Siepel A, Pollard KS, Haussler D. 2006. New Methods for Detecting Lineage-Specific Selection. In: Apostolico A, Guerra C, Istrail S, Pevzner PA, Waterman M, editors. Research in Computational Molecular Biology. Vol. 3909. Berlin, Heidelberg: Springer Berlin Heidelberg. p 190–205. Available from: http://link.springer.com/10.1007/11732990_17

Appendix B

Appendix B.1 Ethnology of the Lebbo and the Kankanaey-Igorot

In order to identify a better representative of the original Austronesian genomic diversity some populations, in particular the Lebbo and the Kankanaey-Igorot, that are thought to have been comparatively isolated in recent history, were sampled. The Lebbo are indigenous hunter gatherers that lived for centuries in higher caves and rock faces in the karst Sangkulirang-Mangkalihat Karstic Range of East Kalimantan (Borneo Island) (Guerreiro, 2015). They are Austronesian speakers and have a matrilocal kinship structure, which is predominant in the Austronesians societies (Jordan et al., 2009).

Perhaps the most interesting of the groups sampled are the Kankanaey-Igorot from the Philippines. The term Igorot is a collective term for the peoples of the *Gran Cordillera Central* (a mountain range in the northern part of Luzon island) of which the Kankanaey form one ethno-linguistic subgroup (Scott, 1970). The archaeological and historical sources tell us very little about their prehistory. The ancestors of the Igorot are known to have resisted assimilation into the Spanish colonial empire since their first contact with the conquistadores in the 1570s (Scott, 1974). Early anthropological work (Jenks, 1905; Antolin, 1971) describes them as rice-farmers and also mentions mining activities. Furthermore, they were thought to lack a centralised political structure.

More recently, detailed archaeological work has been done on the Ifugao, a (geographically) neighbouring group, usually categorised as belonging to the Igorot. They share many characteristics with the Kankanaey. Both groups' languages belong to the South-Central Cordilleran subgroup of the Malayo-Polynesian languages (Hammarström et al., 2018)

and they exhibit a very similar spectrum of maternal genetic lineages (Delfin et al., 2014). They are also known to share subsistence patterns and to have similar traditional religious beliefs (Acabado, 2017). The Ifugao Archaeological Project (IAP) carried out excavations focussed on sites with a long settlement history and multiple occupational layers with corresponding radiocarbon dates. Acabado (2017) found evidence for the cultivation of starchy food sources such as taro, the main crop before rice, yam and breadfruit dating to at least 1000 years BP. The findings in the layers from 1600 onwards indicate intensified wet-rice-cultivation on terraces, which is labour-intensive and requires a considerable amount of social coordination within the

community practising this style of farming. This is accompanied by increased counts of earthenware ceramics and of bones of larger domesticated animals such as pigs and water buffalos. At same time more material suggesting stronger trade links to the Philippine lowlands appears, mainly elaborate foreign Ming style Chinese ceramics. These changes coincided with the establishment of a Spanish colonial presence on northern Luzon Island.

Acabado (2017) interpreted these observations as signs of economic intensification and increasing population sizes accompanied by higher organisational and political complexity. These developments can be contextualised in the framework of "pericolonialism" (Paredes, 2013). It describes cultural changes in groups which were not themselves subjugated politically by a colonial power, but still influenced by the colonisation of neighbouring peoples. It has been argued that they are indicative of an active strategy by the indigenous people to resist absorption into a colonial empire (Dillehay, 2014). Due to their similarities to the Ifugao it can be speculated that comparable developments would have taken place among the Kankanaey. This would mean that from the late 16th century onward they would have been impacted by Spanish colonialism at least indirectly and have been involved more in long distance trade. While this is inconsistent with total isolation, the Kankanaey were most likely still less directly affected by foreign influences than most other groups on Luzon Island.

This makes them intriguing given that according to the current models the last known major migration into the area was the Austronesian expansion (Bellwood, 2007) from Taiwan. Given that the Taiwanese aboriginals are known to have undergone recent admixture from mainland Chinese groups (HUGO Pan-Asian SNP Consortium et al., 2009) it opens the possibility that these mountain people could be better representatives of the original Austronesian stock.

References

Acabado S. 2017. The Archaeology of Pericolonialism: Responses of the "Unconquered" to Spanish Conquest and Colonialism in Ifugao, Philippines. International Journal of Historical Archaeology 21:1–26.

Antolin F. 1971. Notices of the Pagan Igorots in 1789: Part Two. Asian Folklore Studies 30:27–132.

Bellwood PS. 2007. Prehistory of the Indo-Malaysian Archipelago. Canberra: ANU E Press. Available from: http://epress.anu.edu.au/?p=80041

Delfin F, Min-Shan Ko A, Li M, Gunnarsdóttir ED, Tabbada KA, Salvador JM, Calacal GC, Sagum MS, Datar FA, Padilla SG, De Ungria MCA, Stoneking M. 2014. Complete mtDNA genomes of Filipino ethnolinguistic groups: a melting pot of recent and ancient lineages in the Asia-Pacific region. Eur J Hum Genet 22:228–237.

Dillehay TD. 2014. The teleoscopic polity: Andean Patriarchy and materiality. New York: Springer.

Guerreiro A. 2015. The Lebbo' Language and Culture: A Window on Borneo's Ancient Past. In: Wayan Arka I, editor. 12-ICAL Bali, Proceedings volume I. Canberra: ANU. p 149–177.

Hammarström H, Forkel R, Haspelmath M. 2018. Glottolog 3.3. Jena. Available from: https://glottolog.org/ accessed 2018-11-24

HUGO Pan-Asian SNP Consortium, Abdulla MA, Ahmed I, Assawamakin A, Bhak J, Brahmachari SK, Calacal GC, Chaurasia A, Chen C-H, Chen J, Chen Y-T, Chu J, Cutiongco-de la Paz EMC, De Ungria MCA, Delfin FC, Edo J, Fuchareon S, Ghang H, Gojobori T, Han J, Ho S-F, Hoh BP, Huang W, Inoko H, Jha P, Jinam TA, Jin L, Jung J, Kangwanpong D, Kampuansai J, Kennedy GC, Khurana P, Kim H-L, Kim K, Kim S, Kim W-Y, Kimm K, Kimura R, Koike T, Kulawonganunchai S, Kumar V, Lai PS, Lee J-Y, Lee S, Liu ET, Majumder PP, Mandapati KK, Marzuki S, Mitchell W, Mukerji M, Naritomi K, Ngamphiw C, Niikawa N, Nishida N, Oh B, Oh S, Ohashi J, Oka A, Ong R, Padilla CD, Palittapongarnpim P, Perdigon HB, Phipps ME, Png E, Sakaki Y, Salvador JM, Sandraling Y, Scaria V, Seielstad M, Sidek MR, Sinha A, Srikummool M, Sudoyo H, Sugano S, Suryadi H, Suzuki Y, Tabbada KA, Tan A, Tokunaga K, Tongsima S, Villamor LP, Wang E, Wang Y, Wang H, Wu J-Y, Xiao H, Xu S, Yang JO, Shugart YY, Yoo H-S, Yuan W, Zhao G, Zilfalil BA, Indian Genome Variation Consortium. 2009. Mapping human genetic diversity in Asia. Science 326:1541–1545.

Jenks A. 1905. The Bontoc Igorot. Manila: Bureau of Public Printing.

Jordan FM, Gray RD, Greenhill SJ, Mace R. 2009. Matrilocal residence is ancestral in Austronesian societies. Proceedings of the Royal Society B: Biological Sciences 276:1957–1964.

Paredes O. 2013. A Mountain of Difference: The Lumad in Early Colonial Mindanao. Ithaca, New York: Southeast Asia Program Publications, Southeast Asia Program, Cornell University.

Scott W. 1970. Igorot Responses to Spanish Aims: 1576-1896. Philipp Stud 18:695–717.

Scott WH. 1974. The discovery of the Igorots: Spanish contacts with the Pagans of Northern Luzon. Rev. ed. Quezon City: New Day Publ.

Appendix B.2 Samples for analyses of Southeast Asian populations

This file can be found attached to the electronic version of this thesis and contains information on all novel and published genotyped samples included in any analysis in chapter 2.

Appendix B.3 Methodological details on IBD calculations

For the paper published on the genomics of SEA PLINK's IBD algorithm had been applied. In this thesis the Refined IBD was used, mainly because the former approach assumes a homogeneous randomly mating population which can lead to overestimation of IBD as described below.

The PLINK method yielded very high average cross-population IBD values such as 0.1299 between lowland Filipinos and Lebbo (see Appendix B.6). While this value does not represent a precise genealogical kinship coefficient, it is above the approximate threshold for second degree relatedness proposed for this particular method (Thornton et al., 2012). This effect was partly mitigated when the dataset was LD-pruned before the IBD calculations as recommended by the authors of the method (Purcell et al., 2007). However, for the aforementioned example the IBD value still remained at 0.0732. The most plausible explanation for these observations is that PLINK overestimated IBD at least for some population pairs.

A simulation study by Morrison (2013) supports this conjecture and identifies violations of the assumption of a homogeneous randomly mating population as the reason for this behaviour. Under these conditions, PLINK's IBD method of moments technique overstates the degree of kinship between individuals from ancestrally similar groups (i.e. the similarity in allelic states is real but due to older shared ancestors). BEAGLE's IBD method was chosen as it has been shown to outperform PLINK with regards to its power to detect shorter IBD segments and false positive rates, among other reasons because it incorporates LD information explicitly (Browning and Browning, 2010).

The genetic map used for the RefinedIBD calculations was the standard recommended for BEAGLE (http://bochet.gcc.biostat.washington.edu/beagle/genetic_maps/). It was constructed based on a recombination map inferred from the HapMap project and lifted to build GRCh37. The Refined IBD approach was applied using the standard parameters, except for the following changes. The ibdtrim parameter, which species the number of markers trimmed from the end of a shared haplotype when testing for IBD, was adjusted downwards from 40 to 12. This was done to account for the lower SNP density in the data analysed here compared to the parameter values which were originally recommended for a set of ~10^6 SNPs, i.e. a ca. 3.3 times higher coverage. The minimum length of IBD fragments reported was adjusted downwards from 1.5 cM to 0.5 cM as the Refined IBD algorithm has been shown to add power to detect these short segments compared to other methods. Furthermore, here the IBD was used to indicate general

population affinities and not to make definite statements about specific individuals and their underlying pedigrees, which makes some false positives tolerable.

Firstly, shared IBD segments were obtained for pairs of individuals. Their physical length was converted to genetic length using the genetic map mentioned above and the Java archive base2genetic.jar (https://faculty.washington.edu/browning/beagle_utilities/utilities.html).

For pairwise comparisons between populations the following formula was used:

$$IBD_{i,j} = \frac{\sum_{a \in P_i}^{x} \sum_{b \in P_j}^{y} L_{a,b}}{n_i n_j}$$

Here, P_i and P_j denote two populations of interest consisting of n_i and n_j individuals respectively. Accordingly, a and b are indices for individuals in populations P_i and P_j. $L_{a,b}$ is the total length of all IBD segments shared between individuals a and b.

To calculate the average within-population IBD sharing this expression is modified to yield:

$$IBD_i = \frac{\sum_{c=1}^{x} \sum_{d=1}^{y} L_{c,d}}{n_i(n_i - 1)/2}$$

The individuals with the indices c and d both belong to population P_i, with a total of n_i individuals; also note that $c \neq d$, as sharing between the two copies of each chromosome within an individual is not measured here. As above $L_{c,d}$ describes the total shared IBD length between c and d.

Finally, to obtain $IBD_{i,j}$ and IBD_i as rates they were divided by twice the total genetic length (as the data were diploid) of all autosomes in the GRCh37 build (~7092 cM). These values can be interpreted as estimates of the relatedness coefficient.

References

Browning SR, Browning BL. 2010. High-Resolution Detection of Identity by Descent in Unrelated Individuals. The American Journal of Human Genetics 86:526–539.

Morrison J. 2013. Characterization and Correction of Error in Genome-Wide IBD Estimation for Samples with Population Structure: Characterization and Correction of Error in Genome-Wide IBD Estimation for Samples with Population Structure. Genetic Epidemiology 37:635–641.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC. 2007. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am J Hum Genet 81:559–575.

Thornton T, Tang H, Hoffmann TJ, Ochs-Balcom HM, Caan BJ, Risch N. 2012. Estimating Kinship in Admixed Populations. The American Journal of Human Genetics 91:122–138.

Appendix B.4 Methodological details on ADMIXTURE analyses

Dataset I

As recommended by the ADMIXTURE manual the data were pruned for LD using PLINK to remove one SNP from each pair with an r^2 value greater than 0.1 within a 50-SNP window, moving 10 SNPs forward each time. This yielded a dataset of 8,075 SNPs on which the ADMIXTURE 1.23 program was run. This analysis was repeated 100 times at each K = 2 to K = 10 and the cross-validation errors and log-likelihood estimates for each value of K were used to estimate the optimum number of K clusters (Cardona et al., 2014). Here, the ideal value is the one for which the model has the best predicative accuracy. This model is generally indicated by the lowest cross-validation error compared to other Ks.

Dataset II

The genotyping platform for the published and the newly genotyped samples is Illumina, but different genotyping arrays (610K, 650K, 660K and 730K) were used for different sample sets. Before merging the data all A/T and C/G markers were removed to minimise potential strand errors. After filtering for autosomal SNPs with a genotyping success rate of over 97% and a MAF of 1% the intersection of SNPs over the different array types was 305,489 SNPs. Then the data were pruned to minimise the potential effect of LD on the ADMIXTURE results. PLINK was used to remove one SNP from each pair of SNPs in a window of 200 SNPs with an r^2 above 0.4. The window was advanced by 25 SNPs. This resulted in a dataset of 187,519 SNPs which was fed into ADMIXTURE 1.07. ADMIXTURE was run 100 times at K = 3-15. Based on the cross-validation procedure of ADMIXTURE the genetic structure in the verification sample set is best described at K = 9-10 (Appendix B.7A). Furthermore, the convergence of individual ADMIXTURE runs at each K was assessed. For this purpose, the maximum difference in log-likelihood (LL) scores was monitored. It was assumed that a global

LL maximum was reached at a given K if the 10% of the runs with the highest LL score had minimal (< ~2 LL units) variation in LL scores. According to this reasoning, the global LL maximum was reached in runs at K = 2–15.

Differences in pre-filtering of SEA potentially related SEA individuals between Dataset I and Dataset II

In both cases one individual from an intrapopulation pair with a PLINK IBD > 0.125 was removed to minimise the confounding effect of recent genealogical relatedness. Before these IBD analyses were run the SEA individuals were merged with a different set of reference individuals in the two runs (see above). While the marker totals (~ 300k SNPs) were roughly comparable there were some non-overlapping markers specific to each merged dataset which led to subtle differences in the obtained IBD statistics and therefore also affected which individuals were excluded based on IBD cut-offs. Therefore, after filtering for relatives Dataset I contained 185 SEA individuals, whereas Dataset II contained 187 individuals.

Appendix B.5 Methodological details on calculation of selection statistics

Both the iHS and XP-EHH statistics were calculated as in Pickrell et al. (2009), yielding about 10,000-11,000 genomic windows for his and about 13,700 windows for XP-EHH for each SEA population analysed. To pick up only patterns particular to SEA from the top 1% of all iHS signals, the top 5% his windows detected for the CHB population from the HapMap panel were excluded. However, for the analysis of regional sharing based on the iHS this condition did not apply. The use of a reference population is inherent in the XP-EHH method, again CHB was chosen, for similar reasons.

The PBS statistic represents the amount of allele frequency change at a given locus in the history of the test population since it diverged from other populations (Yi et al., 2010). The outgroups chosen for each tested SEA population were the YRI and CHB populations. Pairwise F_{ST} values for the populations of interest and the references were calculated following Weir and Cockerham (Weir and Cockerham, 1984). PBS scores were estimated from pairwise F_{ST} values (Yi et al., 2010). Based on the approach of Pickrell et al. (2009) the genome was divided into windows of a modified size of 100kb sand the maximum PBS score in each window was used

as the test statistic. This resulted in between 26,000 and 27,000 windows for each analysed group.

References

Pickrell JK, Coop G, Novembre J, Kudaravalli S, Li JZ, Absher D, Srinivasan BS, Barsh GS, Myers RM, Feldman MW, Pritchard JK. 2009. Signals of recent positive selection in a worldwide sample of human populations. Genome Research 19:826–837.

Weir BS, Cockerham CC. 1984. Estimating F-Statistics for the Analysis of Population Structure. Evolution 38:1358–1370.

Yi X, Liang Y, Huerta-Sanchez E, Jin X, Cuo ZXP, Pool JE, Xu X, Jiang H, Vinckenbosch N, Korneliussen TS, Zheng H, Liu T, He W, Li K, Luo R, Nie X, Wu H, Zhao M, Cao H, Zou J, Shan Y, Li S, Yang Q, Asan, Ni P, Tian G, Xu J, Liu X, Jiang T, Wu R, Zhou G, Tang M, Qin J, Wang T, Feng S, Li G, Huasang, Luosang J, Wang W, Chen F, Wang Y, Zheng X, Li Z, Bianba Z, Yang G, Wang X, Tang S, Gao G, Chen Y, Luo Z, Gusang L, Cao Z, Zhang Q, Ouyang W, Ren X, Liang H, Zheng H, Huang Y, Li J, Bolund L, Kristiansen K, Li Y, Zhang Y, Zhang X, Li R, Li S, Yang H, Nielsen R, Wang J, Wang J. 2010. Sequencing of 50 Human Exomes Reveals Adaptation to High Altitude. Science 329:75–78.

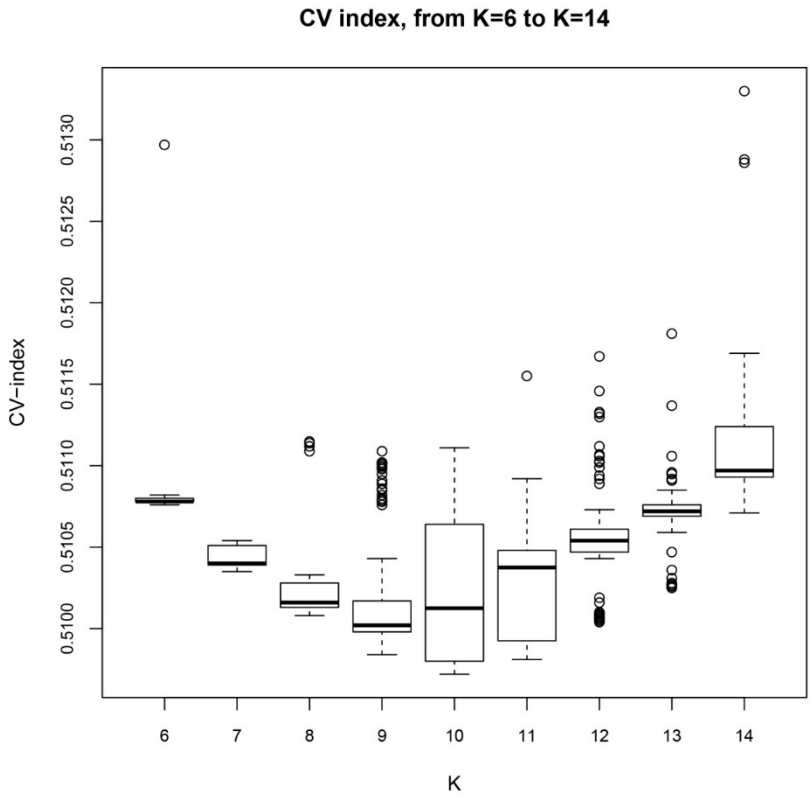
Appendix B.6 Pairwise PLINK IBD values of SEA populations together with a set of reference populations.

Mean between-population IBD calculated with PLINK is reported. The data in the upper triangle of the matrix were pruned for LD using PLINK to remove one SNP from each pair with an r^2 value greater than 0.1 within a 50-SNP window, moving 10 SNPs forward each time.

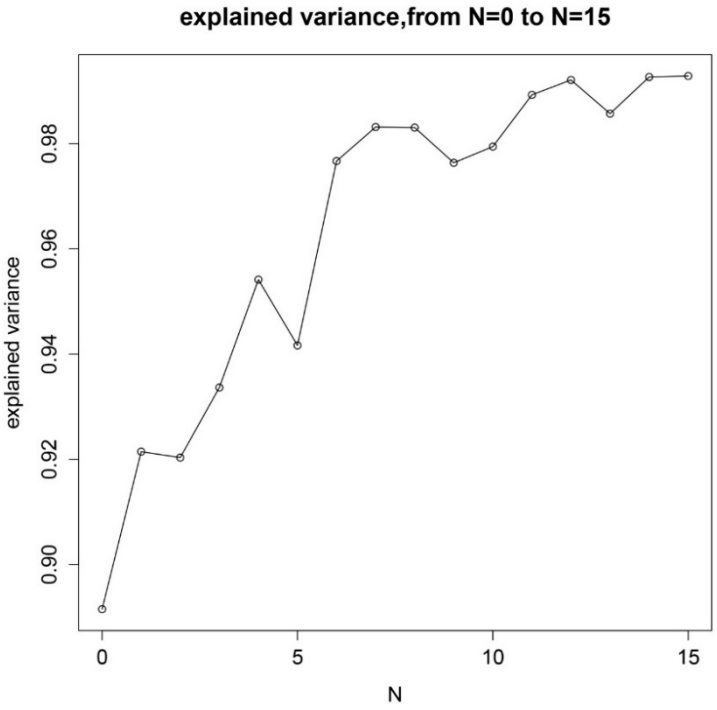
IBD/F_{ST}	Bajo	Burmese	CEU	CHB	Dusun	Filipino	Kankanaey	JPT	Lebbo	Malay	Murut	Papuans	South Indian	Vietnamese	YRI
Bajo		0.0317	0	0	0.0051	0	0.0017	0	0	0	0	0	0	0	0
Burmese	0.0490		0	0	0.0289	0.0023	0.0126	0	0.0057	0.0001	0	0	0	0	0
CEU	0	0		0	0	0	0	0	0	0	0	0	0	0	0
CHB	0.0072	0.0109	0		0	0	0	0	0	0	0	0	0	0	0
Dusun	0.0312	0.0498	0	0.0199		0.0033	0.0488	0	0.0099	0.0001	0	0	0	0	0
Filipino	0.0350	0.0581	0	0.0613	0.0730		0.0253	0	0.0732	0.0592	0.0241	0	0	0.0237	0
Kankanaey	0.0225	0.0377	0	0.0337	0.0770	0.0686		0	0.0421	0.0143	0.0009	0	0	0	0
JPT	0.0024	0.0026	0	0.1072	0.0022	0.0301	0.0086		0	0	0	0	0	0	0
Lebbo	0.0331	0.0627	0	0.0161	0.0746	0.1299	0.0788	0.0001		0.0373	0.0017	0	0	0	0
Malay	0.0233	0.0404	0	0.0298	0.0519	0.1032	0.0582	0.0029	0.0981		0.0273	0	0	0.0133	0
Murut	0.0164	0.0328	0	0.0658	0.0486	0.0754	0.0320	0.0275	0.0563	0.0557		0	0	0.0571	0
Papuans	0	0	0	0	0	0	0	0	0	0	0		0	0	0
South Indian	0	0	0	0	0	0	0	0	0	0	0	0		0	0
Vietnamese	0.0181	0.0298	0	0.0918	0.0518	0.0831	0.0322	0.0636	0.0626	0.0586	0.0847	0	0		0
YRI	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

Appendix B.7 Supporting statistics for ADMIXTURE and TreeMix runs

A) Cross validation errors for ADMIXTURE run performed on the verification/validation dataset (Figure 2.4).



B) Plot of the explained variance statistic for TreeMix over the number of migration edges N.



Appendix B.8 f_3 statistics for all possible combinations from a panel of 45 worldwide populations with the Kankanaey as target

This file can be found attached to the electronic version of this thesis, note that there are no significant (z-score <-2) outcomes.

Appendix B.9 mtDNA Haplotypes observed in a subset of eight Kankanaey individuals.

Individual ID	Haplogroup	Private Mutations
Igorot10	M7b1a2a1	152C
Igorot13	M7b1a2a1	5821A, 13145A
Igorot17	D5b1c1a	4216C, 6260A, 10548C, 16182G, 16183C, 16519C
Igorot23	B4b1a2	1313C, 16182G, 16183C, 16519C
Igorot3	B4a1a6	15784C, 16182G, 16183C, 16519C
Igorot4	B4a1a6	16182G, 16183C, 16213A, 16519C
Igorot5	B4a1a6	11377A, 16182G, 16183C, 16519C
Igorot6	B5b1c1a	9455G, 14757C, 16183C, 16519C

Appendix B.10 Most significant population-specific windows for iHS, XP-EHH and PBS

This file can be found attached to the electronic version of this thesis. The first sheet of the excel file (**Appendix B.10A**) contains the Top 10 iHS windows for each population according to the empirical p-value. The second sheet of the excel file (**Appendix B.10B**) reports the Top 10 XP-EHH windows for each population according to the empirical p-value. The third sheet of the excel file (**Appendix B.10C**) contains the Top 20 PBS windows for each population according to the score of the PBS statistic. In all these appendices windows which appear for multiple populations are bolded.

Appendix B.11 Signal sharing between Southeast Asian populations for iHS and XP-EHH

Appendix B.11A The proportion of 200-kb windows that were detected as 1% selection outliers in the iHS test in one population (indicated in the first column) and also found in the 5% of signals in other populations (in columns 3-12). The order is non-alphabetical to better illustrate the clustering of the Chinese and MSEA groups.

top 1%/top 5%	**average signal-sharing**	CHB	Vietnamese	Burmese	Filipino	Bajo	Malay	Dusun	Murut	Kankanaey	Lebbo
CHB	0.38		0.53	0.61	0.49	0.52	0.49	0.19	0.29	0.13	0.16
Vietnamese	0.37	0.62		0.44	0.48	0.48	0.57	0.18	0.26	0.13	0.15
Burmese	0.35	0.63	0.52		0.39	0.39	0.54	0.18	0.23	0.14	0.13
Filipino	0.39	0.50	0.46	0.39		0.50	0.53	0.31	0.32	0.19	0.29
Bajo	0.36	0.46	0.46	0.46	0.47		0.55	0.22	0.28	0.17	0.20
Malay	0.36	0.54	0.44	0.52	0.47	0.52		0.17	0.24	0.14	0.20
Dusun	0.30	0.35	0.34	0.32	0.42	0.37	0.32		0.24	0.19	0.14
Murut	0.22	0.31	0.18	0.21	0.24	0.26	0.26	0.24		0.15	0.14
Kankanaey	0.19	0.25	0.16	0.18	0.23	0.21	0.20	0.18	0.13		0.13
Lebbo	0.16	0.16	0.14	0.16	0.24	0.18	0.21	0.08	0.15	0.12	

Appendix B.11B The proportion of 200-kb windows that were detected as 1% selection outliers in the XP-EHH test in one population (indicated in the first column) and also found in the 5% of signals in the other populations (in columns 3-12). The order is by the average amount of signal sharing with all other groups.

top 1%/top 5%	**average signal-sharing**	Dusun	Bajo	Filipino	Kankanaey	Malay	Murut	Lebbo	Vietnamese	Burmese
Dusun	0.27		0.32	0.35	0.24	0.27	0.37	0.20	0.23	0.17
Bajo	0.26	0.30		0.31	0.19	0.36	0.27	0.24	0.19	0.19
Filipino	0.26	0.29	0.35		0.24	0.27	0.32	0.25	0.21	0.14
Kankanaey	0.24	0.32	0.29	0.27		0.28	0.26	0.23	0.15	0.12
Malay	0.24	0.28	0.29	0.30	0.16		0.26	0.19	0.22	0.23
Murut	0.21	0.27	0.21	0.29	0.17	0.22		0.22	0.22	0.12
Lebbo	0.20	0.24	0.22	0.23	0.21	0.22	0.17		0.16	0.15
Vietnamese	0.18	0.25	0.19	0.17	0.14	0.18	0.17	0.15		0.18
Burmese	0.15	0.16	0.15	0.14	0.11	0.18	0.17	0.17	0.12	

Appendix C

Appendix C.1 WGS samples for analyses of global patterns of functional and rare variants.

This file can be found attached to the electronic version of this thesis. The first sheet of the excel file is a table containing all samples from the EGDP panel for which whole genome sequence data were generated (analysed in chapters 3 and 4). It contains information on the origin, processing and affiliation of these genomes (modified from Pagani et al. 2016). The second sheet lists the abbreviations used in the samples table.

Appendix C.2 Populations included in MSMC and multiple regression analyses.

Populations included (coded as 1) in MSMC reconstruction on different samples sizes and in multiple regression analyses predicting the load of different classes of deleterious mutations.

Population	MSMC reconstruction for n = 1/2	MSMC reconstruction for n = 4	Inclusion in multiple regression analyses
African – Americans	0	1	0
Altaians	1	1	1
Armenians	0	1	1
Avars	1	0	0
Bajo	0	1	1
Baka – Pygmies	1	0	1
Bashkirs	0	1	1
Belarusians	0	1	1
Brahmin	1	0	1
Burmese	1	1	1
Buryats	0	1	1
Calchaquíes	0	1	1
Chukchi	1	1	1
Colla	1	1	1
Croats	0	1	1
Druze	1	0	1
Dusun	0	1	1
Eskimo	1	1	1
Estonians	1	1	1
Georgians	1	0	0
Germans	1	1	1
Gujaratis	0	1	1
Hadza	0	1	0
Han	1	1	1
Igorot	0	1	1
Iranians	1	1	1

Population	MSMC reconstruction for n = 1/2	MSMC reconstruction for n = 4	Inclusion in multiple regression analyses
Italians	1	1	1
Japanese	0	1	1
Kabardians	0	1	0
Koinanbe	1	0	1
Koryaks	1	1	1
Kyrgyz	0	1	1
Lebbo	1	1	1
Lezgins	0	1	0
Luhya	0	1	0
Malayan	1	0	1
Maris	0	1	1
Mexicans	0	1	0
Mongolians	0	1	1
Murut	0	1	1
Northwest Europeans	1	1	1
Poles	0	1	1
Sandawe	0	1	0
Udmurts	0	1	1
Vepsians	0	1	1
Vietnamese (North)	0	1	1
Vietnamese (South)	0	1	1
Wichi	0	1	1
Yakuts	1	0	1
Yoruba	1	1	1

Appendix C.3 Effective population size over time inferred using MSMC

This file can be found attached to the electronic version of this thesis. It contains the raw data on effective population size (N_e) over time inferred with MSMC. The first sheet of the excel file contains this statistic calculated based on one or two genomes per each of the 22 populations chosen to represent the Diversity Set. The second sheet expands on this and consists of the effective population size (N_e) over time inferred with MSMC using four genomes each from any population with sufficient sample size. These data were generated by Dr Luca Pagani.

Appendix C.4 Population groups used to form the Selection Set (n = 369).

Population group	Abbreviation	Sample size	Population affiliation of samples
Central and West Africa	Afr	26	9 Yoruba, 8 Pygmy, 4 Luhya, 5 African Americans
Middle East	MiE	26	4 Iranians, 7 Arab, 3 Assyrian, 3 Druze, 2 Jordanians, 1 Lebanese, 6 Armenians
South and West Europe	WEu	32	4 Italians, 9 CEPH Europeans, 4 Croats, 4 Germans, 3 Hungarians, 2 Moldavians, 3 Albanians, 2 UK, 1 French
East and North Europe	EEu	53	6 Estonians, 3 Finns, 4 Vepsians, 3 Ingrians, 3 Karelians, 2 Swedes, 3 Latvians, 3 Lithuanians, 5 Poles, 4 Belarussians, 7 Russians, 7 Ukrainians, 3 Cossacks, 1 Mishar Tatar
Volga – Uralic	Vol	23	5 Bashkirs, 3 Chuvashes, 4 Maris, 7 Tatars, 4 Udmurts
South Asia	SoA	28	4 Gujarati, 1 Punjabi, 1 Rajasthani, 2 Madhya Pradesh, 7 Uttar Pradesh, 1 Kerala (Malayalam), 2 Andhra Pradesh, 1 Orissa, 4 Jharkhand, 1 Bengali, 3 Bangladeshi, 1 Nepali Brahmin
West Siberia	WSi	17	3 Forest Nenets, 3 Tundra Nenets, 3 Kets, 3 Khantys, 2 Mansis, 3 Selkups,
South Siberia and Mongolia	SSi	34	6 Altaians, 3 Buryats, 6 Mongolians, 3 Tuvins, 14 Buryats, 2 Shors
Central Siberia	CSi	31	8 Evens, 13 Evenks, 2 Nganasans, 8 Yakuts
Northeast Siberia	NSi	25	5 Chukchi, 4 Eskimos, 16 Koryaks
Mainland East and Southeast Asia	SeM	29	7 Chinese, 4 Japanese, 8 Burmese, 10 Vietnamese
Island Southeast Asia	SeI	45	4 Lebbo, 4 Bajo, 8 Dusun, 8 Murut, 8 Igorot, 2 Luzon, 2 Visayans, 3 Agta, 3 Aeta, 3 Batak

Appendix C.5 Bulk data of variants annotated using Ingenuity Variant Analysis

This file can be found attached to the electronic version of this thesis. It contains a text-formatted repository of 294,360 variants in 382 individuals from 14 worldwide macro-groups with annotations as inferred using Ingenuity Variant Analysis. Additional annotations are described in section 3.1.2. For convenience the global non – reference and derived allele frequencies for all loci have been provided in columns 17 and 18. Columns 19 to 32 contain information on the derived allele frequency in each of the macro-groups. Abbreviations: AN – ancient (hominin) Neanderthal, AD – ancient (hominin) Denisovan.

Appendix C.6 Variants potentially causing nonsense-mediated mRNA decay

This file can be found attached to the electronic version of this thesis. It is a VCF file containing 229 stop – gain variants in 382 individuals from 14 worldwide macro-groups that are annotated as "stop_gained&NMD_transcript" based on information from the GRCh37-compatible version of Ensembl Release 87. The INFO column contains the following abbreviations: AA – ancestral allele inferred from six primate species, AN – ancient (hominin) Neanderthal, AD – ancient (hominin) Denisovan, SING – singleton site, VAR – variant site (i.e. at least two individuals carry at least one non – reference allele). The order of entries in the next subfield CSQ (consequences) is described in the second row of the VCF header.

Appendix C.7 Methodological details on matching of variant annotation categories

The approach that was used to make the variant annotations obtained from different methods as comparable as possible is described in the following. Please note that for two exonic categories highlighted by VEP, "splice region variant" and "stop retained" variant there are no exact equivalents in IVA. Furthermore, IVA only included RefSeq transcripts with the accession prefix "NM_" designating (mostly curated) protein-coding transcripts to determine the relevant translation consequence.

The first high level summary category considered consists of putative LoF variants. Besides sites where stop codons are introduced by the respective single bp mutation it also includes stop loss and severe splicing variants. The "missense equivalent" category contains start loss, missense and general splice region variants. The final category "synonymous equivalent" is comprised of variants changing the base pair sequence but not affecting the translated protein directly that either occur in amino-acid encoding triplets or stop codons. These categories follow the definitions suggested by McCarthy et al.(2014). As many sites can be part of multiple transcripts both IVA and VEP return all possible annotations for a variant. These multiple outcomes were prioritised using a severity ranking (Appendix C.8) following McCarthy et al. and the European Bioinformatics Institute (http://www.ensembl.org/info/genome/variation/predicted_data.html#consequences). For visualisation purposes the outcomes obtained when comparing the annotations assigned by IVA and VEP were normalised as follows. The cells of a 14 x 16 matrix containing the shared counts between IVA and VEP for all annotations $VEP_{1\ldots14}$ and $IVA_{1\ldots16}$ were denoted as $C_{A,B.}$ A constant of one was added to the count, which was then $\log_{10}$-transformed. Z-scores were calculated row-wise, denoted as $ZR_{A,B,}$ and column-wise, denoted as $ZC_{A,B}$, as follows:

$$ZR_{A,B} = \frac{C_{A,B} - \mu_A}{SD_A}$$

μ_A is the mean value of all entries in row A, i.e. of the counts of sites with different IVA annotations corresponding to one particular annotation in VEP. SD_A denotes the row-wise standard deviation.

$$ZC_{A,B} = \frac{C_{A,B} - \mu_{.B}}{SD_{.B}}$$

$\mu_{.B}$ is the mean value of all entries in column B, i.e. of the counts of sites with different VEP annotations corresponding to one particular annotation in IVA. $SD_{.B}$ denotes the column-wise standard deviation.

Appendix C.8 Severity ranking of variant annotations

Matching and assignment into higher level categories of annotations obtained from IVA and VEP that were reported at least once for the dataset presented in this chapter. As both approaches present multiple annotations for some sites, precedence values used to prioritise these are given. Higher values indicate a higher precedence as the reported consequence is thought to be more severe.

Macro category	VEP consequence	IVA consequence	Precedence
LoF equivalent	stop_gained	stop gain	99
	stop_lost	stop loss	95
	splice_acceptor_variant + splice_donor_variant	Splice Site Loss	splice_acceptor_variant 86, splice_donor_variant 87, Splice Site Loss 86.5
Missense equivalent	start_lost	start lost	75
	missense_variant	missense	65
	splice_region_variant	-	63
Synonymous equivalent	stop_retained_variant	-	45
	synonymous_variant	synonymous	40
Not exonic	mature_miRNA_variant	microRNA	30
	5_prime_UTR_variant	5'UTR	26
	3_prime_UTR_variant	3'UTR	25
	intron_variant	Intronic	24
	non_coding_transcript_exon_variant	-	20
	regulatory_region_variant	Promoter region	11

Appendix C.9 The lack of reliability of small indel calls from Complete Genomics data

The unreliability of uncurated small indels called from Complete Genomics sequencing data can be demonstrated using a well characterised 32bp deletion in *CCR5*. It is known to confer partial resistance to HIV in heterozygotes and full resistance in homozygotes (Samson et al., 1996) reaching the highest frequencies worldwide at ~15% in Northeast European Finno-Ugric populations such as the Estonians (Kalev et al., 2000). The larger dataset (n = 730), of which a subset was analysed for this thesis, contains 27 Estonians. *CCR5*-Δ32 was reported for none of these in the MasterVar files provided by Complete Genomics. The probability of this occurring

if these files represent the true genotypes can be approximated by the value of the binomial distribution with p = 0.15 and n = 54 (due to diploidy). This results in Pr[binom(54,0.15) = 0] ~ $1.5*10^{-4}$, i.e. the event in question is very unlikely. It seems more plausible that the absence of *CCR5*-Δ32 in the analysed Estonians results from erroneous inferences from the algorithm used by Complete Genomics for calling indels from raw read data.

References

Kalev I, Mikelsaar AV, Beckman L, Tasa G, Pärlist P. 2000. High frequency of the HIV-1 protective CCR5 delta32 deletion in native Estonians. Eur J Epidemiol 16:1107–1109.

Samson M, Libert F, Doranz BJ, Rucker J, Liesnard C, Farber C-M, Saragosti S, Lapouméroulie C, Cognaux J, Forceille C, Muyldermans G, Verhofstede C, Burtonboy G, Georges M, Imai T, Rana S, Yi Y, Smyth RJ, Collman RG, Doms RW, Vassart G, Parmentier M. 1996. Resistance to HIV-1 infection in Caucasian individuals bearing mutant alleles of the CCR-5 chemokine receptor gene. Nature 382:722–725.

Appendix C.10 Gene lists for purifying selection analyses based on GO-term phenotype associations.

Phenotype	Term used in GO search	Number of genes
Olfactory	*olfact*	435
Virus	*defense response to virus*	858
Bacteria	*bacteria* or *bacterium*	226
Thermoregulation	*thermo*	144

Appendix C.11 Genes reported to be associated with any aspect of normal human pigmentation variation in GWAS studies

Gene	References
APBA2	Beleza et al. 2013
ASIP	Bonilla et al. 2005, Jacobs et al. 2013, Sulem et al. 2008
BNC2	Jacobs et al. 2013
CALCOCO1	Nan et al. 2009
DLGAP1	Nan et al. 2009
DNAH9	Nan et al. 2009
EDNRB	Zhang et al. 2013
EGFR	Quillen et al. 2012
EXOC2	Nan et al. 2009
GRM5	Beleza et al. 2013, Nan et al. 2009

Gene	References
HERC2	Beleza et al. 2013, Eiberg et al. 2008, Jacobs et al. 2013, Han et al. 2008, Nan et al. 2009, Sturm et al. 2008
IRF4	Han et al. 2008, Jacobs et al. 2013, Nan et al. 2009, Sulem et al. 2007
KIF26A	Han et al. 2008
KITLG	Miller et al. 2007, Sulem et al. 2007
KLRG1	Nan et al. 2009
LYST	Liu et al. 2010
M6PR	Nan et al. 2009
MC1R	Cook et al. 2009, Duffy et al. 2004, Han et al. 2008, Jacobs et al. 2013, Nan et al. 2009, Sulem et al. 2007, Valverde et al. 1995
OBSCN	Han et al. 2008
OCA2	Beleza et al. 2013, Cook et al. 2009, Duffy et al. 2004, Eiberg et al. 2008, Han et al. 2008, Jacobs et al. 2013, Nan et al. 2009, Shriver et al. 2003, Sturm et al. 2008, Sulem et al. 2007
OPRM1	Quillen et al. 2012
PLEKHA5	Nan et al. 2009
PPARGC1B	Nan et al. 2009
PRDM15	Nan et al. 2009
RABGGTA	Eiberg et al. 2008
RBP1	Nan et al. 2009
SLC24A4	Han et al. 2008, Sulem et al. 2007
SLC24A5	Beleza et al. 2013, Lamason et al. 2005, Stokowski et al. 2007
SLC45A2	Beleza et al. 2013, Cook et al. 2009, Graf et al. 2005, Han et al. 2008, Nan et al. 2009, Norton et al. 2007, Stokowski et al. 2007
TPCN2	Sulem et al. 2008
TYR	Beleza et al. 2013, Han et al. 2008, Jacobs et al. 2013, Nan et al. 2009, Shriver et al. 2003, Stokowski et al. 2007, Sulem et al. 2007
TYRP1	Kenny et al. 2012, Sulem et al. 2008
VASH2	Zhang et al. 2013
YPEL5	Nan et al. 2009
ZNF18	Nan et al. 2009

Appendix C.12 DIND score assessment of significance

The DIND statistic was calculated in a 100-kb window surrounding each SNP of interest. An empirical significance threshold for comparison to these DIND scores interest was defined by extracting non-genic SNPs for each of the 12 macro-groups from the Selection Set.

Batches (n = 100) of such SNPs were randomly selected from each DAF class in each macro-group. DIND scores were obtained for these subsets and 5 SDs of the resulting DIND score for each DAF class and macro-group were calculated and defined as neutral threshold.

Appendix C.13 Downsampled total exonic SNP counts by macro-group extracted using Ingenuity Variant Analysis

For all groups except the South Americans these represent the mean (rounded to full numbers) counts obtained from repeated downsampling to n = 13 as described in section 3.1.3.

	synonymous	missense	nonsense	Fraction of non-synonymous variants
Afr	35247	31642	283	0.4753
MiE	23828	23148	203	0.4949
WEu	22445	21628	201	0.4930
EEu	22259	21578	198	0.4945
Vol	22920	21729	218	0.4892
SoA	23658	22890	221	0.4941
WSi	21316	20553	176	0.4930
SSi	22039	20975	195	0.4899
CSi	19813	18337	165	0.4829
NSi	19233	17372	163	0.4769
SeM	21597	20303	196	0.4870
SeI	21370	20149	187	0.4876
SAm	18365	18867	159	0.5088

Appendix C.14 Matrix displaying the outcomes of pairwise chi-square tests comparing binned missense and synonymous DAF spectra

Comparisons between each of the 14 macro-groups, given in the form of the negative common logarithm of the p-values. For all groups except the Oceanians these spectra represent the mean (rounded to full numbers) counts obtained from repeated downsampling to n = 8. The upper half contains the values for the DAF spectra of missense the lower half for those of synonymous variants. Cells containing the p-values for comparisons that are significant according to the Bonferroni-corrected threshold ($p = 5.495*10^{-4}$ as there are 91 comparisons each, equal to a logged p of ~3.26) are bolded. Cells containing a value of 323.31 are equivalent to a p-value of $\sim 5*10^{-324}$. The latter is the smallest number that can reliably be represented by the R statistics software used for these calculations (https://stat.ethz.ch/R-manual/R-devel/library/base/html/zMachine.html) (R Core Team, 2017b). Short codes for the macro-groups taken from Table 3.1.

	Afr	**MiE**	**WEu**	**EEu**	**Vol**	**SoA**	**WSi**	**SSi**	**CSi**	**NSi**	**SeM**	**SeI**	**SAm**	**Oce**
Afr		**131.4**	**192.01**	**197.26**	**185.11**	**141.65**	**257.12**	**227.82**	**323.31**	**323.31**	**257.95**	**271.54**	**323.31**	**323.31**
MiE	**279.12**		**7.82**	**8.83**	**7.93**	0.34	**30.12**	**14.58**	**84.86**	**128.39**	**21.84**	**28.88**	**94.91**	**49.72**
WEu	**323.31**	**5.98**		0.03	0.24	**4.92**	**7.41**	1.65	**41.77**	**73.92**	**5.66**	**7.06**	**50.94**	**19.01**
EEu	**323.31**	**9.72**	0.56		0.33	**5.73**	**6.58**	1.28	**39.59**	**71.09**	**5**	**6.15**	**48.53**	**17.52**
Vol	**323.31**	**3.74**	0.85	1.74		**4.99**	**7.23**	2.81	**43.06**	**75.2**	**7.75**	**8.25**	**54.22**	**20.82**
SoA	**282.31**	0.6	**4.7**	**7.45**	**2.04**		**24.04**	**10.74**	**74.75**	**115.91**	**17.46**	**23.19**	**84.89**	**42.26**
WSi	**323.31**	**26.44**	**7.32**	**4.21**	**11.3**	**22.81**		**5.14**	**16.02**	**36.92**	**6.99**	2	**26.63**	**6**
SSi	**323.31**	**10.23**	0.99	1.22	**3.66**	**9.43**	**5.02**		**30.26**	**58.76**	1.24	2.54	**35.5**	**10.54**
CSi	**323.31**	**67.34**	**33.57**	**27.14**	**42.49**	**62.53**	**10.17**	**25.98**		**4.8**	**25.08**	**15.35**	**4.08**	**5.71**
NSi	**323.31**	**102.4**	**59.53**	**50.54**	**70.64**	**95.95**	**25.78**	**49.63**	**3.86**		**51.19**	**37.14**	**8.71**	**20.7**
SeM	**323.31**	**14.65**	**3.48**	**3.76**	**7.71**	**14.61**	**5.69**	0.79	**22.9**	**45.27**		**1.95**	**26.34**	**6.97**
SeI	**323.31**	**23.31**	**6.04**	**4.2**	**10.93**	**21.37**	1.24	2.71	**12.03**	**29.45**	2.04		**20.3**	**2.91**
SAm	**323.31**	**140.8**	**92.44**	**84.47**	**109.48**	**138**	**54.09**	**76.07**	**19.64**	**12.48**	**65.45**	**51.6**		**7.95**
Oce	**323.31**	**38.07**	**16.09**	**14.07**	**24.17**	**37.37**	**6.68**	**9.3**	**9.05**	**23.19**	**5.51**	3.12	**33.23**	

Appendix C.15 Matrix displaying the outcomes of pairwise chi-square tests comparing binned nonsense DAF spectra

Comparisons between each of the 14 macro-groups, for all groups except the Oceanians these spectra represent the mean (rounded to full numbers) counts obtained from repeated downsampling to n = 8. Cells containing the p-values for comparisons that are significant according to the Bonferroni-corrected threshold ($p = 5.495*10^{-4}$ as there are 91 comparisons each, equal to a logged p of ~3.26) are bolded. Cells containing a value of 323.31 are equivalent to a p-value of $\sim 5*10^{-324}$. The latter is the smallest number that can reliably be represented by the R statistics software used for these calculations (https://stat.ethz.ch/R-manual/R-devel/library/base/html/zMachine.html) (R Core Team, 2017b). Short codes for the macro-groups taken from Table 3.1.

	Afr	**MiE**	**WEu**	**EEu**	**Vol**	**SoA**	**WSi**	**SSi**	**CSi**	**NSi**	**SeM**	**SeI**	**SAm**	**Oce**
Afr		0.432	0.383	0.412	0.367	0.547	0.163	0.262	0.121	0.047	0.181	0.098	**$1.4*10^{-4}$**	0.095
MiE			0.996	0.984	0.694	0.899	0.836	0.833	0.754	0.494	0.735	0.717	0.022	0.691
WEu				0.985	0.690	0.878	0.881	0.855	0.807	0.545	0.768	0.772	0.029	0.747
EEu					0.790	0.939	0.847	0.912	0.774	0.470	0.826	0.713	0.023	0.734
Vol						0.897	0.552	0.888	0.506	0.200	0.838	0.393	0.007	0.523
SoA							0.657	0.870	0.581	0.283	0.767	0.497	0.008	0.554
WSi								0.827	0.988	0.788	0.810	0.970	0.102	0.955
SSi									0.793	0.426	0.980	0.684	0.030	0.815
CSi										0.825	0.800	0.981	0.141	0.983
NSi											0.426	0.899	0.281	0.745
SeM												0.679	0.038	0.848
SeI													0.144	0.931
SAm														0.128
Oce														

Appendix C.16 Sharing matrix displaying the count of each combination of annotations by VEP (rows) and IVA (columns) for all 354,396 sites annotated as exonic or splice site altering by at least one of the two approaches.

The abbreviations for the annotations are as follows: 3UTR – 3_prime_UTR_variant, 5UTR – 5_prime_UTR_variant, COV – removed from IVA because of coverage filter, IN – intron_variant, mmiRNA – mature microRNA variant, MIS – missense_variant, NC.EX – non_coding_transcript_exon_variant, PRO – promoter_region, SAL – start_lost, SP – splice_region_variant, SP_SEV – splice_severe_variant, SR – stop_retained_variant, STG – stop_gained, STL – stop_lost, SYN – synonymous_variant, UN – unknown.

	STG	**STL**	**SP_SEV**	**SAL**	**MIS**	**SR**	**SYN**	**mmiRNA**	**5UTR**	**3UTR**	**IN**	**PRO**	**COV**	**NOT_REFSEQ**	**UN**
STG	2281	0	0	1	97	0	120	0	2	3	21	0	39	842	0
STL	7	206	3	1	64	0	25	0	1	3	13	0	6	193	0
SP_SEV	0	0	532	0	110	0	57	0	14	7	32	0	42	4105	0
SAL	0	0	0	305	119	0	6	0	0	0	4	0	4	169	0
MIS	3	0	10	12	153693	1	2301	0	98	86	664	6	1403	23067	9
SP	1	0	3	2	69	1	3032	6	139	91	2506	2	340	31597	1
SR	0	0	0	0	1	102	9	0	0	0	0	0	0	66	0
SYN	3	0	1	1	71	0	108392	0	36	20	197	0	775	11393	6
mmiRNA	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
5UTR	0	0	0	0	126	0	30	0	0	0	0	0	0	0	0
3UTR	5	0	0	0	94	0	26	0	0	0	0	0	0	0	0
IN	7	0	0	0	380	1	179	0	0	0	0	0	0	0	0
NC.EX	30	4	1	2	692	0	546	0	0	0	0	0	0	0	0
NOT_VCF	17	23	3	4	1975	1	599	0	0	0	0	0	0	0	0

Appendix C.17 Matrix displaying the outcomes of Tukey's HSD for each pairwise inter-macro-group comparison of the per individual derived SNP counts for CADD20 variants

The results are reported in the form of the negative common logarithm of the p-values. The lower half contains the values for the total number of SNPs, the upper half for homozygous genotypes only. Cells containing the p-values for comparisons that are significant are bolded. Cells containing a value of 323.31 are equivalent to a p-value of ~5*10-324. The latter is the smallest number that can reliably be represented by the R statistics software used for these calculations (https://stat.ethz.ch/R-manual/R-devel/library/base/html/zMachine.html) (R Core Team, 2017b). Short codes for the macro-groups taken from Table 3.1.

	Afr	**MiE**	**WEu**	**EEu**	**Vol**	**SoA**	**WSi**	**SSi**	**CSi**	**NSi**	**SeM**	**SeI**	**SAm**	**Oce**
Afr		**323.31**	**323.31**	**323.31**	**323.31**	**323.31**	**323.31**	**323.31**	**323.31**	**323.31**	**323.31**	**323.31**	**323.31**	**323.31**
MiE	**323.31**		$1.2*10^{-11}$	0	$3.7*10^{-10}$	$5.5*10^{-9}$	**3.2**	**4.44**	**323.31**	**323.31**	**8.94**	**323.31**	**323.31**	**323.31**
WEu	**323.31**	**1.33**		$8.6*10^{-12}$	0	$3.1*10^{-6}$	**3.92**	**5.59**	**323.31**	**323.31**	**10.67**	**323.31**	**323.31**	**323.31**
EEu	**323.31**	**3.83**	$1.2*10^{-3}$		$3.9*10^{-10}$	$1.3*10^{-7}$	**4.37**	**6.77**	**323.31**	**323.31**	**323.31**	**323.31**	**323.31**	**323.31**
Vol	**323.31**	0.09	$2.2*10^{-3}$	0.28		$5.8*10^{-6}$	**3.37**	**4.49**	**323.31**	**323.31**	**8.75**	**323.31**	**323.31**	**323.31**
SoA	**323.31**	0.04	**4.68**	**8.94**	**1.57**		**2.67**	**3.78**	**323.31**	**323.31**	**8.2**	**323.31**	**323.31**	**323.31**
WSi	**323.31**	**3.88**	0.14	$3.9*10^{-3}$	0.79	**7.59**		$1.1*10^{-11}$	**6.92**	**7.65**	0.02	**11.38**	**323.31**	**323.31**
SSi	**323.31**	0.98	$2.0*10^{-9}$	0.02	$1.1*10^{-4}$	**4.16**	0.31		**323.31**	**323.31**	0.18	**323.31**	**323.31**	**323.31**
CSi	**323.31**	**323.31**	**7.89**	**7.03**	**9.37**	**323.31**	**1.47**	**9.29**		$1.5*10^{-6}$	**5.06**	0.01	**323.31**	**7.95**
NSi	**323.31**	**323.31**	**323.31**	**323.31**	**323.31**	**323.31**	**4.07**	**323.31**	0.08		**5.81**	$7.8*10^{-7}$	**323.31**	**6.5**
SeM	**323.31**	**2.08**	$6.2*10^{-7}$	$4.7*10^{-8}$	0.04	**5.78**	0.02	$1.9*10^{-4}$	**6.15**	**10.28**		**9.6**	**323.31**	**323.31**
SeI	**323.31**	**5.96**	0.24	0.01	1.22	**323.31**	$2.1*10^{-10}$	0.59	**3.91**	**7.95**	0.02		**323.31**	**6.42**
SAm	**323.31**	**323.31**	**323.31**	**323.31**	**323.31**	**323.31**	**323.31**	**323.31**	**5.52**	**2.18**	**323.31**	**323.31**		$1.2*10^{-6}$
Oce	**323.31**	0.07	$6.0*10^{-9}$	$8.6*10^{-4}$	$2.1*10^{-8}$	0.88	0.06	0	**3.37**	**5.82**	$1.8*10^{-5}$	0.06	**323.31**	

Appendix C.18 Matrix displaying the outcomes of Tukey's HSD for each pairwise inter-macro-group comparison of the per individual derived SNP counts for CADD30 variants

The results are reported in the form of the negative common logarithm of the p-values. The lower half contains the values for the total number of SNPs, the upper half for homozygous genotypes only. Cells containing the p-values for comparisons that are significant are bolded. Cells containing a value of 323.31 are equivalent to a p-value of ~5*10-324. The latter is the smallest number that can reliably be represented by the R statistics software used for these calculations (https://stat.ethz.ch/R-manual/R-devel/library/base/html/zMachine.html) (R Core Team, 2017b). Short codes for the macro-groups taken from Table 3.1.

	Afr	**MiE**	**WEu**	**EEu**	**Vol**	**SoA**	**WSi**	**SSi**	**CSi**	**NSi**	**SeM**	**SeI**	**SAm**	**Oce**
Afr		0.80	0.07	0.61	0.93	0.02	**1.64**	**2.18**	**4.09**	**2.88**	**2.73**	**9.61**	**323.31**	**7.29**
MiE	**323.31**		0.00	0.00	0.00	0.02	0.00	0.00	0.13	0.02	0.01	**2.63**	**323.31**	**2.93**
WEu	**323.31**	0.46		0.00	0.02	0.00	0.20	0.28	**1.48**	0.72	0.58	**6.28**	**323.31**	**4.89**
EEu	**323.31**	0.96	0.00		0.00	0.00	0.04	0.05	0.94	0.31	0.22	**6.16**	**323.31**	**4.32**
Vol	**323.31**	0.10	0.00	0.00		0.05	0.00	0.00	0.03	0.00	0.00	**1.67**	**323.31**	**2.38**
SoA	**323.31**	0.00	0.02	0.09	0.00		0.31	0.44	**1.76**	0.94	0.80	**6.60**	**323.31**	**5.18**
WSi	**323.31**	0.44	0.00	0.00	0.00	0.03		0.00	0.00	0.00	0.00	0.44	**9.96**	**1.36**
SSi	**323.31**	0.00	0.06	0.24	0.00	0.00	0.08		0.00	0.00	0.00	1.57	**323.31**	**2.13**
CSi	**323.31**	**3.09**	0.11	0.11	0.20	**1.41**	0.00	**1.94**		0.00	0.00	0.18	**10.44**	0.98
NSi	**323.31**	**3.23**	0.19	0.20	0.30	**1.58**	0.02	**2.10**	0.00		0.00	0.37	**10.76**	1.24
SeM	**323.31**	**1.98**	0.01	0.00	0.03	0.66	0.00	1.00	0.00	0.00		0.71	**323.31**	**1.54**
SeI	**323.31**	1.26	0.00	0.00	0.00	0.21	0.00	0.43	0.02	0.06	0.00		**7.07**	0.09
SAm	**323.31**	0.68	0.00	0.00	0.00	0.12	0.00	0.21	0.00	0.00	0.00	0.00		0.61
Oce	**323.31**	0.48	0.00	0.00	0.00	0.09	0.00	0.15	0.00	0.00	0.00	0.00	0.00	

Appendix C.19 Macro-group-level per individual SNP counts for different groups of derived high confidence LoF variants from LOFTEE

Given for each of the 14 macro-groups in the Variant-Based Analysis Set and all classified as high confidence loss-of-function by the LOFTEE plugin to Ensembl's VEP. For variants with multiple consequences only the most severe (Appendix C.8) is reported here. Abbreviations: hom – homozygote genotypes.

Macro-group	Stop gain	Stop gain hom	Splice site loss	Splice site loss hom
Afr	79.1	10.5	37.5	5.0
MiE	63.4	13.7	34.0	8.0
WEu	62.1	13.8	33.4	7.8
EEu	62.4	13.1	34.2	7.4
Vol	61.4	13.7	36.0	8.6
SoA	65.2	14.6	32.8	8.0
WSi	61.6	14.3	34.6	9.7
SSi	62.8	13.9	35.8	8.7
CSi	59.4	13.6	35.8	10.7
NSi	60.4	15.5	35.2	11.1
SeM	61.8	15.2	33.8	8.6
SeI	60.7	16.4	32.9	9.0
SAm	53.9	17.8	31.6	11.9
Oce	63.3	18.9	31.3	8.7
whole dataset	62.3	14.3	34.7	8.7

Appendix C.20 Matrix displaying the outcomes of Tukey's HSD for each pairwise inter-macro-group comparison of the per individual derived SNP counts for high confidence LoF variants from LOFTEE

The results are reported in the form of the negative common logarithm of the p-values. The lower half contains the values for the total number of SNPs, the upper half for homozygous genotypes only. Cells containing the p-values for comparisons that are significant are bolded. Cells containing a value of 323.31 are equivalent to a p-value of ~5*10-324. The latter is the smallest number that can reliably be represented by the R statistics software used for these calculations (https://stat.ethz.ch/R-manual/R-devel/library/base/html/zMachine.html) (R Core Team, 2017b). Short codes for the macro-groups taken from Table 3.1.

	Afr	**MiE**	**WEu**	**EEu**	**Vol**	**SoA**	**WSi**	**SSi**	**CSi**	**NSi**	**SeM**	**SeI**	**SAm**	**Oce**
Afr		**5.76**	**6.01**	**4.54**	**6.34**	**8.03**	**9.07**	**8.77**	**323.31**	**323.31**	**323.31**	**323.31**	**323.31**	**323.31**
MiE	**323.31**		0.00	0.01	0.00	0.00	0.10	0.00	0.46	**3.40**	0.12	**2.22**	**7.15**	**2.12**
WEu	**323.31**	0.00		0.01	0.00	0.00	0.17	0.00	0.68	**4.03**	0.22	**2.84**	**7.85**	**2.38**
EEu	**323.31**	0.00	0.00		0.10	0.37	**1.37**	0.51	**3.26**	**8.37**	**2.00**	**7.74**	**323.31**	**4.31**
Vol	**323.31**	0.00	0.00	0.00		0.00	0.01	0.00	0.07	**1.99**	0.01	0.96	**5.47**	**1.38**
SoA	**323.31**	0.00	0.01	0.00	0.00		0.00	0.00	0.05	**2.02**	0.00	0.96	**5.63**	1.30
WSi	**323.31**	0.00	0.00	0.00	0.00	0.00		0.00	0.00	0.20	0.00	0.00	**2.51**	0.24
SSi	**323.31**	0.00	0.04	0.00	0.00	0.00	0.00		0.06	**2.29**	0.00	1.17	**6.00**	**1.38**
CSi	**323.31**	0.00	0.00	0.00	0.00	0.01	0.00	0.07		0.24	0.00	0.00	**2.94**	0.24
NSi	**323.31**	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00		0.61	0.00	0.36	0.00
SeM	**323.31**	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00		0.07	**3.68**	0.47
SeI	**323.31**	0.15	0.00	0.12	0.10	0.36	0.00	0.81	0.00	0.00	0.00		**1.75**	0.02
SAm	**323.31**	**3.72**	**2.45**	**3.95**	**3.37**	**4.32**	**2.22**	**5.19**	**2.29**	**2.31**	**2.50**	**1.54**		0.00
Oce	**9.71**	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.59	

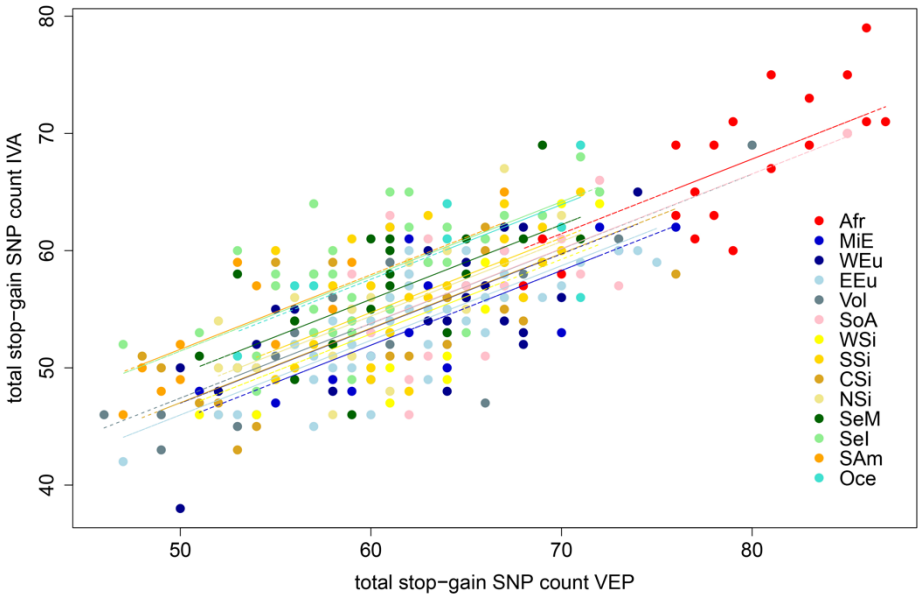
Appendix C.21 Correlation coefficients for the comparisons of stop-gain site counts annotated with IVA and VEP

Different methodologies were used to analyse the relationship between the number of stop-gain SNPs and homozygous derived sites detected by the VEP LOFTEE plugin (high confidence) and IVA.

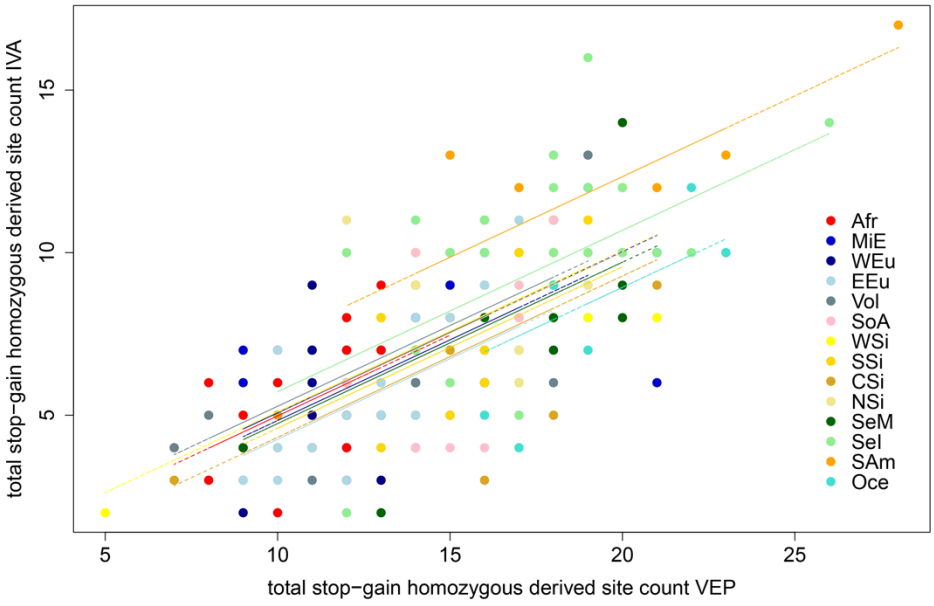
Sites	Treatment of Data	Method	Correlation coefficient and 95% CI	P-value
stop-gain SNPs	pooled	Pearson	0.75 [0.70,0.79]	$< 2.2*10^{-16}$
stop-gain SNPs	pooled	Spearman	0.68 [0.61,0.74]	$< 2.2*10^{-16}$
stop-gain SNPs	averaged by macro-group	Pearson	0.84 [0.57,0.95]	$1.4*10^{-4}$
stop-gain SNPs	averaged by macro-group	Spearman	0.53 [-0.07,0.88]	0.052 (ns)
stop-gain SNPs	intra-macro-group	Repeated measures correlation	0.72 [0.67,0.77]	$5.7*10^{-60}$
stop-gain homozygous derived	pooled	Pearson	0.70 [0.65,0.75]	$< 2.2*10^{-16}$
stop-gain homozygous derived	pooled	Spearman	0.68 [0.62,0.74]	$< 2.2*10^{-16}$
stop-gain homozygous derived	averaged by macro-group	Pearson	0.85 [0.58,0.95]	$1.2*10^{-4}$
stop-gain homozygous derived	averaged by macro-group	Spearman	0.93 [0.71,1.00]	$< 2.2*10^{-16}$
stop-gain homozygous derived	intra-macro-group	Repeated measures correlation	0.64 [0.57,0.69]	$3.8*10^{-43}$

Appendix C.22 Repeated measures correlations between stop-gain site counts annotated with IVA and VEP

A) Number of stop-gain SNPs detected by the LOFTEE VEP plugin (high confidence variants) compared to IVA. Regression lines show the fit of a common (i.e. identical regression slope for each macro-group) repeated-measures correlation design. Short codes for the macro-groups taken from Table 3.1.



B) Number of homozygous derived stop-gain SNPs detected by the LOFTEE VEP plugin (high confidence variants) compared to IVA. Regression lines show the fit of a common (i.e. identical regression slope for each macro-group) repeated-measures correlation design. Short codes for the macro-groups taken from Table 3.1.



Appendix C.23 Multiple regression analyses of the number of homozygous derived LoF sites inferred from IVA

Determination coefficients obtained from multiple regression analyses aiming to predict the number of derived homozygous genotypes for stop gained and splice donor/acceptor disrupting sites inferred by IVA (n = 2590 when only sites for which the reference is ancestral are considered) by the distance from Africa and the long term N_e since the OAA event. The addition of N_e was evaluated by an ANOVA using the chi-square statistic; the respective p-value for the comparison of the model pairs is recorded. Abbreviations: D_AA – distance from Addis Ababa, D_WH – distance from Windhoek

	r^2 (D_AA)	**r^2 (D_AA +N_e)**	**P**	**r^2 (D_WH)**	**r^2 (D_WH +N_e)**	**P**
LoF IVA (n = 382)	0.1994	-	-	0.2177	-	-
LoF IVA (n = 215)	0.2448	0.277	0.002122	0.2695	0.2805	0.07181

Appendix C.24 Matrix displaying the outcomes of pairwise chi-square tests comparing binned CADD DAF spectra

Comparisons between each of the 14 macro-groups, given in the form of the negative common logarithm of the p-values. For all groups except the Oceanians these spectra represent the mean (rounded to full numbers) counts obtained from repeated downsampling to n = 8. The upper half contains the values for the DAF spectra of CADD30 the lower half for those of CADD20 variants. Cells containing the p-values for comparisons that are significant according to the Bonferroni-corrected threshold ($p = 5.495*10^{-4}$ as there are 91 comparisons each, equal to a logged p of ~3.26) are bolded. Cells containing a value of 323.31 are equivalent to a p-value of $\sim 5*10^{-324}$. The latter is the smallest number that can reliably be represented by the R statistics software used for these calculations (https://stat.ethz.ch/R-manual/R-devel/library/base/html/zMachine.html) (R Core Team, 2017b). Short codes for the macro-groups taken from Table 3.1.

	Afr	**MiE**	**WEu**	**EEu**	**Vol**	**SoA**	**WSi**	**SSi**	**CSi**	**NSi**	**SeM**	**SeI**	**SAm**	**Oce**
Afr		0.79	1.5	1.9	1.72	0.76	**3.89**	2.03	**6.6**	**7.63**	2.13	3.33	**11.44**	**8.6**
MiE	**127.48**		0.11	0.23	0.16	0.02	1.48	0.27	3.17	**4.2**	0.35	0.9	**5.75**	**4.23**
WEu	**183.65**	**6.2**		0.05	0.03	0.13	1.2	0.07	2.61	**3.71**	0.07	0.54	**4.4**	**3.37**
EEu	**197.12**	**9.98**	0.51		0.01	0.3	0.77	0	1.94	2.91	0.13	0.25	**3.63**	2.6
Vol	**176.89**	**6.78**	0.42	0.54		0.23	0.89	0.01	2.16	3.17	0.12	0.34	**3.97**	2.88
SoA	**133.89**	0.08	**4.91**	**8.34**	**5.5**		1.81	0.35	**3.62**	**4.76**	0.32	1.1	**6.07**	**4.68**
WSi	**286.42**	**39.18**	**14.27**	**9.55**	**13.79**	**35.91**		0.73	0.34	0.72	1.44	0.21	2.23	0.85
SSi	**231.78**	**14.41**	2.83	2.66	**4.9**	**12.46**	**10.71**		1.87	2.85	0.14	0.22	**3.53**	2.52
CSi	**323.31**	**94.35**	**52.79**	**44.87**	**55.09**	**89.26**	**16.16**	**36.56**		0.15	2.76	0.82	1.25	0.16
NSi	**323.31**	**142.02**	**89.42**	**78.37**	**91.23**	**135.84**	**35.71**	**69.52**	**5.55**		4.01	1.56	1.85	0.4
SeM	**256.31**	**20.07**	**6.83**	**6.7**	**10.06**	**17.9**	**13.26**	0.94	**32.88**	**64.95**		0.6	**3.96**	**3.33**
SeI	**306.59**	**37.72**	**13.86**	**10.44**	**15.8**	**34.46**	**3.5**	**5.94**	**13.31**	**35.44**	**5.31**		2.18	1.24
SAm	**323.31**	**151.16**	**100.82**	**92.28**	**107.21**	**145.13**	**54.81**	**72.97**	**12.93**	**12.09**	**62.41**	**41.63**		0.57
Oce	**323.31**	**76.88**	**43.41**	**38.97**	**48.79**	**72.53**	**22.49**	**24.73**	**8.69**	**24.78**	**18.04**	**10.15**	**14.56**	

Appendix C.25 Matrix displaying the outcomes of pairwise chi-square tests comparing binned DAF spectra for high confidence LoF variants from LOFTEE

Matrix displaying the outcomes of the pairwise chi-square tests comparing the binned HC LoF variant spectra between each of the 14 macro-groups. For all groups except the Oceanians these spectra represent the mean (rounded to full numbers) counts obtained from repeated downsampling to n = 8. Cells containing the p-values for comparisons that are significant according to the Bonferroni-corrected threshold ($p = 5.495*10^{-4}$ as there are 91 comparisons each) are bolded. Short codes for the macro-groups taken from Table 3.1.

	Afr	**MiE**	**WEu**	**EEu**	**Vol**	**SoA**	**WSi**	**SSi**	**CSi**	**NSi**	**SeM**	**SeI**	**SAm**	**Oce**
Afr		0.006	0.005	0.003	0.003	0.007	**$9.0*10^{-5}$**	0.001	**$5.2*10^{-7}$**	**$3.6*10^{-8}$**	**$5.2*10^{-4}$**	**$2.3*10^{-5}$**	**$1.2*10^{-9}$**	**$2.2*10^{-6}$**
MiE			0.993	0.928	0.922	0.756	0.492	0.880	0.080	0.024	0.739	0.390	0.009	0.164
WEu				0.903	0.961	0.804	0.489	0.901	0.082	0.025	0.803	0.406	0.011	0.201
EEu					0.784	0.535	0.695	0.907	0.160	0.057	0.660	0.542	0.017	0.167
Vol						0.887	0.417	0.878	0.066	0.019	0.901	0.381	0.012	0.267
SoA							0.194	0.607	0.019	0.004	0.724	0.175	0.004	0.168
WSi								0.717	0.559	0.298	0.480	0.922	0.104	0.259
SSi									0.186	0.068	0.878	0.665	0.036	0.337
CSi										0.893	0.111	0.627	0.376	0.152
NSi											0.037	0.354	0.414	0.071
SeM												0.520	0.037	0.508
SeI													0.204	0.433
SAm														0.227
Oce														

Appendix C.26 Bulk data of high confidence LoF variants from LOFTEE

This file can be found attached to the electronic version of this thesis. It is a VCF file containing 4,282 high confidence LoF variants in 382 individuals from 14 worldwide macro-groups detected by the LOFTEE Plugin based on information from the GRCh37-compatible version of Ensembl Release 87. The INFO column contains the following abbreviations: AA – ancestral allele inferred from six primate species, AN – ancient (hominin) Neanderthal, AD – ancient (hominin) Denisovan, MOST_SEVERE_CSQ – most severe consequence inferred by VEP based on the precedence criteria given in Table X+7 ,SING – singleton site, VAR – variant site (i.e. at least two individuals carry at least one non – reference allele). The order of entries in the next subfield CSQ (consequences) is described in the second row of the VCF header.

Please note that for five loci (4:663,836; 5:147,207,583; 16:456,363; 16:67,195,842; 17:78,396,004) the most severe consequence provided by VEP was "stop_lost", however these were also called as "splice_acceptor"/"splice_donor" respectively. The latter annotation was reported here as LOFTEE does not evaluate the confidence status of "stop_lost" variants.

Appendix C.27 Gene Ontology categories significantly enriched or depleted in homozygous LoF-containing genes compared to the genome as a whole

The Bonferroni-corrected p-values were obtained using the PANTHER overrepresentation test. Abbreviations: BP – biological process, GO - Gene Ontology, MF – molecular function.

GO category	Type	Homozygous LoF genes	All genes	Enrichment/ Depletion	Corrected P
developmental process (GO:0032502)	BP	55/355	5487/21002	0.59	0.0093
anatomical structure development (GO:0048856)	BP	51/355	5122/21002	0.59	0.0204
metal ion binding (GO:0046872)	MF	107/355	4152/21002	1.52	0.0066
cation binding (GO:0043169)	MF	107/355	4240/21002	1.49	0.0177
protein binding (GO:0005515)	MF	148/355	11174/21002	0.78	0.0276

Appendix C.28 CDS length of a gene in relation to the probability of carrying a homozygous LoF variant

Long genes should be more likely to carry any observable kind of LoF variation assuming that mutation rates and selective constraints are comparable (Lek et al., 2016). To explore this, the CDS for all protein-coding autosomal genes in Ensembl version 75 were extracted ($n_{Ensembl\_Genes}$ = 19,270). On average, the CDS of genes containing at least one homozygous LoF is 2027 bp long, significantly greater than for genes without such a variant at 1755 bp (two-sample-t-test, t = -2.3652, p = 0.01855). One further way to test if gene length is the main determinant for having at least one homozygous LoF is to examine whether it is sufficient as a predictor in a simple logistic regression model. Such a model can be considered a significant improvement over the null model (ANOVA using the chi-square statistic, p = 0.04390). However, the amount of explained variance as given by a form of "pseudo r^2", the Nagelkerke index (Nagelkerke, 1991), is only $1.26*10^{-3}$. While there is still considerable debate over the exact interpretation of this metric (Smith and McKenna, 2013) and how it relates to the r^2 obtained from a "standard" ordinary least squares regression between continuous variables this outcome indicates that CDS length is not the main factor explaining whether a genes carries a homozygous LoF.

References

Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, Tukiainen T, Birnbaum DP, Kosmicki JA, Duncan LE, Estrada K, Zhao F, Zou J, Pierce-Hoffman E, Berghout J, Cooper DN, Deflaux N, DePristo M, Do R, Flannick J, Fromer M, Gauthier L, Goldstein J, Gupta N, Howrigan D, Kiezun A, Kurki MI, Moonshine AL, Natarajan P, Orozco L, Peloso GM, Poplin R, Rivas MA, Ruano-Rubio V, Rose SA, Ruderfer DM, Shakir K, Stenson PD, Stevens C, Thomas BP, Tiao G, Tusie-Luna MT, Weisburd B, Won H-H, Yu D, Altshuler DM, Ardissino D, Boehnke M, Danesh J, Donnelly S, Elosua R, Florez JC, Gabriel SB, Getz G, Glatt SJ, Hultman CM, Kathiresan S, Laakso M, McCarroll S, McCarthy MI, McGovern D, McPherson R, Neale BM, Palotie A, Purcell SM, Saleheen D, Scharf JM, Sklar P, Sullivan PF, Tuomilehto J, Tsuang MT, Watkins HC, Wilson JG, Daly MJ, MacArthur DG, Exome Aggregation Consortium. 2016. Analysis of protein-coding genetic variation in 60,706 humans. Nature 536:285–291.

Nagelkerke NJD. 1991. A note on a general definition of the coefficient of determination. Biometrika 78:691–692.

Smith TJ, McKenna CM. 2013. A comparison of logistic regression pseudo R2 indices. Mult Linear Regres Viewp 39:17–26.

Appendix C.29 Filtering scheme and overlaps with other large genomic datasets for rare homozygous LoF variants

This file can be found attached to the electronic version of this thesis. The excel table displays the pruning of the rhLoF variants initially detected at less than 2% DAF (n = 116) in the Variant-Based Analysis Dataset by a multi-stage filtering process. Firstly, the rhLoF-containing genes were compared to such sets of genes derived from four large scale exome sequencing projects, the overlaps are given in the first sheet of the file. The second sheet provides information on the variants that were already present in a homozygous state in the 1000 Genomes, the Simons Genome Diversity Project or the ExAC browser.

Appendix C.30 Detailed information on 34 LoF variants that have not previously been described as homozygotes

This table contains information on the population they occur in, the type of LoF variant and their co-location with previously described variants. Furthermore, it is recorded whether the genes the rhLoF variants are located in are predicted to be LoF-tolerant based on empirical data from the >60,000 exomes in the ExAC project and how many functionally related tandem duplicated genes described by the Duplicated Genes Database exist. The asterisks mark samples that were first reported in previous studies other than Pagani et al., (2016), i.e. the Pygmies presented by Lachance et al., (2012) or would potentially have been recorded as already existing if the recent projects on Oceanian genomics were included in the filtering scheme (Malaspinas et al., 2016). The abbreviations used are as follows: COSM – IDs from the Catalogue of Somatic Mutations in Cancer (COSMIC) database, CS – HGMD accession ID (splice variant), CSQ – most severe reported consequence, SP_SEV – severe splice site variant, STG – stop gain variant. Short codes for the macro-groups taken from Table 3.1.

CHR	POS	Macro-group	Population	Gene Symbol	CSQ	Known Variant	CADD	Predicted LoF (gene level)	N(tandem duplicates)	Additional information
1	2121850	Afr	Baka Pygmies*	*AL590822.2*	STG	rs145881709	31		0	
1	157772422	EEu	Karelians	*FCRL1*	STG	rs144206423	34	tolerant	2	
1	173567165	MiE	Lebanese	*SLC9C2*	STG	rs773250925	32	tolerant	0	
1	248005039	SeI	Igorot	*OR11L1*	STG	rs376918027, COSM534395	34	tolerant	42	somatic mutation in lung, oesophageal and stomach cancers (1 sample each)
2	11606248	MiE	Jordanians	*AC099344.1*	SP_SEV	rs181333014	1.403		0	
2	98949903	EEu	Vepsas	*AC092675.3*	SP_SEV	rs774391598	0.171		0	
2	231193547	Afr	Congo Pygmies*	*SP140L*	SP_SEV	rs187842782	0.275	tolerant	1	
3	46599404	CSi	Yakuts	*LUZPP1*	STG	rs144667915	23.3		0	

CHR	POS	Macro-group	Population	Gene Symbol	CSQ	Known Variant	CADD	Predicted LoF (gene level)	N(tandem duplicates)	Additional information
3	123699212	CSi	Yakuts	*ROPN1*	SP_SEV	-	23.5	tolerant	1	
5	2752747	SeI	Aeta	*C5orf38*	STG	COSM738175	28.4		0	somatic mutation in lung cancer (1 sample)
5	180661930	WSi	Kets	*TRIM41*	SP_SEV	rs777713854	10.06	intolerant	0	
6	123122552	SeI	Lebbo	*SMPDL3A*	SP_SEV	rs192842064	24.1	tolerant	0	
6	131974033	CSi	Evenks	*ENPP3*	SP_SEV	rs781663942	23.5	tolerant	1	
7	1590620	Oce	Koinanbe*	*TMEM184A*	SP_SEV	rs369699223	22.5	tolerant	0	
7	141954912	SoA	Kapu	*PRSS58*	STG	-	31	tolerant	2	
7	157318751	NSi	Chukchi	*AC006372.1*	STG	rs148144246	32		0	
9	399261	Oce	Koinanbe*	*DOCK8*	SP_SEV	rs756871628	25.6	tolerant	0	
9	36276924	SeI	Bajo	*GNE*	STG	rs200763627	34	tolerant	0	
9	71155642	CSi	Evenks	*TMEM252*	STG	rs199544319	26.9	tolerant	0	
9	137777213	SAm	Calchaquíes	*FCN2*	SP_SEV	rs145512341	22.9	tolerant	1	
11	7694023	NSi	Koryaks	*CYB5R2*	STG	COSM1188159	32	tolerant	0	somatic mutation in lung cancer (1 sample)
11	64858002	NSi	Koryaks (2 individuals)	*VPS51*	STG	rs185287594	29.7	tolerant	0	
12	56308146	SeI	Igorot (2 individuals)	*PYM1*	STG	rs373671473	10.67		0	
12	117484627	MiE	Iranians	*TESC*	STG	rs376764770	7.573		0	
12	120961706	EEu	Vepsas	*COQ5*	STG	rs117192040	0.038	tolerant	0	
14	22133972	SeI	Igorot	*OR4E2*	STG	rs368987419	36	tolerant	23	
14	105849581	CSi	Evenks	*PACS2*	STG	rs782499636	4.704	intolerant	0	

CHR	POS	Macro-group	Population	Gene Symbol	CSQ	Known Variant	CADD	Predicted LoF (gene level)	N(tandem duplicates)	Additional information
16	3447343	SeI	Dusun	*ZSCAN32*	STG	rs367962404	9.721	tolerant	0	
16	84801966	SeI	Igorot	*USP10*	SP_SEV	rs189466547	23.1	intolerant	0	
17	7945810	Oce	Koinanbe* (2 individuals)	*ALOX15B*	SP_SEV	rs139363547	23.2	tolerant	4	
17	47284137	SoA	Low-caste [Madhya Pradesh]	*GNGT2*	SP_SEV	rs534482384	10.49		0	
18	43314238	SeI	Dusun	*SLC14A1*	SP_SEV	rs78937798, CS982369	24	tolerant	0	HGMD predicts that this mutation leads to a urea transport defect.
19	37488331	NSi	Eskimo	*ZNF568*	STG	rs574434686, COSM5636898	37	tolerant	30	somatic mutation in oesophageal cancer (1 sample)
20	31196370	Vol	Udmurts	*RP11-410N8.4*	STG	-	31		0	

Appendix C.31 Loci from a panel of high penetrance mutations underlying Mendelian childhood disorders detected in the Variant-Based Analysis Set

Note that half of the individuals carrying such variants are of European ancestry. The cystic fibrosis-related mutation marked with an asterisk for which one Murut individual is homozygous has received conflicting pathogenicity annotations in different databases (see main text). Abbreviations employed are as follows: AO-age of onset, CHR-Chromosome, CM- missense variant from HGMD, can be assumed for site ID unless stated otherwise, CON-Congenital or < 2 years, CS-splicing-related variant from HGMD, CT-disease category, H- highly severe with some variation, HGMD-Human Genome Mutation Database, M- most severe with significantly reduced mobility or increased mortality in early life, MB-metabolic, NU-neurological, OC-ocular, RP-respiratory, PNT-penetrance, PRV (est)- prevalence estimated based on number of all alleles in Variant-Based Analysis Set related to a particular disease, PRA- pre-adult, i.e. mostly < 18 years, PRV (lit)- prevalence based on literature mined by Chen et al. (2016), SEV-severity, VAR- < 18 years but more variable. Short codes for the macro-groups taken from Table 3.1.

CHR	POS	HGMD-ID [CM]	Gene Symbol	Indivi-dual ID	Population details	Phenotype	CT	PRV (lit)	PRV (est)	PNT (>)	SEV	AO
1	45974647	060075	*MMACHC*	CHB_4	SeM_Han	Methylmalonic Acidaemia	MB	1:50,000-100,000	1:583,700	90%	H	PRA
1	76198337	042915	*ACADM*	TSI_2; UkrnW2	WEu_Italians; EEu_Ukrainians_west	Medium Chain Acyl-CoA Dehydrogenase Deficiency	MB	1:4,900-170,000	1:64,900	90%	H	PRA
1	76215194	910003	*ACADM*	Kapu1	SoA_Kapu	Medium Chain Acyl-CoA Dehydrogenase Deficiency	MB	1:4,900-170,000	1:64,900	90%	H	PRA
2	38298394	-	*CYP1B1*	Assyr4, Armen4	MiE_Assyrians; MiE_Armenians	Primary Congenital Glaucoma	OC	1:5,000-22,000	1:145,900	95%	H	CON

CHR	POS	HGMD-ID [rest CM]	Gene Symbol	Indivi-dual ID	Population details	Phenotype	CT	PRV (lit)	PRV (est)	PNT (>)	SEV	AO
3	15686975	970191	*BTD*	Chuk13	NSi_Chukchi	Biotindidase Deficiency	MB	1:61,067	1:583,700	95%	H	CON
5	138386706	093177	*SIL1*	MPlc1; GIH_4; Khsat1	SoA_Low.caste; SoA_Gujaratis; SoA_Kshatriya	Marinesco-Sjogren syndrome	NU	unknown	1:64,900	90%	M	CON
7	117171029	-	*CFTR*	UkrN1	EEu_ Ukrainians_north	Cystic Fibrosis	RP	1:3,200 in European Americans	1:16,200	95%	H	PRA
7*	117199683*	941976*	*CFTR*	**Murut13 (HOM);** Igrt1; Vizyn3	**SeI_Murut (HOM);** SeI_Igorot; SeI_Vizayan	Cystic Fibrosis	RP	1:3,200 in European Americans	1:16,200	95%	H	PRA
7	117227792	CS900233	*CFTR*	CEPH_13	WEu_NW.Europeans	Cystic Fibrosis	RP	1:3,200 in European Americans	1:16,200	95%	H	PRA
9	34648167	910169	*GALT*	croat16	WEu_Croats	Galactosaemia	MB	1:10,000-30,000	1:145,900	99%	H	PRA
9	34649029	920296	*GALT*	YakS4	CSi_Yakuts_Sakha	Galactosaemia	MB	1:10,000-30,000	1:145,900	99%	H	PRA
9	104189856	880004	*ALDOB*	Selkp1; BelaR2; Lat1; Altai5; CEU_2	WSi_Selkups; EEu_Belarusians; EEu_Latvians; SSi_Altaians; WEu_NW.Europeans	Hereditary Fructose Intolerance	MB	1:20,000-30,000	1:23,300	90%	H	PRA
11	6638385	CS991344	*TPP1*	CEPH_15	WEu_NW.Europeans	Neuronal Ceroid Lipofuscinosis	NU	1.5-9: 1,000,000	1:583,700	90%	H	CON

CHR	POS	HGMD-ID [rest CM]	Gene Symbol	Indivi-dual ID	Population details	Phenotype	CT	PRV (lit)	PRV (est)	PNT (>)	SEV	AO
12	103234271	870016	*PAH*	Est1	EEu_Estonians	Phenylketonuria	MB	1:2,600-200,000 in Europeans	1:583,700	90%	H	PRA
15	80472572	CS930802	*FAH*	Mari1	Vol_Maris	Tyrosinaemia Type I	MB	1:100,000-200,000	1:583,700	90%	M	CON
16	8905010	971228	*PMM2*	Ger3; Ingr3; CEPH_14	WEu_Germans; EEu_Ingrians; WEu_NW.Europeans	Congenital Dis-orders of Glyco-sylation Type Ia	MB	1:20,000	1:64,900	90%	H	CON
17	78190860	971355	*SGSH*	Est4	EEu_Estonians	Mucopolysaccha-ridosis Type IIIA	MB	1:100,000-400,000	1:583,700	90%	M	PRA
19	13008638	960719	*GCDH*	Chuv1	Vol_Chuvashes	Glutaryl-CoA De-hydrogenase De-ficiency / Glutaric Acidaemia Type 1	MB	1:30,000-40,000	1:583,700	90%	H	PRA
21	45709656	970074	*AIRE*	Veps2	EEu_Vepsas	Polyglandular Autoimmune Syndrome	IM	common in Iranian Jews, Sar-dinians, Finns	1:583,700	90%	H	PRA
22	51065404	910051	*ARSA*	Swe1	EEu_Swedes	Arylsulfatase A Deficiency	NU	1:40,000-160,000	1:583,700	90%	H	VAR

Appendix C.32 DAF data for most differentiated missense variants

This file can be found attached to the electronic version of this thesis. The first sheet of the excel file (**Appendix C.32A**) contains the DAFs of the top 20 most differentiated (ΔDAF) missense variants by macro-groups. The second sheet (**Appendix C.32B**) reports these variants excluding all comparisons with Africans, as the 20 most extreme missense DAF differences are exclusively observed between the Africans and non-Africans. Because of the small sample sizes the Southern Americans and the Oceanians were not considered in the ΔDAF calculations; their frequencies are reported for comparative purposes only. * = The asterisk indicates that this site was highlighted as an outlier (>5 SD) by the DIND analyses.

To account for LD if more than two most differentiated SNPs were located in the same 200-kb window only one of the signals is reported here. Note that the maximum difference reported here was calculated from non-rounded values. This accounts for slight discrepancies to the rounded values of the population allele frequencies. Short codes for the macro-groups taken from Table 3.1.

Appendix C.33 DAF data for most differentiated non-African missense variants in five South Asian subclusters

This file can be found attached to the electronic version of this thesis. The first sheet of the excel file contains the DAFs of the top 20 most differentiated (ΔDAF) missense variants observed between all non-African groups for five subclusters of populations from South Asia. The second sheet provides details on the composition of these clusters. Note that there are three polymorphisms (rs1426654 in *SLC24A5*, rs885479 in *MC1R* and rs10497520 in *TTN*) for which ΔDAF > 0.5. Abbreviations employed: SoA – South Asia

Appendix C.34 DAF data for most differentiated nonsense variants

This file can be found attached to the electronic version of this thesis. The first sheet of the excel file (**Appendix C.34A**) contains the DAFs of the top 20 most differentiated (ΔDAF) nonsense variants by macro-groups. The second sheet (**Appendix C.34B**) reports these variants excluding all comparisons with Africans. However, many sites are reported in both tables, so that the total of unique variants is only n = 24. Sites that are unique to Appendix C.32B are

bolded. Because of the small sample sizes the Southern Americans and the Oceanians were not considered in the ΔDAF calculations; their frequencies are reported for comparative purposes only. * = The asterisks indicate that this site was highlighted as an outlier (>5 SD) by the DIND analyses. To account for LD if more than two most differentiated SNPs were located in the same 200-kb window only one of the signals is reported here. Note that the maximum difference reported here was calculated from non-rounded values. This accounts for slight discrepancies to the rounded values of the population allele frequencies. Short codes for the macro-groups taken from Table 3.1.

Appendix C.35 Results of d_i and DIND analyses

This file can be found attached to the electronic version of this thesis. The first sheet of the excel file (**Appendix C.35A**) contains the Top 12 of the most highly divergent SNPs by the d_i score and their DAFs in each of the twelve population groups from the Selection Set. Reported are those most highly differentiated SNPs that are not in high linkage disequilibrium ($r^2 < 0.05$) with any SNP with higher di score. Data were annotated using GRCh37/hg19 Ensembl Genes track from UCSC table browser. In case of SNPs within protein coding genes, all but 'protein_coding' annotations were discarded. Closest 5' and 3' genes are reported within 200 kb distance. If the ancestral allele status is unknown, human reference/alternative (REF/ALT) alleles are shown in the “ancestral/derived allele” column (marked with *).

The second sheet (**Appendix C.35B**) reports the variant findings by DIND analysis in the two highest ranking positive selection signals in each population group. Note for each of the twelve population groups a composite signal based on the empirical p-values of three selection tests was calculated and the two highest ranking windows were subjected to further DIND analyses. These tests were iHS, nSL and Tajima’s D, the results were calculated by colleagues as part of the collaboration leading to the Pagani et al. (2016) paper, however they are not further described here as they are not utilised in any other analysis in this thesis. The content of the other columns is as follows: DINDs - number of significant SNPs from the DIND analysis above a threshold of + 5SD in each of the frequency classes of the SNPs, based on a distribution created from regions without protein or RNA coding genes. Genes - genes with significant DIND SNPs, the gene with the most significant SNP is highlighted in bold SNPs - Build 37 Ensembl Release 75 positions of the most significant SNPs by their DIND score. Short codes for the macro-groups taken from Table 3.1. for both sheets.

Appendix C.36 Overlap of potential positive selection signals with GWAS results

This file can be found attached to the electronic version of this thesis. The first sheet of the excel file contains the overlap of SNPs highlighted either by ΔDAF, di or DIND with results of GWAS ($p < 10^{-6}$) from the GWAS Catalog (MacArthur et al., 2017) and GWAS Central (Beck et al., 2014) databases. The entries in the "Tests" column can be interpreted as follows: d_i indicates that the variant was highlighted in Appendix C.35A, DIND in Appendix C.35B and ΔDAF in Appendices C.32 and C.34. Abbreviations employed: ANC – ancestral allele, CHR- Chromosome, DER – derived allele, POS- Position. Short codes for the macro-groups taken from Table 3.1. The second sheet contains the references for the additional notes given for some loci.

Appendix C.37 Overlap of potential positive selection signals with significant single tissue cis-eQTLs from the GTEx database

This file can be found attached to the electronic version of this thesis. The excel file contains the overlap of SNPs highlighted either by ΔDAF, di or DIND with significant single tissue cis-eQTLs (≤ 1Mb from the transcription start site) across 48 tissues from the GTEx database (Lonsdale et al., 2013).

The significance of the latter was determined by the GTEx Consortium (Aguet et al., 2016) using a multi-step procedure accounting for a) that there are multiple genetic variants in the 1-Mb cis-association window from the transcription start of the genomic element whose mRNA expression levels were quantified and b) that there is a high number of such gene expression phenotypes measured throughout the genome.

The entries in the "Tests" column can be interpreted as follows: d_i indicates that the variant was highlighted in Appendix C.35A, DIND in Appendix C.35B and ΔDAF in Appendices C.32 and C.34. The type of transcript was retrieved from release 27 of the GENCODE database (released August 2017). The IDs used for the original mapping of cis-eQTL were from GENCODE 19 are given as well as their corresponding IDs in GENCODE 27.

Abbreviations employed:

Abs_rank – absolute rank of a particular gene in a tissue among all protein-coding genes (n = 19,822) by total expression,

ANC – ancestral allele, CHR- Chromosome, DER – derived allele, High_tissue_specific_expression – indicates whether a gene highlighted by a particular variant-gene expression level association is among the top 5% protein-coding genes both in terms of a) total mRNA quantified b) relative overexpression in the respective tissue where the relationship was highlighted,

POS- Position, Pval_nominal-Raw p-value from a t-test between the respective variant and the expression of the target gene,

Slope_der – normalised regression coefficient for a simple linear regression of the genotype at a respective site and the expression of the target gene, a positive value indicates an increase in gene expression with the number of derived alleles an individual carries, a negative value designates a decrease respectively,

Pval_beta –permutation-based p-value for the variant-gene pair with the strongest association between genotype and expression levels for each gene (in each tissue), additional FDR adjustment for repeating this procedure across all genes with eQTLs in that tissue, after this adjustment a gene had to have at least one significant eQTL at $\alpha = 0.05$, otherwise it was removed from further analyses,

Tss_dist – distance of the eQTL from the transcription start of the genomic element whose expression levels it influences,

Zscore_rank – z-scores were calculated for each gene across all 48 tissues to detect tissue-specific overexpression, this column indicates the rank of a particular gene among all protein-coding genes (n = 19,822) by this measure of overexpression in the tissue the respective variant-gene pair was detected in. Short codes for the macro-groups taken from Table 3.1.

Appendix C.38 Gene Ontology categories significantly enriched in protein-coding genes regulated by eQTLs that were highlighted as positive selection signals

This table lists genes that in at least one tissue are regulated by eQTLs that were also highlighted as top hits by either the DIND (Appendix C.35A), the d_i (Appendix C.35B) or the ΔDAF (Appendices C.32 and C.34). The Bonferroni-corrected p-values were obtained using the PANTHER overrepresentation test. Abbreviations: BP – biological process, GO - Gene Ontology, MF – molecular function.

GO category	Type	Genes with putatively selected eQTL	All genes	Enrichment/ Depletion	Corrected P
antigen processing and presentation of exogenous peptide antigen via MHC class II (GO:0019886)	BP	9/209	98/21002	9.23	0.00687
antigen processing and presentation of peptide antigen via MHC class II (GO:0002495)	BP	9/209	99/21002	9.14	0.00746
antigen processing and presentation of peptide or polysaccharide antigen via MHC class II (GO:0002504)	BP	9/209	100/21002	9.04	0.0081
activation of immune response (GO:0002253)	BP	19/209	557/21002	3.43	0.034
MHC class II protein complex (GO:0042613)	CC	9/209	18/21002	50.24	$5.2*10^{-10}$
MHC protein complex (GO:0042611)	CC	9/209	27/21002	33.5	$1.79*10^{-8}$
integral component of lumenal side of endoplasmic reticulum membrane (GO:0071556)	CC	7/209	29/21002	24.26	$3.12*10^{-5}$
lumenal side of endoplasmic reticulum membrane (GO:0098553)	CC	7/209	29/21002	24.26	$3.12*10^{-5}$
clathrin-coated endocytic vesicle membrane (GO:0030669)	CC	8/209	51/21002	15.76	$8.23*10^{-5}$
clathrin-coated endocytic vesicle (GO:0045334)	CC	8/209	67/21002	12	$6.38*10^{-4}$
ER to Golgi transport vesicle membrane (GO:0012507)	CC	7/209	60/21002	11.72	0.0039
COPII-coated ER to Golgi transport vesicle (GO:0030134)	CC	7/209	80/21002	8.79	0.0247

GO category	Type	Genes with putatively selected eQTL	All genes	Enrichment/ Depletion	Corrected P
clathrin-coated vesicle membrane (GO:0030665)	CC	8/209	107/21002	7.51	0.0192
Golgi-associated vesicle membrane (GO:0030660)	CC	8/209	108/21002	7.44	0.0206
endocytic vesicle membrane (GO:0030666)	CC	12/209	168/21002	7.18	$2.25*10^{-4}$
integral component of endoplasmic reticulum membrane (GO:0030176)	CC	10/209	148/21002	6.79	0.00397
intrinsic component of endoplasmic reticulum membrane (GO:0031227)	CC	10/209	154/21002	6.53	0.00561
clathrin-coated vesicle (GO:0030136)	CC	10/209	179/21002	5.61	0.0204
endocytic vesicle (GO:0030139)	CC	13/209	276/21002	4.73	0.00678
lytic vacuole membrane (GO:0098852)	CC	14/209	343/21002	4.1	0.0146
lysosomal membrane (GO:0005765)	CC	14/209	343/21002	4.1	0.0146
vacuolar membrane (GO:0005774)	CC	16/209	393/21002	4.09	0.00351
endosome membrane (GO:0010008)	CC	15/209	430/21002	3.51	0.0433
endoplasmic reticulum membrane (GO:0005789)	CC	29/209	1016/21002	2.87	$5.14*10^{-4}$
cytoplasmic vesicle membrane (GO:0030659)	CC	21/209	737/21002	2.86	0.0239
endoplasmic reticulum subcompartment (GO:0098827)	CC	29/209	1021/21002	2.85	$5.67*10^{-4}$
nuclear outer membrane-endoplasmic reticulum membrane network (GO:0042175)	CC	29/209	1038/21002	2.81	$7.91*10^{-4}$
vesicle membrane (GO:0012506)	CC	21/209	755/21002	2.8	0.0339
endoplasmic reticulum part (GO:0044432)	CC	31/209	1288/21002	2.42	0.00681
cytoplasmic vesicle part (GO:0044433)	CC	31/209	1422/21002	2.19	0.0463
vesicle (GO:0031982)	CC	69/209	4241/21002	1.63	0.0128
cytoplasmic part (GO:0044444)	CC	123/209	9373/21002	1.32	0.0336
MHC class II receptor activity (GO:0032395)	MF	6/209	11/21002	54.81	$6.33*10^{-6}$
MHC protein complex binding (GO:0023023)	MF	5/209	20/21002	25.12	0.00658

Appendix C.39 Allele frequencies of relevant *ACTN3* polymorphisms

This table (**Appendix C.39A**) contains the allele frequencies for rs1815739 (*ACTN3* function retained, note that this is the ancestral allele) and rs11227639 (*ACTN3* upregulated, note that this is the derived allele) for 14 macro-groups The East Eurasian clade (including Native Americans) for whom there is a strong correlation between the macro-group allele frequencies at these two loci (r ~0.76) is bolded. Short codes for the macro-groups taken from Table 3.1.

Population	***ACTN3* function retained**	***ACTN3* upregulated**
Afr	0.905	0.167
MiE	0.558	0.154
WEu	0.565	0.161
EEu	0.651	0.17
Vol	0.595	0.19
SoA	0.429	0.054
WSi	**0.676**	**0.794**
CSi	**0.452**	**0.435**
NSi	**0.72**	**0.72**
SSi	**0.515**	**0.382**
SeM	**0.552**	**0.207**
SeI	**0.489**	**0.222**
SAm	**0.077**	**0.115**
Oce	0.438	0.125

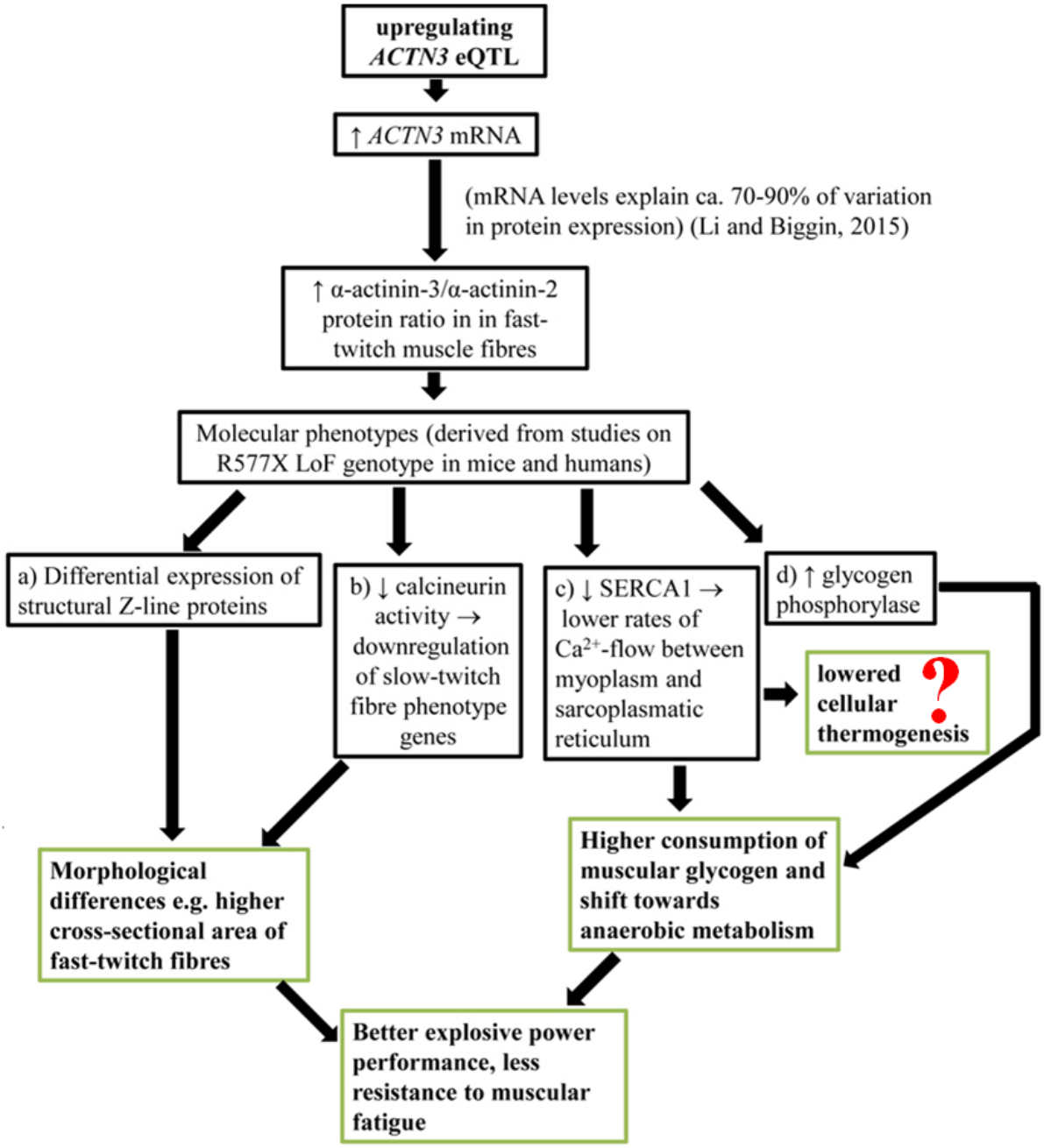
This table (**Appendix C.39B**) contains the allele frequencies for rs1815739 (*ACTN3* function retained, note that this is the ancestral allele) and rs11227639 (*ACTN3* upregulated, note that this is the derived allele) for 21 subpopulations from different parts of Siberia. Short codes for the macro-groups taken from Table 3.1.

Macro-group	**Population**	**Sample size**	***ACTN3* function retained**	***ACTN3* upregulated**
CSi	Nganasans	2	0.75	0.75
CSi	Evenks	13	0.538	0.5
CSi	Yakuts (Magadan)	1	0.5	0.5
CSi	Yakuts (Krasnoyarsk Krai)	4	0.625	0.5
CSi	Evens (Magadan)	4	0.2	0.4
CSi	Yakuts (Sakha)	3	0.5	0.333
CSi	Even (Sakha)	3	0	0
NSi	Eskimo	4	0.75	0.75
NSi	Koryaks	16	0.688	0.75

Macro-group	**Population**	**Sample size**	***ACTN3* function retained**	***ACTN3* upregulated**
NSi	Chukchi	5	0.8	0.6
SSi	Shor	2	1	0.5
SSi	Mongolians	6	0.417	0.417
SSi	Buryats	17	0.529	0.382
SSi	Altaians	6	0.417	0.333
SSi	Tuvinians	3	0.5	0.333
WSi	Forest Nenets	3	0.833	1
WSi	Khanty	3	0.667	0.833
WSi	Selkups	3	0.833	0.833
WSi	Tundra Nenets	3	0.333	0.833
WSi	Kets	3	1	0.667
WSi	Mansis	2	0.25	0.5

Appendix C.40 Summary of the consequences of higher ACTN3 mRNA levels

Mainly based on a recent review of the relevant literature by Lee et al. (2016). The red question mark indicates that it is currently unclear whether the changes in calcium handling and the higher activity of SERCA Ca^{2+}-ATPase inferred a on a cellular level are sufficient to meaningfully impact basal heat generation in humans.



Appendix C.41 Empirical p-values of iHS, nSL and Tajima's D selection tests in 200-kb windows

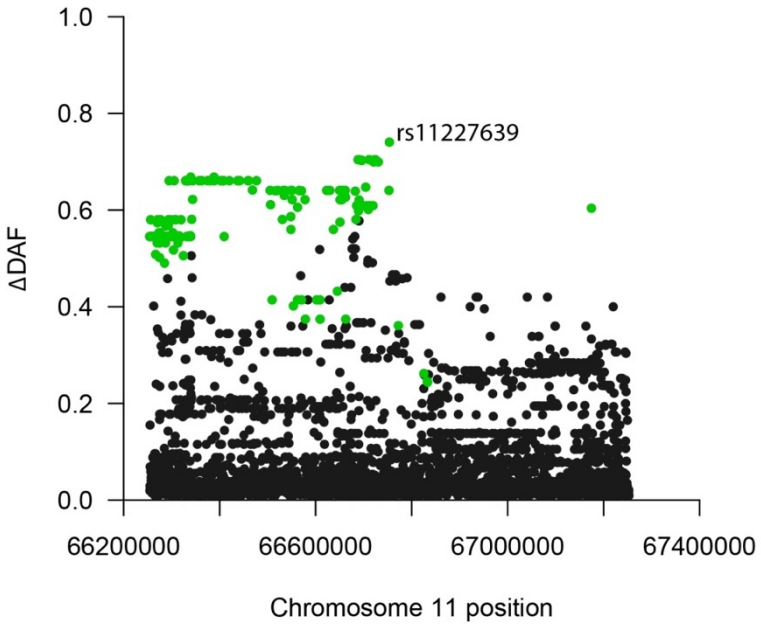
This file can be found attached to the electronic version of this thesis. It contains empirical p-values of three positive selection tests (iHS, nSL, Tajima's D) in 200-kb windows by the twelve macro-groups from the Selection Set (modified from Pagani et al. 2016). These data were generated by Evelyn Jagoda, Dr Guy Jacobs and Dr Charlotte Inchley.

Appendix C.42 Screenshot of the Gene eQTL Visualizer for *ACTN3*

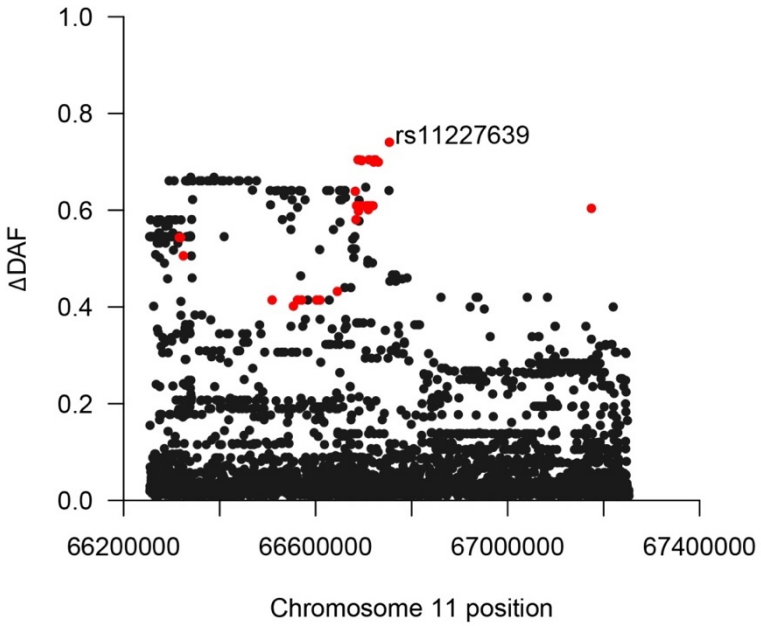
The size of the dots correlates with their nominal p-value, the colour indicates the directionality of the effect of the non-reference allele (red-higher expression, blue-lower expression). Filter settings for cis-eQTLs were $p_{nominal} \leq 10^{-9}$ and a normalised absolute effect size of ≥ 0.3. The sites which are marked with red squares in the LD plot exhibit $r^2 \geq 0.5$ with rs11227639 in the GTEx dense genotype data, all but one of them are significant eQTLs at the specified cut-offs. Underlying data as of 03/01/2018.

Appendix C.43 Manhattan plots of SNP-wise ΔDAF for the 1-Mb window surrounding rs11227639

Appendix C.43A: Manhattan plot of SNP-wise ΔDAF for the 1 Mb window surrounding the potential selection target rs11227639. Sites for which the Northeast Siberians have a MAF >0.5 are highlighted in green. Africans were not included in this ΔDAF calculation.



Appendix C.43B: Manhattan plot of SNP-wise ΔDAF for the 1-Mb window surrounding the potential selection target rs11227639. Sites for which the Northeast and West Siberians both have a MAF >0.5 are highlighted in red. Africans were not included in this ΔDAF calculation.



Appendix C.44 Cis-eQTLs from GTEx in tight LD with the potential selection candidate rs11227639

These Cis-eQTLs have an $r^2 \geq 0.5$ in the GTEx genotype data with the potential selection candidate rs11227639. LD values for the Northeast and West Siberian groups from this thesis are also displayed. The raw p-value for a t-test between the respective variant and the expression of the target gene (nominal P) and the normalised regression coefficient (slope) for a simple linear regression of the genotype at a respective site and the expression of the target gene are given. For the latter a positive value indicates an increase in gene expression with the number of derived alleles an individual carries, a negative value designates a decrease respectively. Abbreviations: Chr – chromosome, der – derived, DNase I hypersensitive site- DHS, Pos – Position. Short codes for the macro-groups taken from Table 3.1.

Chr:Pos	rs_ID	P (nominal)	Slope (der)	DAF (NSi)	DAF (WSi)	r^2 (NSi)	r^2 (WSi)	ΔDAF	DHS
11:66690486	rs79351784	$7.5*10^{-12}$	0.44	0.740	0.735	0.713	0.447	0.704	no
11:66695701	rs60520320	$6.8*10^{-12}$	0.45	0.738	0.654	0.627	0.447	0.702	no
11:66695721	rs57535372	$1.7*10^{-11}$	0.43	0.739	0.700	0.713	0.447	0.703	no
11:66721028	rs2278844	$8.5*10^{-12}$	0.44	0.729	0.735	0.713	0.447	0.700	no
11:66722703	rs58462309	$2.4*10^{-12}$	0.44	0.740	0.735	0.713	0.447	0.704	yes
11:66724371	rs6591228	$6.2*10^{-12}$	0.43	0.740	0.735	0.713	0.447	0.703	yes
11:66730596	rs76756200	$3.8*10^{-11}$	0.44	0.667	0.735	0.826	0.447	0.700	no
11:66753650	rs11227639	$3.6*10^{-10}$	0.4	0.720	0.794	-	-	0.741	yes

Appendix C.45 Reference bias in methods for the classification of deleterious variants

Both the SIFT and PolyPhen2 methods that were combined to indicate whether a variant has deleterious consequences for the IVA output (note that these scores are integrated in the IVA interactive interface) have previously been shown to exhibit a reference-bias (Simons et al., 2014). This effect could be replicated in the Variant-Based Analysis Set. A site where the reference (ref) is derived (der) is much more rarely called as (potentially) deleterious than a site where the reference is ancestral (anc). Where the anc allele is known with certainty (n_{total} = 266,993) the overall percentage of ref = der sites is ca. 5.3%, whereas among the SIFT+ PolyPhen2 sites the fraction is only 0.02%. This also affects the comparisons of the average frequency of deleterious variants between populations. The average DAF for ref = anc is 0.0281 for Africans and 0.0240 for non-Africans respectively, whereas the average DAF when ref ≠ anc is 0.5422 for Africans and 0.6201 for non-Africans. This likely leads to an underestimation

of the average number of deleterious variants in non-Africans as they have a higher DAF at sites where ref = der that are almost never called as deleterious. Among the Africans, this effect is most strongly pronounced for the Pygmies (n = 8, $DAF_{ref = anc} = 0.0290$, $DAF_{ref \neq anc} 0.5286$). This implies that the latter are most divergent from the reference sequence which is also consistent with their known population history of deep splits. Due to this reference bias for SIFT and PolyPhen2 the CADD score was used as an alternative method to describe differences with regards to deleterious variants. This approach represents a considerable improvement, but still exhibits a somewhat less extreme bias. Out of the 176,426 CADD20 sites where the ancestral allele could be inferred with high confidence there are 3,659 cases where the reference is derived, i.e. a total fraction of ca. 2.1%. This is still considerably lower than for all exonic variants, even though it should be noted that these CADD20 sites also consist of ca. 50% non-exonic variants. Furthermore, the average CADD score for this subgroup is 21.22 when ref ≠ anc and 23.42 for ref = anc.

The underlying reasons for these observations are most likely related to the nature of the predictive variables used to build the CADD machine learning model. These are annotations drawn from a) PolyPhen2/SIFT, see above and b) nucleotide-sequence-based conservation-based scores. For the latter a site might appear to be less conserved than it really is when the human reference is derived while other mammals in the alignment carry the ancestral state.

Other ways to address the problem of the reference bias are also discussed by Simons et al. (2014). One simple approach is to assume that the fraction of reference ancestral variants in a particular frequency bin that are classified as deleterious is a valid estimate of the fraction of reference derived sites that are deleterious assuming they fall in the same allele frequency class.

A second method uses a modified version of the position-specific independent count (PSIC) that is part of the PolyPhen-2 algorithm. The approach itself is based on comparing amino acid sequences from multiple species in an alignment. In the modified version the chimpanzee reference genome is used instead of the human reference. However, the results from this approach are currently only publicly available in pre-computed form for sites that were covered in phase 1 of the 1000 Genomes Project (http://genetics.bwh.harvard.edu/shaila/) and no distributable version of a software tool implementing this modified approach is available (Prof Shamil Sunyaev, personal communication, 07/03/2015, to the author's best knowledge this was still true as of 04/11/2017). This means a considerable amount of variation that is private to groups covered exclusively in the EGDP project would not be available for annotation.

References

Simons YB, Turchin MC, Pritchard JK, Sella G. 2014. The deleterious mutation load is insensitive to recent population history. Nature Genetics 46:220–224.

Appendix C.46 Effect of the number of variants per category on differences in the mean number of derived homozygotes

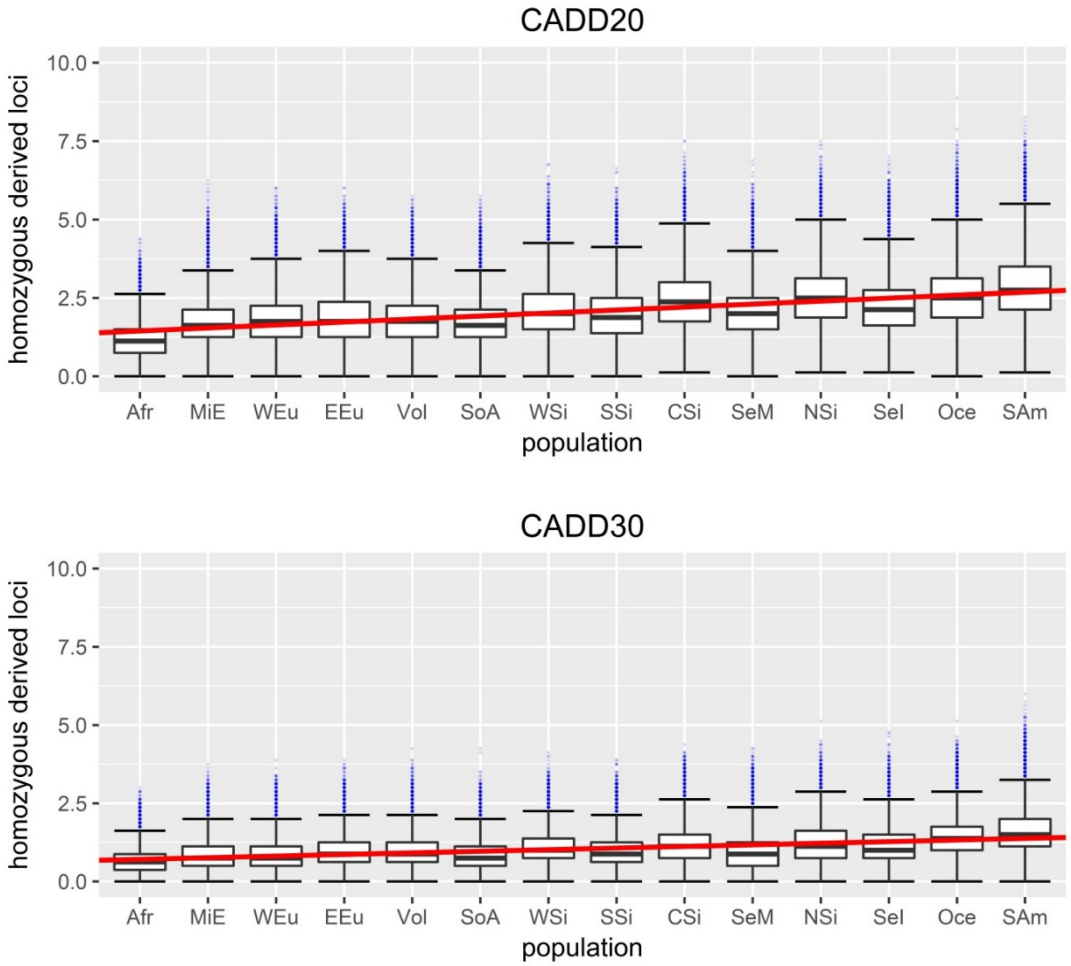
For the CADD20 variant class most macro-groups appear to be significantly differentiated (Figure 3.9) and there is a continuous increase in derived homozygote genotypes per individual with distance from Africa (Figure 3.12). For CADD30 as well as HC LoF variants both patterns are much less pronounced (Figures 3.10-3.12). One explanation is purifying selection to due to highly negative selection coefficients for the latter categories. However, these results could also be caused by a lack of power to find differences between groups due to the small variant counts in these categories. In terms of total variant counts CADD20 sites are observed ~14 times more frequently than CADD30 sites and ~42 times more often than HC LoF variants, however CADD30 variants are shifted more strongly towards rare DAFs than HC LoF (Figure 3.13).One possible way to address this is by choosing the number of derived homozygotes for one of the categories for which little differentiation was observed, followed by drawing the same number of sites randomly from the other categories and testing whether there are detectable differences. This was done using a two-step bootstrap procedure for all three relevant categories of deleterious variants.

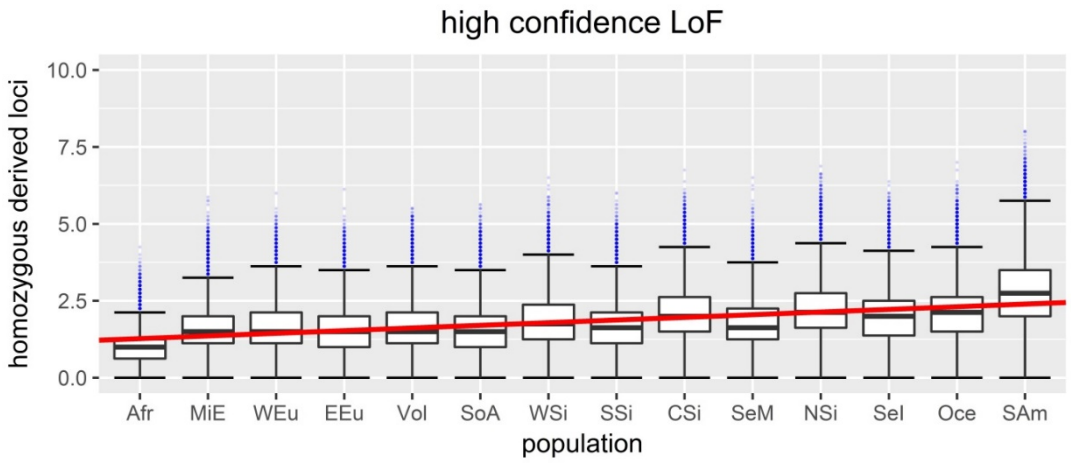
I) Eight individuals were randomly drawn from all macro-groups except the Oceanians who represent the smallest group, this procedure was repeated 100 times.

II) The total number of homozygous derived CADD30 sites was counted for a randomly chosen African individual (CongPy_6/GS000035247-ASM) who carried 19 such variants.

From each of the 100 replicates created in i) 19 sites were drawn 1000 times and the macro-group averages were calculated. This yielded a total of 100,000 data points for each group (for the Oceanians 100,000 samples of size 19 were drawn directly without step I). The outcomes of this approach are plotted below. They demonstrate that even for a small number of sites there is a measurable increase of derived homozygote counts with distance from Africa. However, the slope of the cline for CADD30 variants is flatter than for CADD20 and HC LoF. This can also be demonstrated by running linear regressions of the mean homozygous derived counts for

each group versus the mean per-group distance from Windhoek. Regression coefficients for CADD20 and HCLOF are ~$6.79*10^{-5}$ and $6.60*10^{-5}$ respectively, almost twice as high as for CADD30, where this value is only $3.85*10^{-5}$. This supports the hypothesis that the relative uniformity with regards to CADD30 derived homozygotes results from purifying selection as opposed to low power due to small sample sizes. Interestingly, for the HC LoF category the geographical differentiation is clearer than for the CADD30 which indicates that in this case these differences would be more easily detectable if there were more HC LoF sites in the genome. While this indicates that a substantial fraction of the HC LoF variants has selection coefficients that are closer to neutrality than for the CADD30 the low totals per each genome for the former class still suggest that the HC LoF variation is constrained by natural selection.



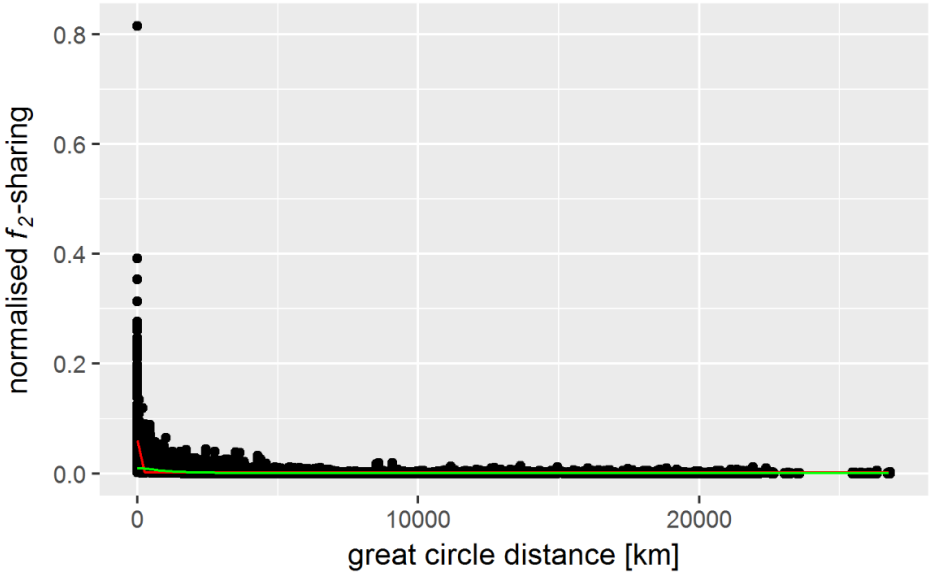


100,000 bootstrap replicas of size n = 19 variants were sub-sampled from different variant classes using a two-step procedure and displayed in a boxplot. The red lines indicate regressions of the macro-group means over all bootstraps versus an integer variable increasing in steps of 1 with distance from Africa. The order of macro-groups is based on mean distance from Windhoek and the short codes for the macro-groups are taken from Table X+5.

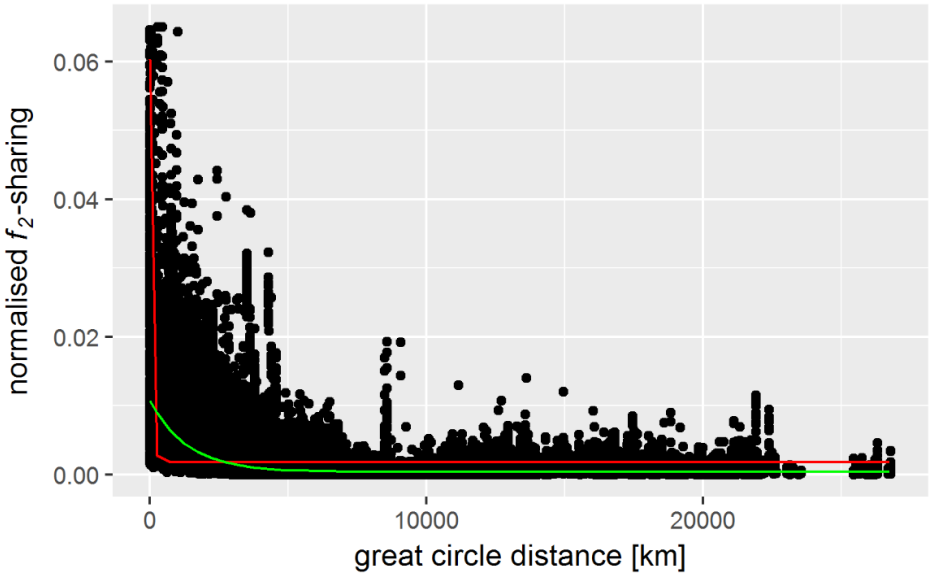
Appendix D

Appendix D.1 Normalised f_2 sharing as of a function of pairwise great circle distance

Appendix D.1A A normalised measure of the sharing of f_2 variants between all individuals from the Diversity Set (n = 447) as a function of the great circle distance (via plausible waypoints) between each pairs of individuals. Two exponential models were fit to this dataset using nonlinear least squares regression. The red line indicates a model fit on the raw data, the green line a simplified fit based on median f_2 sharing in 500 km bins.



Appendix D.1B This graph displays the same data as Appendix D.1A, however the f_2 sharing beyond 0.06 is not shown to better illustrate the fits for lower values of the normalised f_2 sharing metric.



Appendix D.1C

The geographical distance data and normalised f_2 sharing were fitted to the following negative exponential model using a nonlinear least squares regression approach implemented in the R function nls (R Core Team, 2017).

$$f_2 = a * e^{-bd} + c$$

To facilitate the finding of plausible starting parameter estimates for model fitting the f_2 sharing data were summarised by 500-km bins. The means of these bins were then plotted against the bin-wise median f_2 and first starting values were inferred using the trendline feature in Microsoft Excel 2013 (a = 0.0011, b = $6*10^{-5}$, L was not estimated at this stage). These initial values were then utilised as starting parameters for the nls function in R (the scalar L, in the following denominated as c, was set to a starting value of 0, see Eq. 4.3). This fitting was done for a) a simplified exponential model of distance bin means and bin-wise median f_2 and b) the parameter estimates from a) were in turn used for a nonlinear regression on the raw distance and f_2 data. The final exponential model was formulated as follows:

The independent variable d and the parameters a and b are the same as in equation 4.2, whereas the scalar correction factor L has been summarised to a single constant c (if the original formula is used for parameter inference L differs from c only by 10^{-4} which is less than 1/500th of the total parameter value and the residual standard errors are identical). The solution to the outlier detection problem outlined above proposed here is a normalised residual metric based on the best fitting model obtained using equation (4.3).

$$residual(norm) = \frac{y_i - \hat{y}_i}{\max(\hat{y}_i) - \min(\hat{y}_i)}$$

Here, y_i designates the observed f_2 sharing for the ith pair of individuals whereas $\hat{y}_i$ represents the predicted f_2 sharing between this individual pair given its geographical distance based on the exponential model. The difference of observed and expected doubleton sharing is divided by the range of predicted values. The latter divisor is introduced to take the error of the prediction into account. The maxima and minima were inferred with the propagate (Spiess, 2017) package in R which uses a Monte Carlo simulation approach to estimate the error of predictions derived from nonlinear models.

References

R Core Team. 2017. R: A language and environment for statistical computing. Available from: https://www.R-project.org/

Spiess A-N. 2017. propagate: Propagation of Uncertainty. Available from: https://CRAN.R-project.org/package=propagate

Appendix D.2 Testing whether RVCs represent continuous haplotypes

The overarching goal of this pipeline was to test whether regions in VCFs corresponding to RVCs shared by two individuals exhibit inconsistencies under the assumption that an RVC should represent a continuous haplotype unbroken by recombination events.

It consists of bash and R scripts, commented versions of which together with most input/output files can be found in the folder AppendixD2_RVC_consistency_pipeline attached to the electronic version of this thesis. The main exception are the files containing every variable position belonging to each of the 241 RVC runs, however one set of example files for the RVC with the genomic coordinates chr1:183,212,668-184,089,833 has been included.

Steps 1-4 were run in a UNIX server environment using R (version 3.4.0) with the additional package stringr (version 1.2.0) (Wickham, 2017). Step 5 was run locally on a Windows 8 machine with R (version 3.4.1) and includes functions from the packages data.table (version 1.11.2) (Dowle and Srinivasan, 2018), plyr (version 1.8.4) (Wickham, 2011) and stringr (version 1.2.0).

1. **create_vcf_multiple_chr.sh**: this script takes chromosome-specific text files (chrXsites_cac_afr.txt, groups here Africans and Calchaquíes) describing each RVC as CHR:START(BP)-END(BP) as input. It calls tabix 0.2.6 (Li, 2011) to extract region-specific VCF files spanning each RVC region from phased VCFs for the Diversity Set.
2. **vcf_name_content_change.sh**: this script edits the name and the content of the VCF files generated in step 1 so that they can be read into R.
3. **vcf_format_haplotype.sh**: this script is a wrapper that reads a tab-separated file list_vcf_cac_afr_unix line by line. This tab-separated file contains in its first column all VCF names generated in step 2 and in its second column the name of the target VCF for the output of this step. The contents of each line of this file are arguments for the

script **VCF_to_haplotype_unix_cac_afr.R** which splits the columns in the VCF containing phased genotype information into two haplotypes for each individual.

4. **vcf_haplotype_score.sh**: this script is a wrapper that reads a tab-separated file list_for_hap_score_cac_afr line by line. The tab-separated file consists of six columns and beginning from the first column contains the following entries: a) the names of VCF-like files with haplotypes, b)-e) assembly IDs of the two individuals sharing the RVC with V.1/V.2 added to distinguish the two haplotypes each individual carries, f) the names of files the output is recorded to.

   The contents of each line of this file are arguments for the script **VCF_hap_consistency_score_unix.R.** The latter compares the four haplotypes of the two individuals sharing this RVC region and assigns different states describing the consistency of each site with the hypothesis that the RVCs represent segments unbroken by recombination.
5. **process_report_data.R**: this script consists of three parts. In the first part all reports describing whether the sites forming the RVCs fulfil the consistency criteria from step 4 are summarised in one table (output: afr_cac_runs_annotated.txt). The second part filters out RVCs with at least one inconsistent homozygote, which indicates a run break regardless of phasing accuracy. These inconsistent homozygotes are used as break-points. Then it is tested whether any genomic chunks between these boundaries fulfil the original RVC definition, i.e. whether they contain at least five doubletons. In the final part of the script the genetic positions for the boundaries of these RVCs are imported (retrieved by applying base2genetic.jar) and the consistency information across these shortened runs is again summarised (output: afr_cac_fixedruns_annotated.txt).

References

Dowle M, Srinivasan A. 2018. data.table: Extension of `data.frame`. Available from: https://CRAN.R-project.org/package=data.table

Wickham H. 2011. The Split-Apply-Combine Strategy for Data Analysis. Journal of Statistical Software [Internet] 40. Available from: http://www.jstatsoft.org/v40/i01/

Wickham H. 2017. stringr: Simple, Consistent Wrappers for Common String Operations. Available from: https://CRAN.R-project.org/package=stringr

Appendix D.3 Example parameter files for cosi2 demographic simulations

Command to run coalescent simulations with cosi2 in UNIX environment:

mmg/alexmo/cosi-2.0/coalescent -p
/mmg/alexmo/cosi-2.0/examples/1000_genomes/params_scenarioX_chrN
-n 20 -m -P 15 > F2_scenarioX_chrN_n20.txt

Commented example of cosi2 parameter file for scenario A and chromosome 22

in bp

length 51304566

per bp per generation

mutation_rate 1.25e-8

gene_conversion_relative_rate 0.45

the recombination model is based on HapMap2 data

recomb_file /mmg/alexmo/cosi-
2.0/examples/1000_genomes/recom/Pagani2016_recomrates_reduced_chr22.txt

pop_define 1 YRI

pop_define 2 EAF

pop_define 3 CEU

pop_define 4 EEU

pop_define 5 SAM

YRI (W Africa)

#number of diploid individuals in population

#number of sampled haploid chromosomes

pop_size 1 14474

sample_size 1 18

EAF (East African, Hadza/Sandawe like)

pop_size 2 12229

sample_size 2 60

CEU (NW Europe) (also OoA population before West/East Eurasian split)

pop_size 3 338000

sample_size 3 418

EEU (East Eurasian, Han Chinese like, also encompasses Andeans before they split off)

pop_size 4 454000

sample_size 4 388

SAM (Andeans)

pop_size 5 25000

sample_size 5 10

#interpretation as follows, e.g. for first line European expansion from bottlenecked N_e of 1,032 #to 338,000 starting 1,080 generations ago and continuing through the present

pop_event exp_change_size "eur expansion" 3 0 1080 338000 1032

pop_event exp_change_size "as expansion" 4 0 880 454000 554

pop_event exp_change_size "Native American expansion" 5 0 880 25000 554

#for these low-level migrations the third number indicates the endpoint from which the
#migration rate is specified pastward

pop_event migration_rate "Wafr->Eafr migration" 1 2 0. .000025

pop_event migration_rate "Eafr->Wafr migration" 2 1 0. .000025

pop_event migration_rate "Wafr->eur migration" 1 3 0. .000025

pop_event migration_rate "eur->Wafr migration" 3 1 0. .000025

pop_event migration_rate "Wafr->as migration" 1 4 0. .0000078

pop_event migration_rate "as->Wafr migration" 4 1 0. .0000078

pop_event migration_rate "Wafr->SAM migration" 1 5 0. .00000078

pop_event migration_rate "SAM->Wafr migration" 5 1 0. .00000078

pop_event migration_rate "Eafr->eur migration" 2 3 0. .000025

pop_event migration_rate "eur->Eafr migration" 2 3 0. .000025

pop_event migration_rate "Eafr->as migration" 2 4 0. .0000078

pop_event migration_rate "as->Eafr migration" 4 2 0. .0000078

pop_event migration_rate "Eafr->SAM migration" 2 5 0. .00000078

pop_event migration_rate "as->Eafr migration" 5 2 0. .00000078

pop_event migration_rate "eur->as migration" 3 4 0. .0000311

pop_event migration_rate "as->eur migration" 4 3 0. .0000311

```
pop_event migration_rate "eur->SAM migration" 3 5 0. .00000311
pop_event migration_rate "SAM->eur migration" 5 3 0. .00000311

pop_event migration_rate "as->SAM migration" 4 5 0. .0000311
pop_event migration_rate "SAM->as migration" 5 4 0. .0000311

pop_event split "EEU/Native American split" 4 5 880

pop_event split "as/eur split" 3 4 1080

pop_event change_size "set OoA size" 3 1080 1861

pop_event migration_rate "Wafr->OoA migration" 1 3 1080 .00015
pop_event migration_rate "OoA->Wafr migration" 3 1 1080 .00015
pop_event migration_rate "Eafr->OoA migration" 2 3 1080 .00015
pop_event migration_rate "OoA->Eafr migration" 3 2 1080 .00015

pop_event split "West East Africa split" 1 2 1160

pop_event migration_rate "Afr->OoA migration" 1 3 1160 .00015
pop_event migration_rate "OoA->Afr migration" 3 1 1160 .00015

pop_event split "out of Africa" 1 3 2400

pop_event change_size "african expansion" 1 5920 7310
```

For scenario B the parameters chosen were identical to those above except adjustments to "Wafr<->SAM migration"/"Eafr<->SAM migration". The following line was added for scenarios C-F to simulate a single pulse admixture event:

pop_event admix "transatlantic_slave_trade" 5 1 18 0.005

The fields after the label designate a) the admixed target population (Pop5); b) the source population (Pop1) followed by c) the admixture date t and d) the admixture fraction a. The latter two vary between runs.

Appendix D.4 MSMC split time estimates

This file can be found attached to the electronic version of this thesis. The excel files contains MSMC split time estimates in kya, inferred when the relative cross coalescence rate equals 0.5, for all pairs of populations.

Appendix D.5 Average counts of f_2 variants per haploid genome for a "model world" depleted for Eastern Eurasians

Data displayed for a subset of individuals (n = 87) similar in sample size to the "model world" but with fewer individuals from East Asia/SEA and more Western Eurasians. Abbreviations except the following taken from Table 34: Eas- East Asians and SEA populations, Eur- Europeans.

Whole dataset	f_2 **counts**
Afr	47257.0
MiE	13087.1
Vol	13169.1
Eur	12952.9
SoA	11439.8
Sib	11480.1
Eas	11124.4
Ame	10599.9

Appendix D.6 Macro-group level summary of f_2 variant sharing between all individuals from the Diversity Set

Population abbreviations are taken from Table 34.

	Afr	MiE	WEu	EEu	Vol	SoA	CeA	WSi	SSi	CSi	NSi	SeM	SeI	Ame	Oce
Afr	1618485	80006	19463	15842	5702	13318	8499	2794	5330	2476	1753	9400	14985	40412	5167
MiE	80006	129255	63523	76846	30805	45670	42286	12108	18978	7938	2987	10965	9373	17501	3223
WEu	19463	63523	44423	88560	20983	19592	16248	8037	7827	2623	1556	4341	4506	14095	2446
EEu	15842	76846	88560	125257	49530	20470	26422	18434	15550	7692	4838	6222	8615	17893	3576
Vol	5702	30805	20983	49530	22090	9201	13522	13423	11959	6764	2715	6387	3575	4597	1063
SoA	13318	45670	19592	20470	9201	121268	30030	4720	10419	3188	1264	29738	16615	5303	2767
CeA	8499	42286	16248	26422	13522	30030	18185	9166	31903	13021	2829	23190	9362	5480	1126
WSi	2794	12108	8037	18434	13423	4720	9166	34300	13258	13874	3098	2927	2546	3149	592
SSi	5330	18978	7827	15550	11959	10419	31903	13258	35005	26418	4477	30287	10685	3881	1060
CSi	2476	7938	2623	7692	6764	3188	13021	13874	26418	33686	9690	8175	3313	2089	434
NSi	1753	2987	1556	4838	2715	1264	2829	3098	4477	9690	46192	2261	1245	2819	310
SeM	9400	10965	4341	6222	6387	29738	23190	2927	30287	8175	2261	67034	67494	2317	2023
SeI	14985	9373	4506	8615	3575	16615	9362	2546	10685	3313	1245	67494	296887	3999	26288
Ame	40412	17501	14095	17893	4597	5303	5480	3149	3881	2089	2819	2317	3999	45918	973
Oce	5167	3223	2446	3576	1063	2767	1126	592	1060	434	310	2023	26288	973	108599

Appendix D.7 Matrices containing f_2 totals and normalised f_2 sharing statistics for the Diversity Set

This file can be found attached to the electronic version of this thesis. The first sheet of the excel file (**Appendix D.7A**) is a raw sharing matrix of f_2 variant counts between all individuals from the Diversity Set (n = 447). The row and column names are formatted as "macrogroupID_individualID", the individual IDs are identical to those listed in the "ID" column of the master samples table Appendix C.1. The second sheet of the excel file contains the same sharing patterns (**Appendix D.7B**) for a normalised f_2 metric (rounded to the fifth decimal) that corrects for differences in overall genomic diversity. The third sheet reiterates the abbreviations for the macro-groups.

Appendix D.8 Matrices containing f_2 totals and normalised f_2 sharing statistics for the "model world" sampling scheme

This file can be found attached to the electronic version of this thesis. Note that these matrices represent the mean of 20 randomly sampled subsets of the Diversity Set to reduce the overrepresentation of non-African populations (n = 87). All Africans were retained and the remaining 48 non-Africans were randomly drawn from eight macro-groups (see Table 4.2).

The first sheet of the excel file (**Appendix D.8A**) is the mean of raw sharing matrices of f_2 variant counts The row and column names are formatted as "macrogroupID_individualID", the individual IDs are identical to those listed in the "ID" column of the master samples table Appendix C.1. The second sheet of the excel file contains the same sharing patterns (**Appendix D.8B**) for a normalised f_2 metric that corrects for differences in overall genomic diversity.

Appendix D.9 Matrices containing chunk numbers and total genetic length shared detected by ChromoPainter

This file can be found attached to the electronic version of this thesis. The first sheet of the excel file (**Appendix D.9A**) is a matrix containing the number of ChromoPainter/fineSTRUCTURE chunks shared between all individuals from the Diversity Set (n = 447). The row and column names are formatted as "macrogroupID_individualID", the

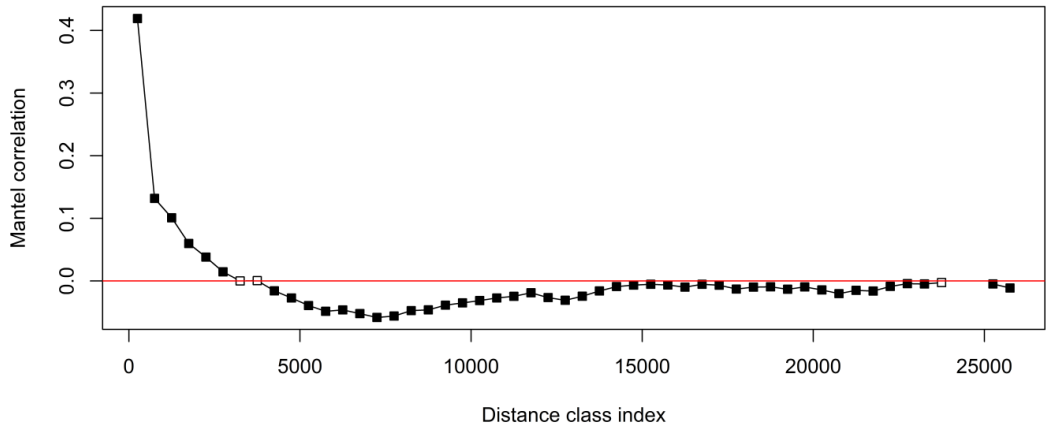
individual IDs are identical to those listed in the "ID" column of the master samples table Appendix C.1. The second sheet of the excel file (**Appendix D.9B**) contains the sharing of total genetic length of DNA in cM according to ChromoPainter/fineSTRUCTURE. Note that both matrices are not symmetrical and columns represent donor individuals and rows receptors respectively.

Appendix D.10 Outliers from a linear regression of f_2 and CP sharing

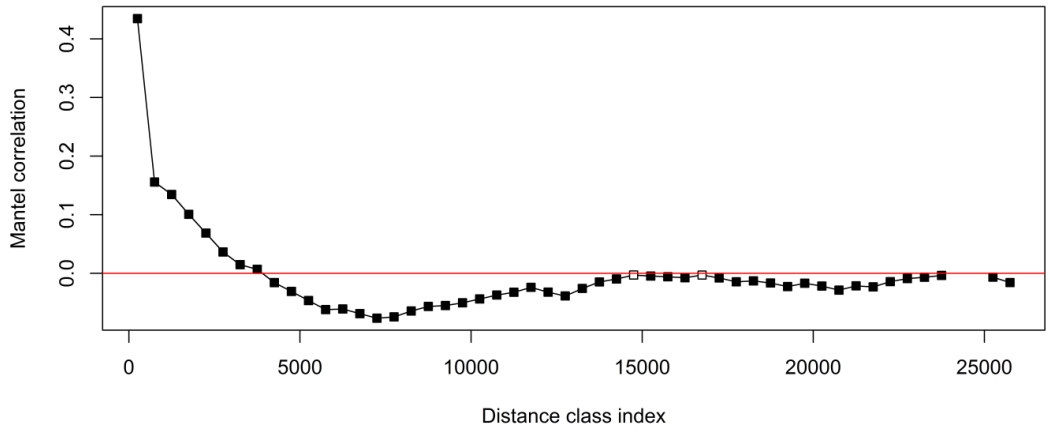
This file can be found attached to the electronic version of this thesis. It contains 1,741 pairs of individuals with a standardised residual with an absolute value greater than 3 obtained from the linear regression of f_2 and CP sharing respectively. Information on whether the individual has a known relative in the dataset and whether unusual metrics were reported by Complete Genomics (both taken from Appendix C.1) is also given. Note that the order of individuals inside each pair follows the matrices in Appendices D.7 and D.9. Abbreviations: ind-individual.

Appendix D.11 Mantel correlograms of great circle distance and i) f_2 sharing / ii) CP sharing

Appendix D.11A) Mantel correlogram of the great circle distance in km (in 500km bins) and a normalised f_2 sharing metric. Each point represents the Mantel correlation between the f_2 similarity matrix and a matrix denoting the absence/presence (0/1) of an individual pair in the respective distance class. Unfilled points correspond to a non-significant ($p > 0.05$) correlation after 1000 permutations and an additional correction for multiple testing (due to the high number of bins). The gap in the plot indicates distance classes without any pairs of individuals.



Appendix D.11B) Mantel correlogram of the great circle distance in km (in 500km bins) and total length of ChromoPainter sharing. Each point represents the Mantel correlation between the ChromoPainter similarity matrix and a matrix denoting the absence/presence (0/1) of an individual pair in the respective distance class. Unfilled points correspond to a non-significant ($p > 0.05$) correlation after 1000 permutations and an additional correction for multiple testing (due to the high number of bins). The gap in the plot indicates distance classes without any pairs of individuals.



Appendix D.12 Inter-population pairs that represent outliers (empirical top 1%) of a normalised residual metric of excess f_2 sharing relative to an exponential fit

Results are ordered by different categories of great circle distances between individual pairs. Abbreviations employed are as follows: dist_cat: great circle distance between relevant pairs of individuals, given in thousand km, N: number of individuals pairs from respective populations that are outliers, pop: population. The short codes for the macro-groups are the same as in Table 3.1.

Dist_cat	Pop long 1	Pop long 2	Macro-group 1	Macro-group 2	N
1000-4999	Kyrgyz	Mongolians	CeA	SSi	18
1000-4999	Buryats	Yakuts	SSi	CSi	18
1000-4999	Kyrgyz	Buryats	CeA	SSi	12
1000-4999	Vietnamese	Dusun	SeM	SeI	11
1000-4999	Bajo	Koinanbe	SeI	Oce	11
1000-4999	Mongolians	Han	SSi	SeM	10
1000-4999	Kazakhs	Buryats	CeA	SSi	9
1000-4999	Kazakhs	Evens	CeA	CSi	9
1000-4999	Vietnamese	Murut	SeM	SeI	9
1000-4999	Kyrgyz	Altaians	CeA	SSi	8
1000-4999	Tamang	Burmese	SoA	SeM	7
1000-4999	Mongolians	Yakuts	SSi	CSi	7
1000-4999	Bajo	Kosipe	SeI	Oce	7
1000-4999	Kazakhs	Mongolians	CeA	SSi	6
1000-4999	Mongolians	Japanese	SSi	SeM	5
1000-4999	NW-Europeans	Tatars	WEu	Vol	5
1000-4999	Tamang	Han	SoA	SeM	4
1000-4999	Evens	Koryaks	CSi	NSi	4
1000-4999	Buryats	Evens	SSi	CSi	4
1000-4999	Vietnamese	Lebbo	SeM	SeI	4
1000-4999	Selkups	Evens	WSi	CSi	3
1000-4999	Tundra-Nenets	Yakuts	WSi	CSi	3
1000-4999	Altaians	Yakuts	SSi	CSi	3
1000-4999	Kazakhs	Altaians	CeA	SSi	3
1000-4999	Ho	Burmese	SoA	SeM	3
1000-4999	Tuvinians	Yakuts	SSi	CSi	3
1000-4999	Mongolians	Chinese_PGP	SSi	SeM	3
1000-4999	Kazakhs	Tuvinians	CeA	SSi	3
1000-4999	Kyrgyz	Tuvinians	CeA	SSi	3
1000-4999	Burmese	Lebbo	SeM	SeI	3
1000-4999	Shor	Yakuts	SSi	CSi	3
1000-4999	Yakuts	Koryaks	CSi	NSi	3

Dist_cat	Pop long 1	Pop long 2	Macro-group 1	Macro-group 2	N
1000-4999	NW-Europeans	Russians	WEu	EEu	3
1000-4999	Tamang	Chinese_PGP	SoA	SeM	2
1000-4999	Santhal	Burmese	SoA	SeM	2
1000-4999	Selkups	Yakuts	WSi	CSi	2
1000-4999	Dhaka-mixed-population	Burmese	SoA	SeM	2
1000-4999	Kol	Burmese	SoA	SeM	2
1000-4999	Kapu	Burmese	SoA	SeM	2
1000-4999	Udmurts	Khantys	Vol	WSi	2
1000-4999	Croats	Belarusians	WEu	EEu	2
1000-4999	Tuvinians	Nganasans	SSi	CSi	2
1000-4999	Buryats	Han	SSi	SeM	2
1000-4999	Uyghurs	Buryats	CeA	SSi	2
1000-4999	Kyrgyz	Han	CeA	SeM	2
1000-4999	Kets	Shor	WSi	SSi	2
1000-4999	Selkups	Shor	WSi	SSi	2
1000-4999	Kyrgyz	Yakuts	CeA	CSi	2
1000-4999	Germans	Cossacks	WEu	EEu	2
1000-4999	Chinese_PGP	Murut	SeM	SeI	2
1000-4999	Belarusians	Bashkirs	EEu	Vol	2
1000-4999	Saudi-Arabians	Belarusians	MiE	EEu	2
1000-4999	Iranians	Bashkirs	MiE	Vol	2
1000-4999	Lebanese	Bashkirs	MiE	Vol	2
1000-4999	Saudi-Arabians	Udmurts	MiE	Vol	2
1000-4999	Iranians	Udmurts	MiE	Vol	2
1000-4999	Saudi-Arabians	Bashkirs	MiE	Vol	2
1000-4999	Iranians	Belarusians	MiE	EEu	1
1000-4999	Iranians	Komis	MiE	EEu	1
1000-4999	Karelians	Maris	EEu	Vol	1
1000-4999	Komis	Bashkirs	EEu	Vol	1
1000-4999	Evenks	Chukchi	CSi	NSi	1
1000-4999	Mongolians	Vietnamese	SSi	SeM	1
1000-4999	Marwadi-Middle-caste	Burmese	SoA	SeM	1
1000-4999	Kabardians	NW-Europeans	MiE	WEu	1
1000-4999	Assyrians	Italians	MiE	WEu	1
1000-4999	Armenians	Italians	MiE	WEu	1
1000-4999	Udmurts	Selkups	Vol	WSi	1
1000-4999	Udmurts	Yakuts	Vol	CSi	1
1000-4999	Balija-Middle-caste	Shugnan	SoA	CeA	1
1000-4999	Balija-Middle-caste	Ishkashim	SoA	CeA	1
1000-4999	Tamang	Turkmens	SoA	CeA	1
1000-4999	Balija-Middle-caste	Burmese	SoA	SeM	1
1000-4999	Asur	Burmese	SoA	SeM	1

Dist_cat	Pop long 1	Pop long 2	Macro-group 1	Macro-group 2	N
1000-4999	Low-caste (Madhya Pradesh)	Burmese	SoA	SeM	1
1000-4999	Ukrainians	Baptised-Tatars	EEu	Vol	1
1000-4999	Cossacks	Mansis	EEu	WSi	1
1000-4999	Russians	Udmurts	EEu	Vol	1
1000-4999	Latvians	Maris	EEu	Vol	1
1000-4999	Saami	Chuvashes	EEu	Vol	1
1000-4999	Belarusians	Khantys	EEu	WSi	1
1000-4999	Belarusians	Mansis	EEu	WSi	1
1000-4999	Cossacks	Tajiks	EEu	CeA	1
1000-4999	Russians	Bashkirs	EEu	Vol	1
1000-4999	Thakur	Ishkashim	SoA	CeA	1
1000-4999	Saudi-Arabians	Khantys	MiE	WSi	1
1000-4999	Armenians	Altaians	MiE	SSi	1
1000-4999	Lezgins	Mongolians	MiE	SSi	1
1000-4999	Albanians	Belarusians	WEu	EEu	1
1000-4999	Selkups	Evenks	WSi	CSi	1
1000-4999	Jordanians	Altaians	MiE	SSi	1
1000-4999	Albanians	Poles	WEu	EEu	1
1000-4999	Bashkirs	Tundra-Nenets	Vol	WSi	1
1000-4999	Mansis	Evens	WSi	CSi	1
1000-4999	Kumyks	Ishkashim	MiE	CeA	1
1000-4999	Iranians	Khantys	MiE	WSi	1
1000-4999	Iranians	Ishkashim	MiE	CeA	1
1000-4999	Iranians	Gond	MiE	SoA	1
1000-4999	Iranians	Gujaratis	MiE	SoA	1
1000-4999	Gujaratis	Burmese	SoA	SeM	1
1000-4999	Vietnamese	Bajo	SeM	SeI	1
1000-4999	Iranians	Poles	MiE	EEu	1
1000-4999	Iranians	Dhaka-mixed-population	MiE	SoA	1
1000-4999	Jordanians	Asur	MiE	SoA	1
1000-4999	Lebanese	Gujaratis	MiE	SoA	1
1000-4999	Gond	Burmese	SoA	SeM	1
1000-4999	Kshatriya	Aeta	SoA	SeI	1
1000-4999	Kurmi	Dusun	SoA	SeI	1
1000-4999	Shor	Evens	SSi	CSi	1
1000-4999	Buryats	Burmese	SSi	SeM	1
1000-4999	Mongolians	Burmese	SSi	SeM	1
1000-4999	Uzbek	Han	CeA	SeM	1
1000-4999	Kets	Mongolians	WSi	SSi	1
1000-4999	Turkmens	Mongolians	CeA	SSi	1
1000-4999	Kyrgyz	Burmese	CeA	SeM	1
1000-4999	Yaghnobi	Buryats	CeA	SSi	1

Dist_cat	Pop long 1	Pop long 2	Macro-group 1	Macro-group 2	N
1000-4999	Kyrgyz	Vietnamese	CeA	SeM	1
1000-4999	Forest-Nenets	Evens	WSi	CSi	1
1000-4999	Forest-Nenets	Yakuts	WSi	CSi	1
1000-4999	Kets	Nganasans	WSi	CSi	1
1000-4999	Kets	Yakuts	WSi	CSi	1
1000-4999	Bajo	Wongatha	SeI	Oce	1
1000-4999	Vietnamese	Luzon	SeM	SeI	1
1000-4999	Burmese	Luzon	SeM	SeI	1
1000-4999	Burmese	Dusun	SeM	SeI	1
1000-4999	Tuvinians	Evens	SSi	CSi	1
1000-4999	Christian-Arabs-Israel	Albanians	MiE	WEu	1
1000-4999	Christian-Arabs-Israel	Roma	MiE	WEu	1
1000-4999	Druze	Albanians	MiE	WEu	1
1000-4999	Germans	Estonians	WEu	EEu	1
1000-4999	Hungarians	Finnish	WEu	EEu	1
1000-4999	NW-Europeans	Mordvins	WEu	EEu	1
1000-4999	Croats	Latvians	WEu	EEu	1
1000-4999	Croats	Lithuanians	WEu	EEu	1
1000-4999	Germans	Karelians	WEu	EEu	1
1000-4999	Hungarians	Ingrians	WEu	EEu	1
1000-4999	Hungarians	Russians	WEu	EEu	1
1000-4999	NW-Europeans	Belarusians	WEu	EEu	1
1000-4999	NW-Europeans	Estonians	WEu	EEu	1
1000-4999	NW-Europeans	Karelians	WEu	EEu	1
1000-4999	Roma	Komis	WEu	EEu	1
1000-4999	United-Kingdom	Swedes	WEu	EEu	1
1000-4999	NW-Europeans	Maris	WEu	Vol	1
1000-4999	Roma	Udmurts	WEu	Vol	1
1000-4999	Germans	Mansis	WEu	WSi	1
1000-4999	Mongolians	Evenks	SSi	CSi	1
1000-4999	Chinese_PGP	Visayan	SeM	SeI	1
1000-4999	Han	Murut	SeM	SeI	1
1000-4999	Germans	Russians	WEu	EEu	1
1000-4999	Hungarians	Swedes	WEu	EEu	1
1000-4999	Mongolians	Nganasans	SSi	CSi	1
1000-4999	Marwadi-Middle-caste	Rushan-Vanch	SoA	CeA	1
1000-4999	Bengali	Rushan-Vanch	SoA	CeA	1
1000-4999	Brahmin (Central India and Nepal)	Ishkashim	SoA	CeA	1
1000-4999	Gujaratis	Ishkashim	SoA	CeA	1
1000-4999	Kshatriya	Rushan-Vanch	SoA	CeA	1
1000-4999	Batak	Koinanbe	SeI	Oce	1
1000-4999	Batak	Wongatha	SeI	Oce	1

Dist_cat	Pop long 1	Pop long 2	Macro-group 1	Macro-group 2	N
1000-4999	Visayan	Koinanbe	SeI	Oce	1
1000-4999	Belarusians	Maris	EEu	Vol	1
1000-4999	Belarusians	Udmurts	EEu	Vol	1
1000-4999	Cossacks	Udmurts	EEu	Vol	1
1000-4999	Russians	Mansis	EEu	WSi	1
1000-4999	Bashkirs	Selkups	Vol	WSi	1
1000-4999	Bashkirs	Khantys	Vol	WSi	1
1000-4999	Latvians	Mansis	EEu	WSi	1
1000-4999	Vietnamese	Visayan	SeM	SeI	1
1000-4999	Lebanese	Belarusians	MiE	EEu	1
1000-4999	Lebanese	Komis	MiE	EEu	1
1000-4999	Lezgins	Belarusians	MiE	EEu	1
1000-4999	Lebanese	Udmurts	MiE	Vol	1
1000-4999	Lezgins	Komis	MiE	EEu	1
1000-4999	Lezgins	Udmurts	MiE	Vol	1
1000-4999	Kumyks	Baptised-Tatars	MiE	Vol	1
1000-4999	Lebanese	Forest-Nenets	MiE	WSi	1
1000-4999	Lebanese	Khantys	MiE	WSi	1
5000-9999	Kyrgyz	Japanese	CeA	SeM	28
5000-9999	Kyrgyz	Evens	CeA	CSi	17
5000-9999	Kazakhs	Japanese	CeA	SeM	12
5000-9999	Kazakhs	Evens	CeA	CSi	9
5000-9999	Buryats	Vietnamese	SSi	SeM	9
5000-9999	Sandawe	Saudi-Arabians	Afr	MiE	9
5000-9999	Roma	Dhaka-mixed-population	WEu	SoA	9
5000-9999	Roma	Gujaratis	WEu	SoA	9
5000-9999	Roma	Brahmin (Central India and Nepal)	WEu	SoA	7
5000-9999	Yoruba	Iranians	Afr	MiE	6
5000-9999	Kazakhs	Vietnamese	CeA	SeM	6
5000-9999	Bashkirs	Evens	Vol	CSi	6
5000-9999	Roma	Middle-caste (Odisha)	WEu	SoA	6
5000-9999	Roma	Bengali	WEu	SoA	5
5000-9999	Russians	Chukchi	EEu	NSi	5
5000-9999	Uzbek	Japanese	CeA	SeM	5
5000-9999	Turkmens	Japanese	CeA	SeM	5
5000-9999	NW-Europeans	Tatars	WEu	Vol	5
5000-9999	NW-Europeans	Rushan-Vanch	WEu	CeA	5
5000-9999	Uyghurs	Japanese	CeA	SeM	4
5000-9999	Tamang	Japanese	SoA	SeM	4
5000-9999	Uzbek	Evens	CeA	CSi	4
5000-9999	African-Americans	Saudi-Arabians	Afr	MiE	3
5000-9999	Maasai	Iranians	Afr	MiE	3

Dist_cat	Pop long 1	Pop long 2	Macro-group 1	Macro-group 2	N
5000-9999	Yakuts	Burmese	CSi	SeM	3
5000-9999	Sandawe	Iranians	Afr	MiE	3
5000-9999	African-Americans	Iranians	Afr	MiE	3
5000-9999	Lebanese	Yakuts	MiE	CSi	3
5000-9999	Estonians	Chukchi	EEu	NSi	3
5000-9999	Saudi-Arabians	Yakuts	MiE	CSi	3
5000-9999	Roma	Kapu	WEu	SoA	3
5000-9999	Karelians	Chukchi	EEu	NSi	3
5000-9999	Mordvins	Chukchi	EEu	NSi	3
5000-9999	Roma	Low-caste (Madhya Pradesh)	WEu	SoA	3
5000-9999	Roma	Malayali	WEu	SoA	3
5000-9999	Roma	Santhal	WEu	SoA	3
5000-9999	Turkmens	Chinese_PGP	CeA	SeM	3
5000-9999	Roma	Gond	WEu	SoA	3
5000-9999	Roma	Gupta	WEu	SoA	3
5000-9999	Roma	Ho	WEu	SoA	3
5000-9999	Roma	Marwadi-Middle-caste	WEu	SoA	3
5000-9999	Yoruba	Saudi-Arabians	Afr	MiE	3
5000-9999	Azerbaijanis	Burmese	MiE	SeM	2
5000-9999	United-Kingdom	Chinese_PGP	WEu	SeM	2
5000-9999	Ukrainians	Chukchi	EEu	NSi	2
5000-9999	Balkars	Evens	MiE	CSi	2
5000-9999	Udmurts	Chukchi	Vol	NSi	2
5000-9999	Bashkirs	Vietnamese	Vol	SeM	2
5000-9999	Vepsas	Chukchi	EEu	NSi	2
5000-9999	Bashkirs	Chinese_PGP	Vol	SeM	2
5000-9999	Turkmens	Evens	CeA	CSi	2
5000-9999	Chuvashes	Chukchi	Vol	NSi	2
5000-9999	Turkmens	Yakuts	CeA	CSi	2
5000-9999	Roma	Balija-Middle-caste	WEu	SoA	2
5000-9999	Kazakhs	Chinese_PGP	CeA	SeM	2
5000-9999	Roma	Kol	WEu	SoA	2
5000-9999	Roma	Asur	WEu	SoA	2
5000-9999	Roma	Kshatriya	WEu	SoA	2
5000-9999	NW-Europeans	Ishkashim	WEu	CeA	2
5000-9999	Roma	Kurmi	WEu	SoA	2
5000-9999	Roma	Thakur	WEu	SoA	2
5000-9999	Italians	Shugnan	WEu	CeA	2
5000-9999	United-Kingdom	Uzbek	WEu	CeA	2
5000-9999	Iranians	Portuguese	MiE	WEu	2
5000-9999	Luhya	Iranians	Afr	MiE	1
5000-9999	Bedzan-Pygmies	Iranians	Afr	MiE	1

Dist_cat	Pop long 1	Pop long 2	Macro-group 1	Macro-group 2	N
5000-9999	Italians	Vietnamese	WEu	SeM	1
5000-9999	Ukrainians	Middle-caste (Odisha)	EEu	SoA	1
5000-9999	Belarusians	Evens	EEu	CSi	1
5000-9999	Mishar-Tatars	Evens	EEu	CSi	1
5000-9999	Altaians	Visayan	SSi	SeI	1
5000-9999	Mongolians	Visayan	SSi	SeI	1
5000-9999	Chukchi	Vietnamese	NSi	SeM	1
5000-9999	Chukchi	Mexicans	NSi	Ame	1
5000-9999	Burmese	Wongatha	SeM	Oce	1
5000-9999	Buryats	Lebbo	SSi	SeI	1
5000-9999	Mongolians	Murut	SSi	SeI	1
5000-9999	Cossacks	Buryats	EEu	SSi	1
5000-9999	Eskimo	Mexicans	NSi	Ame	1
5000-9999	Croats	Buryats	WEu	SSi	1
5000-9999	Circassians	Chinese_PGP	MiE	SeM	1
5000-9999	Moldavians	Brahmin (Central India and Nepal)	WEu	SoA	1
5000-9999	United-Kingdom	Mongolians	WEu	SSi	1
5000-9999	Croats	Chukchi	WEu	NSi	1
5000-9999	Italians	Chinese_PGP	WEu	SeM	1
5000-9999	United-Kingdom	Tatars	WEu	Vol	1
5000-9999	Druze	Middle-caste (Odisha)	MiE	SoA	1
5000-9999	Lebanese	Bengali	MiE	SoA	1
5000-9999	Muslim-Arabs-Israel	Luzon	MiE	SeI	1
5000-9999	Croats	Gond	WEu	SoA	1
5000-9999	Iranians	Yakuts	MiE	CSi	1
5000-9999	Saudi-Arabians	Evens	MiE	CSi	1
5000-9999	Udmurts	Koryaks	Vol	NSi	1
5000-9999	Saami	Mexicans	EEu	Ame	1
5000-9999	Bashkirs	Japanese	Vol	SeM	1
5000-9999	Udmurts	Evens	Vol	CSi	1
5000-9999	Maris	Japanese	Vol	SeM	1
5000-9999	Santhal	Koinanbe	SoA	Oce	1
5000-9999	Estonians	Murut	EEu	SeI	1
5000-9999	Komis	Mexicans	EEu	Ame	1
5000-9999	Bashkirs	Burmese	Vol	SeM	1
5000-9999	Tatars	Dusun	Vol	SeI	1
5000-9999	Tamang	Evens	SoA	CSi	1
5000-9999	Gujaratis	Japanese	SoA	SeM	1
5000-9999	Gond	Bajo	SoA	SeI	1
5000-9999	Low-caste (Madhya Pradesh)	Bajo	SoA	SeI	1
5000-9999	Ho	Bajo	SoA	SeI	1

Dist_cat	Pop long 1	Pop long 2	Macro-group 1	Macro-group 2	N
5000-9999	Malayali	Visayan	SoA	SeI	1
5000-9999	Tatars	Evens	Vol	CSi	1
5000-9999	Dhaka-mixed-population	Koinanbe	SoA	Oce	1
5000-9999	Maris	Evens	Vol	CSi	1
5000-9999	Middle-caste (Odisha)	Bajo	SoA	SeI	1
5000-9999	Maris	Chukchi	Vol	NSi	1
5000-9999	Tamang	Kosipe	SoA	Oce	1
5000-9999	Ishkashim	Evens	CeA	CSi	1
5000-9999	Rushan-Vanch	Evens	CeA	CSi	1
5000-9999	Portuguese	Maris	WEu	Vol	1
5000-9999	French	Brahmin (Central India and Nepal)	WEu	SoA	1
5000-9999	NW-Europeans	Dhaka-mixed-population	WEu	SoA	1
5000-9999	Moldavians	Kurmi	WEu	SoA	1
5000-9999	NW-Europeans	Malayali	WEu	SoA	1
5000-9999	Circassians	Dhaka-mixed-population	MiE	SoA	1
5000-9999	Muslim-Arabs-Israel	Bengali	MiE	SoA	1
5000-9999	Circassians	Mongolians	MiE	SSi	1
5000-9999	Druze	Bengali	MiE	SoA	1
5000-9999	Georgians	Buryats	MiE	SSi	1
5000-9999	Druze	Kapu	MiE	SoA	1
5000-9999	Portuguese	Russians	WEu	EEu	1
5000-9999	French	Baptised-Tatars	WEu	Vol	1
5000-9999	Tatars	Japanese	Vol	SeM	1
5000-9999	Druze	Vietnamese	MiE	SeM	1
5000-9999	Portuguese	Chuvashes	WEu	Vol	1
5000-9999	Albanians	Buryats	WEu	SSi	1
5000-9999	Ingrians	Chukchi	EEu	NSi	1
5000-9999	Uyghurs	Evens	CeA	CSi	1
5000-9999	Poles	Gujaratis	EEu	SoA	1
5000-9999	Armenians	Chukchi	MiE	NSi	1
5000-9999	Saudi-Arabians	Koryaks	MiE	NSi	1
5000-9999	Komis	Middle-caste (Odisha)	EEu	SoA	1
5000-9999	Russians	Middle-caste (Odisha)	EEu	SoA	1
5000-9999	Latvians	Chukchi	EEu	NSi	1
5000-9999	Poles	Chukchi	EEu	NSi	1
5000-9999	Roma	Burmese	WEu	SeM	1
5000-9999	Germans	Kyrgyz	WEu	CeA	1
5000-9999	Germans	Shugnan	WEu	CeA	1
5000-9999	Germans	Uyghurs	WEu	CeA	1
5000-9999	NW-Europeans	Kazakhs	WEu	CeA	1
5000-9999	NW-Europeans	Shugnan	WEu	CeA	1

Dist_cat	Pop long 1	Pop long 2	Macro-group 1	Macro-group 2	N
5000-9999	NW-Europeans	Uyghurs	WEu	CeA	1
5000-9999	United-Kingdom	Rushan-Vanch	WEu	CeA	1
5000-9999	Kyrgyz	Bajo	CeA	SeI	1
5000-9999	Kyrgyz	Dusun	CeA	SeI	1
5000-9999	Kyrgyz	Visayan	CeA	SeI	1
5000-9999	French	Ishkashim	WEu	CeA	1
5000-9999	Italians	Rushan-Vanch	WEu	CeA	1
5000-9999	United-Kingdom	Yaghnobi	WEu	CeA	1
5000-9999	Portuguese	Mansis	WEu	WSi	1
5000-9999	Kyrgyz	Igorot	CeA	SeI	1
5000-9999	Tundra-Nenets	Burmese	WSi	SeM	1
5000-9999	French	Yaghnobi	WEu	CeA	1
5000-9999	United-Kingdom	Shugnan	WEu	CeA	1
5000-9999	Komis	Chukchi	EEu	NSi	1
5000-9999	Maasai	Portuguese	Afr	WEu	1
5000-9999	Yoruba	French	Afr	WEu	1
5000-9999	Yoruba	Moldavians	Afr	WEu	1
5000-9999	Iranians	United-Kingdom	MiE	WEu	1
5000-9999	Finnish	Chukchi	EEu	NSi	1
5000-9999	Italians	Tatars	WEu	Vol	1
5000-9999	Saudi-Arabians	Portuguese	MiE	WEu	1
10000-19999	NW-Europeans	Mexicans	WEu	Ame	15
10000-19999	NW-Europeans	Puerto-Ricans	WEu	Ame	11
10000-19999	Italians	Mexicans	WEu	Ame	11
10000-19999	Italians	Puerto-Ricans	WEu	Ame	7
10000-19999	United-Kingdom	Mexicans	WEu	Ame	5
10000-19999	French	Mexicans	WEu	Ame	4
10000-19999	Germans	Mexicans	WEu	Ame	4
10000-19999	Germans	Puerto-Ricans	WEu	Ame	4
10000-19999	Muslim-Arabs-Israel	Puerto-Ricans	MiE	Ame	3
10000-19999	Portuguese	Mexicans	WEu	Ame	3
10000-19999	Hungarians	Puerto-Ricans	WEu	Ame	3
10000-19999	Swedes	Mexicans	EEu	Ame	3
10000-19999	Albanians	Puerto-Ricans	WEu	Ame	3
10000-19999	Portuguese	Puerto-Ricans	WEu	Ame	3
10000-19999	Assyrians	Mexicans	MiE	Ame	2
10000-19999	Christian-Arabs-Israel	Mexicans	MiE	Ame	2
10000-19999	African-Americans	Mexicans	Afr	Ame	2
10000-19999	United-Kingdom	Puerto-Ricans	WEu	Ame	2
10000-19999	Belarusians	Mexicans	EEu	Ame	2
10000-19999	Latvians	Mexicans	EEu	Ame	2
10000-19999	Saudi-Arabians	Mexicans	MiE	Ame	2

Dist_cat	Pop long 1	Pop long 2	Macro-group 1	Macro-group 2	N
10000-19999	Croats	Mexicans	WEu	Ame	2
10000-19999	Hungarians	Mexicans	WEu	Ame	2
10000-19999	Swedes	Puerto-Ricans	EEu	Ame	2
10000-19999	Ukrainians	Mexicans	EEu	Ame	2
10000-19999	Muslim-Arabs-Israel	Mexicans	MiE	Ame	2
10000-19999	Druze	Mexicans	MiE	Ame	1
10000-19999	Jordanians	Puerto-Ricans	MiE	Ame	1
10000-19999	Eskimo	Colla	NSi	Ame	1
10000-19999	Assyrians	Puerto-Ricans	MiE	Ame	1
10000-19999	Yoruba	Mexicans	Afr	Ame	1
10000-19999	Armenians	Mexicans	MiE	Ame	1
10000-19999	Moldavians	Mexicans	WEu	Ame	1
10000-19999	Moldavians	Puerto-Ricans	WEu	Ame	1
10000-19999	Roma	Puerto-Ricans	WEu	Ame	1
10000-19999	Belarusians	Puerto-Ricans	EEu	Ame	1
10000-19999	Estonians	Calchaquíes	EEu	Ame	1
10000-19999	Estonians	Puerto-Ricans	EEu	Ame	1
10000-19999	Latvians	Puerto-Ricans	EEu	Ame	1
10000-19999	Mordvins	Mexicans	EEu	Ame	1
10000-19999	NW-Europeans	Luzon	WEu	SeI	1
10000-19999	Albanians	Mexicans	WEu	Ame	1
10000-19999	Saudi-Arabians	Puerto-Ricans	MiE	Ame	1
10000-19999	French	Puerto-Ricans	WEu	Ame	1
10000-19999	Kets	Colla	WSi	Ame	1
10000-19999	Kumyks	Puerto-Ricans	MiE	Ame	1
10000-19999	Altaians	Calchaquíes	SSi	Ame	1
10000-19999	Mongolians	Mexicans	SSi	Ame	1
10000-19999	Armenians	Puerto-Ricans	MiE	Ame	1
>20000	Yoruba	Puerto-Ricans	Afr	Ame	9
>20000	African-Americans	Puerto-Ricans	Afr	Ame	8
>20000	Luhya	Puerto-Ricans	Afr	Ame	4
>20000	Baka/Bakola Pygmies	Puerto-Ricans	Afr	Ame	3
>20000	Maasai	Puerto-Ricans	Afr	Ame	3
>20000	Sandawe	Puerto-Ricans	Afr	Ame	2
>20000	Lezgins	Calchaquíes	MiE	Ame	2
>20000	Yoruba	Calchaquíes	Afr	Ame	1
>20000	Druze	Calchaquíes	MiE	Ame	1
>20000	NW-Europeans	Calchaquíes	WEu	Ame	1

Appendix D.13 Evaluation of the fit of an exponential model relating pairwise great circle distance and normalised f_2 sharing

This file can be found attached to the electronic version of this thesis. It contains a table of pairs of individuals (ordered as in the matrices giving similarity/dissimilarity statistics presented, e.g. in **Appendices D.7** and **D.9**). For these pairs the great circle distances (via plausible waypoints) and the normalised f_2 sharing are given. It also contains the predicted normalised f_2 sharing based on geographical distance according to an exponential fit. Plausible upper and lower boundaries for these predictions based on a Monte Carlo simulation approach are also given. The final metric, the normalised residual, characterises the deviation of observed and predicted normalised f_2 sharing. Its directionality is intuitive, a positive value indicates a greater observed similarity of rare variants than expected based on geography and a negative value implies the opposite. Abbreviations employed are as follows: ind-individual, min-minimum, max-maximum, pop-population. The short codes for the macro-groups are the same as in Table 3.1.

Appendix D.14 List of all rare variant clusters (RVCs) detected in the Diversity Set

This file can be found attached to the electronic version of this thesis. RVCs were defined based on individual-level rare variant databases. Therefore, there can be an asymmetry in the length of a particular run shared between pairs of individuals depending on which individual-level database was used for run detection ("Individual 1" denotes this individual). Short codes for the macro-groups taken from Table 3.1.

Appendix D.15 Different metrics describing the properties of RVCs shared within populations with three or more individuals in the Diversity Set

The short codes for the macro-groups are the same as in Table 3.1.

Macro-group	Population	Number of RVCs per pair	Total length of RVCs per pair [cM]	Mean lengths of RVCs [cM]	Median length of RVCs [cM]
Afr	African-Americans	206.5	95.59	0.46	0.1
Afr	Baka/Bakola Pygmies	643	465.85	0.72	0.18
Afr	Congo-Pygmies	786	513.82	0.65	0.19
Afr	Hadza	341.9	924.32	2.7	0.82
Afr	Luhya	305.67	163.42	0.53	0.18
Afr	Maasai	261.67	93.14	0.36	0.13

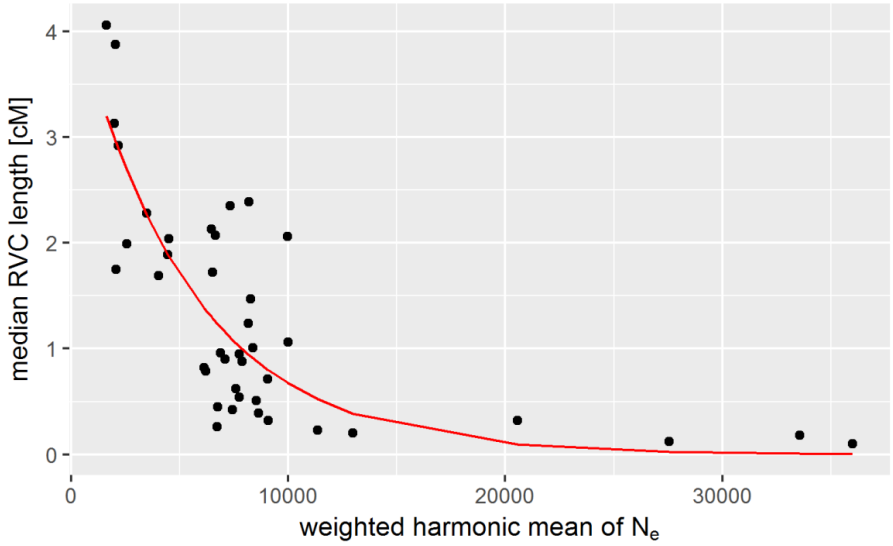
Macro-group	Population	Number of RVCs per pair	Total length of RVCs per pair [cM]	Mean lengths of RVCs [cM]	Median length of RVCs [cM]
Afr	Sandawe	374.9	480.69	1.28	0.32
Afr	Yoruba	348.97	122.16	0.35	0.12
MiE	Abkhazians	31.67	83.23	2.63	1.37
MiE	Armenians	5.93	4.42	0.75	0.32
MiE	Assyrians	25.67	87.92	3.43	2.24
MiE	Avars	28	41.31	1.48	1.02
MiE	Azerbaijanis	3.33	0.99	0.3	0.27
MiE	Balkars	18.33	60.91	3.32	1.52
MiE	Christian-Arabs-Israel	18.33	39.01	2.13	1.29
MiE	Circassians	10.67	16.84	1.58	1.04
MiE	Druze	30	72.14	2.4	1.56
MiE	Iranians	5	1.52	0.3	0.2
MiE	Kabardians	11.33	27.65	2.44	0.71
MiE	Kumyks	13	24.01	1.85	1.02
MiE	Lezgins	19.67	77	3.92	1.24
MiE	Tabasarans	27	40.58	1.5	0.86
WEu	Albanians	31	73.26	2.36	1.9
WEu	Croats	10.67	35.23	3.3	1.47
WEu	Germans	5.33	7.24	1.36	0.51
WEu	Hungarians	1.33	1.05	0.79	0.79
WEu	Italians	7.83	12.26	1.57	1.01
WEu	NW-Europeans	7.58	15.32	2.02	0.42
WEu	Roma	141.33	1225.35	8.67	2.92
EEu	Belarusians	6	6.54	1.09	0.96
EEu	Cossacks	6	8.35	1.39	0.48
EEu	Estonians	8.67	16.19	1.87	0.9
EEu	Finnish	21.33	48.01	2.25	1.3
EEu	Ingrians	20	29.12	1.46	1.21
EEu	Karelians	26.67	64.61	2.42	1.55
EEu	Latvians	16.67	27.54	1.65	0.74
EEu	Lithuanians	9	6.38	0.71	0.43
EEu	Mordvins	11.67	29.95	2.57	1.01
EEu	Poles	6.2	5.65	0.91	0.62
EEu	Russians	7.19	9.76	1.36	0.76
EEu	Saami	120.67	839.27	6.96	4.14
EEu	Ukrainians	5.43	6.17	1.14	0.58
EEu	Vepsas	28.67	104.55	3.65	2.35
Vol	Baptised Tatars	3	4.3	1.43	0.99

Macro-group	Population	Number of RVCs per pair	Total length of RVCs per pair [cM]	Mean lengths of RVCs [cM]	Median length of RVCs [cM]
Vol	Bashkirs	17.6	50.11	2.85	2.06
Vol	Chuvashes	22.33	86.57	3.88	3.38
Vol	Maris	39	139.78	3.58	2.13
Vol	Tatars	4.33	11.57	2.67	2.19
Vol	Udmurts	61.17	186.46	3.05	2.07
SoA	Brahmin (Central India and Nepal)	16.67	6.17	0.37	0.15
SoA	Dhaka-mixed-population	23.67	10.57	0.45	0.36
SoA	Gujaratis	26.83	14.48	0.54	0.23
CeA	Kazakhs	8	5.11	0.64	0.56
CeA	Kyrgyz	10.29	18.66	1.81	1.06
CeA	Turkmens	33.33	114.54	3.44	2.45
CeA	Uzbek	6.67	1.52	0.23	0.11
WSi	Forest-Nenets	88.67	1140.96	12.87	7.1
WSi	Kets	94.67	806.09	8.51	4.91
WSi	Khantys	54.67	171.63	3.14	1.86
WSi	Mansis	26	221.15	8.51	4.74
WSi	Selkups	41.33	158.95	3.85	2.43
WSi	Tundra-Nenets	39.33	150.07	3.82	1.8
SSi	Altaians	26.67	106.32	3.99	2.39
SSi	Buryats	27.27	74.34	2.73	1.72
SSi	Mongolians	11.67	12.23	1.05	0.54
SSi	Tuvinians	41.67	113.17	2.72	1.79
CSi	Evenks	28.33	97.24	3.43	2.47
CSi	Evens	46.27	227.16	4.91	2.76
CSi	Yakuts	39.27	138.76	3.53	2.28
NSi	Chukchi	69.5	389.17	5.6	2.28
NSi	Eskimo	119	874.68	7.35	3.88
NSi	Koryaks	85.6	469.9	5.49	2.92
SeM	Burmese	21.54	15.17	0.7	0.39
SeM	Chinese_PGP	9.67	6.78	0.7	0.14
SeM	Han	14.17	6.15	0.43	0.26
SeM	Japanese	74.17	79.06	1.07	0.79
SeM	Vietnamese	24.64	24.51	0.99	0.45
SeI	Aeta	348.33	836.43	2.4	1.29
SeI	Agta	364	1249.16	3.43	1.49
SeI	Bajo	49	150.46	3.07	0.88
SeI	Batak	295.67	910.62	3.08	1.49

Macro-group	Population	Number of RVCs per pair	Total length of RVCs per pair [cM]	Mean lengths of RVCs [cM]	Median length of RVCs [cM]
SeI	Dusun	78.04	258.94	3.32	1.89
SeI	Igorot	113.79	395	3.47	1.99
SeI	Lebbo	163.17	588.23	3.61	2.04
SeI	Murut	90.18	268.05	2.97	1.69
Ame	Calchaquíes	31.1	120.11	3.86	1.75
Ame	Colla	56	306.65	5.48	3.13
Ame	Mexicans	10.2	32.97	3.23	0.95
Ame	Puerto-Ricans	29	155.14	5.35	4.96
Ame	Wichi	204.33	1275.04	6.24	4.06
Oce	Koinanbe	498.67	1201.25	2.41	1.37
Oce	Kosipe	526.67	1644.54	3.12	1.66

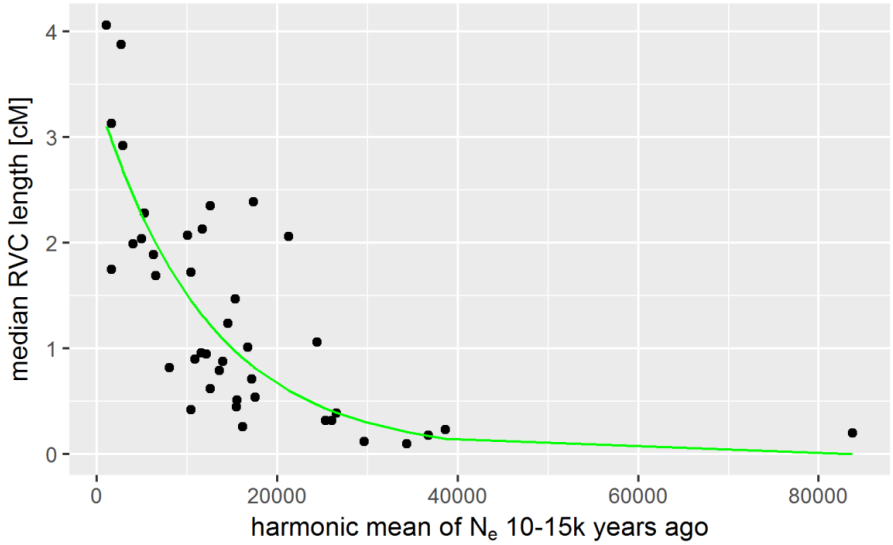
Appendix D.16 Different model fits for median RVC length as a function of N_e

Appendix D.16A) Median RVC length within populations as a function of the weighted harmonic mean of N_e over the last 30 kya. The red line indicates the fit of an exponential model using nonlinear least squares regression.



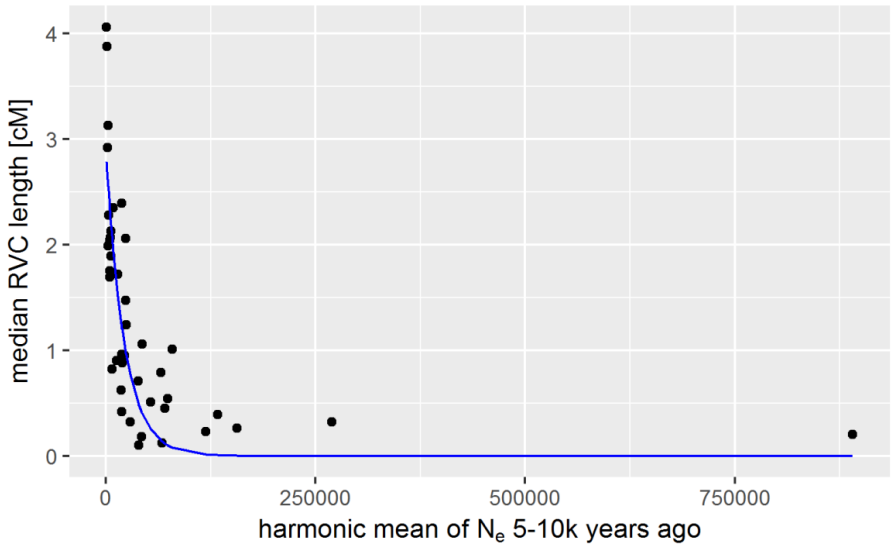
Appendix D.16B) Median RVC length within populations as a function of the weighted harmonic mean of N_e between 10-15 kya. The green line indicates the fit of an exponential model using nonlinear least squares regression. The equation for this model is also given.

$$\tilde{L} = 3.397 * e^{-8.113*10^{-5}N_e(10-15kya)}$$



Appendix D.16C) Median RVC length within populations as a function of the weighted harmonic mean of N_e between 5-10 kya. The blue line indicates the fit of an exponential model using nonlinear least squares regression. The equation for this model is also given.

$$\tilde{L} = 2.892 * e^{-4.568*10^{-5}N_e(5-10kya)}$$



Appendix D.17 Rank correlation coefficients for median RVC length and N_e within different intervals

Appendix D.17A) Spearman's rank correlation coefficients for N_e inferred by MSMC within 5,000-year right closed intervals during the last 30,000 years vs median length of RVCs by population

N_e	Spearman's ρ
0-5k	-0.670
5-10k	-0.812
10-15k	-0.750
15-20k	-0.561
20-25k	-0.478
25-30k	-0.372

Appendix D.17B) P-values for comparisons of Fisher Z transformed rank correlation coefficients between N_e and median RVC length for different time slices. Significant values bolded.

	5-10k	10-15k	15-20k	20-25k	25-30k
0-5k	0.156	0.474	0.444	0.207	0.068
5-10k		0.481	**0.029**	**0.007**	**0.001**
10-15k			0.139	**0.048**	**0.011**
15-20k				0.619	0.288
20-25k					0.572

Appendix D.18 Power law model for the relationship between N_e and median RVC length

Appendix D.18A)

When a power law is fitted to long-term population-level N_e (0-30 kya, weighing as described in section 3.1.3) vs median RVC length $\tilde{L}$ the resulting model equation is:

$$\tilde{L} = \frac{1674.108}{N_e^{0.826}}$$

The Vuong test suggests that this model provides an approximately equally good fit to the observed data compared to the exponential model (Eq.4.6) (Z = 0.685, p = 0.247). The power law using N_e from 5-10 kya in turn represents a significant improvement over N_e from both the 0-5 kya (Z = 3.067, p = 0.001) and 10-15 kya intervals (Z = 2.767, p = 0.003). Furthermore, it even seems to be a better fit than an exponential model for the 5-10 kya interval (Z = 2.157, p = 0.015).

This outcome can be linked to theoretical work on the relationship between IBD segment lengths and demographic history. More complex expressions which have been derived to describe the expected length distribution of IBD segments as a function of N_e also take the form of power laws (Palamara et al., 2012). Therefore, it is perhaps not surprising that the medians of the population-specific empirical distributions of a metric (RVC) that is related to IBD can be expressed as a function of N_e using a power law equation.

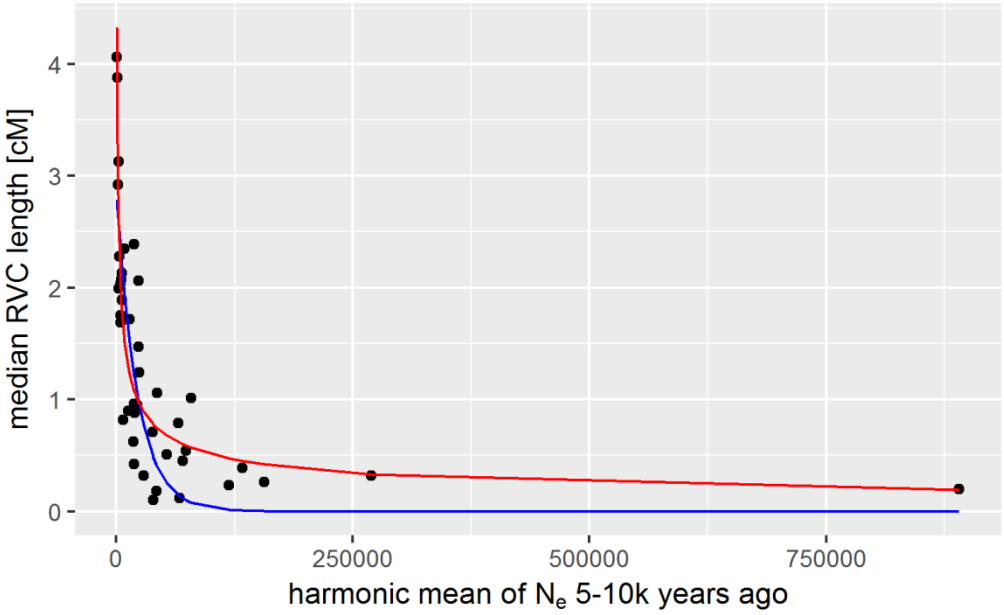
In biological terms, the decline in median segment length for very high recent N_e according to the power law fit to this particular dataset (Appendix D.18B) is less steep than would be expected. However, the lack of populations with $N_e > 2.5*10^5$ does not allow for good inference of fit for this range of N_e.

References

Palamara PF, Lencz T, Darvasi A, Pe'er I. 2012. Length Distributions of Identity by Descent Reveal Fine-Scale Demographic History. The American Journal of Human Genetics 91:809–822.

Appendix D.18B) Median RVC length within populations as a function of the weighted harmonic mean of N_e between 5-10 kya. The blue line indicates the fit of an exponential model (see Appendix D.14C) using nonlinear least squares regression whereas the red line represents a power law model. The equation for the latter is given below.

$$\tilde{L} = \frac{86.385}{N_e^{0.445}}$$



Appendix D.19 Summary of the average number of RVCs shared between macro-groups across all individuals from the Diversity Set

The short codes for the macro-groups are the same as in Table 3.1.

	Afr	MiE	WEu	EEu	Vol	SoA	CeA	WSi	SSi	CSi	NSi	SeM	SeI	Ame	Oce
Afr		1.39	0.38	0.09	0.09	0.14	0.16	0.02	0.03	0.01	0.01	0.04	0.13	2.30	0.25
MiE	1.39		0.73	0.43	0.45	0.63	0.76	0.19	0.33	0.13	0.04	0.10	0.02	0.28	0.08
WEu	0.38	0.73		1.27	0.76	0.47	0.43	0.30	0.21	0.04	0.07	0.05	0.02	0.42	0.16
EEu	0.09	0.43	1.27		1.22	0.18	0.40	0.42	0.22	0.12	0.16	0.04	0.01	0.22	0.11
Vol	0.09	0.45	0.76	1.22		0.35	0.79	1.21	0.74	0.46	0.17	0.20	0.02	0.16	0.09
SoA	0.14	0.63	0.47	0.18	0.35		1.32	0.14	0.38	0.08	0.02	1.23	0.38	0.09	0.33
CeA	0.16	0.76	0.43	0.40	0.79	1.32		0.58	2.31	1.19	0.16	0.88	0.12	0.12	0.07
WSi	0.02	0.19	0.30	0.42	1.21	0.14	0.58		1.14	1.86	0.29	0.07	0.02	0.07	0.02
SSi	0.03	0.33	0.21	0.22	0.74	0.38	2.31	1.14		2.81	0.24	1.14	0.16	0.07	0.05
CSi	0.01	0.13	0.04	0.12	0.46	0.08	1.19	1.86	2.81		1.59	0.40	0.04	0.03	0.01
NSi	0.01	0.04	0.07	0.16	0.17	0.02	0.16	0.29	0.24	1.59		0.08	0.00	0.13	0.02
SeM	0.04	0.10	0.05	0.04	0.20	1.23	0.88	0.07	1.14	0.40	0.08		1.54	0.02	0.22
SeI	0.13	0.02	0.02	0.01	0.02	0.38	0.12	0.02	0.16	0.04	0.00	1.54		0.01	2.85
Ame	2.30	0.28	0.42	0.22	0.16	0.09	0.12	0.07	0.07	0.03	0.13	0.02	0.01		0.04
Oce	0.25	0.08	0.16	0.11	0.09	0.33	0.07	0.02	0.05	0.01	0.02	0.22	2.85	0.04	

Appendix D.20 Summary of the average total length [cM] of RVCs shared between macro-groups across all individuals from the Diversity Set

The short codes for the macro-groups are the same as in Table 3.1.

	Afr	MiE	WEu	EEu	Vol	SoA	CeA	WSi	SSi	CSi	NSi	SeM	SeI	Ame	Oce
Afr		0.20	0.06	0.02	0.01	0.01	0.02	0.00	0.00	0.00	0.01	0.00	0.01	0.45	0.00
MiE	0.20		0.24	0.17	0.19	0.18	0.31	0.07	0.16	0.05	0.02	0.03	0.00	0.07	0.02
WEu	0.06	0.24		0.77	0.38	0.17	0.16	0.14	0.08	0.02	0.03	0.01	0.00	0.16	0.07
EEu	0.02	0.17	0.77		1.00	0.05	0.19	0.30	0.11	0.06	0.15	0.01	0.00	0.08	0.04
Vol	0.01	0.19	0.38	1.00		0.12	0.52	1.03	0.57	0.30	0.10	0.06	0.00	0.05	0.03
SoA	0.01	0.18	0.17	0.05	0.12		0.43	0.05	0.14	0.02	0.01	0.44	0.09	0.02	0.02
CeA	0.02	0.31	0.16	0.19	0.52	0.43		0.35	2.16	0.95	0.04	0.33	0.02	0.03	0.02
WSi	0.00	0.07	0.14	0.30	1.03	0.05	0.35		1.01	2.31	0.16	0.02	0.00	0.02	0.01
SSi	0.00	0.16	0.08	0.11	0.57	0.14	2.16	1.01		2.97	0.08	0.44	0.03	0.02	0.01
CSi	0.00	0.05	0.02	0.06	0.30	0.02	0.95	2.31	2.97		2.03	0.13	0.01	0.00	0.00
NSi	0.01	0.02	0.03	0.15	0.10	0.01	0.04	0.16	0.08	2.03		0.02	0.00	0.03	0.00
SeM	0.00	0.03	0.01	0.01	0.06	0.44	0.33	0.02	0.44	0.13	0.02		0.54	0.00	0.01
SeI	0.01	0.00	0.00	0.00	0.00	0.09	0.02	0.00	0.03	0.01	0.00	0.54		0.00	0.70
Ame	0.45	0.07	0.16	0.08	0.05	0.02	0.03	0.02	0.02	0.00	0.03	0.00	0.00		0.02
Oce	0.00	0.02	0.07	0.04	0.03	0.02	0.02	0.01	0.01	0.00	0.00	0.01	0.70	0.02	

Appendix D.21 Summary of the mean segment length [cM] of RVCs shared between macro-groups across all individuals from the Diversity Set

The short codes for the macro-groups are the same as in Table 3.1.

	Afr	MiE	WEu	EEu	Vol	SoA	CeA	WSi	SSi	CSi	NSi	SeM	SeI	Ame	Oce
Afr		0.15	0.15	0.18	0.11	0.06	0.14	0.10	0.11	0.07	0.79	0.05	0.04	0.20	0.02
MiE	0.15		0.34	0.39	0.42	0.29	0.40	0.34	0.50	0.43	0.49	0.27	0.11	0.26	0.21
WEu	0.15	0.34		0.61	0.50	0.35	0.36	0.46	0.36	0.49	0.49	0.24	0.17	0.38	0.45
EEu	0.18	0.39	0.61		0.82	0.27	0.47	0.71	0.50	0.54	0.93	0.30	0.10	0.39	0.40
Vol	0.11	0.42	0.50	0.82		0.35	0.65	0.86	0.77	0.64	0.59	0.32	0.15	0.31	0.33
SoA	0.06	0.29	0.35	0.27	0.35		0.33	0.34	0.36	0.32	0.26	0.36	0.23	0.17	0.05
CeA	0.14	0.40	0.36	0.47	0.65	0.33		0.61	0.94	0.80	0.26	0.38	0.19	0.28	0.29
WSi	0.10	0.34	0.46	0.71	0.86	0.34	0.61		0.89	1.24	0.56	0.21	0.22	0.25	0.33
SSi	0.11	0.50	0.36	0.50	0.77	0.36	0.94	0.89		1.06	0.34	0.39	0.18	0.32	0.14
CSi	0.07	0.43	0.49	0.54	0.64	0.32	0.80	1.24	1.06		1.28	0.31	0.15	0.11	0.23
NSi	0.79	0.49	0.49	0.93	0.59	0.26	0.26	0.56	0.34	1.28		0.23	0.01	0.22	0.04
SeM	0.05	0.27	0.24	0.30	0.32	0.36	0.38	0.21	0.39	0.31	0.23		0.35	0.13	0.05
SeI	0.04	0.11	0.17	0.10	0.15	0.23	0.19	0.22	0.18	0.15	0.01	0.35		0.04	0.25
Ame	0.20	0.26	0.38	0.39	0.31	0.17	0.28	0.25	0.32	0.11	0.22	0.13	0.04		0.49
Oce	0.02	0.21	0.45	0.40	0.33	0.05	0.29	0.33	0.14	0.23	0.04	0.05	0.25	0.49	

Appendix D.22 Summary of the median segment length [cM] of RVCs shared between macro-groups across all individuals from the Diversity Set

The short codes for the macro-groups are the same as in Table 3.1.

	Afr	MiE	WEu	EEu	Vol	SoA	CeA	WSi	SSi	CSi	NSi	SeM	SeI	Ame	Oce
Afr		0.04	0.05	0.05	0.04	0.01	0.04	0.01	0.02	0.04	0.06	0.01	0.01	0.06	0.00
MiE	0.04		0.17	0.21	0.22	0.15	0.20	0.20	0.23	0.27	0.19	0.13	0.03	0.14	0.05
WEu	0.05	0.17		0.35	0.30	0.19	0.19	0.23	0.21	0.23	0.23	0.17	0.05	0.22	0.26
EEu	0.05	0.21	0.35		0.46	0.16	0.28	0.41	0.29	0.37	0.54	0.14	0.05	0.21	0.24
Vol	0.04	0.22	0.30	0.46		0.23	0.39	0.53	0.42	0.40	0.37	0.18	0.07	0.16	0.32
SoA	0.01	0.15	0.19	0.16	0.23		0.18	0.22	0.19	0.25	0.16	0.20	0.08	0.09	0.03
CeA	0.04	0.20	0.19	0.28	0.39	0.18		0.43	0.55	0.44	0.15	0.21	0.09	0.12	0.05
WSi	0.01	0.20	0.23	0.41	0.53	0.22	0.43		0.51	0.67	0.29	0.17	0.05	0.20	0.13
SSi	0.02	0.23	0.21	0.29	0.42	0.19	0.55	0.51		0.60	0.17	0.20	0.10	0.13	0.06
CSi	0.04	0.27	0.23	0.37	0.40	0.25	0.44	0.67	0.60		0.65	0.19	0.09	0.05	0.23
NSi	0.06	0.19	0.23	0.54	0.37	0.16	0.15	0.29	0.17	0.65		0.09	0.01	0.10	0.04
SeM	0.01	0.13	0.17	0.14	0.18	0.20	0.21	0.17	0.20	0.19	0.09		0.19	0.07	0.03
SeI	0.01	0.03	0.05	0.05	0.07	0.08	0.09	0.05	0.10	0.09	0.01	0.19		0.02	0.06
Ame	0.06	0.14	0.22	0.21	0.16	0.09	0.12	0.20	0.13	0.05	0.10	0.07	0.02		0.12
Oce	0.00	0.05	0.26	0.24	0.32	0.03	0.05	0.13	0.06	0.23	0.04	0.03	0.06	0.12	

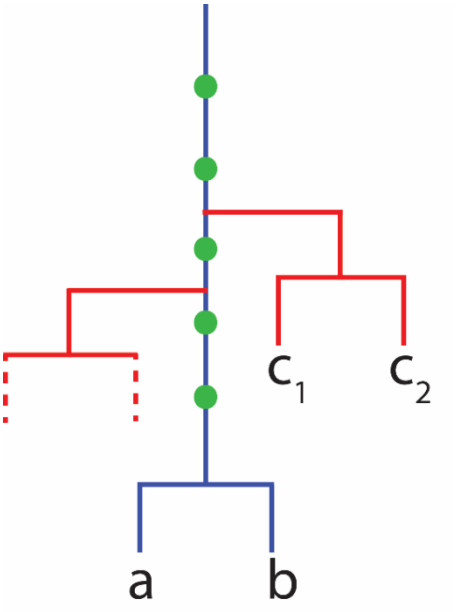
Appendix D.23 RVCs shared between Northeast Siberians and Africans from the Diversity Set

Note that the RVCs were defined based on individual-level rare variant databases. Therefore, there can be an asymmetry in the length of a particular run shared between pairs of individuals depending on which individual-level database was used for run detection ("Individual 1" denotes this individual). Abbreviations: Chr-Chromosome, Ind-Individual, Macro-Macro-group, No-Number. Population abbreviations are taken from Table 34.

Chr	Start (bp)	Length (bp)	Start (cM)	Length (cM)	No f_2	Ind1	Macro1	Ind2	Macro2
1	65017047	125034	95.07824	0.2299	17	YRI_3	Afr	Chuk8	NSi
1	65017047	125034	95.07824	0.2299	17	Chuk8	NSi	YRI_3	Afr
3	33306876	9325	57.4759	0.01294	19	BedzPy1	Afr	Chuk2	NSi
3	33306876	9325	57.4759	0.01294	19	Chuk2	NSi	BedzPy1	Afr
6	107535376	14338	113.1142	0.0064	6	YRI_11	Afr	Chuk2	NSi
6	107535376	14338	113.1142	0.0064	6	Chuk2	NSi	YRI_11	Afr
10	47100598	1553192	65.31547	2.81011	5	Chuk8	NSi	MKK_2	Afr
13	102708747	239653	98.19849	0.21125	21	BakaPy2	Afr	Chuk8	NSi
13	102708747	239653	98.19849	0.21125	21	Chuk8	NSi	BakaPy2	Afr
13	103438093	117489	99.08057	0.06491	7	ASW_1	Afr	Chuk8	NSi
13	103438093	117489	99.08057	0.06491	7	Chuk8	NSi	ASW_1	Afr
17	1963596	3088909	6.424974	6.355056	6	Chuk8	NSi	MKK_4	Afr
17	1963596	167945	6.424974	0.035928	5	MKK_4	Afr	Chuk8	NSi

Appendix D.24 Schematic illustrating how large reference datasets can be informative about the directionality of gene flow underlying RVC sharing

Example of an idealised tree underlying an RVC. Two individuals from the Diversity Set (a and b) share ≥5 f_2 variants (green dots) forming a unique branch (blue). If they do not belong to the same population this segment could either result from gene flow from either of their populations of origin into the other or from a third source which admixed with both ancestries in the past. Integration of information from the large gnomAD dataset can make additional branches of this phylogeny visible (red) where other individuals share at least one of these f_2 variants. If the vast majority of these individuals $c_1 \ldots c_n$ belong to a particular population this can support inferences about the directionality of gene flow.



Appendix D.25: gnomAD allele frequencies from selected populations for f_2 variants from the Diversity Set which constitute the two longest RVCs shared between Chukchi and Maasai

The genomic coordinates of the runs are 10:47,100,598-48,653,790 and 17:1,963,596-5,052,505 respectively. Abbreviations: Alt-Alternative Allele, Chr-Chromosome, Pos-Position, Ref-Reference Allele

Chr	Pos	dbSNP ID	Ref	Alt	African (includes African-Americans)	East Asian	European (Non-Finnish)	Finnish
10	47100598	-	A	G	-	-	-	-
10	47117234	-	T	C	-	-	-	-
10	47142516	-	A	G	-	-	-	-
10	47667655	rs530704389	G	A	0.0003	0	0.0009	0
10	48653790	rs536497458	T	C	0.0001	0	0.0005	0
17	1963596	rs149527790	A	G	0.0052	0	0	0
17	1964247	rs141530811	G	C	0.0050	0	0	0
17	2036391	rs146985355	T	C	0.0055	0	0	0
17	2085752	rs147655211	G	A	0.0056	0	0	0
17	2131541	rs146521642	C	T	0.0052	0	0	0.0001
17	5052505	rs543891949	C	T	0.0009	0	0.0038	0.0258

Appendix D.26 Symmetric matrix of the number of RVCs shared between individuals from the Diversity Set

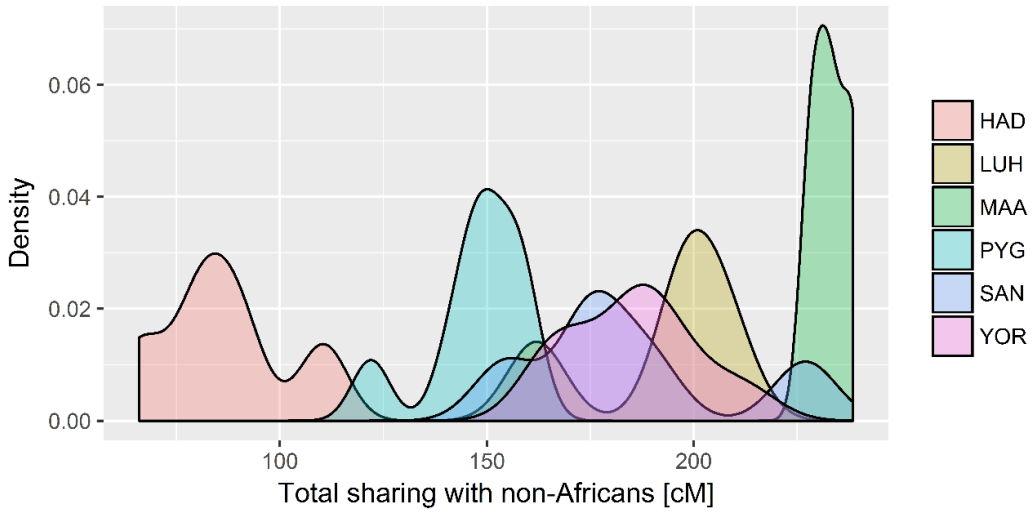
This file can be found attached to the electronic version of this thesis.

Appendix D.27 Symmetric matrix of total length [cM] of RVCs shared between individuals from the Diversity Set

This file can be found attached to the electronic version of this thesis.

Appendix D.28 RVC sharing of African populations with non-Africans

Appendix D.28A) Kernel density estimates of the distributions of RVC sharing with all non-Africans for different African populations from the Diversity Set. Abbreviations: HAD-Hadza, LUH-Luhya, MAA-Maasai, PYG-Pygmy, SAN-Sandawe, YOR-Yoruba.

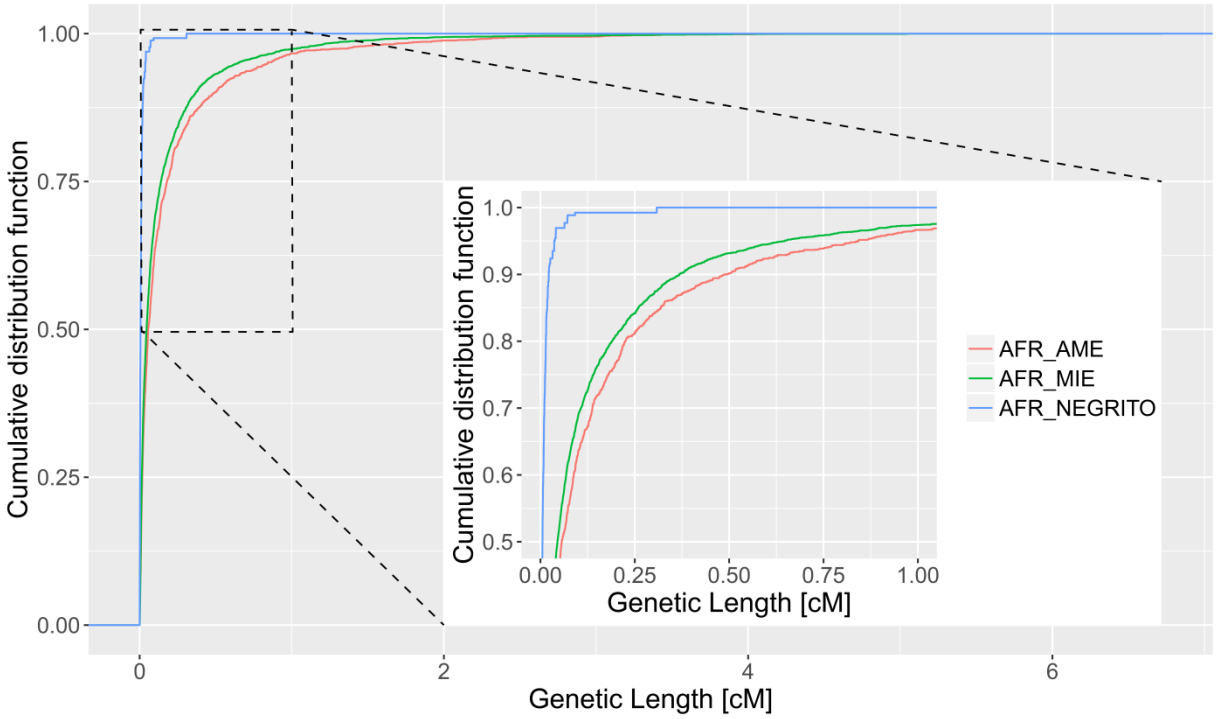


Appendix D.28B) P-values for t-tests comparing the total amount of RVC sharing with non-Africans for different African populations. Significant values are bolded.

	Hadza	Luhya	Maasai	Pygmies	Sandawe	Yoruba
Hadza		<0.001	<0.001	<0.001	<0.001	<0.001
Luhya			0.023	0.016	0.651	0.595
Maasai				<0.001	0.014	<0.001
Pygmies					0.034	<0.001
Sandawe						0.946
Yoruba						

Appendix D.29 Empirical cumulative distribution functions of the length of RVCs shared between Africans and various non-African groups

Note that African-Americans were excluded. Sharing is plotted between Africans (AFR) and a) admixed American populations (AME), b) Middle Easterners (Arabs and Iranians) (MIE) and c) Philippine Negritos. Inset: Empirical cumulative distribution functions for RVCs between 0-1 cM above the 0.5 quantile.



Appendix D.30 Sharing of RVCs between Africans and Calchaquíes

This file can be found attached to the electronic version of this thesis. The first sheet of the excel file **(Appendix D.30A)** lists all RVCs detected in the Diversity Set shared between the Calchaquíes and any African population. Here, the vrGV databases for Calchaquíes individuals were used to define the RVCs. The last three columns contain information on the consistency of the assumption that the RVC regions correspond to continuous haplotypes based on comparison to the phased VCF files. RVCs that are potentially derived from European introgression into both Africans and Native Americans are marked with an asterisk. The second sheet **(Appendix D.30B)** contains RVCs between the same interpopulation pairs, however the vrGV databases of the African individuals were utilised as the reference. The third sheet **(Appendix D.30C)** lists RVCs (with the Calchaquíes vrGV databases as reference) that result

when inconsistent homozygotes are used as breakpoints and RVCs are defined based on the resulting subruns.

Appendix D.31 Chromosome-wise f_3(C;A,B) statistics to investigate whether the Calchaquíes (C) can be modelled as a mixture of the Wichi (A) and the Yoruba (B)

All statistics calculated using the threepop program which is part of the TreeMix software. Significant results are bolded.

Chromosome	f_3(C;A,B)	Standard error	Z
1	-0.000644	0.000495	-1.30
2	0.000638	0.000699	0.91
3	0.000416	0.000604	0.69
4	-0.000473	0.000618	-0.77
5	0.001184	0.000690	1.71
6	-0.001268	0.000774	-1.64
7	0.002668	0.000667	4.00
8	-0.000180	0.000749	-0.24
9	-0.000459	0.000752	-0.61
10	0.000271	0.000727	0.37
11	0.000161	0.000631	0.26
12	0.000287	0.000783	0.37
13	-0.003460	0.000795	-4.35
14	0.000483	0.001322	0.37
15	-0.001048	0.000886	-1.18
16	0.001084	0.000969	1.12
17	0.002653	0.001089	2.44
18	0.000666	0.000721	0.92
19	0.000743	0.000797	0.93
20	0.003112	0.000739	4.21
21	-0.000059	0.001022	-0.06
22	-0.001132	0.001218	-0.93

Appendix D.32 gnomAD continental frequencies of 3,290 doubletons that form the 241 RVCs detected in Calchaquíes with African populations

This file can be found attached to the electronic version of this thesis. gnomAD allele frequencies were retrieved for all f_2 variants constituting RVCs shared between Calchaquíes and Africans. Abbreviations: AF-allele frequency, AFA-African Americans, AFR-Africans, Alt-Alternative Allele, ASJ-Ashkenazi Jewish, AMR-admixed Americans, BAK-Baka/Bakola Pygmies, BED-Bedzan Pygmies, CAC-Calchaquíes, Chr-Chromosome, CON-Congo Pygmies, EAS-East Asian, FIN-Finnish, HAD-Hadza, Ind-Individual, LUH-Luhya, MAA-Maasai,

MNFE-Non-Finnish Europeans, OTH-"Other", here individuals that did not unambiguously cluster with any of the other major groups in gnomAD, Pop-Population, POPMAX-Population with highest allele frequency (at a particular f_2 site), Ref-Reference Allele, SAN-Sandawe, YOR-Yoruba.

Appendix D.33 gnomAD continental frequencies for the three RVCs shared between Calchaquíes and Africans consistent with West Eurasian gene flow into both.

The genomic coordinates of these runs are as follows: 1:39,561,117-39,910,651 (Cachi1-MKK_1), 3:157,703,340-159,568,153 (Cachi5-ASW_3) and 13:56,396,163-57,098,291 (Cachi2-MKK_2). Abbreviations: AFR-Africans, Alt-Alternative Allele, Chr-Chromosome, NFE-Non-Finnish Europeans, Pos- Position, Ref-Reference Allele

Chr	Pos	dbSNP ID	Ref	Alt	African (includes African-Americans)	European (Non-Finnish)	AFR/NFE ratio
1	39561117	rs564890372	C	T	0	0.0003	absent AFR
1	39597888	rs143670217	G	A	0	0.0003	absent AFR
1	39774975	rs140746926	G	T	0	0.0003	absent AFR
1	39872363	rs548984889	C	T	0	0.0003	absent AFR
1	39910651	rs538087937	A	G	0	0.0003	absent AFR
3	157703340	rs141705264	A	T	0	0.0011	absent AFR
3	158536301	rs143747506	A	G	0.0008	0.0061	0.13
3	158738061	rs191498197	A	G	0.0001	0.0018	0.06
3	158885333	rs563413390	A	T	0.0001	0.0015	0.07
3	158950510	rs148677849	G	A	0	0.0004	absent AFR
3	158978685	rs142679158	C	T	0.0002	0.0023	0.1
3	159271261	rs573182905	T	C	0	0.0007	absent AFR
3	159327873	rs749008938	G	A	0	0.0009	absent AFR
3	159370951	rs562725561	T	C	0	0.0007	absent AFR
3	159503758	rs189025521	A	G	0.0003	0.0007	0.52
3	159568153	rs183074864	G	A	0.001	0.0038	0.27

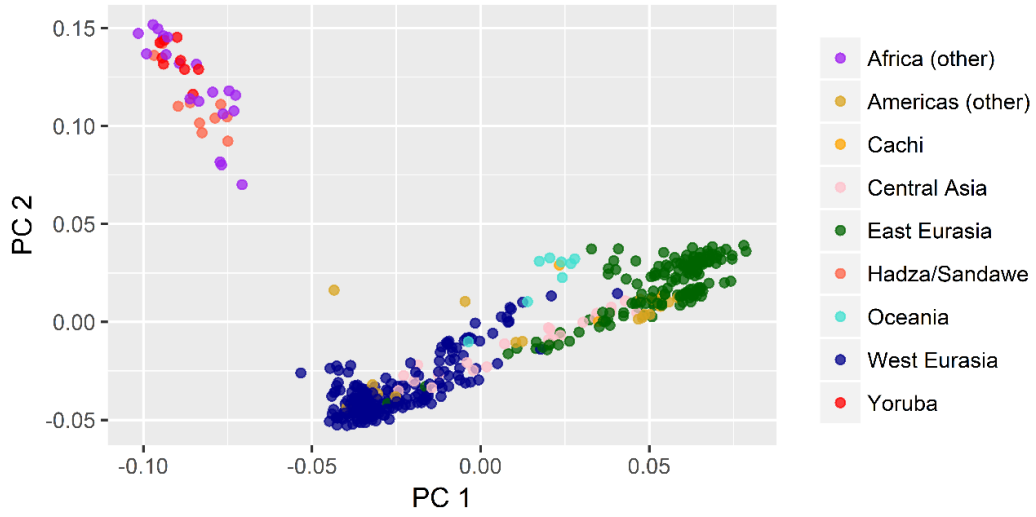
Chr	Pos	dbSNP ID	Ref	Alt	African (includes African-Americans)	European (Non-Finnish)	AFR/NFE ratio
13	56396163	rs145067755	G	A	0.0011	0.0004	2.85
13	56396652	rs150148519	T	A	0.0011	0.0004	2.86
13	56399040	rs139389896	C	G	0.0011	0.0004	2.86
13	56403346	rs140197586	A	G	0.0011	0.0004	2.87
13	56439088	rs148720765	G	T	0.0001	0.0004	0.29
13	56447236	rs146654046	T	G	0.0017	0.0004	4.3
13	56457191	rs142669527	T	G	0.0007	0.0004	1.71
13	56470315	rs117154680	A	G	0.0007	0.0005	1.47
13	56487048	rs144349306	G	A	0.0007	0.0005	1.47
13	56487662	rs143385065	G	T	0.0007	0.0005	1.47
13	56502316	rs144856974	C	T	0.0001	0.0005	0.24
13	56508281	rs141440049	G	A	0.0007	0.0005	1.47
13	56521267	rs140917381	G	A	0.0007	0.0005	1.47
13	56524220	rs116869588	A	T	0.0007	0.0005	1.47
13	56525765	rs139055259	C	T	0.0001	0.0005	0.25
13	56533265	rs138779141	A	G	0.0007	0.0005	1.47
13	56534516	rs147954492	T	C	0.0007	0.0005	1.47
13	56536157	rs140995812	C	T	0.0007	0.0005	1.47
13	56540982	rs146729977	G	A	0.0007	0.0005	1.47
13	56546891	rs143341931	G	C	0.0007	0.0005	1.47
13	56550107	rs186986045	T	C	0.0007	0.0005	1.48
13	56551150	rs138499391	C	G	0.0007	0.0005	1.47
13	56561142	rs561328841	G	A	0.0007	0.0005	1.46
13	56573625	rs151303820	A	C	0.0007	0.0005	1.47
13	56594852	rs117650089	C	A	0.0008	0.0005	1.72
13	56598270	rs181410931	G	A	0.0007	0.0005	1.47
13	56603574	rs147776426	A	G	0.0103	0.0005	22.13
13	56609453	rs115448965	T	C	0.0103	0.0005	22.08
13	56623766	rs117647948	G	A	0.0008	0.0005	1.71
13	56626034	rs144083545	G	A	0.0008	0.0005	1.71
13	56628994	rs148991207	A	G	0.0008	0.0005	1.72
13	56646504	rs116893948	C	A	0.0001	0.0005	0.24
13	56707432	rs144932424	A	G	0.0001	0.0005	0.25
13	56746556	rs146760953	C	T	0.0001	0.0005	0.25
13	56746878	rs117636048	C	T	0.0001	0.0005	0.25
13	56869120	rs188121655	G	T	0.0001	0.0005	0.24
13	56891028	rs141721579	T	A	0.0015	0.0005	3.19
13	56922177	rs117053304	A	G	0.0001	0.0005	0.25
13	56953860	rs552101655	A	G	0.0001	0.0005	0.24
13	57098291	rs142771263	A	T	0.0001	0.0005	0.25

Appendix D.34 gnomAD continental frequencies of all doubletons constituting RVCs shared between Calchaquíes and Europeans

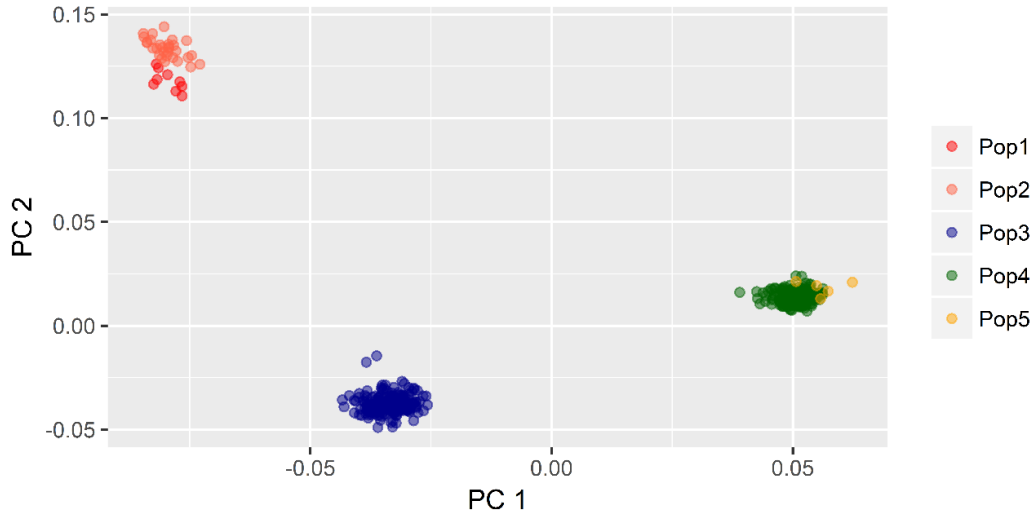
This file can be found attached to the electronic version of this thesis**.** gnomAD allele frequencies were retrieved for all f_2 variants constituting RVCs shared between Calchaquíes and Europeans. Abbreviations: AFR-Africans, AF-allele frequency, Alt-Alternative Allele, ASJ-Ashkenazi Jewish, AMR-admixed Americans, Chr-Chromosome, EAS-East Asian, FIN-Finnish, Ind-Individual, NFE-Non-Finnish Europeans, OTH-"Other", here individuals that did not unambiguously cluster with any of the other major groups in gnomAD, Pop-Population, POPMAX-Population with highest allele frequency (at a particular f_2 site), Ref-Reference Allele. The second sheet of the excel file explains the short codes used for the European populations ("pop2"-column).

Appendix D.35 PCA plots for chromosome 22 based on real and simulated data

Appendix D.35A PCA of the Diversity Set based on chromosome 22 (n_{SNPs} = 5,937, filtered as described in section 4.1.4).



Appendix D.35B PCA of genomic data simulated under scenario A based on chromosome 22 (n_{SNPs} = 7,401). Note that the number of SNPs is higher than in the real data as the short arm of the acrocentric chromosome 22 is treated like accessible sequence in the simulations.



Appendix D.35C

The PCA plots above demonstrate that known continental clusters based on genomic diversity in extant human populations can be reproduced very well with the data generated using cosi2 under the baseline scenario (Figure 4.1) proposed here.

The main differences between the empirical and the simulated datasets are briefly described in the following. Firstly, in the simulations Pop3 and Pop4 mimicking Northwest Europeans and Han Chinese respectively appear as discrete clusters whereas in real genomic data represent the geographical ends of a continuum across Eurasia. Secondly, the empirical data suggest a shift in some Africans towards West Eurasians which most plausibly indicates previously described gene flow back into Africa (e.g. Pickrell et al., 2014). Furthermore, in the simulated data the Native American-like Pop5 falls very narrowly within the range of Pop4 whereas in the real data this clustering is less tight. Besides post-Columbian admixture this hints at a more complex ancient history of Native Americans with an earlier split from their common ancestor with East Asians than previously thought and gene flow 25-20 kya from a group labelled "ancient North Eurasians" which has been revealed by aDNA studies (Moreno-Mayar et al., 2018; Raghavan et al., 2014).

Including these aspects of population history would require increasing the number of parameters as additional populations and events would need to be incorporated which increases the probability of model misspecification. Furthermore, none of these events are crucial to understanding whether and how low-level admixture from and into populations which are as deeply diverged as West Africans and the Calchaquíes is detectable from rare variant sharing patterns.

References

Moreno-Mayar JV, Potter BA, Vinner L, Steinrücken M, Rasmussen S, Terhorst J, Kamm JA, Albrechtsen A, Malaspinas A-S, Sikora M, Reuther JD, Irish JD, Malhi RS, Orlando L, Song YS, Nielsen R, Meltzer DJ, Willerslev E. 2018. Terminal Pleistocene Alaskan genome reveals first founding population of Native Americans. Nature 553:203–207.

Pickrell JK, Patterson N, Loh P-R, Lipson M, Berger B, Stoneking M, Pakendorf B, Reich D. 2014. Ancient west Eurasian ancestry in southern and eastern Africa. Proceedings of the National Academy of Sciences 111:2632–2637.

Raghavan M, Skoglund P, Graf KE, Metspalu M, Albrechtsen A, Moltke I, Rasmussen S, Stafford TW, Orlando L, Metspalu E, Karmin M, Tambets K, Rootsi S, Mägi R, Campos PF, Balanovska E, Balanovsky O, Khusnutdinova E, Litvinov S, Osipova LP, Fedorova SA, Voevoda MI, DeGiorgio M, Sicheritz-Ponten T, Brunak S, Demeshchenko S, Kivisild T, Villems R, Nielsen R, Jakobsson M, Willerslev E. 2014. Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. Nature 505:87–91.

Appendix D.36 Sharing of RVCs between Pop5 (Calchaquíes-like) and Pops1/2 (African-like) in simulated data.

This file can be found attached to the electronic version of this thesis. The first sheet of the excel file lists all RVCs shared between Pop5 (Calchaquíes-like) and Pops1/2 (African-like) detected in simulated whole genome data based on 25 scenarios describing different demographic histories. Here, the vrGV databases for Pop5 individuals were used to define the RVCs. The second sheet contains RVCs between the same interpopulation pairs, however the vrGV databases of Pops1/2 were utilised as the reference. Abbreviations (in individual IDs): EAF – East African-like, SAM – South American-like, YOR – Yoruba-like.

Appendix D.37 Anderson Darling Tests comparing RVC length distributions between empirical and simulated data

Appendix D.37A P-values for the Anderson Darling Tests comparing the distributions of RVC lengths based on the empirical sharing of Calchaquíes and Africans and of the analogous Pop5 vs Pops1/2 from several different simulated demographic histories. Significant values are bolded.

	Scenario A	Scenario B	Scenario C	Scenario D	Scenario E	Scenario F
Empirical	0.673	0.103	0.272	0.613	0.163	**0.003**
Scenario A		0.343	0.338	0.725	0.223	**0.010**
Scenario B			0.071	0.199	**0.030**	**0.003**
Scenario C				0.908	0.329	0.116
Scenario D					0.438	0.082
Scenario E						0.624

Appendix D.37B P-values for the Anderson Darling Tests comparing the distributions of RVC lengths based on the empirical sharing of Mexicans/Puerto-Ricans and Africans and of the analogous Pop5 vs Pops1/2 from several different simulated demographic histories. Significant values are bolded.

	Mexican-African RVCs	Puerto Rican-African RVCs
Scenario A	0.128	0.541
Scenario B	0.160	0.171
Scenario C	**0.012**	**0.040**
Scenario D	0.143	0.267
Scenario E	**0.013**	**0.045**
Scenario F	**$1.38*10^{-6}$**	**$5.04*10^{-5}$**

Appendix D.38 Effect of mutability on f_2 variant and run densities

Appendix D.38A

Fedorova et al. (2016) used the following formula to argue that the RVCs identified by them indicate true IBD instead of only IBS.

$$P(random\ match) = C_k^n * p^k * (1-p)^{n-k}$$

In the above equation the length and undisrupted nature of the RVC depend on the number of shared rare variants k and the scanning window size n which both then determine the total number of non-shared variants (singletons and doubletons shared with different individuals) allowed in each window (see Section 4.3.1). However, it also assumes that the null hypothesis, i.e. the sharing of one particular rare variant between two individuals by chance has the same probability across the whole genome and that the chance of observing recurrent mutations is practically zero. While the assumptions of the infinite site model support this view, it is still worth exploring much f_2 variation is influenced by genome-wide variability in mutation rates and whether there is any evidence for the occurrence of recurrent mutations.

A genome wide map of *de novo* mutations (DNMs) based on WGS data from 1,548 Icelandic parent-child trios was downloaded on 29/05/2018 from the supplementary information given by Jónsson et al. (2017). After excluding non-autosomal variants and mapping the positions originally given in hg38 to hg19 using the implementation of the UCSC liftover tool provided in the R/Bioconductor package rtracklayer (Lawrence et al., 2009) 105,443 of originally 108,778 DNMs were retained. To investigate the relationship of f_2 variants and statistics derived from them to mutability the autosomes were binned into 2,897 non-overlapping 1-Mb windows. Of these 2,732 contained at least one doubleton. To reduce the impact of potential confounders windows with < 100 f_2 variants were excluded so that 2,703 1-Mb windows were kept.

A linear regression approach was used to determine how well the 1-Mb window-specific densities of f_2 variants and related properties are correlated with DNM density. For the 4,206,155 f_2 variants contained in the relevant windows this correlation of their density with DNM density is only moderate (r_{f2_DNM} = 0.536) (Appendix D.38B). However, Smith et al. (2018) demonstrated that in simulations of genetic diversity data with a priori known mutation rates a DNM map of the same resolution as provided by Jónsson et al. (2017) has an expected correlation of 0.58 with SNP density. Assuming that overall SNP density and f_2 density are

closely related the observed correlation therefore implies that the vast majority of variation in f_2 numbers at the 1 Mb scale can be explained by variation in mutation rates.

In their analysis of the ExAC samples Lek et al. (2016) showed that in this large dataset f_2 variants resulting from CpG transitions are 3-4 times more likely than other substitution types to be shared between two populations. Under a null hypothesis of structured populations with negligible recent long-distance gene flow this implies that the high mutability CpG transitions are much more likely to occur independently and therefore create doubletons which are in IBS but not indicative of IBD. In the Diversity Set there exists a weak negative correlation between DNM density and the fraction of f_2 variants in a window that are shared between different macro-groups ($r_{f2inter\_DNM}$ = -0.133). This argues against the notion that high mutability windows contain an excess of doubletons representing independent mutation events at the given sample size (894 haploid genomes).

The general picture based on RVCs is comparable. Only RVCs with a length between 30kb and 2 Mb were included as fragments of these lengths are used in the main text to argue for inter-macro-group gene flow. This led to a reduction of the set of RVCs to 237,393 (81.3% of the original total). There is a positive, though somewhat weaker, correlation between DNM and RVC densities (r_{RVC_DNM} = 0.331). On the other hand, there is almost no measurable correlation between DNM density and the proportion of inter-macro-group RVCs ($r_{RVCinter\_DNM}$ = -0.043).

In conclusion, while mutability drives rare variant density, there is no evidence for an excess of doubletons that are in IBS for high mutation rate windows in the Diversity Set. This is in agreement with theoretical expectations and suggests that almost all doubletons that are part of RVCs with high certainty represent true IBD and that the formula for estimating the probability of RVCs occurring due to chance can be applied in a similar manner across the whole genome.

References

Fedorova L, Qiu S, Dutta R, Fedorov A. 2016. Atlas of Cryptic Genetic Relatedness Among 1000 Human Genomes. Genome Biology and Evolution 8:777–790.

Jónsson H, Sulem P, Kehr B, Kristmundsdottir S, Zink F, Hjartarson E, Hardarson MT, Hjorleifsson KE, Eggertsson HP, Gudjonsson SA, Ward LD, Arnadottir GA, Helgason EA, Helgason H, Gylfason A, Jonasdottir A, Jonasdottir A, Rafnar T, Frigge M, Stacey SN, Th. Magnusson O, Thorsteinsdottir U, Masson G, Kong A, Halldorsson BV, Helgason A, Gudbjartsson DF, Stefansson K. 2017. Parental influence on human germline de novo mutations in 1,548 trios from Iceland. Nature 549:519–522.

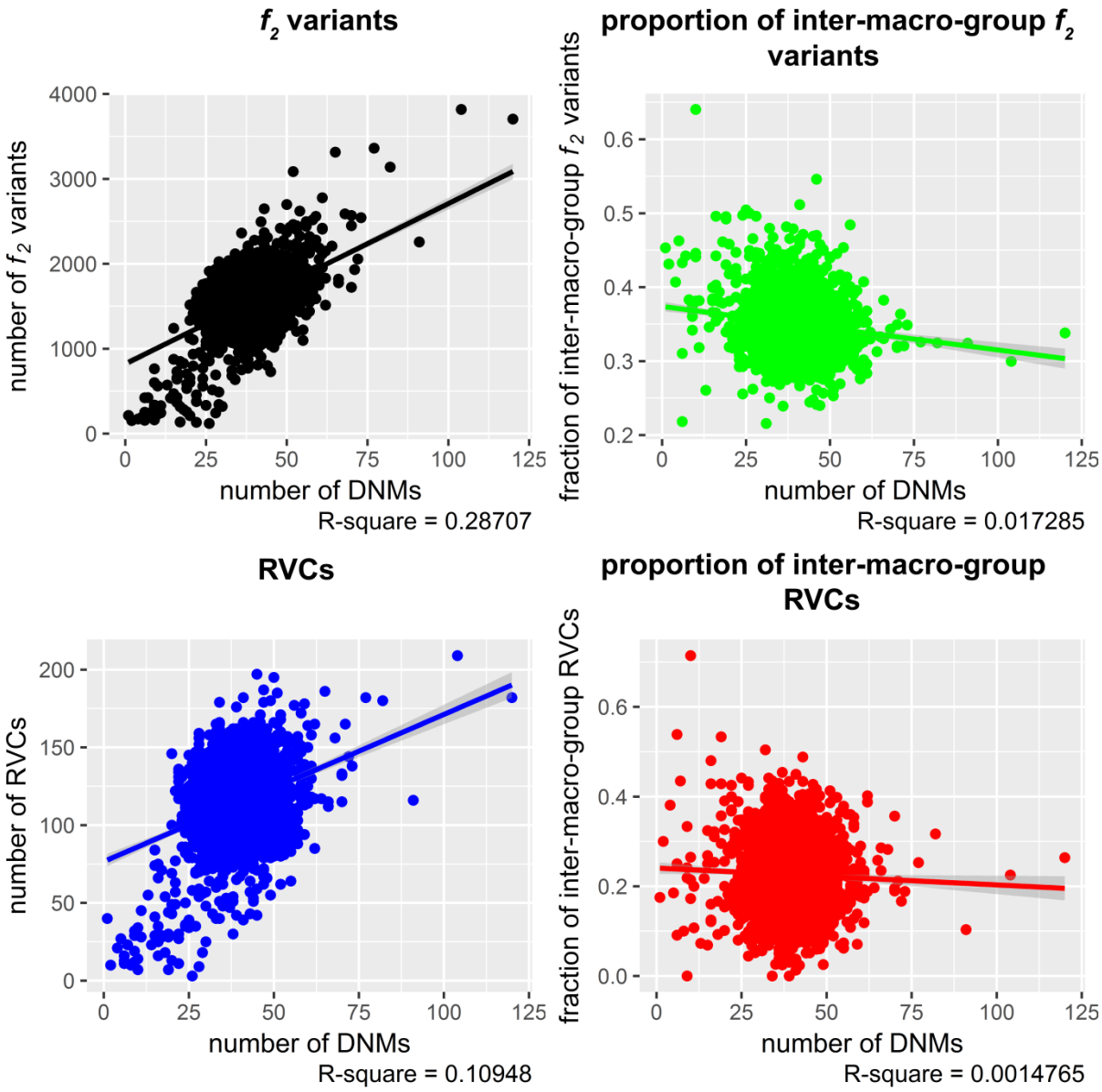
Lawrence M, Gentleman R, Carey V. 2009. rtracklayer: an R package for interfacing with genome browsers. Bioinformatics 25:1841–1842.

Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, Tukiainen T, Birnbaum DP, Kosmicki JA, Duncan LE, Estrada K, Zhao F, Zou J, Pierce-Hoffman E, Berghout J, Cooper DN, Deflaux N, DePristo M, Do R, Flannick J, Fromer M, Gauthier L, Goldstein J, Gupta N, Howrigan D, Kiezun A, Kurki MI, Moonshine AL, Natarajan P, Orozco L, Peloso GM, Poplin R, Rivas MA, Ruano-Rubio V, Rose SA, Ruderfer DM, Shakir K, Stenson PD, Stevens C, Thomas BP, Tiao G, Tusie-Luna MT, Weisburd B, Won H-H, Yu D, Altshuler DM, Ardissino D, Boehnke M, Danesh J, Donnelly S, Elosua R, Florez JC, Gabriel SB, Getz G, Glatt SJ, Hultman CM, Kathiresan S, Laakso M, McCarroll S, McCarthy MI, McGovern D, McPherson R, Neale BM, Palotie A, Purcell SM, Saleheen D, Scharf JM, Sklar P, Sullivan PF, Tuomilehto J, Tsuang MT, Watkins HC, Wilson JG, Daly MJ, MacArthur DG, Exome Aggregation Consortium. 2016. Analysis of protein-coding genetic variation in 60,706 humans. Nature 536:285–291.

Smith TCA, Arndt PF, Eyre-Walker A. 2018. Large scale variation in the rate of germ-line de novo mutation, base composition, divergence and diversity in humans. PLOS Genetics 14:e1007254.

Appendix D.38B

Plots of various statistics describing the properties of rare variation in 1-Mb windows vs. DNM density as a proxy for general mutability. Grey shaded areas around regression lines indicate the 95% confidence interval for linear predictions of the different statistics from a linear model.



Appendix D.39 Investigation of excess rare variant sharing between a British and two Chinese individuals

Appendix D.39A

One of the outcomes of the analyses aiming to detect outlier pairs who exhibit higher f_2 variant affinities than expected based on geographical distance is that UK1 shares at total of 352 doubletons with Chn2 ($n_{f2\_ UK1 \cap Chn2} = 202$) and Chn3($n_{f2\_ UK1 \cap Chn2} = 150$). In terms of magnitude this is comparable to the Vietnamese individual VietN3 who shares 353 doubletons with these two Chinese.

The purpose of this note is to examine whether this result indicates true relatedness or can plausibly be explained as a methodological artefact. A strong argument for the latter possibility is that UK1 only has a total of 26 f_2 variants in common with the other five Han Chinese in the dataset. Intuitively, all individuals from a population which shares rare variation with a target sample should exhibit a signal of a similar magnitude. This is because the target individual should be related approximately equally close to all of them. The empirical data support this, there is a very good correlation (r~0.94) between the number of f_2 variants shared with Chn2/Chn3 and similarity with regards to rare variants with the other five Chinese individuals (Appendix D.39B). UK1 represents an extreme outlier with a standardised residual of 13.99. While there are twelve more outliers for which the absolute value of the standardised residual exceeds two these are more biologically plausible (Appendix D.39C). In particular, Chn2 and Chn3 were sampled from a location close to Shanghai whereas all but one of the other Chinese are from Beijing. In consequence, individuals from the SEA region are more similar than expected to the first cluster and Northeast Asians less similar.

An apparently contradictory observation is that at least some of the shared f_2 variants occur in close proximity as part of RVCs that should indicate true IBD. The total length of all RVCs shared by UK1 with Chn2 and Chn3 is 1.55 Mb, almost all of it due to the pair UK1-Chn2 (Appendix D.39D). The latter value in turn represents the mean length of RVCs detected separately based on the individual-level f_2 databases for UK1 and Chn2. As the RVCs are defined based on a minimum cut-off of shared rare variants in a subset of consecutive rows (each row represents one singleton or doubleton in the reference individual for the respective database) the method can result in an asymmetry. This pattern is very distinct for this particular

case. There are genomic segments with a total length of ~2420 kb that only belong to RVCs if the UK1-f_2 database is used as reference.

The main biological explanation for such a pattern would be a very deep distinct lineage (perhaps introgressed from an archaic species) that is only shared by these three individuals. This would explain why some of the f_2 variant clusters comprise only very short SNP-dense segments (Appendix D.39D), as over time, recombination would have broken down almost all of the continuous IBD segments these f_2 variants were originally located on. Given that these three individuals are from well-studied populations and not from (understudied) isolated groups it does, however, not seem intuitively plausible that such a lineage would not have been detected already. Given these RVCs, either a fraction of the shared f_2 variants mark true common ancestry or the processes underlying the error are at least partly non-random. There are three main steps of NGS data generation which could have caused errors a) the sequencing itself, b) the alignment of the reads to the human reference genome, c) the variant calling from the mapped reads.

In absence of the read data for the PGP free genomes the main aspect that can be examined is how the doubletons shared with Chn2 and Chn3 are distributed across the genome of UK1. One important confounder that could be revealed are alignment artefacts caused by reads from duplicated sequences that only differ by a single or a few bases. If one copy of this repetitive element is misaligned it could appear to support substitutions at the respective differing positions in another copy (Koboldt et al., 2010).

The raw output from Complete Genomics sequencing has a complex structure consisting of two paired-end reads typically representing the ends of fragment a few hundred bp long. Each of these reads has a length of 35 bp and in turn consists of four read sections with a tightly defined relative spacing (Carnevali et al., 2012). An unusual clustering of doubletons in the genome of UK1 shared with Chn2/Chn3 within this read length could therefore potentially indicate such alignment errors.

This condition applies to 72/352 f_2 variants, i.e. the closest doubleton adjacent to them is located within 35 bp (these variants consist of 34 pairs and one four site cluster). To contextualise how unusual this spacing of doubletons is trios of individuals similar to UK1-Chn2/Chn3 were extracted from the whole Diversity Set. The conditions were as follows:

a) Target individual A_I shares >300 and <400 f_2 variants with two other individuals A_J and A_K

b) Both $n_{f2\_AI \cap AJ}$ and $n_{f2\_AI \cap AK}$ are >100 and <300

c) A_J and A_K are located in adjacent rows/columns in the raw f_2 sharing matrix (Appendix D.7A).

A total of 966 trios fulfil these requirements encompassing a total of 312,341 f_2 variants (representing 179,674 unique sites, duplicates occur because doubletons can be part of multiple three individual combinations). Of all these f_2 variants only 3477 (~1.1%) are located within 35 bp of another doubleton, i.e. there is an 18-fold increase of such closely clustered f_2 variants for UK1-Chn2/Chn3 compared to the genome-wide average for similar trios of individuals. This excess is still more than 7-fold even if only variants within 1 kb of each other are considered, ca. 2/3 of these are within 35 bp for UK1-Chn2/Chn3 whereas they are almost uniformly distributed if the data from all comparable trios are considered (Appendix D.11E). This suggests that most of the 72 f_2 variants located within one read length of each other should be considered potentially suspect. If they were all removed, only one RVC with a mean length of ~347 kb would remain (Appendix D.39D).

The most plausible technical reason for an artefact lies in how the samples were processed. After the sequencing as such Complete Genomics employs a proprietary pipeline to map and call variants from the read data (Drmanac et al., 2010). The samples in the Diversity Set were analysed with different versions of this pipeline and only the three individuals of interest were processed with version 2.0.3.x of this software (see column "CG-software version" in Appendix C1). This suggests that the majority of the rare variants shared between UK1 and Chn2/Chn3 are erroneous findings resulting from a batch effect. Interestingly, this unusual pattern is limited to how the samples relate to each other. This is further supported by 63 tripletons shared between them (data not shown). Otherwise, the common variant-based analyses that were conducted for the Pagani et al. (2016) paper would have revealed them as problematic. Furthermore, their f_2 totals per haploid genome fall within the range that is expected for their respective macro-groups ($n_{f2\_WEu\_without\_Roma}$ = 4,742±375, $n_{f2\_UK1 \cap x}$ = 5,138; n_{f2_SeM} = 5,858±469, $n_{f2\_Chn2 \cap x}$ = 5,726, $n_{f2\_Chn3 \cap x}$ = 5,804).

References

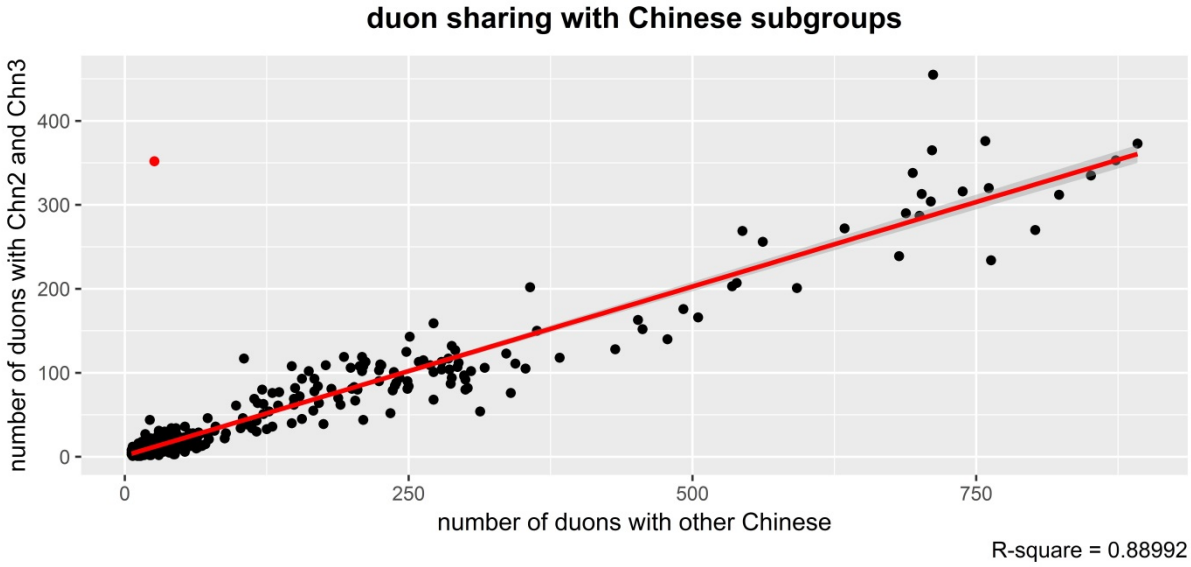
Carnevali P, Baccash J, Halpern AL, Nazarenko I, Nilsen GB et al. 2012. Computational Techniques for Human Genome Resequencing Using Mated Gapped Reads. Journal of Computational Biology 19:279–292.

Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL et al. 2010. Human Genome Sequencing Using Unchained Base Reads on Self-Assembling DNA Nanoarrays. Science 327:78–81.

Koboldt DC, Ding L, Mardis ER, Wilson RK. 2010. Challenges of sequencing human genomes. Briefings in Bioinformatics 11:484–498.

Pagani L, Lawson DJ, Jagoda E, Mörseburg A, Eriksson A et al. 2016. Genomic analyses inform on migration events during the peopling of Eurasia. Nature 538:238–242.

Appendix D.39B: Plot of f_2 sharing with Chn2/Chn3 relative to affinities with other Chinese individuals including a linear regression line. UK1 is highlighted in red.



Appendix D.39C: Standardised residuals with an absolute value >2 representing the difference between observed sharing of an individual with Chn2 and Chn3 and a prediction of this sharing based on affinities to the other five Chinese individuals in the dataset.

Macro-group	Individual ID	Standardised residual
WEu	UK1	13.99
SeM	VietS4	6.96
SeM	VietS3	3.22

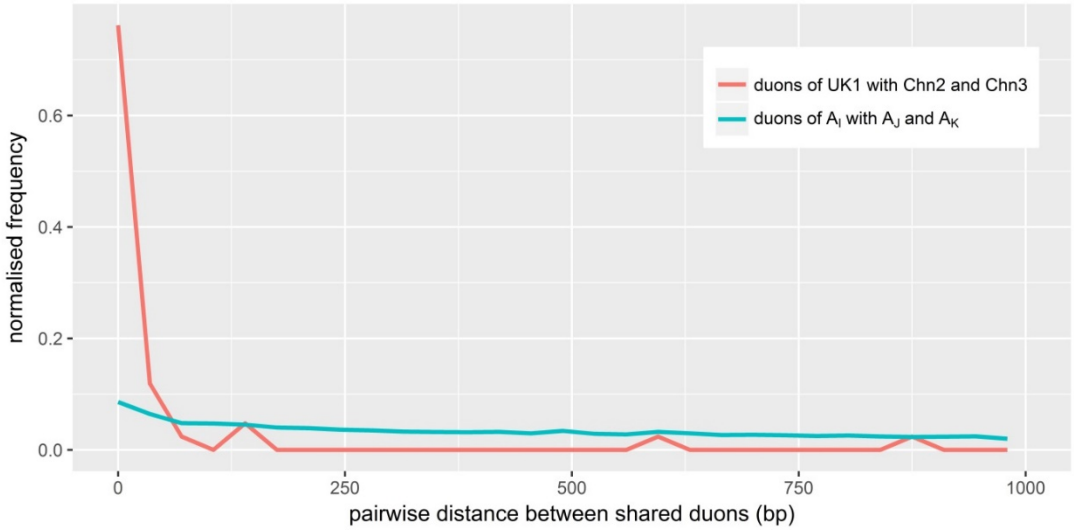
Macro-group	Individual ID	Standardised residual
SeI	Agta3	3.02
SeM	VietS5	2.90
SeM	VietC2	2.38
SeI	Murut6	2.34
SeM	Burm15	2.01
SSi	Mong5	-2.23
SeM	JPT_3	-2.28
CeA	Kaz3	-2.57
SSi	Altai2	-3.02
SeM	JPT_4	-3.12

Appendix D.39D: Table describing the composition of RVCs shared between UK1 and Chn2/Chn3. Bolded f_2 variants are located within 35 bp of each other and potentially suspect. Due to the methodology used to define RVCs based on individual-level f_2 databases there is an asymmetry in the "detected in" column. Abbreviations: Chr – chromosome, Ind - Individual

Chr	RVC start	RVC end	RVC length	f_2 site	Detected in	Ind1	Ind2	Distance to previous f_2
2	96416501	98253679	1837179	96416501	UK1	Chn2	UK1	-
2	96416501	98253679	1837179	96468072	UK1	Chn2	UK1	51571
2	**96416501**	**98253679**	**1837179**	**96496578**	**UK1**	**Chn2**	**UK1**	**28506**
2	**96416501**	**98253679**	**1837179**	**96496594**	**UK1**	**Chn2**	**UK1**	**16**
2	96416501	98253679	1837179	96650960	UK1	Chn2	UK1	154366
2	96416501	98253679	1837179	97726160	UK1	Chn2	UK1	9035
2	**96416501**	**98253679**	**1837179**	**98253671**	**UK1**	**Chn2**	**UK1**	**527511**
2	**96416501**	**98253679**	**1837179**	**98253679**	**UK1**	**Chn2**	**UK1**	**8**
2	112062161	112607041	544881	112062161	UK1	Chn2	UK1	-
2	112062161	112607041	544881	112555968	UK1	Chn3	UK1	493807
2	**112062161**	**112607041**	**544881**	**112596922**	**UK1**	**Chn2**	**UK1**	**40954**
2	**112062161**	**112607041**	**544881**	**112596929**	**UK1**	**Chn2**	**UK1**	**7**
2	**112062161**	**112607041**	**544881**	**112596941**	**UK1**	**Chn2**	**UK1**	**12**
2	**112062161**	**112607041**	**544881**	**112596955**	**UK1**	**Chn2**	**UK1**	**14**
2	112062161	112607041	544881	112607041	UK1	Chn2	UK1	10086
2	**112596922**	**112607041**	**10120**	**112596922**	**Chn2**	**Chn2**	**UK1**	**40954**
2	**112596922**	**112607041**	**10120**	**112596929**	**Chn2**	**Chn2**	**UK1**	**7**
2	**112596922**	**112607041**	**10120**	**112596941**	**Chn2**	**Chn2**	**UK1**	**12**
2	**112596922**	**112607041**	**10120**	**112596955**	**Chn2**	**Chn2**	**UK1**	**14**
2	112596922	112607041	10120	112607041	Chn2	Chn2	UK1	10086

Chr	RVC start	RVC end	RVC length	f_2 site	Detected in	Ind1	Ind2	Distance to previous f_2
4	**9790586**	**9795624**	**5039**	**9790586**	**Chn3, UK1**	**Chn3**	**UK1**	**-**
4	**9790586**	**9795624**	**5039**	**9790596**	**Chn3, UK1**	**Chn3**	**UK1**	**10**
4	**9790586**	**9795624**	**5039**	**9791718**	**Chn3, UK1**	**Chn3**	**UK1**	**1122**
4	**9790586**	**9795624**	**5039**	**9791730**	**Chn3, UK1**	**Chn3**	**UK1**	**12**
4	9790586	9795624	5039	9795624	Chn3, UK1	Chn3	UK1	3894
6	57825835	58197261	371427	57825835	UK1	Chn2	UK1	-
6	57825835	58197261	371427	57873706	UK1	Chn2	UK1	47871
6	57825835	58197261	371427	58153180	UK1	Chn2	UK1	1575
6	57825835	58197261	371427	58153314	UK1	Chn2	UK1	134
6	57825835	58197261	371427	58158436	UK1	Chn2	UK1	5122
6	57825835	58197261	371427	58163482	UK1	Chn2	UK1	5046
6	57825835	58197261	371427	58197261	UK1	Chn2	UK1	33779
6	57873706	58197261	323556	57873706	Chn2	Chn2	UK1	47871
6	57873706	58197261	323556	58153180	Chn2	Chn2	UK1	1575
6	57873706	58197261	323556	58153314	Chn2	Chn2	UK1	134
6	57873706	58197261	323556	58158436	Chn2	Chn2	UK1	5122
6	57873706	58197261	323556	58163482	Chn2	Chn2	UK1	5046
6	57873706	58197261	323556	58197261	Chn2	Chn2	UK1	33779

Appendix D.39E: Plot of the normalised frequency distribution of pairwise distances between shared doubletons that are within 1 kb of each other. The bin size is 35 bp, there is a clear excess of inter-f_2 distances falling into the closest distance bin for UK1-Chn2/Chn3 compared to similar trios of individuals from the whole Diversity Set. For simplicity, this graph does not separately highlight cases where more than two doubletons fall within 35 bp.



Appendix D.40 Effect of sample composition on per-individual f_2 variant counts

Appendix D.40A

The purpose of this technical note is to explore which factors explain the almost three-fold higher per-individual count of f_2-variants in the Diversity Set compared to phase 3 of the 1000 Genomes Project. The two not mutually exclusive main hypotheses are that this divergence is due to i) differences in sample sizes and strategies and ii) technical factors, i.e. the higher sequencing coverage for the EGDP samples.

To test hypothesis i) in principle data from one chromosome from the 1000 Genomes Project data should be sufficient. Chromosome 2 was chosen, under the simplifying assumption that variant density is approximately equal across all chromosomes it represents ca. 8.44 % of all autosomal variants. The relevant genotype data was retrieved from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/ALL.chr2.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz (downloaded on 15/03/2018). Doubletons were extracted as described in section 4.1.1. To create a sub-dataset comparable to the "model world" generated for the Diversity Set six individuals were randomly sampled from 24/26 populations that are part of the 1000 Genomes Project phase 3 data. Note that the British and Southern Han Chinese were not included in the "model world" as the original analyses of the 1000 Genomes data suggest that they are so similar to the Han Chinese from Beijing and the CEPH Northwest Europeans respectively that their inclusion would mean an oversampling of very specific subpopulations.

Regardless of whether they were down-sampled or not the general patterns of population differences in overall f_2-diversity in the 1000 Genomes Project data are largely consistent with those observed for the Diversity Set (see section 4.3.1). Furthermore, as theoretically expected while the total number of f_2-variants on chromosome 2 is higher when all individuals are included (n = 758,644) than for the reduced set (n = 232,415) for the per-individual f_2 counts this relationship is reversed. This is because in the latter case many variants occurring in a heterozygous state in three or more individuals in a larger sample become doubletons, this shift is strongest in Sub-Saharan Africans.

Extrapolating the observations from chromosome 2 to the whole genome for the "model world" yields an individual mean f_2-count of ~39,000 for Sub-Saharan Africans (Appendix D.40B)

compared to an average of ~47,300 doubletons in individuals from the same region in the Diversity Set. However, the actual populations representing this part of the world are quite different in the two analyses. Notably, the Diversity Set contains Pygmy groups with very high levels of genome-wide diversity who at least partly drive the observed effect ($\overline{f_2}_{\text{Pygmy\_model\_world}}$ = 65,309.1). However even for the Yoruba, the only group that is part of both datasets, higher f_2-counts are observed in the Diversity Set ($\overline{f_2}_{\text{Yoruba\_model\_Diversity}}$ = 44,185.2 vs. $\overline{f_2}_{\text{Yoruba\_model\_1000Genomes}}$ = 39,492.89). The per-individual counts for non-Africans from the 1000 Genomes dataset under the "model world" sampling scheme are ~10,000-12,000 (Appendices D.40B-C). These totals are in a very comparable range to the results from the relevant subset of the Diversity Set. For the latter the f_2-counts in non-Africans vary between ~10,000 and ~14,000 which is on average slightly higher than for the 1000 Genomes data, except for the East Asians/mainland Southeast Asians. They potentially have a slightly lower average f_2-count in the Diversity Set ($\overline{f_2}_{\text{EastandMainlandSoutheastAsians\_model\_Diversity}}$ = 10,077.9).

In conclusion, most of the divergence in per-individual f_2 counts between the Diversity Set and the 1000 Genomes phase 3 data can be accounted for by differences in sample size and composition. The remaining more modest trend towards higher f_2–numbers for most populations in the Diversity Set could perhaps be attributable to the higher average sequencing coverage in the samples from the EGDP panel.

Appendix D.40B Average per individual counts of f_2 variants on chromosome 2 in global superpopulations displayed a) for the unadjusted sample sizes in the 1000 Genomes phase 3 dataset, b) on a subset of individuals (n = 144) designated as the "model world" which was designed to make the data comparable to a similarly treated subset of the Diversity Set analysed in this thesis.

Superpopulation	f_2 **counts full dataset**	f_2 **counts "model world"**
African	397.6	3303.6
American	247.4	968.7
East Asian	306.9	864.7
European	189.7	859.2
South Asian	326.9	968.0

Appendix D.40C: Average per individual counts of f_2 variants on chromosome 2 in 26 global populations displayed a) for the unadjusted sample sizes in the 1000 Genomes phase 3 dataset, b) on a subset of individuals (n = 144) randomly downsampled and designated as the "model world". It was designed to make the data comparable to a similarly treated subset of the Diversity Set analysed in this thesis. Note that for the latter British and Southern Han Chinese were not analysed as they are so similar to closely related groups (Northern Han, Northwest Europeans) that their inclusion would cause an overrepresentation of the respective subpopulation.

Population	f_2 **counts full dataset**	f_2 **counts "model world"**
African-Caribbean	340.0	2852.8
African-Americans Southwest	376.7	2913.3
Esan	306.3	3494.3
Gambian Mandinka	413.7	3188.8
Luhya	623.2	3575.1
Mende	460.4	3767.9
Yoruba	271.5	3333.2
Colombian	266.3	1045.5
Mexican-American	218.3	931.4
Peruvian	191.6	798.0
Puerto Rican	293.9	1099.9
Dai Chinese	344.4	928.4
Han Chinese	276.5	835.9
Southern Han Chinese	266.8	-
Japanese	311.5	880.8
Kinh Vietnamese	341.2	813.6
Utah residents (CEPH) with Northern and Western European ancestry	184.4	892.5
Finnish	153.7	872.0
British	179.3	-
Spanish	189.8	850.4
Tuscan	236.5	822.0
Bengali	318.4	982.9
Gujarati	275.2	1018.8
Indian	338.8	965.7
Punjabi	327.2	918.3
Sri Lankan	373.9	954.2

Appendix D.41 Matrices containing f_2 sharing totals on chromosome 2 for all individuals from phase 3 of the 1000 Genomes Project

This file can be found attached to the electronic version of this thesis. The first sheet of the excel file (**Appendix D.41A**) is a raw sharing matrix of f_2 variant counts on chromosome 2 between all individuals from phase 3 of the 1000 Genomes Project (n = 2,504). The row and column names are formatted as "populationID_individualID". The second sheet of the excel file contains the same sharing patterns (**Appendix D.41B**) for a reduced set of individuals (n = 144) where a sample of size six was randomly drawn from each group. The third sheet explains the population abbreviations.

Appendix D.42 Correlation between ChromoPainter and f_2 sharing using Mathieson's transformations

Appendix D.42A

The purpose of this technical note is to briefly explore whether the high correlation between ChromoPainter affinities and rare variant sharing is robust to additional transformations of such data that have been applied by other researchers.

Mathieson (2013) used the following statistic to describe f_2 sharing between two individuals i and j: $\widetilde{d_{ij}} = \log_{10}\left(\frac{d_{ij}}{d_{*j}}\right)$ where d_{ij} is the number of f_2 variants shared by individuals i and j whereas d_{*j} is the number of f_2 variants individual j shares with any other individual. The metric $\widetilde{d_{ij}}$ was then converted into z-scores and the same approach was applied to total IBD sharing length inferred with fastIBD. Finally, the correlation between the two normalised metrics calculated for all individuals from phase 1 of the 1000 Genomes Project was obtained (r ~0.58). Besides the added transformations this approach differs from Eq. (4.1) in that the diversity-level normalisation is only done with the f_2 totals in one individual and not with the mean from the totals from both individuals. Therefore, $\widetilde{d_{ij}}$ and $\widetilde{d_{ji}}$ have different values, i.e. the resulting rare variant sharing matrix is non-symmetric whereas the one obtained from Eq. (4.1) is symmetric.

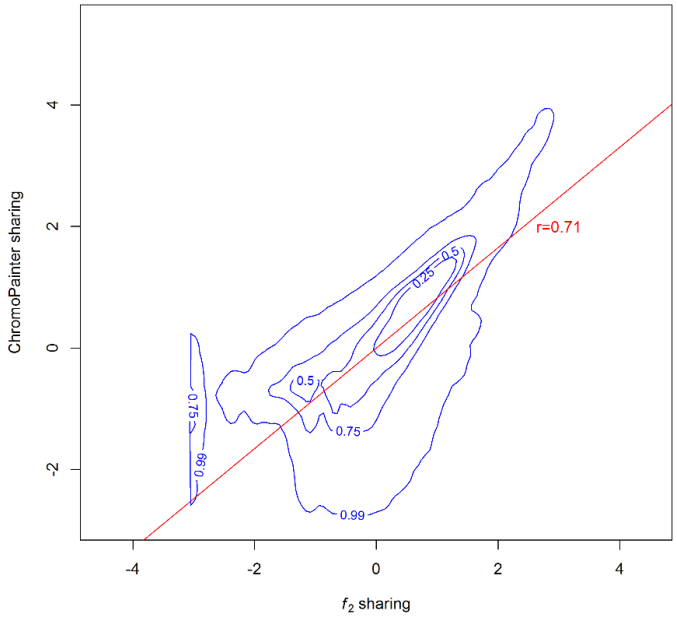
The log transformation should reduce the rightward skew of the pairwise f_2 and IBD sharing, who both follow long-tailed distributions, but preserve the fundamental underlying relationship between the two variables. To test this expectation i) f_2 sharing was calculated as proposed by Mathieson and ii) the metric $d(i1 \cap i2)$ resulting from Eq. (4.1) was transformed in the same way as Mathieson's sharing statistic. Finally, the ChromoPainter sharing matrix (Appendix D9B) was treated similarly. Consistent with the relationship between the untransformed variables (Figure 4.7) both Mathieson's f_2 sharing (r ~0.71) (Appendix D.42B) and the transformed $d(i1 \cap i2)$ (r ~0.83) (Appendix D.42C) exhibit a higher correlation with ChromoPainter sharing than originally observed by Mathieson for IBD.

References

Mathieson I. 2013. Genes in space: selection, association and variation in spatially structured populations. PhD dissertation. University of Oxford

Appendix D.42B

Contour plot displaying the joint empirical cumulative distribution derived from a bivariate kernel density estimate of ChromoPainter and f_2 sharing in the Diversity Set. Both metrics were transformed following Mathieson (2013). This graph was chosen to avoid overplotting given the high number of data points (199,362). The concentric contoured areas labelled with "x" contain x*100 % of the total dataset, beginning at 25% up to 99%.



Appendix D.42C

The methodology is the same as in Appendix D.42B. However, d(i1∩i2) was transformed instead of Mathieson's f_2 metric and compared to similarly treated ChromoPainter sharing.