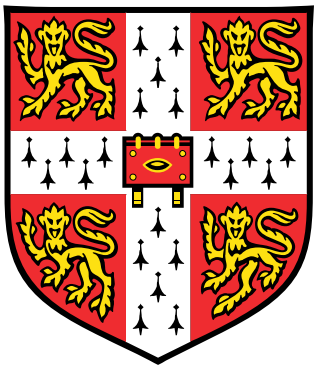
Human population history and genetic adaptation in the Himalayan region

University of Cambridge

Corpus Christi College



A thesis submitted for the degree of
Doctor of Philosophy

Elena Arciero

The Wellcome Sanger Institute
Wellcome Genome Campus
Hinxton, Cambridge
CB10 1SA, UK

September 2018

To my parents, Cristina and Virgilio

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification at the University of Cambridge or any other University or similar institution. This dissertation is my own work carried out under the supervision of Prof. Chris Tyler-Smith at the Wellcome Sanger Institute, while a member of Corpus Christi College, University of Cambridge, and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation does not exceed the prescribed word limit of 60,000 words excluding bibliography, figures, tables, equations and appendices.

Elena Arciero

September 2018

Acknowledgements

First and foremost I would like to express my sincere gratitude to my supervisor Dr Chris Tyler-Smith for giving me the opportunity to carry out this project and for his continuous support, advice and mentorship during my PhD. I would like to thank Dr Yali Xue and Dr Qasim Ayub for their guidance and for their critical and constructive assessment of my work. I would also like to thank Dr Marc Haber for his help in the day to day work. Many thanks to the rest of my thesis committee, Dr Toomas Kivisild and Dr Daniel Gaffney for their insightful comments and ideas on my work. I am also very grateful to the Wellcome Sanger Institute Graduate Programme and the Committee of Graduate Studies, as well as the Wellcome Trust for my PhD studentship. My sincere thanks to all my collaborators including Dr Thirsa Kraaijenbrink, Prof. Peter de Knijff, Dr Asan, Prof. Mark Jobling and Prof. George van Driem for providing their expertise and useful insight on Himalayan populations. Thank you to other people who contributed to this project including Dr Michał Szpak, Dr Massimo Mezzavilla, Dr Laurits Skov, Dr Shane McCarthy and Dr Yuan Chen for the fruitful discussions and their contribution with some of the analyses of this project. Many thanks also to Dr Andrew Knights, Dr Sebastian Lukasiak and Dr Irene Gallego-Romero for their expertise and advice in the design of the functional work on *EPAS1*. Thank you also to Dr Nikolaos Panousis for his assistance with the RNA-seq data and to Dr Fengtang Yang and the Cytogenetic Core Facility for the karyotyping and support with the cell lines. I would also like to thank all current and former Human Evolution team members, in particular Dr Pille Hallast, Mohamed Almarri, Dr Javi Prado Martinez and Dr Michał Szpak for their support, stimulating discussions over lunch break and all the fun we had together over the years. Thank you to my fellow PhD students, especially Mari Niemi for all the days we spent together with endless hours of work before deadlines. Also thank you to all other friends at campus for the coffee breaks and interesting chats and the gym clubbers for the good time training together. Many thanks to my all my friends in Cambridge, especially María for her positivity and encouragement over the years and the Italian community, including Claudia, Cinzia, Luca, Annalisa, Giuseppe and Gianni for their support, love and for making me always feel at home. I also thank my best friends Francesca, Martina, Chiara, Marina and Chiara for being always with me regardless of our physical distance and for believing

in me every time. Thanks to my friend Simone for making me smile every time even in very stressful situations. Many thanks to my boyfriend Chris for his love, support and patience in a difficult moment of my life.
Last but not the least, I would like to thank all my family: my parents Cristina and Virgilio, my grandparents Clara e Mario, my grandmother Giovanna that is no longer with us, my cousin Silvia and my aunt Anna, my uncle Arturo and his wife Barbara, and all the others for supporting me throughout my life.

Abstract

The Himalayan mountain range contains the highest peaks on Earth and has provided a diversity of environments for humans, some of which have required substantial genetic adaptation. I have used a combination of SNP-chip data, genome sequences and functional studies to explore the demographic history, genetic structure and signatures of adaptation in Himalayan populations. Eight hundred and eighty three individuals from 49 different autochthonous groups from Nepal, Bhutan, North India, and the Tibetan Plateau in China were genotyped for ~600,000 genome-wide SNPs. High-coverage whole-genome sequences of 87 individuals from a subset of these populations plus three additional ones were also generated. Himalayan populations share a common genetic component derived from a single ancestral population, followed by the development of local fine structure correlating with language and geographical distribution. I find higher genetic diversification within the Himalayan populations than in the surrounding regions which correlates with the distribution of Indo-European and Tibeto-Burman speakers, suggesting that both language and geography have influenced the genetic structure of these populations. I refined Himalayan population demographic history, using both autosomal and uniparental sequences. Himalayan populations display different proportions of gene flow with neighbouring populations and diverse effective population sizes and split times. The Y-chromosome lineages identified are common in South and East Asia and Tibet, but mostly form distinct clusters in the Himalayas. High altitude adaptation seems to have originated in a single ancestral population and then spread widely in the Himalayan region in the last 5,000 years. Genetic signatures of adaptation to high altitude are observed in the Endothelial PAS Domain Protein 1 gene (*EPAS1*) and several other known and novel candidates. *EPAS1* has previously been reported to be under selection and involved in adaptation to living at high altitudes and was suggested to result from introgression of DNA from an extinct hominin species (Denisovans) into Tibetans. However, functional studies of *EPAS1* variants have not been systematically carried out and it is still unknown which variant(s) are responsible for high altitude adaptation and their mechanism of action. I used both *in silico* and *in vitro* studies to explore these topics and validate *EPAS1* candidate regulatory variants. The introgressed haplotype extends for over 300 kb spanning six

genes (*EPAS1, TMEM247, ATP6V1E2, RHOQ, CRIPT, PIGF)*. I optimised a protocol to induce hypoxia in cell lines with and without the Denisovan introgressed haplotype. Preliminary results show that in cell lines without the introgressed haplotype, *EPAS1* expression increases under hypoxic conditions, whereas in the cell lines with the introgressed haplotype the expression of *EPAS1* remains constant. The most likely functional candidate variants fall in a ~32.7kb region within *EPAS1*. High altitude adaptation thus seems to be driven by *EPAS1* as well as coordinated by other genes involved in the hypoxic response.

Table of Contents

List of figures 13
List of tables 17
1. Introduction 19
1.1 Human origins, ancient migrations and mixing 19
1.2 Genetic variation and positive selection in human populations 22
1.2.1 Human genetic variation 22
1.2.2 Natural selection 29
1.3 The Himalayan region: an overview 32
1.4 High altitude adaptation and the hypoxia molecular pathway 35
1.4.1 Adaptation at high altitude 35
1.4.2 Hypoxia molecular pathway 38
1.5 Aims and objectives 42
1.6 Outline of the thesis 43
2. Population genetic analyses of Himalayan samples using SNP-genotype data 45
2.1 Introduction 45
2.2 Materials and Methods 49
2.2.1 Samples and dataset 49
2.2.2 Methods used for the analyses 52
2.3 Results 60
2.3.1 Himalayan samples show distinct patterns of population structure 60
2.3.2 Complex demographic history in the Himalayas 76
2.3.3 Signatures of adaptation in the Himalayan region 83
2.4 Discussion 91
2.4.1 Population structure and demography 91
2.4.2 High altitude adaptation 93
3. Fine-scale demographic history of Himalayan populations using whole-genome sequencing data 97
3.1 Introduction 97
3.2 Materials and Methods 98
3.2.1 Samples and datasets 98
3.2.2 Variant calling and filtering 99
3.2.3 QC 108
3.2.4 Datasets 117
3.2.5 Genetic variation analyses 118

3.2.6 Demographic analyses 120
3.2.7 Fine mapping of positive selection 123
3.3 Results 125
3.3.1 A large high-quality sequencing dataset of Himalayan populations 125
3.3.2 Recent demographic history – genetic structure 132
3.3.3 Long term demographic history 156
3.3.4 Fine-scale positive selection signals 166
3.4 Discussion 175
4. Functional investigation of *EPAS1* in high altitude adaptation 179
4.1 Introduction 179
4.2 Material and methods 181
4.2.1 Samples 181
4.2.2 Methods and protocols used for the analyses 183
4.3 Results 191
4.3.1 The *EPAS1* introgressed haplotype in 1000 Genomes Project data 191
4.3.2 Re-analysis of *EPAS1* expression from publicly available data 195
4.3.3 Sequence conservation patterns in the *EPAS1* locus 200
4.3.4 Cell line karyotyping 202
4.3.6 *EPAS1* expression profiles 203
4.4 Discussion 219
5. General discussion 221
5.1 Summary 221
5.2 Next steps 228
5.3 Future research 229
References 233
APPENDIX A 253
APPENDIX B 254
APPENDIX C 261
APPENDIX D 273

List of figures

Figure 1. 1 Candidate genes for high altitude adaptation in Andeans, Ethiopians and Tibetans....37
Figure 1. 2 The molecular hypoxic pathway....39

Figure 2. 1 Population samples analysed in this project....60
Figure 2. 2 PCA and ADMIXTURE analysis using the world dataset....63
Figure 2. 3 PCA and ADMIXTURE analysis using the Himalayan dataset....66
Figure 2. 4 Positive correlation between the high altitude-specific genetic component and altitude....67
Figure 2. 5 Long-term effective population size (Ne) and divergence times of Himalayan populations....68
Figure 2. 6 Runs of homozygosity (ROHs) in the Himalayan populations....70
Figure 2. 7 Genetic structure of the Himalayan populations from haplotype analysis using fineSTRUCTURE, and comparison with language....72
Figure 2. 8 Statistical analysis of genetic sharing between pairs of modern high altitude Himalayan populations and Han individuals....75
Figure 2. 9 Admixture history of five Himalayan populations....77
Figure 2. 10 TreeMix results from the worldwide dataset....79
Figure 2. 11 Relative genetic similarity of the Himalayan region and other populations to four ancient DNA samples....80
Figure 2. 12 Principal component analysis of the world dataset with ancient Himalayans projected onto the plot....81
Figure 2. 13 Principal component analysis of modern human populations, Denisova and Neanderthal....82
Figure 2. 14 Signals of positive selection (adaptation) in the Himalayan populations....87
Figure 2. 15 SLC52A3 protein homology model....89

Figure 3. 1 Population samples included in this project....98
Figure 3. 2 The distribution of SNP parameters after filtering the Himalayan dataset....100
Figure 3. 3 The distribution of INDEL parameters after filtering the Himalayan dataset....101
Figure 3. 4 The distribution of singleton SNPs in the Himalayan dataset....112
Figure 3. 5 Site frequency spectrum (SFS) of the Himalayan SNP dataset....113
Figure 3. 6 INDEL frequency distribution in the Himalayan dataset....115
Figure 3. 7 Observed folded site frequency spectrum (SFS) distribution of the Himalayan INDEL dataset....116
Figure 3. 8 Functional consequences of INDELs in the coding sequence....125
Figure 3. 9 Principal component analyses (PCA)....129
Figure 3. 10 ADMIXTURE (K value of 2 and 3) analysis of the Himalayan samples....129
Figure 3. 11 PCA on CNVs....130
Figure 3. 12 TreeMix results from the worldwide dataset....133
Figure 3. 13 Heatmap of pairwise F_{ST} values for the worldwide dataset....134
Figure 3. 14 Migration edges for the worldwide dataset from the Treemix analysis....135

Figure 3. 15 Common variant outgroup *f3* statistics for the Himalayan samples 137
Figure 3. 16 Rare variant outgroup *f3* statistics for the Himalayan samples.....138
Figure 3. 17 Local ancestry PCAdmix analysis.....139
Figure 3. 18 Heterozygosity rate in the Himalayan populations.....140
Figure 3. 19 Heterozygosity rate in the Himalayan populations compared to other worldwide populations.....142
Figure 3. 20 Heatmap of shared IBD tracts 143
Figure 3. 21 Heatmap of singleton sharing between Himalayan and 1000 Genomes Project individuals 145
Figure 3. 22 Heatmap of *f2* variant sharing in the worldwide dataset 146
Figure 3. 23 Zoom showing the Himalayan samples from the heatmap of *f2* variants in Figure 3.22 147
Figure 3. 24 MSMC2 Ne plot 148
Figure 3. 25 MSMC2 Ne plot 149
Figure 3. 26 SMC++ Ne plot 150
Figure 3. 27 Map of Y chromosome haplogroups 151
Figure 3. 28 Phylogenetic tree of the Y chromosomes of Himalayan and worldwide (SGDP) samples 153
Figure 3. 29 Map of mtDNA haplogroups 154
Figure 3. 30 Phylogenetic tree of mtDNA of Himalayan samples and other worldwide individuals from the 1000 Genomes Project.....155
Figure 3. 31 MSMC2 analysis of split times 157
Figure 3. 32 MSMC2 analysis of split times between Himalayan populations.....158
Figure 3. 33 MSMC2 analysis of split times between Himalayan populations.....159
Figure 3. 34 MSMC2 analysis of split times between Himalayan populations.....160
Figure 3. 35 MSMC2 analysis of split times between Himalayan populations.....161
Figure 3. 36 Comparison of archaic introgressed segments in different populations 163
Figure 3. 37 Archaic segments near *EPAS1* 165
Figure 3. 38 Manhattan plots of genome-wide F_{ST} between low and high altitude Himalayan populations.....167
Figure 3. 39 Manhattan plot of single-locus F_{ST} values around *EPAS1* between Tibetans and Han Chinese 168
Figure 3. 40 Manhattan plot of genome-wide Fine*MAV* scores in Himalayans.....170
Figure 3. 41 Manhattan plot of genome-wide Fine*MAV* scores in Himalayans.....172
Figure 3. 42 HIF-2α Position Weight Matrix Logo 173

Figure 4. 1 Workflow of RNA sequencing experiment. 190
Figure 4. 2 Haplotype network of the ~32.7 kb *EPAS1* introgressed region 194
Figure 4. 3 *EPAS1* expression across different tissues in the GTEx Portal 195
Figure 4. 4 *EPAS1* expression across different tissues in the ENCODE database.....196
Figure 4. 5 Comparison of the hypoxic factors expression levels in different human tissues.....198
Figure 4. 6 Analysis of publicly-available Tibetan placental RNA-seq data 199
Figure 4. 7 *EPAS1* evolutionarily conserved regions (ECRs) 201
Figure 4. 8 qPCR amplification plot of target genes 204
Figure 4. 9 Primer melting temperature curves 205
Figure 4. 10 C_T differences between normoxia and hypoxia at 4 hours 206
Figure 4. 11 *EPAS1* expression at 4 hours hypoxic exposure 207

Figure 4. 12 C_T differences between normoxia and hypoxia at 24 hours.....209
Figure 4. 13 *EPAS1* expression at 24 hours hypoxic exposure.....210
Figure 4. 14 Total number of reads sequenced in the first run (26520).....212
Figure 4. 15 Total number of reads sequenced in the second run (26660).....214
Figure 4. 16 Proportion of exonic reads in the samples 214
Figure 4. 17 PCA on whole-genome expression in LCL samples.....215
Figure 4. 18 Normalised read count between normoxia and hypoxia.....216
Figure 4. 19 Differential gene expression between normoxia and hypoxia and introgressed and non introgressed samples.....218

Figure 5. 1 Genomic location of potential drivers of selection in *EPAS1* region.....227

List of tables

Table 2. 1 List of populations analysed in this study 51
Table 2. 2 Datasets used for the analyses.....53
Table 2. 3 Genomic regions showing the strongest signals of positive selection in the Himalayan populations.....86

Table 3. 1 Read depth estimated from BAM files.....106
Table 3. 2 Genotype concordance between whole-genome sequencing and SNP array genotypes in 26 Himalayan samples 109
Table 3. 3 SNP call set QC metrics.....110
Table 3. 4 INDEL call set QC metrics.....114
Table 3. 5 Different datasets used for the different analyses 118
Table 3. 6 Y chromosome split times within Himalayan individuals.....152

Table 4. 1 *EPAS1* genotypes of the LCLs used in this study 182
Table 4. 2 Primer sequences used for RT-qPCR analyses. 188
Table 4. 3 The 1000 Genomes Project samples carrying the *EPAS1* Denisovan-core introgressed haplotype.....193
Table 4. 4 Cell line karyotypes 202

Table 5. 1 Potential drivers of selection in the *EPAS1* region.....225

1. Introduction

1.1 Human origins, ancient migrations and mixing

The origin of the genus *Homo* in Africa was proposed for the first time by Charles Darwin in *"The Descent of Man and Selection in Relation to Sex"* (1871) (1). Founded on embryological and Thomas Huxley's comparative anatomical studies that illustrated how modern humans and apes share a common ancestor (2), Darwin came to the conclusion that the last shared ancestor of humans and apes probably lived in Africa. The discovery of relevant fossils and, more recently, the advent of genetic studies, confirmed Darwin and Huxley's suggestions. The earliest known human species, *Homo habilis*, lived around 2.5 million years ago in Africa, and the closely related species *Homo erectus* and *Homo ergaster* appeared after 1.9 million years ago. *Homo erectus* was the first species with a body size and shape resembling modern humans and was also the first to expand outside Africa. Between 700,000–300,000 years ago, at least four main human lineages were present in different part of the world: *Homo erectus* in Asia, *Homo neanderthalensis* in Eurasia, *Homo sapiens* in Africa and the Denisovans in Asia, a group closely related to Neanderthals (3, 4). Anatomically modern humans (AMH) seem to have evolved in Africa in the last 300 thousand years (ka) where the earliest AMH fossils known thus far have been found at Jebel Irhoud in Morocco dating to ~310 thousand years ago (5), and at the Ethiopian sites of Omo Kibish dating to ~195 ka (6), and Herto dating to ~160 ka (7). Nowadays, although there is general acceptance of the African origin of modern humans, where specifically within this continent modern humans originated, and how and when the expansion of our ancestors from Africa to the rest of the world occurred, are still debatable. The traditional location for modern human origins has been Eastern Africa, where the human fossils of Omo Kibish and Herto were discovered, which were for many years the oldest known AMH remains. Genetic studies supporting this hypothesis reported that non-African populations carry part of the genetic diversity present in Eastern African individuals (8, 9). Nevertheless, this hypothesis has been challenged in recent years by new archaeological and genetic discoveries. The recent report of human fossils dating back to ~300 thousand years

ago in Morocco raise the possibility of a Northern African origin. Genetic evidence from the Y chromosome reported the deepest-rooting Y chromosome clade (called A00) in West Africa (10, 11). On the other hand, the genetic analysis of sub-Saharan African hunter-gatherer communities suggested the possibility of a Southern African origin. The analysis of population differentiation and linkage disequilibrium (LD) patterns suggested that these populations are the most genetically differentiated in the world and thus in agreement with a South African modern human origin (12). New genetic and archaeological evidences could confirm one of these hypotheses or even suggest a more complex origin of our species from multiple regions (13).

Another topic of debate is the route of the modern human dispersal out of Africa (OoA). Currently there are two main hypotheses, both with dispersal from Eastern Africa, but via a northern or southern route (14, 15). The northern path passes across northern Egypt into the Sinai Peninsula through a land corridor. Modern human remains found at Skhul (16) and Qafzeh (17) in Israel dating between 120 and 90 ka support this route of dispersal into the Middle East. More recently, a maxilla with associated dentition was discovered at Misliya Cave in Israel and dated to around 177-194 ka, suggesting that modern humans might have left Africa even earlier than previously thought through a northern route (18, 19). Conversely, the southern route implies crossing the Bab al Mandab Strait between the current countries of Eritrea and Yemen, and a rapid dispersal to South-East Asia and Oceania. This path involves crossing of a body of water, although during glacial periods it was less wide than the current state and easier to cross (14, 15, 20). Nevertheless, this coastal route is supported by some genetic evidence, although it lacks of clear archaeological evidence (21-23).

The timing of the OoA expansion and the involvement of single or multiple waves of dispersal are also topics with little consensus. Currently, there are two main hypotheses. In the first scenario, modern humans left Africa in a single major dispersal around 60-50 ka, and earlier dispersals in the Levant (Middle East) represented by individuals such as Skhul, Qafzeh and Misliya Cave did not contribute to present-day populations. The second hypothesis includes multiple waves of dispersal starting around 100-130 ka. The discovery of human remains and artefacts across Asia and Australia dated before 60 ka raises the possibility that the earlier dispersal into the Middle East (Skhul, Qafzeh, Misliya Cave) extended to East and South-East Asia (24-

28). Genetic studies have generally favoured a single dispersal (23), but some have reported evidence for a small contribution (<2%) from an earlier migration (29).

Modern humans during their dispersals into different parts of the world subsequently encountered and intermixed with at least other two human species, the Neanderthals and the Denisovans. The Denisovans were a population closely related to Neanderthals that split from the latter after the separation of both from the modern human lineage around 400–700 ka (4, 30, 31). Neanderthals contributed genetic material to all non-African individuals, with slightly more in East Asians than West Asians/Europeans. One hypothesis proposes two events of introgression: a first event in the Levant close in time to the modern human out-of-Africa migration around 50,000–60,000 years ago (32), followed by a second smaller introgression that only contributed genetically to Asians (33). An alternative hypothesis suggests dilution of the Neanderthal contribution in Europeans by mixing with a population lacking Neanderthal introgression. Denisovans contributed genetic material to more geographically-specific groups of modern humans including populations from South-East Asia (including Himalayans), and dwellers of Papua New Guinea and Aboriginal Australians (34, 35). A recent study outlined the possibility of two pulses of Denisovan introgression, one including both South-East Asian and Oceanian individuals, and a second one only into the former group (36).

1.2 Genetic variation and positive selection in human populations

1.2.1 Human genetic variation

The eukaryotic cell is composed of cytoplasm, and a nucleus enclosed within membranes that contains the nuclear DNA. The cytoplasm has organelles, such as the mitochondria, that are essential for the functioning of the cell. These organelles are responsible for cellular metabolism and energy production, and contain a circular molecule of DNA that is maternally inherited and independent from the DNA present in the nucleus. In each cell, there are many mitochondria. The human nuclear genome is organised into 23 chromosome pairs, of which 22 are autosomes and the last pair are sex chromosomes. Females carry two copies of the X chromosome, whereas males have one copy of the X and one of the Y chromosome. The human genome contains around 20,000 genes that code for proteins and together make up 1–2% of the genome. In the remaining ~99% of the DNA, many genomic regions play important roles in regulating gene function (37-39).

Genetic variation is the term used to describe differences in the DNA sequence between individuals. These genetic differences are mostly neutral but can result in differences in phenotype. Genetic variation can arise in both the germline (sperm and egg cells) and in somatic tissues, but only the variation occurring in the germline can be transmitted from one generation to the next and thus affect population dynamics. Two major sources of genetic variation are recombination and mutation.

Recombination is the process by which the genetic material inherited from both parents mixes when homologous chromosomes align and exchange genetic material, crossing over, during prophase I of meiosis. The crossing over shuffles paternal and maternal DNA in the offspring. Recombination between the X and Y chromosomes, except as described below, disrupts sex determination resulting in males with the absence of essential genes on the Y chromosome and females with possibly damaging genes. Hence, 95% of the human Y chromosome does not recombine and just two pseudoautosomal regions at the extremities of these chromosomes recombine. The rest of the Y chromosome is passed patrilineally from one generation to the other intact except for mutations (40). Likewise, mtDNA does

not recombine and passes matrilineally from one generation to the other intact, unless a mutation occurs (41).

Mutations are changes to the DNA sequence and occur at different scales from entire chromosomal aberrations, such as translocation of chromosomal segments or changes in chromosome numbers, to single nucleotides. They can involve the insertion, deletion or duplication of up to millions of base-pairs (Mb) (structural variant, SV, or copy-number variant, CNV), or the insertion or deletions of a few bases (INDEL). Insertion of transposable elements and change in number of tandem repeats of DNA motifs are also mutational events. The smallest mutations are single base substitutions resulting in single nucleotide polymorphisms, SNPs. Most new mutations are neutral, and do not affect the fitness of an organism. However, some of them can be beneficial or deleterious (42). CNVs are widespread in human populations and, although most do not have any phenotypic effect, others can be advantageous or deleterious (43, 44). For example, the duplication of the salivary amylase gene has undergone positive selection in populations with high-starch diets, whereas the alpha-thalassemia deletion in the alpha-globin gene cluster is considered protective against malaria (45, 46). On the other hand, CNVs have been associated with the increase risk of developing diseases such as cancer and a series of Mendelian diseases and complex traits (43, 47). Likewise, small INDELs can be deleterious especially when they occur within an open reading frame (ORF) because they can lead to a different protein sequence or to shifting the reading frame and premature termination. INDELs have been associated with diseases such as cystic fibrosis, acute myeloid leukaemia and other types of cancer (48). On the other hand, INDELs can bring beneficial effects such as the insertion-deletion variant (c.34delCinsTCCT) in *HTRA1* that has been reported as protective against age-related macular degeneration in the Chinese population (49). Another example is the 4 bp deletion located in the second intron of *EPAS1* that has been associated with high altitude adaptation in Tibetans (50).

SNPs are the most abundant and most studied type of genetic variation within human populations. One of the main causes for base substitution is incorrect base incorporation during DNA replication. New (*de novo*) mutations arise when incorrect base incorporation during DNA replication is not corrected by DNA repair enzymes and the new base gets fixed in the DNA. However, replication is a high-fidelity process

and errors occur at very low frequency (10^{-9}-10^{-11} per nucleotide per replication). Nevertheless, the substitution rate is not uniform across the genome. The CpG islands are genomic regions rich in dinucleotides where a cytosine is followed by a guanine in the 5' → 3' direction, and are mutational hotspots. Most CpG dinucleotides (~75%) are targets of DNA methylation with the formation of 5-methylcytosine by the addition of a methyl (-CH_3) group to the 5-carbon of the cytosine ring by a methyltransferase enzyme. On the other strand, the complementary bases form a CpG site as well, thus methylation also affects the C base on this strand. When 5-methycytosine is deaminated, it yields a uracil and subsequent mispairing with the guanosine on the other strand. These can lead to two scenarios: the U-G mispair can be fixed by replacing G with an A producing (after DNA replication) a T-A pair, or by changing the U to a C re-forming a C-G pair. (42, 51). Similarly to CNVs and INDELs, SNPs in different parts of the genome can have different functional impacts. Substitutions that do not alter any amino acid are called synonymous, and are frequently assumed to be selectively neutral. Mutations that lead to an amino acid change are called nonsynonymous or missense substitutions and can have different effects on function according to their position in the protein (42). Loss-of-function (LoF) mutations are predicted to reduce or completely abolished protein coding-genes function. On the other hand, gain-of-function (GoF) mutations can confer an abnormal activity on a protein (52, 53).

Overall, the term 'variant' is used to refer to any genetic difference between two DNA sequences, and different versions of the same variant are called 'alleles'. Conventionally, the reference allele denotes the base that is found in the reference genome assembly (https://www.ncbi.nlm.nih.gov/grc). Conversely, the alternative allele indicates any variant that differs from the reference. It is important to note that the reference allele is not necessarily the more frequent allele (major allele) or the alternative the less frequent (minor allele) in a population. Moreover, the reference allele does not necessarily represent the ancestral state of an allele. The ancestral allele is defined as the ancestral state of a variant, and the derived allele has arisen subsequently by mutation. In practice, the ancestral allele for humans is inferred from the alignment of different primates from Ensembl compara and thus takes into account the alleles of several different apes (54).

Alleles that are physically close on the same chromosome tend to be inherited together more often than expected by chance. These blocks of alleles form relatively stable haplotypes. LD is a measure of how often two alleles are inherited together, and the coefficient (*D*) of LD is defined as the difference between the observed frequency of a pair of alleles at two loci (p_{AB}) and the product of the frequencies of those two alleles expected for random association $(p_A p_B)$(55):

$$D_{AB} = p_{AB} - p_A p_B$$

Each pair of alleles is characterised by their own coefficient of disequilibrium. The extent of LD across the genome is variable within different human populations and LD maps are population specific (56, 57). African individuals generally show lower levels of LD compared to populations outside Africa, probably due to the larger size of human populations in Africa over time and greater time depth, compared with the relatively small number of modern humans that exited Africa during their expansion into the rest of the world (58).

In population genetics, the variation in allele frequencies within populations is a valuable metric to understand population dynamics over time. According to the Hardy-Weinberg equilibrium theory, in the absence of any evolutionary force, the allele and genotype frequencies in an infinitely large and panmictic population will remain constant through time (59, 60). Factors that can change the allele frequency in a population and violate the Hardy-Weinberg equilibrium include mutations, genetic drift, non-random mating, gene flow and natural selection, for example. Mutations introduce new alleles into a population and, therefore, change the allele frequency spectrum. The frequency of the original allele will decrease because mutation changes it into a different allele (42). Genetic drift is the change in the frequency of an existing allele due to stochastic sampling of individuals in creating the next generation. The effect of genetic drift over time relates to the population size, with less genetic drift in large populations and vice versa (42, 61): in a small population, an allele can rapidly get either lost (allele frequency =0) or fixed (allele frequency =1), whereas in a big population alleles tend to vary less and stay near their starting frequencies. To summarise the effect of genetic drift in a population, the concept of effective population size (Ne) is often used. Ne differs from the census size (N) of a population and in genetics represents the number of individuals on an

idealised population that undergoes the same amount of genetic drift as the respective natural population. The effective population size is usually smaller than the census size of a population (42, 62-64). Ne is most commonly measured with respect to coalescence time. In an idealised diploid population with no selection, the pairwise nucleotide diversity is equal to 4μNe, where μ is the mutation rate. Hence, the effective population size can be calculated by dividing the nucleotide diversity by the mutation rate (65). Similarly, gene flow is another cause of departure from Hardy-Weinberg equilibrium. It consists of the movement of alleles from one population to another through interbreeding. This may be coupled with physical migration of individuals from one population to the other. The higher the gene flow, the lower the genetic difference between two populations. Inversely, isolation can increase the genetic differentiation between two populations (42). Lastly, natural selection is another mechanism leading to changes in allele frequencies over time. Natural selection occurs when individuals with a certain genotype have an advantage over other individuals with a different genotype in their ability to pass their alleles to their offspring. Positive selection increases the frequency of beneficial alleles in a population and the probability that these variants get fixed within the population, while negative selection decreases the frequency of deleterious alleles (42). However, both will skew local variation spectrum towards an excess of rare alleles.

Advances in SNP genotyping and high throughput sequencing technologies over the last ten years have allowed the genotyping and sequencing of large numbers of samples in a cost-effective way, leading to the availability of large modern human population datasets. Datasets such as the 1000 Genomes Project (66), the Human Genome Diversity Project (HGDP) (67), the Simons Diversity Project (SGDP) (23) and the Estonian Biocentre Human Genome Diversity Panel (EGDP) (29) contain individuals from worldwide populations. They have allowed the description of worldwide patterns of human genetic variation and are a valuable resource for comparison when new populations are sequenced. In the last decade, the advent of new protocols for the extraction of ancient DNA (42, 68, 69) have resulted in an increasing number of ancient human DNA sequences and a better understanding of worldwide prehistorical and historical migrations, and how they have influenced the genetic variation we see today (70-72). Different approaches have been developed to

summarise genetic variation, in particular to measure genetic distances and genetic relationships between individuals. Widely used methods to estimate and display genetic distances are principal component analysis (PCA), STRUCTURE-like plots (particularly ADMIXTURE) (73, 74) and the Fixation index (F_{ST}). PCA is an unsupervised statistical approach used to reduce dimensionality in large datasets while minimising the information loss. In genetics, it clusters individuals based on their ancestry creating new uncorrelated variables that sequentially maximize variance. It is widely used in population genetics to identify population structure, individual outliers and potential genotyping errors (75, 76). STRUCTURE and ADMIXTURE are model-based clustering approaches also used to summarise population structure. However, in contrast to PCA, they require the specification of the number of ancestral components to use for assigning the ancestry of the individuals in the dataset. F_{ST} is an allele frequency based measure of genetic differentiation between populations and, thus, informative about genetic distance between them. Phylogenetic tree and network methods are used to compare genetic relationships between individuals and populations. They can be based on allele frequency differences such as TreeMix (77) or genetic linkage information (haplotypes) such as ChromoPainter and fineSTRUCTURE (78). The latter methods use phased haplotypes to assess genetic similarity between individuals. Moreover, methods such as PCAdmix (79), RFmix (80) and HAPMIX (81) use haplotype information to infer local ancestry in admixed populations using reference panels related to the admixture sources. More recently, new approaches to study genetic affinity within modern human populations, and between them and ancient individuals, such as *f3* and *f4* statistics and *D* statistics, have been commonly used (82). These methods assay genetic affinity by computing tests of allele frequency correlations within individuals. They normally test if the allele frequencies in the populations analysed are consistent with a simple tree topology or if admixture events happened between them. Finally, coalescent methods such as PSMC or MSMC are used to reconstruct population demographic histories (42, 83, 84) and, because they require information from both variant and non-variant sites, can be only performed on whole-genome sequencing data. These methods infer the history of effective population size over time using the distribution of coalescence times within a genome. When multiple genomes from different populations are used, it is possible

to infer the divergence time between two populations comparing the relative rates of within and between-population coalescences.

1.2.2 Natural selection

Positive selection, as described in the previous section, is the mechanism through which beneficial alleles rise in frequency in populations over time. A trait undergoes positive selection if it is advantageous for the individual to survive and reproduce, and if it is heritable so that it can pass on to the next generation. The availability of whole-genome data for many different modern human populations has allowed the study of genome-wide patterns of positive selection in humans adapted to live in different environments around the world (42, 85). More recently, the increased number of genome sequences from archaic and ancient humans and other species has permitted a better understanding of the time depth of many adaptive signals in modern populations. At the molecular level, when alleles under positive selection become more frequent in a population, they leave specific signatures of genetic variation in the DNA sequences of the population. These patterns can be identified by comparison with the overall background distribution of human genetic variation that, according to the neutral theory, has evolved mainly under neutrality. Under this assumption, most of the population changes in genetic variation over time are due to genetic drift (86, 87). One of the biggest challenges is being able to tell if a change in genetic variation is due to selection or to demographic history confounder effects, such as substructure, bottlenecks (period of reduced sample size) or expansion.

The most studied selective sweeps, hard or classic sweeps, happen when a new advantageous allele rapidly increases in frequency and carries with it the haplotype on which it occurred, reducing levels of variation in the neighbouring area (88). This effect is reduced with the increase of genetic distance from the allele under selection due to recombination. A soft sweep occurs when a neutral variant segregating in the population becomes beneficial, due to an environmental change for example. This is also called a sweep from standing variation (variants that are polymorphic in a population), and the advantageous allele already present in the population rises in frequency. A second case is a soft sweep from recurrent mutation that occurs when the sweep includes multiple distinct beneficial alleles appearing at the same locus increasing their frequencies. A partial sweep arises when a beneficial allele raises from low frequency to intermediate, but is still polymorphic (89-91). Hard sweeps are

relatively straightforward to detect and have been extensively studied. On the other hand, soft sweeps are more complicated to detect and distinguish from neutrality (42). Polygenic adaptation represents an alternative mode where the adaptive change involves small frequency changes in possibly hundreds or thousands of alleles in different genes (91, 92). Many human traits and diseases are highly polygenic, such as height, skin pigmentation, body mass index and type 2 diabetes (93). Balancing selection is another kind of selective process where multiple alleles are actively maintained in the gene pool of a population by selection favouring heterozygotes or the rare allele (whichever it is). The most famous example of heterozygote advantage is the sickle-cell anaemia phenotype at the β-haemoglobin locus. Individuals heterozygous for the sickle-cell (S) allele show resistance to malaria (94). Finally, background selection is a process by which weakly deleterious mutations are purged from the population. This can cause reduction in genetic diversity, particularly around functionally-important regions (91).

Over the decades, many methods for detecting positive selection and selective sweeps have been proposed. Most of them use information about ancestral and derived alleles, instead of reference and alternative. These methods can be classified into three main categories 1) SFS-based; 2) LD-based and 3) Population allele frequency based. The site frequency spectrum (SFS) is the distribution of the allele frequencies at many loci, such as SNPs or INDELs, in a sample or population (95, 96). SFS-based methods rely on the departure of the allele frequency spectrum in the region of a selective sweep from the neutral expectation. In particular, an increase of high- and low-frequency derived variants is expected in the vicinity of the beneficial allele under a classic sweep model. Examples of these methods are Tajima's *D* and Fay and Wu's *H* (42, 97). LD-based approaches rely on the extent of linkage disequilibrium in the selected region. Selection on one allele can rapidly increase the frequency of the surrounding haplotype and, thus, LD can remain high at each side of the advantageous allele (42, 97). These methods are generally called extended haplotype tests, and examples are the extended haplotype homozygosity (EHH) test, the cross-population extended haplotype homozygosity (XP-EHH) test and the integrated haplotype score (iHS) (42). The last category of methods relies on measures of inter-population differentiation. They assume that selection has acted on one population, but not on the other. If two populations have an extreme difference in the allele frequency of a

variant compared with the allele frequencies of other variants, it is possible that selection is the cause of such a difference. Methods that measure differences in derived allele frequencies between populations, such as ΔDAF, single locus fixation index (F_{ST}) and population branch statistics (PBS), are widely used to detect signals of positive selection in human populations (42, 98, 99), and are applicable to both classic and soft sweeps. DeltaDAF (ΔDAF) is the measure of the allele frequency differences between a pair of populations (42). PBS is the measure of distance between three populations using F_{ST} and it estimates if there is an allele with extreme frequency difference compared to two other populations (99). Exploiting all these approaches, strong candidates of positive selection have been found in modern humans including SNPs in the Ectodysplasin A receptor (*EDAR*) gene, associated with hair, tooth and skin development in Asians, the solute carrier family 24 member 5 (*SLC24A5*) and the solute carrier family 45 member 2 (*SLC45A2*), both involved in light skin pigmentation in Europeans (100). Variants in proximity to the lactase gene (*LCT*) have been linked to the ability to digest lactose as an adult in Europe and Western Asia. Moreover, several candidates for selection seem to be involved in malaria resistance, including the Duffy antigen protein (*DARC*) and Glucose-6-phosphate dehydrogenase (*G6PD*), and mutations in the α- and β-globin genes that can lead to thalassemias or sickle cell anemia (42, 46, 101, 102). Finally, another signal of adaptation to extreme environments is the high altitude adaptation in Ethiopians, Andeans and Himalayans (103). In particular, Himalayans show permanent adaptation to high altitude, and variants lying in the Endothelial PAS Domain Protein 1 (*EPAS1*) gene show the strongest signature of selection (99). Interestingly, it has been discovered that this gene has a highly unusual haplotype structure in Tibetans compared to other modern human populations, probably due to introgression of DNA from Denisovans into the ancestors of Tibetans (104).

1.3 The Himalayan region: an overview

The Greater Himalayan Region is a geographical area containing the world's highest mountain peaks and a diversity of environments that have required substantial genetic adaptations by the humans who live there. This mountain barrier has also shaped the genetic, cultural and ethnolinguistic mosaic of South and East Asia. At present, the area falls into the countries of Nepal, Bhutan, India, Pakistan and the Tibetan Plateau in China. Opinions are divided about whether the Himalayas were used as a corridor that facilitated human migrations from the Tibetan plateau to South Asia in ancient times, or alternatively remained uninhabited due to their inhospitality until more recent times (105-108). Archaeological data suggest that the central Tibetan Plateau was populated during the Neolithic period (109), and there is evidence of earlier human occupation in the north-eastern Qinghai region (110).

The Himalayan region is also one of the most complex linguistic areas in the world, containing two main language families, Indo-European and Sino-Tibetan, with Sino-Tibetan divided into the Tibeto-Burman (TB) and Chinese subfamilies. Tibeto-Burman speakers represent the majority of the Himalayan populations whereas the Indo-European speakers are restricted to the south of the mountain range. In all, the region contains six linguistic phyla with multiple languages within each phylum, and at least two language isolates (Burushaski and Kusunda) (111, 112). However, the region has not been fully represented in earlier genetic studies. Previous analyses have mainly focused on populations residing to the north or south of this area, or on small numbers of populations (105, 113-116). In fact, most of the studies have focused on high altitude Tibetans and Sherpa, and a few other Nepalese populations, leaving unpresented most the genetic variation in the region (50, 99, 116, 117). Key aims of these initial studies were to understand the population separation between Han Chinese and Tibetans, and to understand the genetic relationship between Tibetans and Sherpa. There is little agreement on the split time between Han Chinese and Tibetans, with estimates ranging from around 2,700 years ago, via an intermediate time around 7,000-9,000 years ago to a deep split time around 15,000-16,000 years ago (50, 118, 119). Another topic with little consensus is the divergence time between

Tibetans and the Nepalese Sherpa, and indeed whether or not they are genetically distinct groups.

Some studies support the hypothesis of a very recent split (~ 1,500 years ago) and propose that Sherpa are a recently derived Tibetan sub-lineage. Other studies report a deeper split time between 3,200–11,300 years ago, and that Tibetans and Sherpa are genetically separate groups. Another open question is the origin of the different Himalayan populations and if they share a common genetic component. Studies on modern and, more recently, ancient individuals from the Himalayan region have shown that Himalayan populations have a possible East Asian origin with subsequent gene flow from South Asia (105, 120-122). However, the genetic relationships between highlanders and lowlanders in the Himalayan region, and the genetic background and the amount of gene flow in these populations, needs further investigation.

The first systematic genetic survey of Himalayan populations, which used autosomal microsatellite markers (STRs) (111), showed that there is higher genetic diversification among the Himalayans compared with the populations from the surrounding regions, and observed genetic differentiation between Indo-European and Tibeto-Burman speakers, suggesting that both language and geography have influenced the genetic structure of these populations.

Himalayan populations reside in a broad range of altitudes and environments. At one extreme, the Terai region at the foothills of Himalayas is a tropical forest with endemic mosquito-borne diseases, in particular malaria (123), dengue fever (124) and parasitic infections (125). In the early 1990s it was reported that a population of this region, the Tharu, carried an alpha-thalassaemia deletion that was almost fixed and probably adaptive against malaria, but this possibility has not been extensively explored from the genetic point of view (46). In contrast, other Himalayan populations reside at high altitudes, and have developed extraordinary adaptations to this environment. Genomic scans in the Himalayas have previously identified genomic regions associated with high altitude adaptation. In particular, a Denisovan-derived *EPAS1* haplotype, whose frequency is strongly correlated with altitude in the Himalayan populations (99, 104, 126, 127) has been identified, but the functional

variants within the locus and the molecular mechanisms of high altitude adaptation in Himalayans remain largely unknown.

1.4 High altitude adaptation and the hypoxia molecular pathway

1.4.1 Adaptation at high altitude

High altitude adaptation is an evolutionary mechanism that allows a species to survive permanently at high altitudes (>2500 meters above sea level). This adaptation involves long-term physiological responses to hypoxia that are often associated with heritable genetic modifications. Hypobaric hypoxia is caused by the decrease of barometric pressure with progressively increasing altitude, and therefore with fewer oxygen molecules in the air compared to sea level. This means that each human breath contains less oxygen than at low altitude. Hypobaric hypoxia is a source of severe stress as a constant supply of oxygen is required for mitochondrial metabolism (the mitochondrial respiratory chain). The oxygen transport system in the mitochondria probably evolved under normoxic conditions at sea level (103).

Modern human populations residing in the Himalayas in Asia, in the Andes in South America, and in Ethiopia in Africa, have all evolved an ability to live in hypoxic environments. These populations have adapted so that they maintain appropriate oxygen levels in the tissues that permit survival, growth, reproduction and development. Individuals living at low altitudes can suffer from acute mountain sickness (AMS), a serious condition arising from exposure to extreme hypoxia, that can progress to high altitude pulmonary edema (HAPE) or high altitude cerebral edema (HACE) (128, 129). Indigenous high altitude populations do not suffer from AMS because they have undergone drastic physiological and genetic changes. From the physiological point of view, the main changes are at the levels of haemoglobin concentration and oxygen saturation (the percentage of haemoglobin binding sites occupied by oxygen) in the bloodstream that together control the arterial oxygen content. Andean dwellers show a higher haemoglobin concentration compared to lowlanders, mirroring in some ways the response of lowland individuals moving to high altitudes (acclimatisation). When Andeans move to low altitudes, their haemoglobin concentration decreases to normal levels. Nevertheless, Andeans have an increased arterial oxygen content compared to lowlanders that provides more efficient transportation of oxygen throughout the body and lower oxygen saturation

(103, 130). Similarly to Andeans, the Ethiopian highlanders have elevated haemoglobin levels and increased arterial oxygen. Although few studies have been conducted on Ethiopian highlanders, these individuals are adapted to live at high altitude and do not show any signs of illness (131-133). Finally, Tibetans do not show increased haemoglobin concentration. Instead, they show lowered oxygen saturation and arterial oxygen content. They also exhibit enlarged lung volumes, higher training capacity and better oxygenation at birth (134, 135).

Together with these physiological adjustments, high altitude populations also show genetic signatures of convergent evolution to living at high altitudes (Figure 1.1). Numerous genome-wide scans for positive selection have been performed in these populations and several genes highlighted. In Andeans, the hypoxic inducible factor (HIF) genes *EGLN1*, *PRKAA1*, *NOS2*, *TGFA*, *CXCR4* and *VEGFB*, as well as the non-HIF genes *ELTD1* and *PRKG1* have implicated the nitric oxide pathway as a target of positive selection (133, 136). Genomic scans of Ethiopian highlanders have reported genes shared with Andeans and Himalayans, plus novel candidates. Examples of these are *BHLHE41*, *ARNT2*, *THRB*, *CXCL17*, *PAFAH1B3* and a gene-rich region on chromosome 19 containing genes implicated in vascular physiology (137). Additional genes have been associated with angiogenesis (*VAV3*), and calcium uptake (*CBARA1*) (137). Himalayans are the most studied populations for high altitude adaptation. The first genome-wide positive selection scans identified several genes in the HIF pathway, including *EGLN1* and *EPAS1*. In particular, their *EPAS1* haplotype has been suggested to be the result of introgression of DNA from an extinct hominin species (Denisovans), as mentioned in the previous section (104). The widespread distribution of this haplotype was confirmed by genotyping the *EPAS1* locus in the large dataset of Himalayan populations used in this thesis. High altitude populations show the highest proportion of these variants and they may have spread through the region by gene flow (127). More recently, additional candidates of selection have been proposed in the Himalayans including *HMOX2* (a modifier of haemoglobin metabolism), *PTGIS*, *VDR* and *KCTD12* (50, 138).

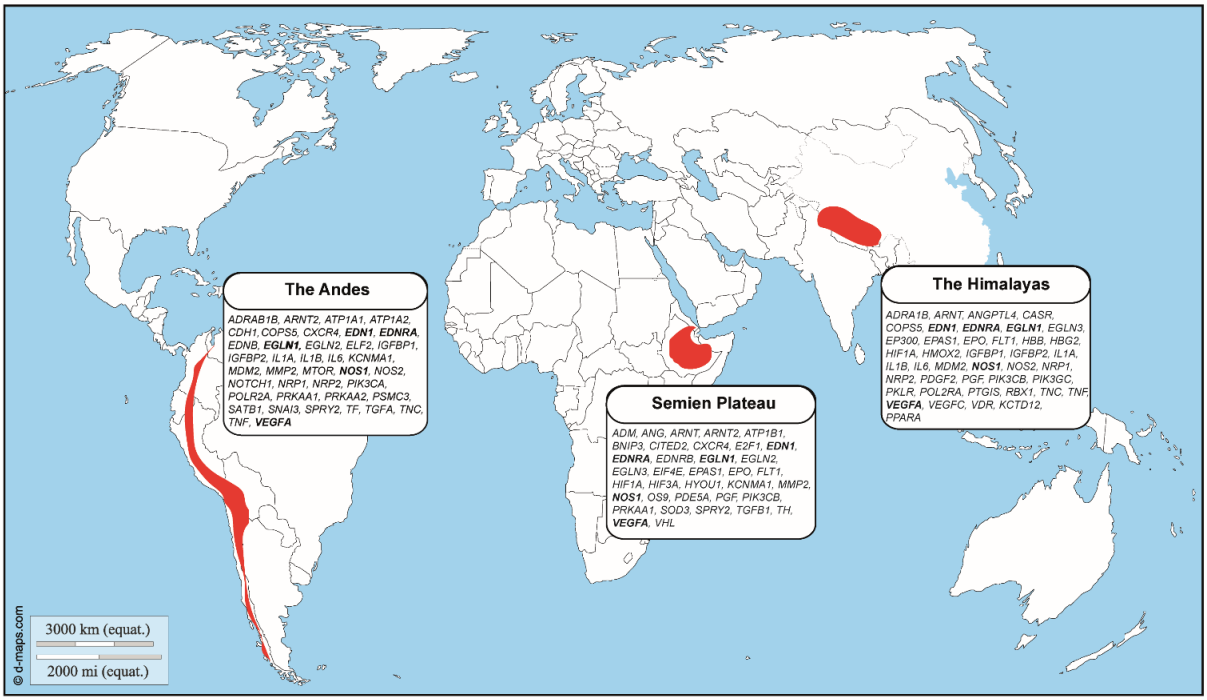


Figure 1. 1 Candidate genes for high altitude adaptation in Andeans, Ethiopians and Tibetans. The three geographic regions where these populations reside are highlighted in red. A list of selection candidates for each population is contained in the boxes (50, 133, 136, 137). The shared candidates are shown in bold.

Many candidate genes have thus been proposed and doubtless many more will be found associated with high altitude adaptation. From previous studies and the new results presented in this thesis, it seems that no single variant or gene is responsible for high altitude adaptation. To understand the roles of these genes, it is now important to move in the direction of linking genotype to phenotype by identifying the causal variants responsible for high altitude adaptation, as has been done for the functional validation of *EGLN1* variants in Tibetans (126).

1.4.2 Hypoxia molecular pathway

Hypoxia inducible factor (HIF) is the master transcriptional regulator that detects and coordinates cellular responses to hypoxia in most mammalian cells. It is a heterodimer that consists of one of three α subunits (HIF-1α, HIF-2α (*EPAS1*), or HIF-3α) and a common β subunit (HIF-β, also known as ARNT). The HIF family belong to the basic–helix–loop–helix-PAS (bHLH-PAS) superfamily and both α and β subunits contain a Per Arnt Sim (PAS) and N-terminal helix–loop–helix (HLH) domain. Furthermore, the α subunits contain a transcriptional activation and a C-terminal oxygen-dependent degradation domain. The HIF-*α* subunits are specifically induced by hypoxia whereas the HIF-*β* subunit is constitutively expressed. HIF activity is tightly regulated by oxygen concentration through site-specific prolyl hydroxylation of the different α subunits. Under normoxia, one of the members of the prolyl hydroxylase protein family, PHD1 (*EGLN2*), PHD2 (*EGLN1*), or PHD3 (*EGLN3*) hydroxylates the HIF-α subunit in its oxygen-dependent degradation domain. The PHDs are members of the 2-oxoglutarate-dependent dioxygenase family. The primary site of hydroxylation in HIF-1α is on Pro564, Pro531 in HIF-2α and Pro490 in HIF-3α. The hydroxylated α subunit then becomes a target for the von Hippel-Lindau (VHL) protein, a component of an E3 ubiquitin ligase complex that triggers the degradation of HIF-α by the ubiquitin–proteasome pathway (Figure 1.2). During hypoxia, the oxygen-dependent hydroxylation of HIF-α is prevented, leading to the stabilization of HIF-α and its dimerization with HIF-β (132, 139). It is important to note that the activity of PDH enzymes is also regulated by the concentration of reactive oxygen species (ROS), Krebs cycle intermediates, and ascorbate and iron concentrations (140). Moreover, HIF activity is mostly modulated by hydroxylation, but other post-translational modifications, such as phosphorylation and acetylation, can also regulate its activity. An additional regulation level of HIF-α is provided by the factor inhibitor HIF (FIH), another member of the 2-oxoglutarate-dependent dioxygenases family. FIH targets an asparagine residue in the HIF activation domain and hydroxylates it, specifically Asn803 in HIF-1α, and Asn851 in HIF-2α. In normoxic conditions, this hydroxylation blocks the interaction with the transcriptional coactivator CBP/p300. In contrast, under hypoxic condition, the reduction of this modification allows the functional interaction of HIF-α with CBP/p300 and thus

activates a transcriptional cascade. Therefore, hydroxylation plays a key role in regulating HIF activity by both conferring protein stability and transcriptional activation. Once HIF-α is stabilized, it forms a heterodimer with HIF-β via their HLH and PAS domains and then regulates the expression of hundreds of genes involved in cellular and systemic responses to hypoxia, by binding to specific sequence motifs on the target genes termed the hypoxia responsive elements (HRE) (132, 139, 141).

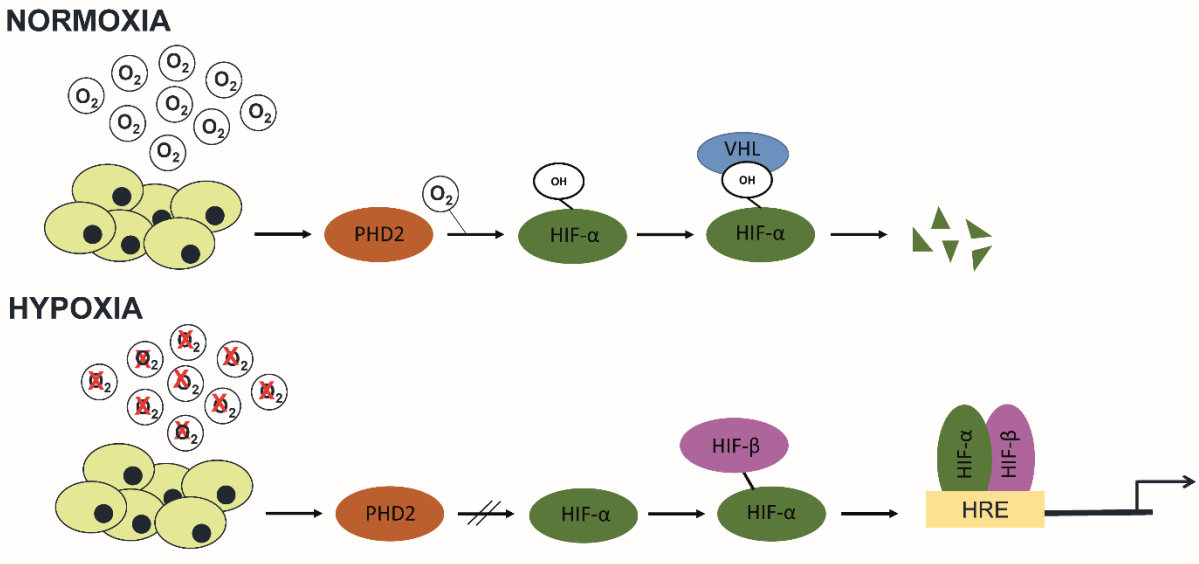


Figure 1. 2 The molecular hypoxic pathway. In normoxia, PHD2 prolyl hydroxylates HIF-α subunits, targeting them for degradation via VHL-dependent ubiquitination. In hypoxia, the hydroxylation is blocked, allowing HIF-α stabilization, dimerization with HIF-β, and binding to target genes through the hypoxia response element (HRE).

HIF-1α and HIF-2α can regulate both common and distinct genes and cellular responses. Both can regulate the activity of the Vascular Endothelial Growth Factor A (*VEGFA*) that codes for a protein essential for angiogenesis. In genetically engineered mice with deletion of one or two copies of the gene, it has also been shown that haploinsufficiency of either HIF-1α or HIF-2α can prevent or delay hypoxia-induced pulmonary hypertension (132, 142). HIF-1α is the key regulator of glycolytic enzyme genes, mitochondrial biogenesis and oxidative phosphorylation. It is also important in the regulation of hematopoietic stem cell activity. In contrast, HIF-2α tightly regulates the erythropoietin (*EPO*) gene in kidney cells. *EPO* is responsible for erythropoiesis (red blood cell production). Evidence in both mice and humans

showed that HIF-2α LoF leads to anaemia, whereas GoF leads to erythrocytosis (increased number of circulating red blood cells). Furthermore, it has been shown that HIF-1α and HIF-2α can also display antagonistic effects: the former triggers cell cycle arrest whereas the latter stimulates its progression. In murine endothelial cells, HIF-1α induces nitric oxide (NO) production through the activation of the NO synthase gene (*NOS2*), whereas HIF-2α inhibits NO production, inducing the activity of the arginase gene (*ARG*) (132). Little is known about HIF3-α, but it seems to act both as an activator and repressor of transcription (132).

Like the HIF factors, the PHDs can have common and individual functions. PHD2 is the most studied and, possibly, the most important. Mouse knockout of *PHD2* leads to embryonic lethality, whereas *PHD1* and *PHD3* knockouts are viable and fertile (143). Moreover, PHD2 plays an important role in the regulation of erythropoiesis through the regulation of HIF-2α and the transcription of *EPO* (126, 132). Studies of PHD1 and PHD3 have been limited compared to PHD2, but it seems that *PHD3* is a HIF target and down-regulates HIF activity through a negative feedback loop (144).

Himalayan populations consistently show strong signals of positive selection at the *EGLN1* (PHD2) and *EPAS1* (HIF-2α) genes. A key question is how the genetic changes in the two antagonists lead to the final phenotype of adaptation. Many different scenarios can be proposed from the different regulation of these two factors in highlanders versus lowlanders. In lowlanders, levels of HIF-1α and HIF-2α are kept low through hydroxylation and degradation. At high altitude, PHD2 activity is reduced and the activity of HIF-1α and HIF-2α triggered. In highlanders, it is possible that the PHD2 allele is either a LoF with attenuated hydroxylase activity, or a GoF with enhanced activity. Likewise, HIF-2α can show increased activity (GoF) or reduced activity (LoF). This indicate the possibility of four different models of adaptation in Himalayans (132):

1. PHD2 activity is reduced (LoF) and, thus, HIF-2α activity in enhanced (GoF). However, this model seems unlikely as it would predict high haemoglobin levels, which are not observed in Tibetans and Sherpa (145).
2. GoF for both PHD2 and HIF-2α: This combination would be possible if the PHD2 allele was stronger than the HIF-2α one.

3. GoF for PHD2 and LoF for HIF-2α: This would lead to reduced activity of HIF-α factors. In particular, it would reduce EPO with a consequently diminished response to hypoxia.
4. LoF for both PHD2 and HIF-2α: The LoF of PHD2 would imply that the HIF-2α allele is a very strong LoF allele that could lead to the inhibition of HIF-2α and simultaneous activation of HIF-1α.

Finally, oxygen deprivation leads to the production of reactive oxygen species (ROS) in the mitochondria. ROS are a variety of free radicals, chemical species with one unpaired electron, and molecules derived from molecular oxygen. Most of them derive from their precursor, the superoxide. The latter are formed when the redox centres in the mitochondrial electron transport chain leak electrons to oxygen (139). ROS lead to the stabilisation of HIF-α isoforms during hypoxia, and activation of the downstream pathway. There is increasing evidence that mitochondria are one of the most important oxygen sensors, and hence they have a central role in the hypoxic response. Under hypoxic conditions, mutations in the mitochondrial respiratory chain or complex 1 inhibitors block the stabilisation of HIF-α (139). Hence, the hypoxic response involves tight regulation steps both at the cytoplasmic and nuclear levels of the cell. Functional studies in Himalayan individuals are necessary to understand which putative model of adaptation they went through and which genes are involved in this mechanism.

1.5 Aims and objectives

The aim of this thesis was to provide a fine-scale description of the human genetic variation in the Himalayas. In particular, my goal was to describe populations that had not previously been genetically characterised, exploring their population structure and demographic history. My intent was to understand when Himalayans communities separated from each other, and from the neighbouring populations in South and East Asia, and the genetic relationships between them. Furthermore, my objective was to understand how geographical isolation, genetic drift and selection have affected the genetic landscape of Himalayan populations. Similarly, I wanted to explore if there was correlation between genes and languages in this region, and if both of them participated in shaping the variability we see today in the Himalayas. Finally, because Himalayan populations live in a broad range of environments and altitudes, my intention was to look for signals of positive selection at both low and high altitudes. Moreover, due to the great number of high altitude populations in the dataset, I wanted to explore the genetic relationships between them and look for known and possibly new candidates of high altitude adaptation. Finally, because of the unknown molecular mechanism of action, I wished to explore candidate functional variants in *EPAS1* both *in silico* and *in vitro*. Using cell lines heterozygous for the Denisovan introgressed and non-introgressed haplotypes from the same genetic background, I wanted to understand if differential expression in *EPAS1* was detectable between them. My hypothesis was a scenario in which there was LoF for HIF-2α leading to reduced activity of the HIF-α factor, and subsequent suppression of *EPO* expression in introgressed cell lines. Hence, to address these scientific questions, I used a combination of SNP-genotype data, whole-genome sequences, functional work and expression data.

1.6 Outline of the thesis

This thesis presents the results of the largest genome-wide population-genetic analysis of humans residing in the Himalayan region thus far reported, and a preliminary biological and functional analysis of positive selection of the *EPAS1* gene in high altitude Himalayan populations. The first results chapter (Chapter 2) describes the characterisation of the populations, their structure, and genome-wide signatures of high altitude adaptation using SNP-chip genotype data. The second results chapter (Chapter 3) uses a subset of samples with high coverage whole-genome sequencing data to refine demographic history using autosomal and uniparentally inherited sequences (mtDNA and Y chromosomes) with both common and rare variants. The third results chapter (Chapter 4) describes the experimental exploration of *EPAS1* molecular function under laboratory-induced hypoxia, simulating high altitude, and the analyses of its expression profile using quantitative PCR and RNA-sequencing. The general introduction at the beginning is expanded in the individual chapters, giving more detailed information about chapter-specific topics, and general discussion points are covered in the final chapter (Chapter 5).

2. Population genetic analyses of Himalayan samples using SNP-genotype data

2.1 Introduction

The genetic characterisation of human populations requires a large number of samples and genetic markers. In the past, analyses of population history and the genetic relationships between populations were mostly performed using the uniparentally inherited mtDNA and Y chromosome and single autosomal locus markers, such as the human leukocyte antigen (HLA) (146). More recently, technological advances such as next-generation sequencing methods have allowed us to perform genome-wide analyses in a reasonably short time period and in a cost-effective way, improving drastically researchers' possibility to describe human populations and detect alleles associated with a specific disease. In particular, single nucleotide polymorphisms (SNP) genotyping arrays have become one of the most used methods for assessing human genetic variation. High-throughput SNP-genotyping is a technique that allows us to assay SNPs across the genome of an individual on a large scale. There is a plethora of commercially available SNP-chips designed for different purposes and for diverse species with the aim of testing hundreds of thousands of SNPs on a single chip. Human SNP-chips from Illumina and Affimetrix are used most frequently and, even though there are differences between the two platforms, they both can assess hundreds of thousands to millions of markers. Both rely on DNA base pair complementarity where A binds to T and C binds to G (147). Short nucleotide probes are designed to bind specific DNA target sequences and directly tag SNPs. Genomic DNA is amplified by PCR, fragmented, labelled with a fluorescent marker and bound to the chip. After the hybridisation of the DNA to the probes, all the non-specific and unbound DNA is washed away. The occurrence of the target allele in the DNA sample is assessed by evaluating the fluorescent signal at each probe. Genome-wide genotyping has the advantage of being an affordable and reliable platform for describing genetic variation across the entire genome of many individuals and for identifying genomic regions linked to a specific phenotype or

disease susceptibility. Although whole-genome sequencing provides a finer base-by-base resolution of the genome, it is still too expensive to perform for the big cohorts commonly used in genome-wide association studies (GWAS).

SNP-chips have been widely used in population genetics to characterise human populations, describing their genetic structure, and to identify loci undergoing positive selection (148). SNP-chip data produced by different studies are easy to combine together and many bioinformatic tools have been developed to handle such data. Nevertheless, genome-wide genotyping has some drawbacks. One of them is the ascertainment bias. This is the systematic bias of statistical results as a result of the choice of SNPs included in the chip. These have to be polymorphic in at least one population and, since European populations are the most studied, many of the commercially available SNP-chips are constructed using markers known to be polymorphic in Europeans. This bias can have important effects on the analysis results. Firstly, the allele frequency distribution can be skewed towards intermediate frequency alleles and the heterozygosity in Europeans can be over-estimated compared with non-Europeans individuals. In addition, the SNP distribution across the genome is not uniform with certain regions of the genome being enriched for SNPs whereas others are depleted. SNP ascertainment bias can affect linkage disequilibrium (LD) inferences such as the ability to tag specific genomic regions by SNPs, which relies on the strength of LD. Therefore, it is important to take into account the ascertainment bias when performing population genetics analyses and demographic inferences. In particular, it is critical to acknowledge this bias when performing demographic history analyses or positive selection scans that use allele frequency distributions and heterozygosity estimates (148-150). Nevertheless, despite these caveats, many of the computational methods developed to analyse genome-wide genotype data are relatively robust to ascertainment bias and the design of some of the SNP-chips captures variation across multiple worldwide populations. For example, the Affymetrix Human Origins array has been developed specifically for investigating human migrations, history and natural selection (82). It contains SNPs from eleven modern human populations from the Human Genome Diversity Cell Line Panel (HGDP-CEPH), Neanderthals, Denisovans and chimpanzees (82, 151). Markers were chosen to avoid confounding biases and to allow robust statistical analyses in population genetics. Similarly, the Illumina

HumanOmniExpress arrays have been designed selecting genomic variants from the HapMap project to represent the best amount of common SNP variation (MAF > 5%) in individuals with European, Asian and African ancestry (152). Overall, SNP ascertainment bias is an important concern when performing analyses, but it should not only be interpreted as problematic and SNP-chips are an invaluable and inexpensive platform for population genetics and GWAS.

Finally, genome–wide SNP genotyping has also been chosen as one of the approaches for capturing genetic variability from human ancient genomes. To enrich for the small amount of human DNA versus the abundant microbial sequences in most post-mortem human bone remains, researchers at the Max Plank Institute in Leipzig designed a method to fish human DNA out of ancient samples. This targeted sequencing approach involved the use of fifty-two-bases long DNA sequences designed to capture over a million variable positions across the human genome (153). This approach proved to be very cost-effective and sensitive for capturing genetic variation, although for a limited number of genomic positions. To have a complete overview of genetic variability a better approach is to perform whole-genome sequencing of the ancient remains, but this is expensive and time-consuming and usually results in very low coverage of the sample, due to the low endogenous ancient DNA content.

In this chapter, I will describe the analyses performed using genome-wide SNP-genotype data from the biggest modern Himalayan population dataset analysed thus far. It comprises individuals from 49 diverse populations from Nepal, Bhutan, Tibet and North India, representing the two major language families, Tibeto-Burman and Indo-European. The main aim of this project was to genetically characterise human populations that had not been extensively studied before and give a better picture of the genetic variation in the Himalayan region. Another focus was on understanding the genetic relationships within the Himalayan populations and, due to the linguistic richness in the Himalayan region, to examine whether there was a correlation between genes and languages. I also explored the genetic relationship of the Himalayan individuals with other worldwide populations, in particular South and East Asians, and investigated possible admixture events with those populations. I also investigated the genetic affinity of modern Himalayans with ancient Eurasian individuals. Finally, due the harsh environment in which many Himalayan

populations live, I screened for variants underlying positive selection in association with high altitude adaptation. The text and figures of this chapter have been mostly taken from Arciero et al, 2018 (154).

2.2 Materials and Methods

2.2.1 Samples and dataset

Eight hundred and eighty-three individuals belonging to 49 Himalayan populations were genotyped and analysed after obtaining informed consent. The dataset included 26 populations from Nepal, 16 from Bhutan, two from North India sampled in Bhutan, and five from Tibet in China (Figure 2.1; Table 2.1). The samples represent the two major linguistic families in the area: Indo-European and Tibeto-Burman (also known as Trans-Himalayan). Specifically, 44 populations comprise Tibeto-Burman speakers from Tibet, Bhutan, North India or Nepal, and five comprise Indo-European speakers from Nepal (Chetri, Damai, Majhi, Sarki and Sonar). The Bhutanese, North Indian and Nepalese samples were collected as part of the 'Language and Genes of the Greater Himalayan Region' project, a genetic survey of Tibeto-Burman and Indo-European speakers from these Himalayan countries, and have been described previously (111). Tibetan samples were selected from participating members of an epidemiological study in the Tibet Autonomous Region, China, in 2007, that was approved by the institutional ethics review board of BGI-Shenzhen. Samples were collected from healthy unrelated Tibetans from five villages based on their medical records and a comprehensive medical examination during sampling. Peripheral venous blood or saliva was collected for DNA extraction and genotyping. All participants had a self-reported family history of at least three generations living at the sampling site. The samples were genotyped using three Illumina SNP-chips: 1) HumanOmniExpress-12 v1.0 BeadChip (741k SNPs) at the Wellcome Sanger Institute; 2) HumanOmni1-Quad BeadChip (~1M SNPs) at the Leiden University Medical Center; and 3) HumanOmniExpress-24 BeadChip (~713k SNPs) at BGI-Shenzhen. Genotype calling and QC on all samples were performed using the Sanger Institute's variant calling pipelines, and SNP positions were mapped to the human reference assembly GRCh37. Genotypes from the three arrays were merged using PLINK 1.92 (155), resulting in a dataset of 600,838 SNPs.

Code	Population	Sample size	Language family	Latitude (^{0})	Longitude (^{0})	Altitude (m)
Bhutan						
KAT	Brokkat	16	Tibeto-Burman	27.73	90.43	3750
BRP	Brokpa	14	Tibeto-Burman	27.40	91.72	1608
BUM	Bumthang	19	Tibeto-Burman	27.67	90.55	3250
CHL	Chali	19	Tibeto-Burman	27.38	91.02	3201
DAK	Dakpa	2	Tibeto-Burman	27.47	91.52	2169
DZA	Dzala	14	Tibeto-Burman	27.90	91.15	2705
GNG	Gongduk	17	Tibeto-Burman	27.08	90.93	1437
KHG	Khengpa	18	Tibeto-Burman	27.13	90.68	808
KUR	Kurtöp	19	Tibeto-Burman	27.82	90.82	3850
LHK	Lakha	18	Tibeto-Burman	27.68	90.15	2985
LAY	Layap	12	Tibeto-Burman	28.07	89.68	4115
MNG	Mangde	18	Tibeto-Burman	27.42	90.22	2980
MON	Black Mountain Mönpa	20	Tibeto-Burman	27.22	90.22	2027
NGA	'Ngalop	18	Tibeto-Burman	27.53	89.48	2880
NUP	Nup	18	Tibeto-Burman	27.58	90.33	2385
TSH	Tshangla	19	Tibeto-Burman	27.18	91.32	2240
India						
BOD	Bodo	18	Tibeto-Burman	26.67	90.33	99
TOT	Toto	46	Tibeto-Burman	26.67	89.00	86
Nepal						
BHI	Bahing	1	Tibeto-Burman	27.23	86.33	1561
BTW	Bantawa	2	Tibeto-Burman	26.53	87.06	1500
BAR	Barâm	20	Tibeto-Burman	28.07	84.67	1080
CHM	Chamling	7	Tibeto-Burman	28.24	83.22	1550
CHN	Chantyal	20	Tibeto-Burman	28.40	83.37	1932
CHP	Chepang	18	Tibeto-Burman	27.58	84.70	241
CHE	Chetri	19	Indo_european	29.17	81.20	987

DAM	Damai	8	Indo_european	27.33	83.03	987
DHI	Dhimal	20	Tibeto-Burman	26.50	87.70	86
DUM	Dumi	2	Tibeto-Burman	27.19	86.51	1650
GUR	Gurung	17	Tibeto-Burman	28.30	84.12	2182
KUL	Kulung	2	Tibeto-Burman	27.35	86.56	1400
LIM	Limbu	19	Tibeto-Burman	27.19	87.83	1094
MGR	Magar	17	Tibeto-Burman	28.08	83.83	844
MAJ	Majhi	17	Indo_european	27.83	83.67	969
NAC	Nachiring	1	Tibeto-Burman	27.23	86.51	2350
NWR	Newar	17	Tibeto-Burman	27.62	85.40	1665
PUM	Puma	1	Tibeto-Burman	26.54	86.56	1400
SAM	Sampang	3	Tibeto-Burman	27.15	87.08	1300
SAR	Sarki	6	Indo_european	27.37	83.07	987
SHE	Sherpa	10	Tibeto-Burman	27.73	86.58	4550
SON	Sonar	2	Indo_european	27.39	83.09	987
SUN	Sunwar	4	Tibeto-Burman	27.34	86.16	1750
TMG	Tamang	19	Tibeto-Burman	27.88	85.42	1900
THK	Thakali	16	Tibeto-Burman	28.82	83.75	3035
WAM	Wambule	13	Tibeto-Burman	27.07	86.36	1350
Tibet						
LHA	Lhasa	36	Tibeto-Burman	29.65	91.23	3700
NAG	Nagqu	31	Tibeto-Burman	32.00	91.13	4500
NYI	Nyingchi	20	Tibeto-Burman	28.42	94.43	2900
SHA	Shannan	26	Tibeto-Burman	29.03	91.69	3600
TIN	Tingri	19	Tibeto-Burman	29.47	87.03	4500

Table 2. 1 List of populations analysed in this study. The table is subset into regions and reports the three letter population codes, the full population name, the population sample sizes, language family, geographical coordinates (latitude, longitude) and altitude (m).

2.2.2 Methods used for the analyses

2.2.2.1 Data quality control and datasets

The assessment of SNP-chip data quality plays an essential role for the precision and accuracy of downstream analyses. I used tools implemented in PLINK 1.92 and EIGENSOFT 6.0 for performing quality control (QC) of my dataset (156, 157). Firstly, I set the genotyping success rate and sample missingness thresholds to 99% and 10%, respectively, to exclude variants and individuals with an excess of missing genotypes. Sex-linked and mitochondrial SNPs as well as autosomal ones with Hardy-Weinberg Equilibrium p-value <0.0000001 were removed. SNPs departing to this extent from Hardy-Weinberg equilibrium are likely to be genotype and variant call errors (158). Another important assumption for population genetic studies is that the samples be unrelated, as including duplicated samples or first-second degree relatives can skew the allele frequency distribution and not be a good representation of the entire population. I used estimates of identity by state (IBS) for identifying pairs of samples that looked genetically more similar than would be expected by chance in a random sample. IBS was calculated for each pair of individuals as the number of shared alleles at the autosomal genotyped SNPs. From the genome-wide IBS values it is possible to infer estimates of identity by descent (IBD), the degree of recent shared ancestry for each pair of samples. IBD values of 1 indicate duplicated or monozygotic twin individuals, IBD values of 0.5, 0.25 and 0.125 are for first, second and third-degree relatives, respectively. I filtered out individuals with an IBD score > 0.35 (PI_HAT > 0.35): I chose this cut-off as it was good for removing related individuals that could affect the analyses without removing a big proportion of the samples.

I also removed outliers (individuals that show allele frequency that differ from the distributions of the remaining individuals of the same population) using EIGENSOFT 6.0 (157, 159). These filters resulted in a final dataset of 738 individuals and 583,011 SNPs. For comparison with worldwide populations, the final Himalayan dataset was merged with published datasets (Table 2.2) (67, 160, 161) resulting in 1,962 individuals and 268,861 SNPs. Two additionally pruned datasets were generated from these by filtering out SNPs in high LD ($r^2 > 0.5$) to reduce the occurrence of highly correlated pairs of SNPs

that could skew allele frequency distributions. The pruned Himalayan dataset consisted of 256,506 SNPs, and the pruned worldwide dataset included 190,287 SNPs. For comparison with ancient samples, I generated two further datasets: 1) I merged my dataset with the Human Origins data, a dataset comprising both modern and ancient individuals including archaic genomes from Denisovans and Neanderthals, and a chimpanzee (82), resulting in 82,647 SNPs in common; and 2) we merged our Himalayan and worldwide datasets with published ancient Himalayan genomes from the Annapurna Conservation Area in Nepal (121). From the published ancient BAMs a single sequence with a minimum quality of ≥20 to represent each SNP in our Himalayan dataset was randomly sampled (162), trimming 5 bp from both ends of reads to reduce the effect of ancient DNA deamination. The calling from the Himalayan ancient samples was performed by Dr Marc Haber. This resulted in 582,810 SNPs in the Himalayan dataset being covered by at least one of the ancient samples (Table 2.2).

Dataset	Number of individuals	Number of SNPs
Modern Himalaya	738	583,011
Modern Himalaya LD pruned	738	256,506
World (Himalaya *+ Li et al., 2008* + South Asians: *Metspalu et al., 2011, Chaubey et al., 2010*)	1,962	268,861
World LD pruned (Himalaya + *+ Li et al., 2008* + South Asians: *Metspalu et al., 2011, Chaubey et al., 2010*)	1,962	190,287
Himalaya + The 1000 Genomes Project Consortium., 2015	3,274	579,640
Modern Himalaya + Ancient Himalaya (*Jeong et al., 2016*)	743	582,810
Himalaya + Human Origins (*Patterson et al., 2012*)	2,809	82,647

Table 2. 2 Datasets used for the analyses. The table reports the number of individuals and number of genomic variants contained in each dataset.

2.2.2.2 Population characterization and demography

The genetic structure of the Himalayan populations was examined using several statistical approaches. Firstly, I performed principal component analysis (PCA) using EIGENSOFT 6.0 (157, 159) on the LD-pruned datasets. For the worldwide dataset, the eigenvectors were calculated using the global diversity and the Himalayan individuals were projected onto the plot. I ran ADMIXTURE v1.2(74) on the pruned datasets for cluster analysis and the cross validation (CV) error for identifying the best K value, which represents the best number of ancestral populations to describe the dataset. ADMIXTURE is a programme designed to estimate individual ancestries from dense SNP genotype dataset using a maximum likelihood approach. Estimation of long-term effective population size (Ne) for each Himalayan population and population divergence time was performed by Dr Massimo Mezzavilla using the NeON R package (163), which calculates the harmonic mean of the population size at each generation and the time of divergence between populations in generations. More specifically, using LD information (r^2) and recombination distance (c) the effective population size is estimated using the nonlinear regression model: $y_i = 1/(\alpha+\beta c_i)+e_i$, with $y_i = (r^2\ 1/n)$ (r^2 adjusted for chromosome sample size) for SNP pair i at recombination distance c_i (in Morgans) .The change in population size over time is then estimated, as LD between loci with a recombination rate of c that reflects the ancestral effective population size 1/(2c) generations ago (164). The model is based on the assumption of linear growth or decline. However, some populations might depart from the assumed model characteristics and LD patterns in these will be affected, so the relationship t = 1/(2c) should be viewed only as an approximate, but useful indication of timeframes (165). Furthermore, the time of divergence estimates are based on the assumption of a ‘‘clean’’ population split, and migration will create a stronger correlation of LD (larger values of r^2), thereby biasing the estimate of divergence time downwards. Nevertheless, this method is still useful to assess isolation and difference in Ne between populations (166, 167). Only populations with sample size ≥10 were used, as the harmonic mean is sensitive to sample size. For all analyses we assumed a generation time of 29 years (168).

Runs of homozygosity (ROHs) are long stretches of the genome where identical haplotypes are inherited from each parent that in turn they inherited from a recent

common ancestor. This produces a region of homozygous variants and the length of the run is indicative of when the inbreeding event happened. I identified ROHs using PLINK 1.92 (155) with specific thresholds to maximize the detection of autozygous segments in the Himalayan populations (169) and other worldwide populations: a pruned dataset (LD, $r^2 > 0.5$) with only common variants (MAF >0.05) was used. The minimum number of SNPs to call an ROH was set to 100, the heterozygote allowance was set to zero, the missing SNP allowance was set to 5 (5% of the SNP threshold), and the window threshold to call an ROH was set to 0.05. I also calculated the coefficient of inbreeding (F) with PLINK 1.92 to test for positive correlation between the total length of ROHs and F in each individual (155).

A worldwide dataset using a maximum of 10 individuals from every population was used in the ChromoPainter and fineSTRUCTURE-2.0.6 (78) analyses to study the genetic relationship between Himalayan populations. ChromoPainter is a tool for inferring the ancestry of each individual by reconstructing their haplotype segments from other individuals in the dataset. FineSTRUCTURE uses the co-ancestry matrix inferred from ChromoPainter to construct a model-based Bayesian population-relationship tree and was run with 10,000,000 burn-in steps and 10,000,000 iterations. The software requires phased data as input. The haplotypes were phased with SHAPEIT (66) using the 1000 Genomes Project Phase 3 dataset (66) as a reference panel. I also performed a PCA using the co-ancestry matrix generated by fineSTRUCTURE. To assess the robustness of the results from the above dataset, I additionally ran fineSTRUCTURE on a dataset with fewer samples (the Himalayans and 1000 Genomes Project Phase 3 populations), but more markers (579,640 SNPs) using the same parameters.

The Himalaya dataset comprises sixteen extreme high altitude populations (altitude of 2500 meters or more above sea level) (135), residing at different geographical locations across the Himalayas. To understand whether there is a correlation between genetic similarity among high altitude populations and their geographical location, I used *D*-statistics, a tool for detecting genetic similarity and mixture events within individuals of pairs of populations. I used YRI (Yoruba in Ibadan, Nigeria) as an outgroup and calculated *D*-statistics (qpDstat function in ADMIXTOOLS v3 package (82)) using the following phylogeny: *D*(Yoruba, Han; high altitude Himalayan 1,

high altitude Himalayan 2) where high altitude Himalayan 1 and high altitude Himalayan 2 are pairs of Sherpa, Tibetan or Bhutanese populations (82) (120). I computed *D*-statistics with the above phylogeny using the worldwide dataset for comparison (supplementary tables S1 and S3). Then, I tested the correlation between values of *D*-statistics with pairwise differences in longitude and latitude for each pair of populations using the Mantel test implemented in the "Ade4" R package (mantel.rtest function) (170).

Population admixture was studied using ALDER v1.03 (171), three-population statistics (*f3*) (82, 172) and TreeMix 1.12 (77). The ALDER software computes a LD-based test for admixture that can estimate the time of the admixture event, whereas three-population *f3* statistics and TreeMix are allele frequency based methods. Only populations with at least six individuals were included in these tests. ALDER was used with the default parameters and the threshold of LD in the reference groups was inferred by the program. A test was considered positive when both the 2-ref weighted LD curve was significant and the decay rates between the 2-ref and 1-ref curves were consistent. For *f3*-statistics analyses I considered a jack-knife block of 500 SNPs. I also estimated the shared genetic drift between modern populations and ancient samples using outgroup *f3*-statistics (ancient genome, X, Yoruba) (82) with the Yoruba as an outgroup. I used different ancient genomes in this investigation:

1. Eurasian hunter-gatherer (MA1, 24,000-year-old Upper Palaeolithic Siberian) genetically related to Native American lineages and to modern Europeans (173).
2. Bronze Age Yamnaya population (3500-2700-year-old) from the Pontic-Caspian steppe that expanded into Eurasia impacting the genetic landscape during the Bronze Age. The Yamnaya population expansion parallels the expansion of Indo-European languages, making the Yamnaya a good candidate for the putative spread of Indo-European languages across Eurasia (174).
3. Neolithic European farmer from the Linear Pottery culture (LBK_EN, 5,500-4,800-year-old) (175).
4. Mesolithic hunter-gatherer (La Braña, 7,000-year-old) (176) that shares genetic ancestry with MA1 and some modern-day northern European populations.
5. Eurasian hunter-gatherer (Ust'-Ishim, 45,000-year-old Upper Palaeolithic Siberian) deriving from an ancestral population that lived before or concomitantly to the separation of western and eastern Eurasians and has slightly more genetic affinity to East Asian populations than to Western Eurasians (153);

6. Five ancient Himalayan genomes 3,150-1,250-years-old (C1, M63, S10, S35 and S41) from the Annapurna Conservation Area, Nepal (121).

I also used the archaic Denisovan and Neanderthal genomes and the chimpanzee to study genetic affinity of Himalayan samples to these archaic individuals: I calculated principal components using Denisovan, Neanderthal and chimpanzee, and projected modern samples onto them (35). I also computed *D*-statistics using (Yoruba, X; Denisovan, Chimpanzee) where X were different modern human populations from the worldwide dataset.

2.2.2.3 Positive selection

Signals of positive selection were evaluated in four ways to try and reduce, as much as possible, the spurious signals of selection due to the extensive population structure and relatedness of the Himalayan samples. First, I considered the Spearman's correlation between derived allele frequency and the residence altitude of each population (127), adjusting the p-value for multiple tests by applying the Bonferroni correction (requiring <0.05/number of tests). Spearman's correlation is a non-parametric test that measures the strength of correlation between two variables where the value rho = 1.0 is perfect positive correlation and the value rho = -1 is a perfect negative correlation. It does not, however, take into account the population structure. Second, I calculated a genome-wide association between allele frequency and altitude using a mixed model approach implemented in the Efficient Mixed-Model Association eXpedited program (EMMAX) (177). EMMAX detects variants where the observed allele frequency is significantly divergent from the expected frequency, and accounts for population stratification and sample relatedness through a variance component approach. A kinship matrix was constructed to account for population structure and implemented in a linear mixed model. Variants with p-value $<5x10^{-8}$ were considered significantly associated with altitude. Although these methods detect associations between allele frequency and altitude, they are not able to distinguish between high- and low altitude selection signals. For this reason I also calculated the Fixation Index (F_{ST}) (178) for each SNP position between Himalayan, European (CEU; Utah Residents (CEPH) with Northern and Western European Ancestry) and East Asian (CHB; Han Chinese in Beijing, China) from the 1000 Genomes Project Phase 3 populations, and searched for unusual values using the Population Branch Statistic (PBS) (99). To reduce noise due to population structure within the Himalayan populations and differences in sample sizes, I ran PBS assuming that the CHB is the most closely related population to Tibetans and looked specifically for signals of high altitude adaptation (99). For this analysis, I only used the populations from Bhutan and Tibet that clustered together in the fineSTRUCTURE analysis compared with CEU and CHB. Variants above the 99.99^{th} percentile of the empirical distribution were considered statistically significant (179, 180). The top hits from each method were assessed and the overlap collated. LD estimations for the regions containing the top candidates were calculated and plotted

using Haploview (181). I also used Fisher's method (182) for combining p-values of the three statistics used for detecting positive selection. Firstly, I calculated a rank p-value of the PBS values (values were ranked in decreasing order from the most significant value and divided by the total number of SNPs used in the analysis); I then combined the p-values of the three statistics genome-wide and adjusted the p-value for multiple tests by applying the Bonferroni correction (requiring < 0.01/number of tests). Finally, to further validate the selection signals, I calculated genome-wide associations between allele frequency and altitude using BayEnv v2 (183, 184), a Bayesian framework specifically designed to detect correlation between allele frequencies and environmental factors taking population structure into account. I generated the input files for BayEnv2 v2 from a LD ($r^2 > 0.5$) pruned SNP file using PGDSpider (185) and I standardised altitude (the environmental variable) according to the BayEnv2 v2 manual. BayEnv v2 was run with the default parameters and the Bayes Factors interpreted according to previous recommendations (186): only candidate variants falling into the category "Decisive" (Bayes Factors (BF) > 100, $\log_{10}(BF) > 2$) were considered significant. Where possible, allele frequencies in the five Himalayan ancient genomes for our top candidates of selection were also calculated.

I generated a protein homology model for SLC52A3 using Phyre2 software (187) and mapped the missense variant found in *SLC52A3* onto the protein structure using PyMOL (The PyMOL Molecular Graphics System, Version 1.8 Schrödinger, LLC). I predicted protein-protein interaction networks using the STRING software (v. 10.5) (188) for our top selection candidates. Finally, I used the Ensembl Variant Effect Predictor (VEP) (189) to predict the consequences of variants of interest on gene expression and protein sequence. I retrieved the Combined Annotation Dependent Depletion v.1.2 (CADD) scores (190) of our top candidates and also overlapped the results with the Genotype-Tissue Expression (GTEx) database (191).

2.3 Results

2.3.1 Himalayan samples show distinct patterns of population structure

I first investigated the population history and demography of the Himalayan region (Figure 2.1, Table 2.1) by determining the genetic relationships among the Himalayan populations, and comparing them with other worldwide populations from published datasets such as the Human Genome Diversity Project (HGDP), 1000 Genomes Project and other South Asian specific datasets (Table 2.2).

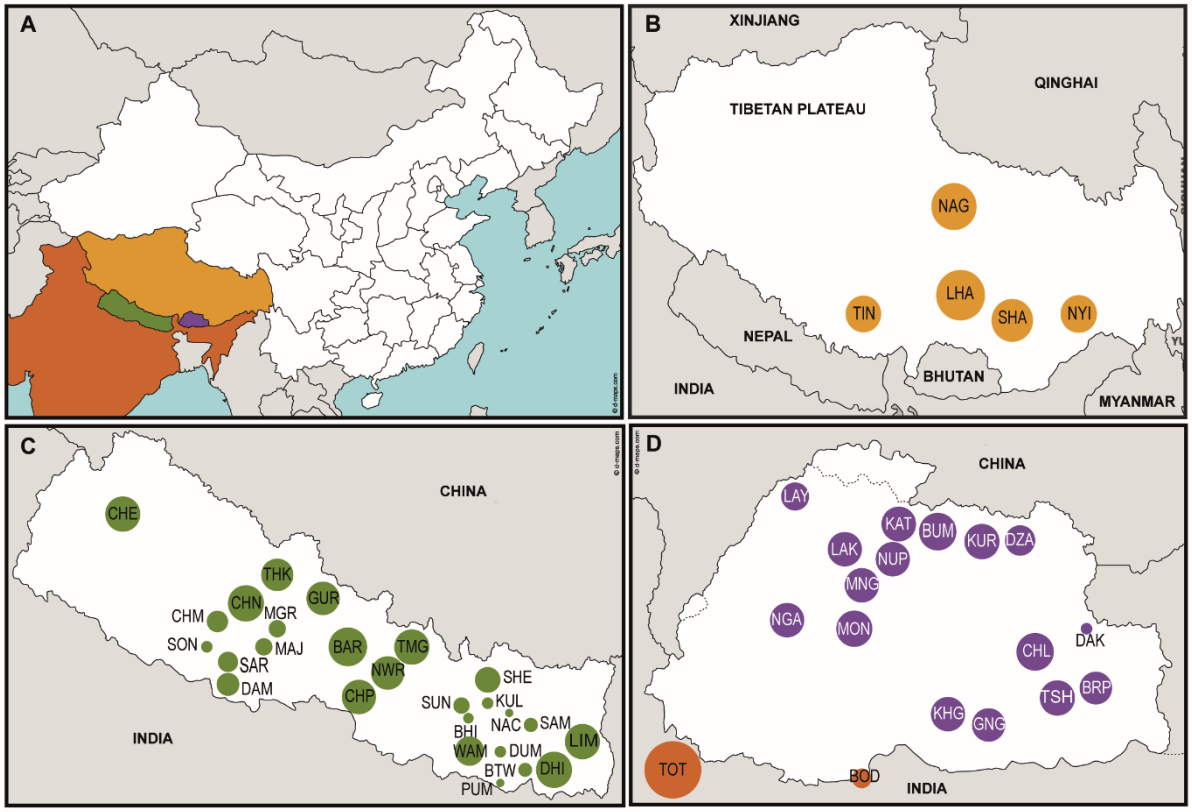
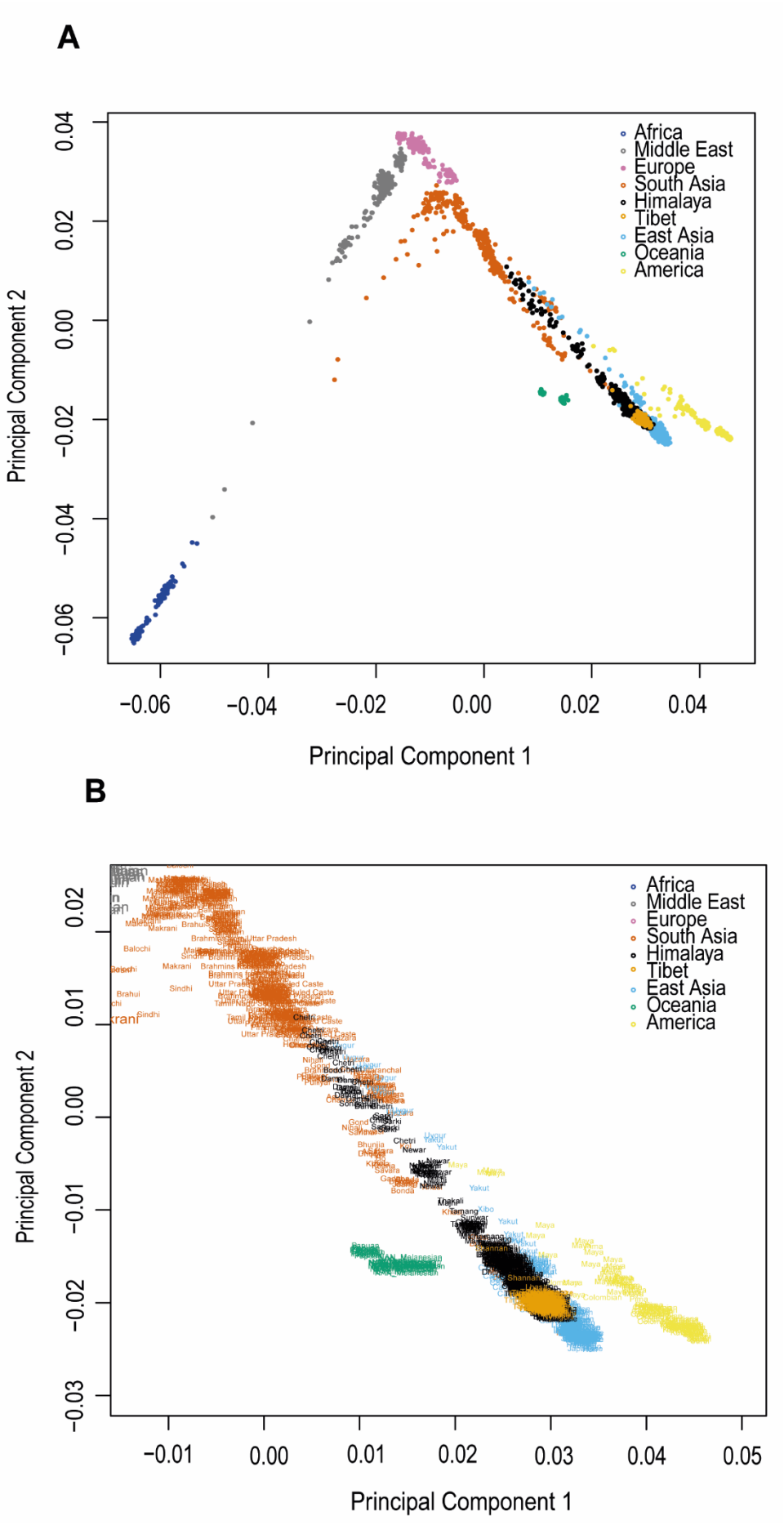


Figure 2. 1 Population samples analysed in this project. A. Map of South and East Asia, highlighting the four regions examined, and the colour assigned to each. B. Samples from the Tibetan Plateau. C. Samples from Nepal. D. Samples from Bhutan and India. The circle areas are proportional to the sample sizes. The three letter population codes in B-D are defined in Table 2.1.

Principal Components Analysis (PCA) shows that the Himalayan populations form a cline, lying between the South and East Asian samples. Populations from Nepal are close to Indians, whereas those from Bhutan and Tibet are closer to East Asians (Figure 2.2A, B). This pattern of genetic affinity to South and East Asian populations is also supported by an ADMIXTURE analysis at K=10 of worldwide populations (Figure 2.2C), where the genetic component from South Asia (orange) is observed particularly in the Nepalese, and the East Asian (gold) component in the Nepalese, as well as the Bhutanese and Tibetans. However, except for the Toto, all other Himalayan populations are mainly characterised by their own ancestral component (blue). I also found some detectable European and Middle Eastern ancestral components (off-white and green) in some Nepalese.


A
Africa
Middle East
Europe
South Asia
Himalaya
Tibet
East Asia
Oceania
America
Principal Component 2
Principal Component 1
0.04
0.02
0.00
−0.02
−0.04
−0.06
−0.06
−0.04
−0.02
0.00
0.02
0.04
B
Africa
Middle East
Europe
South Asia
Himalaya
Tibet
East Asia
Oceania
America
Principal Component 2
Principal Component 1
0.02
0.01
0.00
−0.01
−0.02
−0.03
−0.01
0.00
0.01
0.02
0.03
0.04
0.05

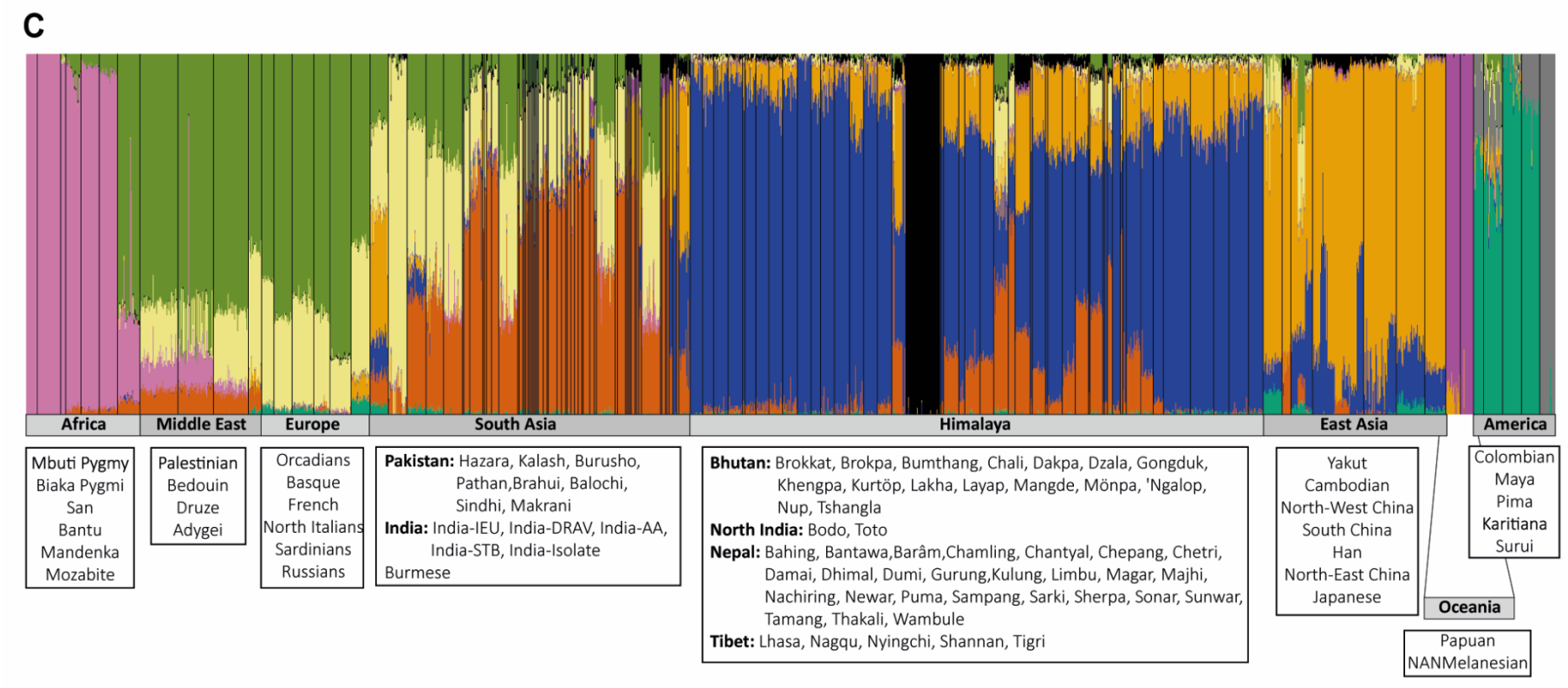
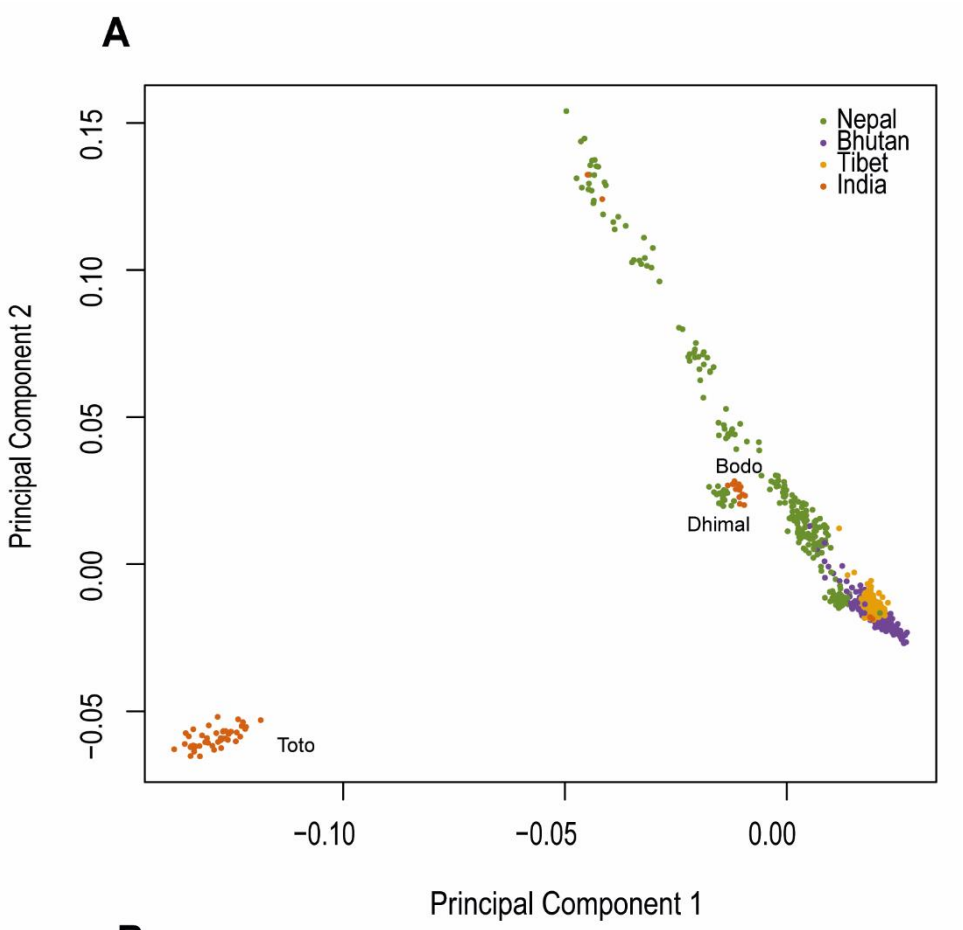
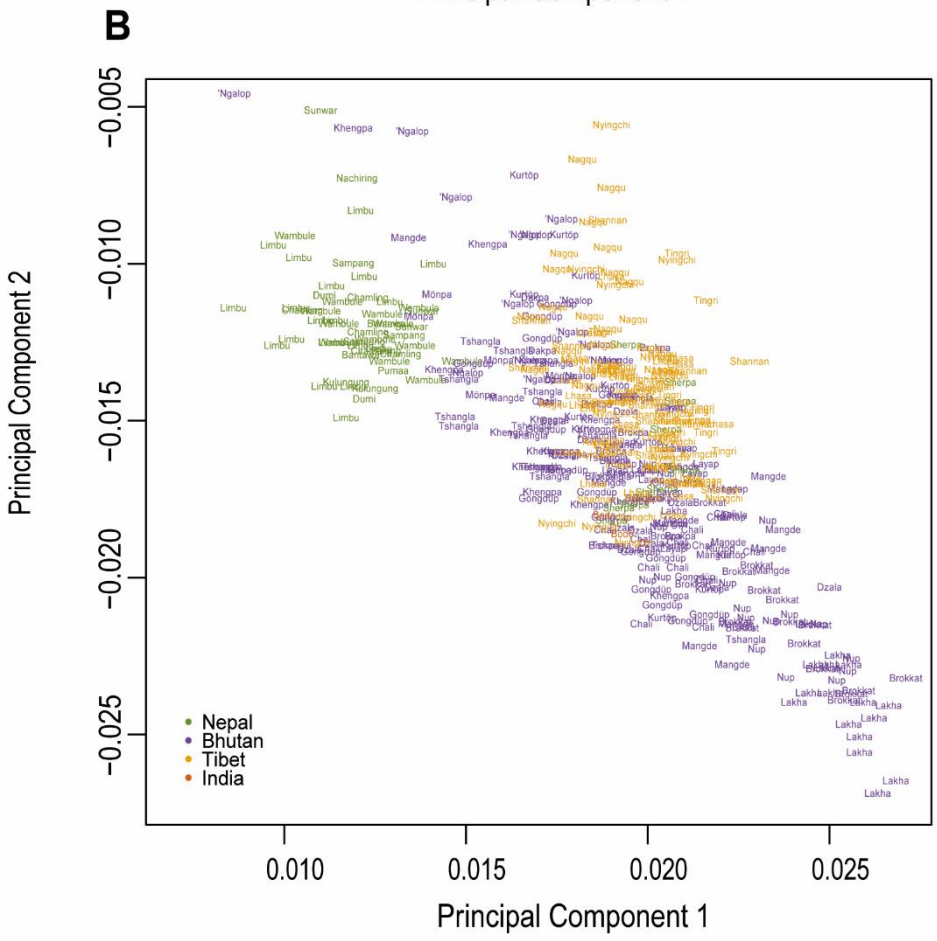


Figure 2. 2 PCA and ADMIXTURE analysis using the world dataset. A. PCA of the world dataset. Each dot represents a sample, coded by region as indicated. The Himalayan region samples lie between other East Asian and South Asian samples. B. The plot displays a zoom of Himalayan populations in their worldwide context. Nepalese samples lie close to Indians, whereas Bhutanese and Tibetans are close to East Asian populations. C. ADMIXTURE results for K=10. Himalayan populations display their own ancestral component (blue) that is also visible in some of the Pakistani and East Asian populations. Himalayan populations are also characterised by the presence of South Asian (orange) and East Asian (gold) components.

On a finer scale, the first component of a PCA using only the Himalayan populations shows strong geographical clustering with the Toto population forming an outlier, while the second principal component identifies substructure within the Himalayan populations (Figure 2.3A). Individuals from Nepal lie in several dispersed clusters, whereas those from Bhutan and Tibet group together. Interestingly, the Nepalese Sherpa cluster with the Tibetans and some Bhutanese populations from high altitude. A distinct cluster is formed by Dhimal and Bodo individuals from Nepal and North India, respectively (Figure 2.3A)


A
Nepal
Bhutan
Tibet
India
Bodo
Dhimal
Toto
Principal Component 2
Principal Component 1
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.05
0.00


B
Nepal
Bhutan
Tibet
India
Principal Component 2
Principal Component 1
−0.005
−0.010
−0.015
−0.020
−0.025
0.010
0.015
0.020
0.025

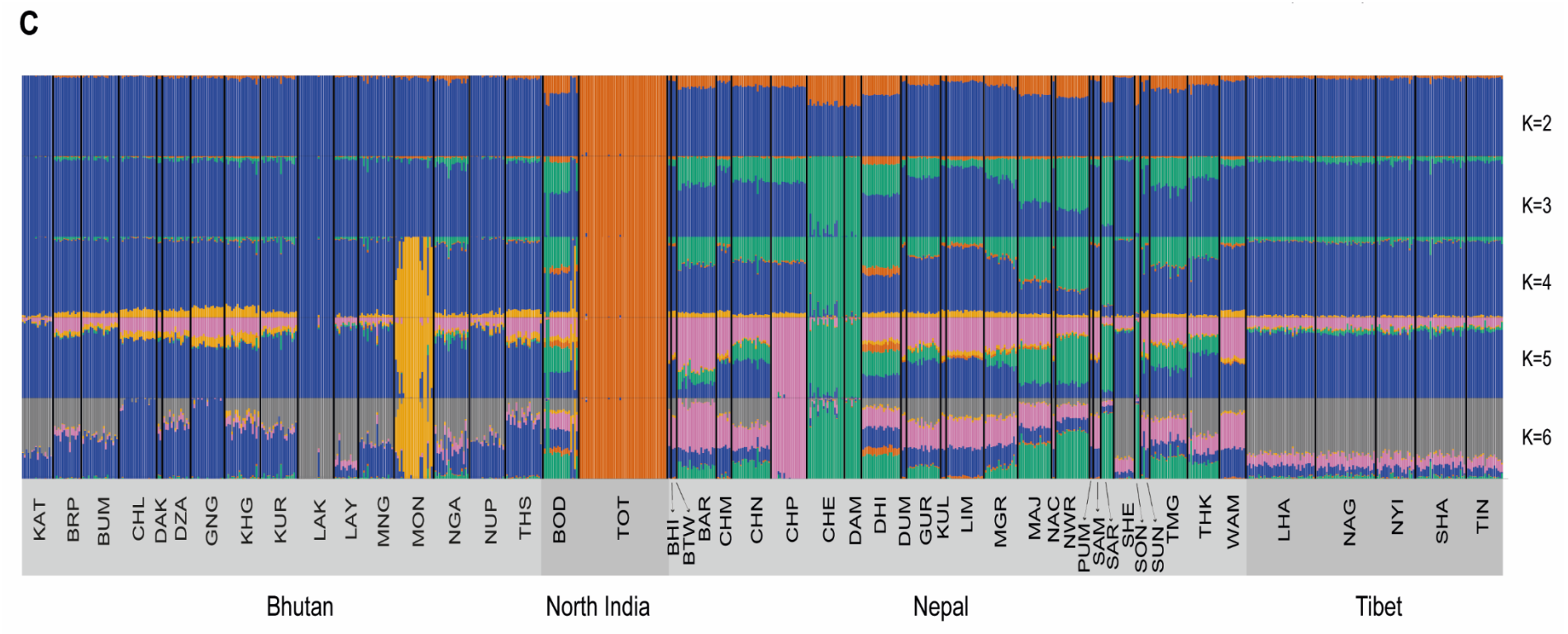


Figure 2. 3 PCA and ADMIXTURE analysis using the Himalayan dataset. A. PCA of the Himalayan populations alone. Each dot represents a sample, coded by country or region as indicated. Most samples lie on an arc between Bhutanese and Nepalese samples; Toto (India) are seen as extreme outlier in the bottom left corner, while Dhimal (Nepal) and Bodo (India) also form outliers. B. Zoom of PCA using the Himalayan dataset only. The plot displays Bhutanese and Tibetans clustering together. Sherpa from Nepal cluster with other high altitude populations from Bhutan and Tibet. C. ADMIXTURE (K values of 2 to 6, as indicated) analysis of the Himalayan samples. Note that most increases in the value of K result in single population being distinguished.

The ADMIXTURE analysis using only the Himalayan populations shows patterns consistent with the PCA, with different proportions of ancestral components between Nepal, Bhutan, North India and Tibet (Figure 2.3C). Each increase in the value of K between 2 and 5 usually leads to a single population being distinguished, suggesting extensive genetic isolation and drift. Toto, an outlier in the PCA, is also characterised by an independent ancestral component even at a K value of 2 (Figure 2.3C). By contrast, the five Tibetan populations do not show any substructure in this analysis. The lowest CV error was at a K value of 6, where we observe a single widespread ancestral component (grey) which is shared among all the high altitude populations and is significantly positively correlated with altitude (rho = 0.79; p=2.2 x 10^{-18}) (Fig 2.4).

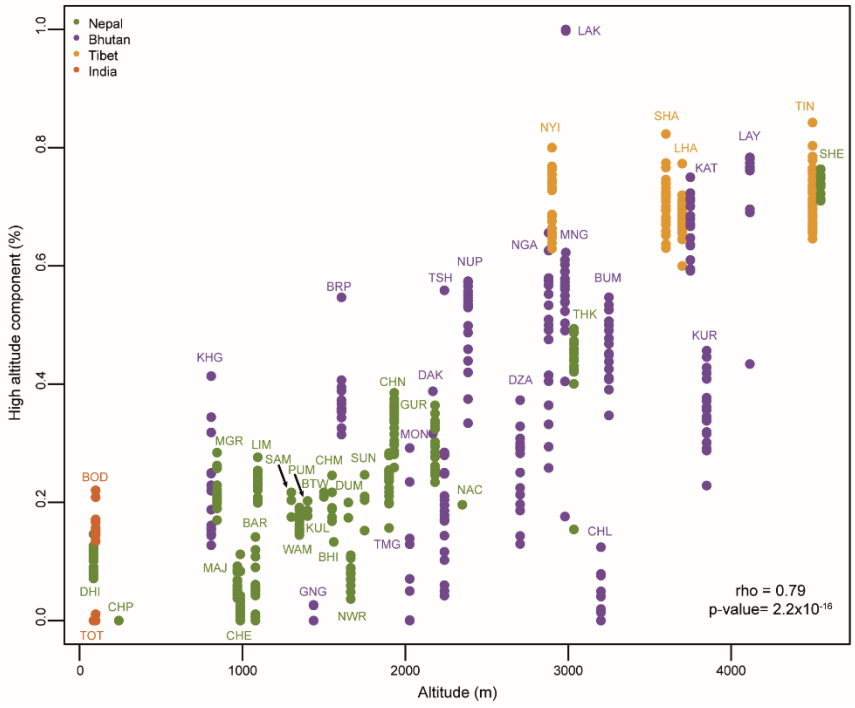


Figure 2. 4 Positive correlation between the high altitude-specific genetic component and altitude. The plot shows the correlation between the percentage of the grey component from the ADMIXTURE analysis (K=6) in each Himalayan population (Figure 2.3C) and the altitude at which the population resides.

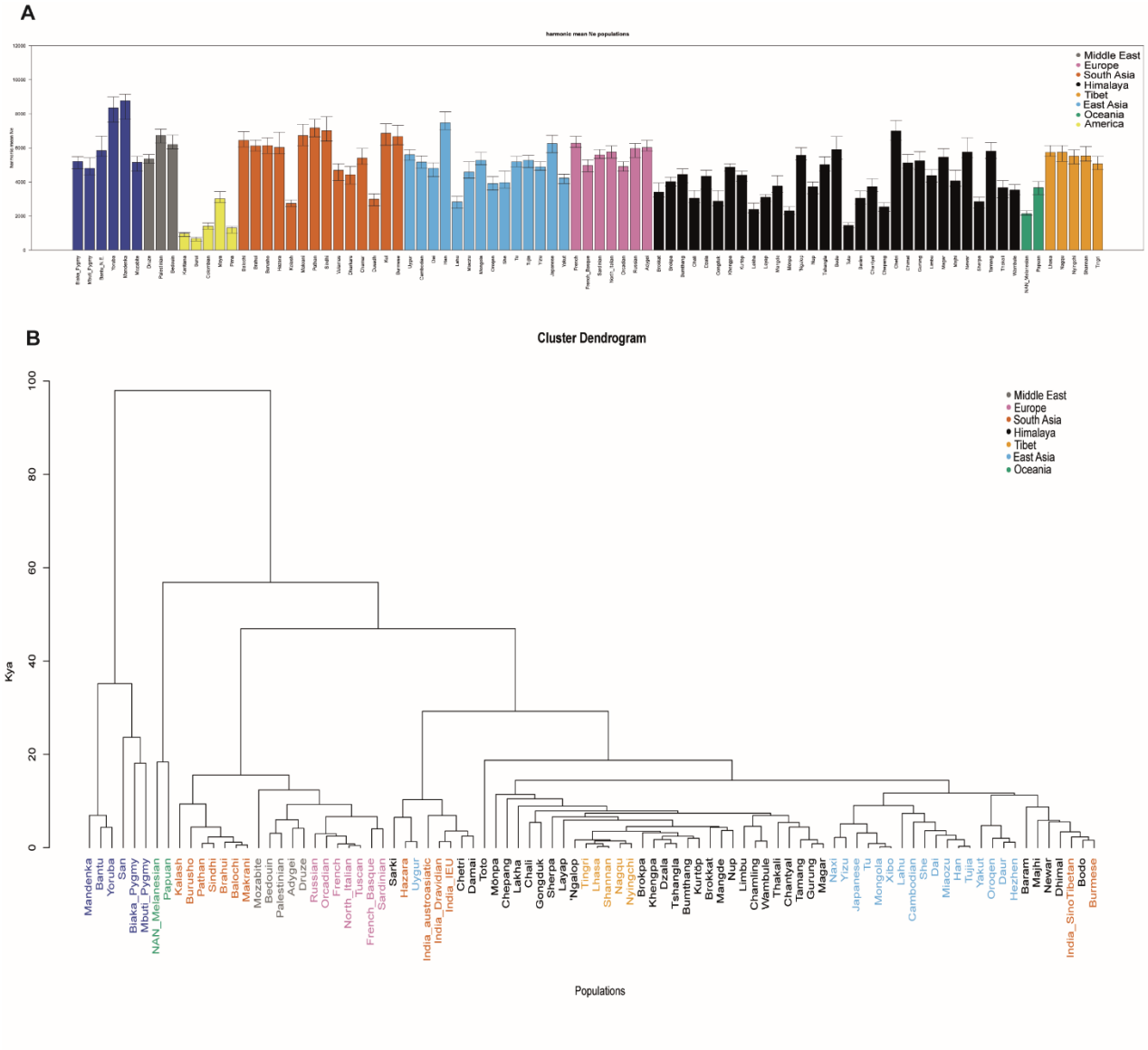


Figure 2. 5 Long-term effective population size (Ne) and divergence times of Himalayan populations. A. Long-term effective population size (Ne) of Himalayan populations. The plot displays the harmonic mean of the Ne for each population. Only populations with sample size ≥ 10 were used for this analysis. Within the Himalayan populations, Chetri have the highest Ne, whereas Toto have the lowest. B. UPGMA tree based on estimated time of divergence of the Himalayan populations. The populations analysed are shown on the x-axis and the time of divergence (Kya) on the y-axis. Most of the Himalayan populations show a split time with Indian populations around 25,000-20,000 years ago and with East Asians around 15,000-10,000 years ago. Chetri, Damai and Sarki are outliers to other Himalayan populations as they show divergence times similar to Indians. Tibetan populations display a split time among them of around 1,000 years ago.

Long-term Ne values can be estimated using SNP genotyping data, but these have limitations and can only be used as a proxy for the variability of their effective population sizes and thus the overall genetic diversity. Nevertheless they permit some informative comparisons. The Chetri have the highest long-term Ne, whereas Toto have the lowest (Figure 2.5A), suggesting that the low genetic variation in Toto could be due to genetic drift or endogamy (192). All Tibetan populations display similar population sizes (Figure 2.5A) (165). The sequence of splits suggests that the Himalayans separated first from Indian populations (with possible exceptions of Chetri, Damai and Sarki), then from East Asians and finally among themselves. Interestingly, all of the high altitude populations in this dataset display a similar differentiation time from other Himalayans, and place this at around 6,000-5,000 years ago (Fig 2.5B). Despite the limitations of this approach, this estimate is in line with several previous genetic and linguistic estimates (50, 118), but differs from others that estimated older (~15,000–9,000 years ago) or more recent split times (~2,700 years ago) respectively (99, 108, 110, 119). The various Tibetan populations display very recent split times from each other, which is consistent with the lack of substructure within these populations.

I also explored whether or not Himalayan populations show extended runs of homozygosity (ROHs), which may arise from endogamy. Overall, Himalayan populations are characterized by a high number of autozygous segments of different lengths across the genome (119). Nepalese and Bhutanese populations show the most numerous ROHs, and these are also the longest, up to ~80 and ~90 Mb in length, respectively. Toto from India are characterized by the highest number of individual ROHs up to ~50 Mb in length. On the other hand, Tibetan populations show the lowest number and length of ROHs (Fig 2.6A). The total length of ROHs per sample correlates positively with the coefficient of inbreeding (F) (Fig 2.6B). Bhutanese, Indian and Nepalese populations show the highest coefficient of inbreeding values and have total lengths of ROHs around 300-400 Mb. Tibetans show a very low coefficient of inbreeding associated with low numbers of ROHs. Overall, the number and length of ROHs in Himalayan populations are in line with those in other worldwide populations: in such a comparison, Toto show the highest numbers, followed by American and Middle Eastern populations, while Bhutanese populations show a total length and number of ROHs similar to populations from South Asia.

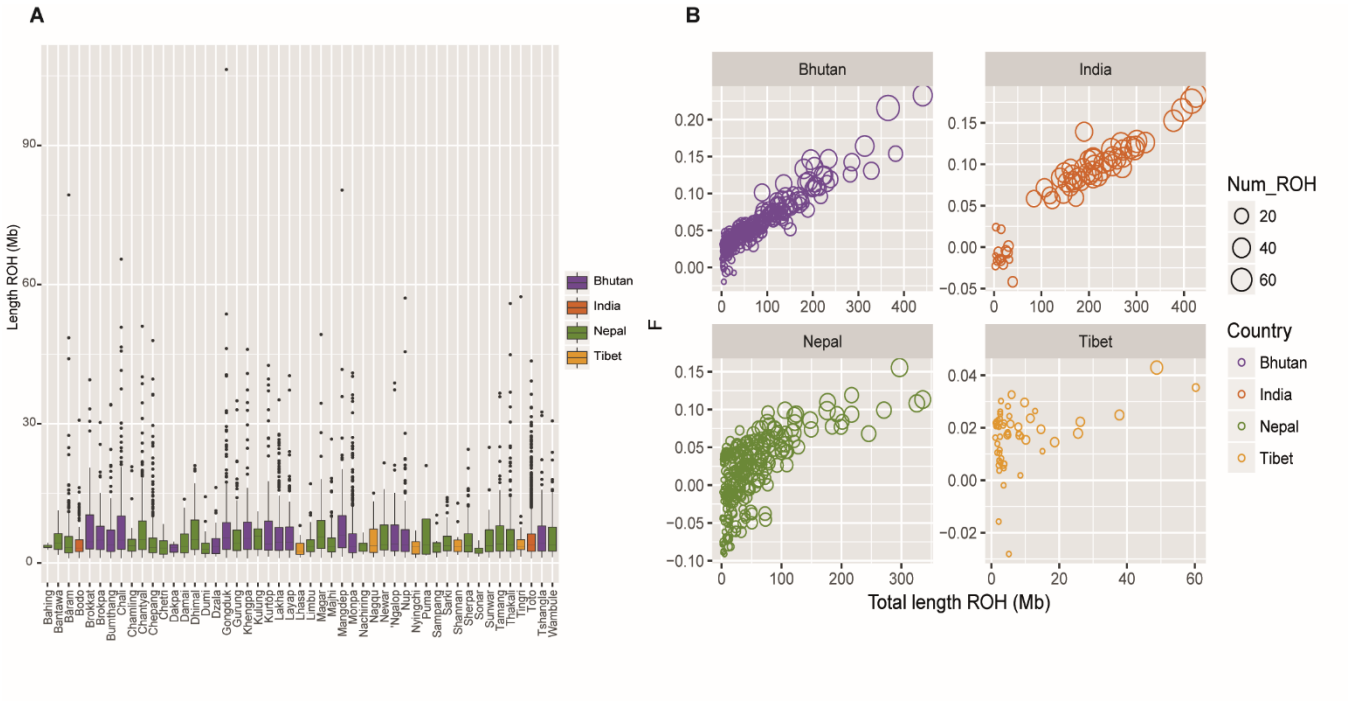


Figure 2. 6 Runs of homozygosity (ROHs) in the Himalayan populations. A. Distribution of runs of homozygosity (ROHs) in the Himalayan populations. The x-axis shows the populations analysed and the y-axis represents the median length of ROHs in each population. Nepalese and Bhutanese populations show the highest number and longest ROHs, whereas Tibetan populations show the lowest number and length of ROHs. B. Positive correlation between total length of ROHs per sample and coefficient of inbreeding (F). The total length (Mb) of ROHs per sample is shown on the x-axis, and the coefficient of inbreeding (F) on the y-axis. Note the different scale in each section. Bhutanese, Indian and Nepalese populations display the highest coefficient of inbreeding values and the highest total length of ROHs. Tibetan populations show the lowest coefficient of inbreeding and the lowest number of ROHs.

I also used the phased Himalayan and worldwide population data to reconstruct phylogenetic relationships between the samples and to identify population structure through a Bayesian clustering algorithm implemented in fineSTRUCTURE. The inferred phylogenetic tree shows two main branches splitting Nepalese from Bhutanese plus Tibetans (Figure 2.7A). All the Himalayan high altitude populations, including the Tibetans, cluster together, with the exception of the Thakali population from Nepal, which clusters with its Nepalese neighbours. Within genetic clusters of the Nepalese and Bhutanese it is possible to recognize substructure based on population and linguistic features. This tree topology was replicated when fineSTRUCTURE was applied to a dataset comprising only Himalayan and 1000 Genomes Project Phase 3 populations, which allowed a higher number of SNPs to be used. PCA was also calculated from the co-ancestry matrix generated by fineSTRUCTURE confirming that the Himalayan

populations are distributed along a cline with the Sherpa, Bhutanese and Tibetans clustering together.

The Himalayan region is one of the richest areas from the linguistic point of view with many different language families and linguistic isolates spoken. To test if there was correlation between genetic and linguistic variability in the Himalayas, I compared the genetic tree with the linguistic affiliation of each Himalayan population (Figure 2.7B). I observe, particularly in Bhutan, that there is agreement between genetic and linguistic sub-divisions. Speakers of Kiranti languages from Nepal form a separate cluster, and their languages constitute a distinct linguistic subgroup within the Tibeto-Burman language family. Dhimal from Nepal and Bodo from North India form a separate branch, supporting the PCA result, but not the traditionally-accepted language affiliation, and also correspond well with a new linguistic hypothesis which groups Dhimal and the Bodo-Koch languages together within a “Brahmaputran” subgroup (193).

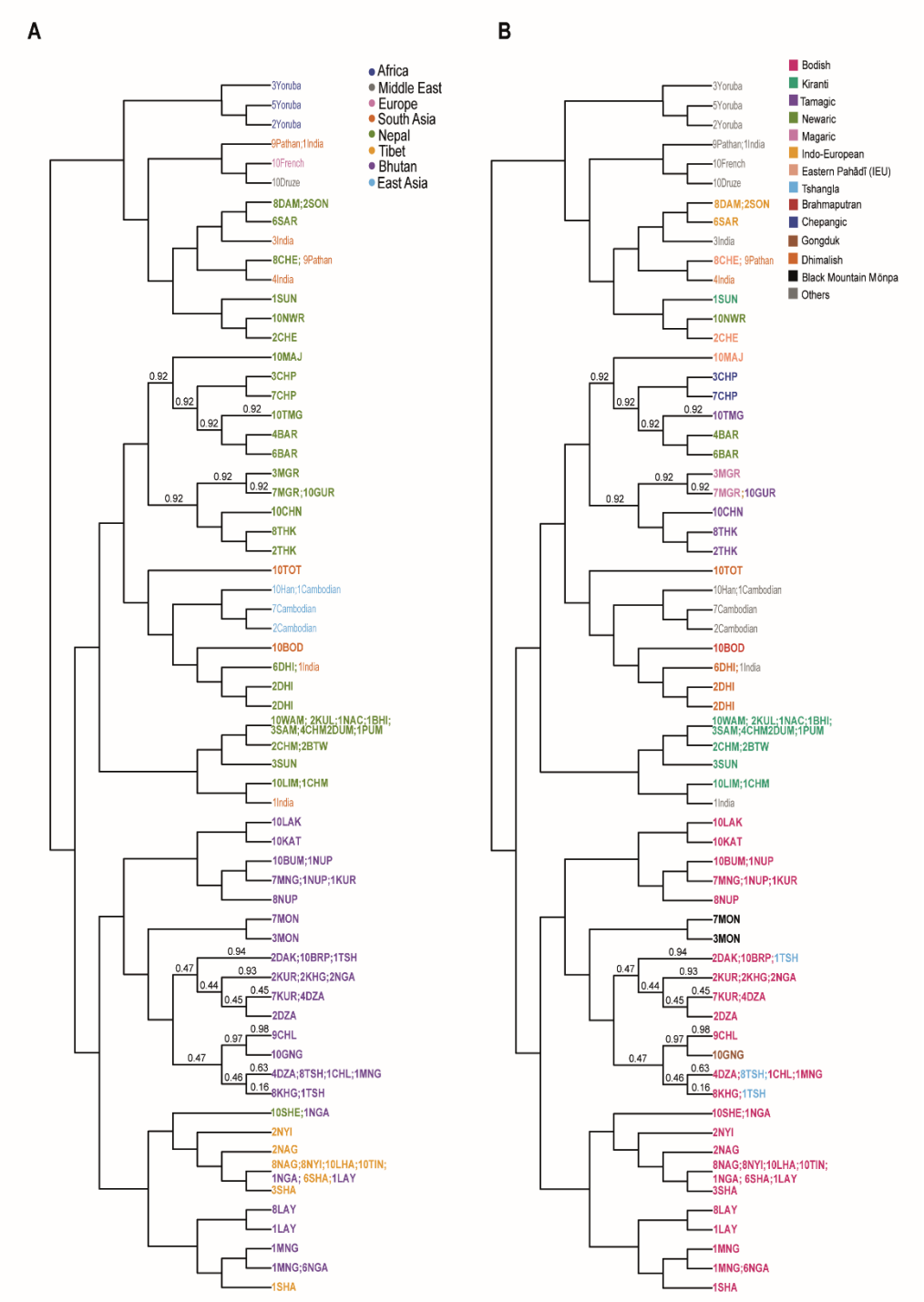
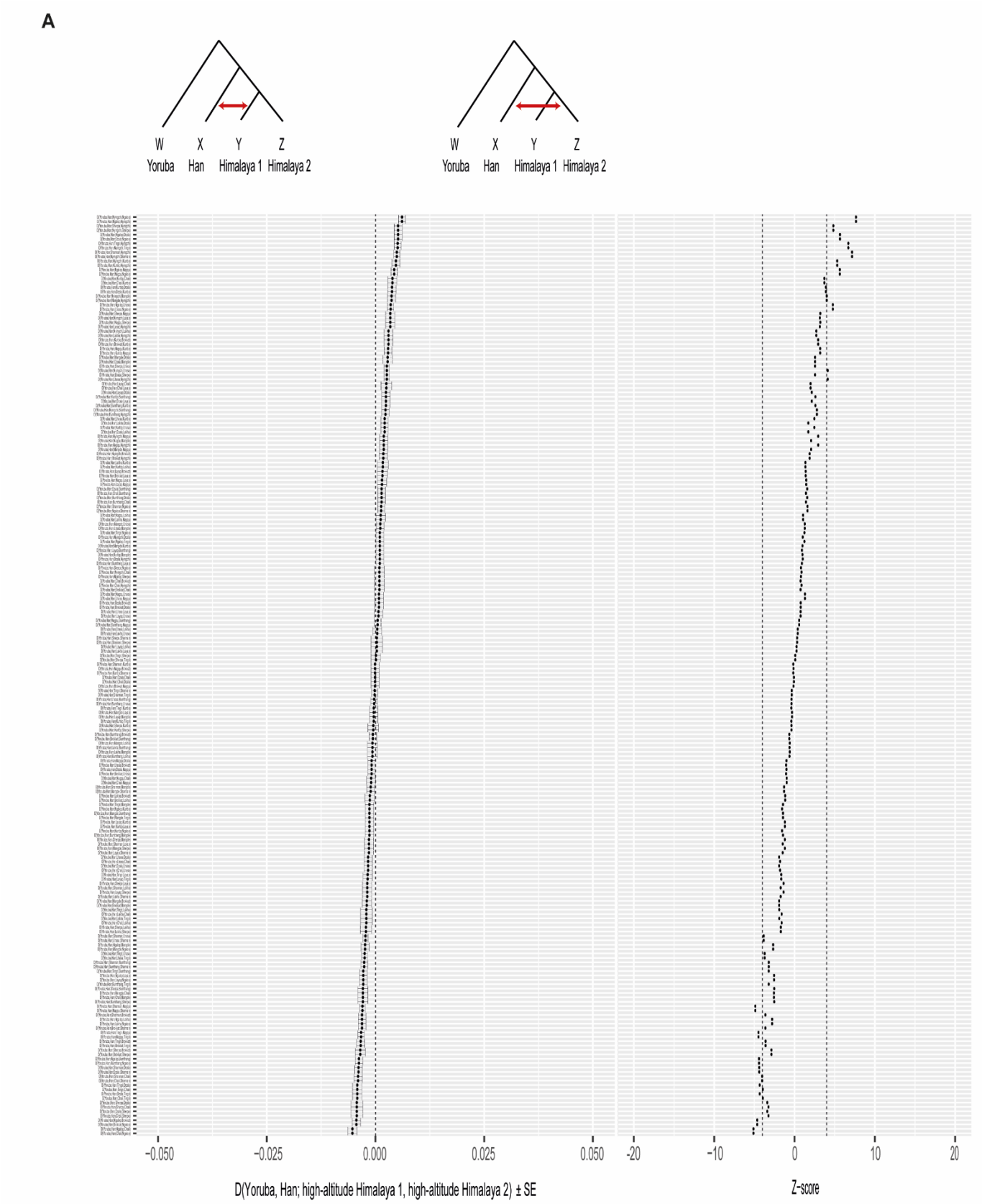


Figure 2. 7 Genetic structure of the Himalayan populations from haplotype analysis using fineSTRUCTURE, and comparison with language. A. Populations are clustered according to haplotype sharing; the branching pattern represents this hierarchy, but the branch lengths have no meaning. Note the geographical clustering of populations, particularly the Bhutanese. B. Language family annotation of the genetic clusters revealing correspondences between genetics and language. Population codes are defined in table 2.1.Indicated on the tree are branch confidence values. Branches without values have a confidence of 1.

Finally, I computed *D*-statistics (Yoruba, Han; high altitude Himalayan 1, high altitude Himalayan 2) for pairs of Sherpa, Tibetan and Bhutanese populations using our worldwide dataset to test for the presence of correlation between genetic similarity and geographical location of the high altitude populations. Jeong et al 2017 found an east-west longitudinal cline within Tibetan samples with individuals from more eastern locations having higher genetic affinity with lowland East Asians (120). I tested a similar hypothesis using the high altitude populations available to me. *D*-statistics values were close to zero for most of the pairs ($0.0001 \leq |D\text{-statistic}| \leq 0.0061$), with just 36 out of 210 tests statistically significant at a Z score ≥4 (values $0.072 \leq |Z| \leq 7.656$), showing that some high altitude Himalayan populations have increased genetic affinity with the low altitude East Asians (Fig 2.8A). However, unlike the Tibetan samples in Jeong et al 2017, our Himalayan populations do not follow a longitudinal cline (or a latitudinal one) related to their genetic affinity to low altitude East Asians (Mantel test r=0.15 and p-value = 0.18 for longitude, r=0.11 and p-value=0.18 for correlation with latitude). This difference may reflect the smaller range of longitude of our samples (Fig 2.8B).


A
W
X
Y
Z
Yoruba
Han
Himalaya 1
Himalaya 2
W
X
Y
Z
Yoruba
Han
Himalaya 1
Himalaya 2
−0.050
−0.025
0.000
0.025
0.050
D(Yoruba, Han; high-altitude Himalaya 1, high-altitude Himalaya 2) ± SE
−20
−10
0
10
20
Z-score

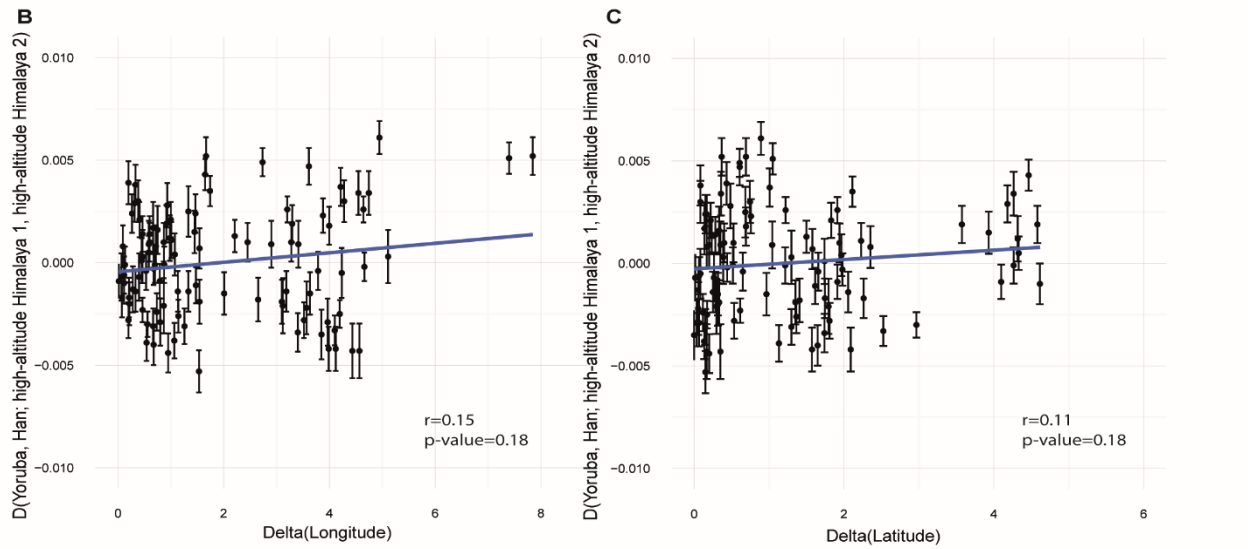


Figure 2. 8 Statistical analysis of genetic sharing between pairs of modern high altitude Himalayan populations and Han individuals. A. The plot shows the results for the *D*-statistic test (x-axis) in the form *D*(Yoruba, Han; high altitude Himalaya1, high altitude Himalaya Himalaya2). Left: value of the *D*-statistic; right: associated Z-score for statistical significance. Values of *D*-statistics departing from zero and with Z-score in the extreme tail of the distribution (vertical dashed line in the plot) are indicative of more genetic sharing of one of the Himalayan populations with Han. B. Correlation plot of genetic and spatial distance between pairs of modern high altitude Himalayan populations. The plot shows the results for *D*-statistic test in the form of *D*(Yoruba, Han; high altitude Himalaya1, high altitude Himalaya2) (y-axis) compared with the difference in longitude (A) or latitude (B) for pairs of Himalayan populations (x-axis). The correlation coefficient (r) and the p-value are the results of the Mantel test between genetic distance (*D*-statistics values) and spatial distance (longitude and latitude).

2.3.2 Complex demographic history in the Himalayas

The Himalayan region separates South Asia from East Asia and is ideal to study admixture between highlanders residing in the Himalayas and their neighbouring lowland populations. To test this, I studied gene flow and admixture between Himalayan and nearby populations through three approaches: *f3*-statistics, ALDER and TreeMix. All the tests provide evidence of admixture between Himalayan and other populations (Figure 2.9, 2.10). Overall, Himalayan populations are characterized by gene flow within the region and with neighbouring populations from South and East Asia. The *f3*-statistics and ALDER show significant admixture events between the Nepalese, North Indians, Tibetans from China, South Asians, Middle Eastern and European populations (Figure 2.9). ALDER also detected extra, although limited, admixture events between the Bhutanese and populations from South and East Asia around 800 and 900 years ago. Furthermore, Chetri, Majhi, Newar, Dhimal, Bodo and Lhasa show gene flow from Europe and the Middle East that might be attributed to the presence of these western components as part of the Ancestral North Indian component in South Asians (71, 160, 172, 194). Chetri, Bodo, Majhi and Dhimal show a signature of admixture dated to between 1000 and 200 years ago. Newar and Lhasa display older signatures of gene flow dated between 1000 and 2000 years ago (Figure 2.9).

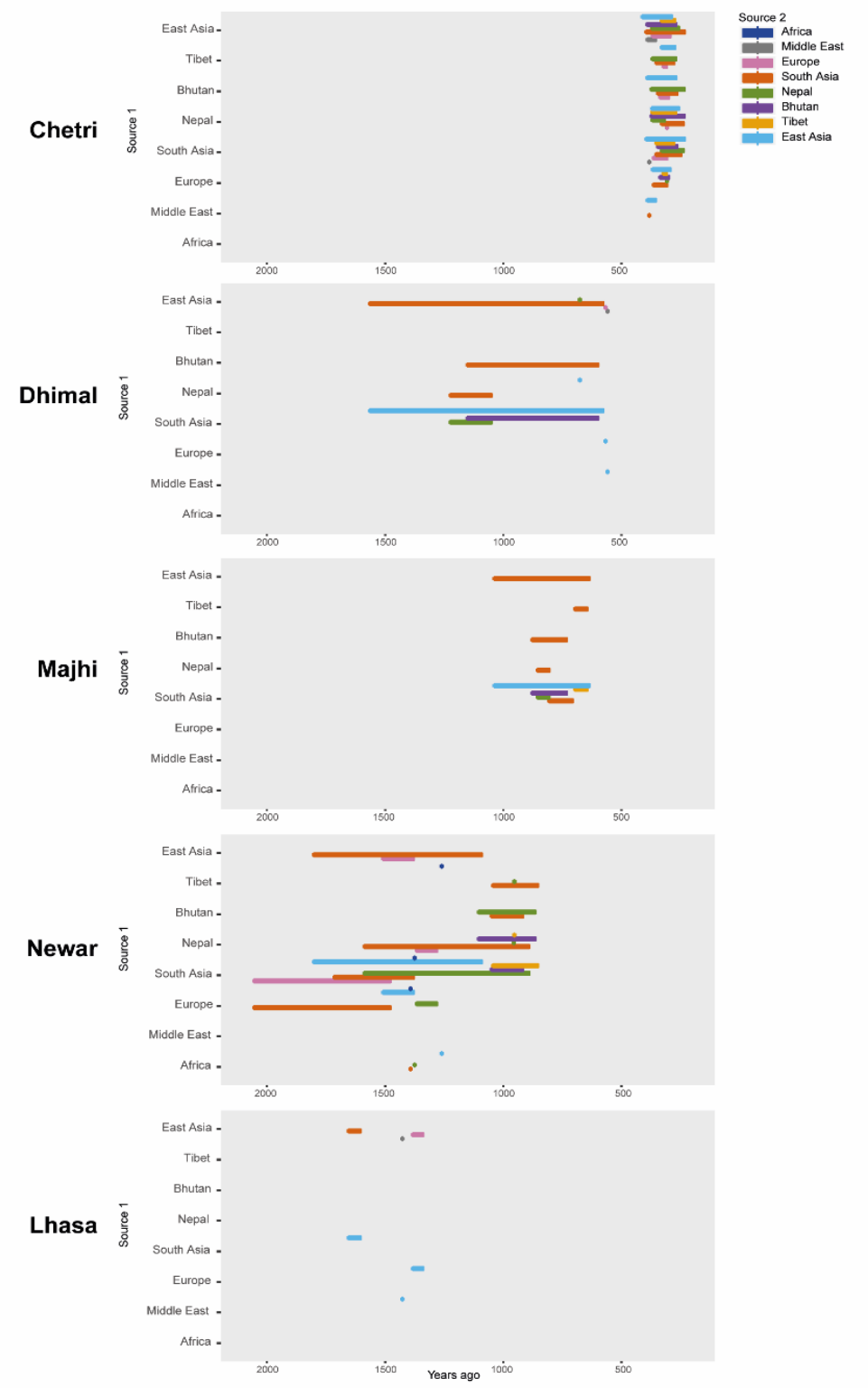


Figure 2. 9 Admixture history of five Himalayan populations. The five populations, each named on the left, could be modelled as a mixture between different source populations from two regions. One of these is shown on the vertical axis, while the second is indicated by the colour of the horizontal bar. The position of this bar represents the inferred time of admixture, and the length in time of these admixture events, according to the scale on the horizontal axis. Thus, the Chetri, for example, can be modelled as a mixture of a large number of Asian and European pairs of populations, occurring around 200-400 years ago.

TreeMix analysis shows long branches for the Toto, Mönpa and Chepang populations in agreement with the genetic drift patterns (Figure 2.10). This is supported by the lack of detectable admixture events for these populations with *f3*-statistics and only a few significant results for Toto with ALDER, showing an admixture event around 600-800 years ago with Chinese and Indian populations. Migration edges involving populations from South and East Asia are detectable (Figure 2.10).

The genetic affinity between the Himalayan populations and five ancient Eurasian genomes was examined using the *f3*-outgroup statistics. Himalayans show greater affinity to Eurasian hunter-gatherers (MA-1, a 24,000-year-old Upper Palaeolithic Siberian), and the related Bronze Age Yamnaya, than to European farmers (5,500-4,800 years ago; Figure 2.11A, C and D) or European hunter-gatherers (La Braña, 7,000 years ago; Figure 2.11B), like other South and East Asian populations. I also explored the affinity of Himalayan populations by comparing them with the 45,000-year-old Upper Palaeolithic hunter-gatherer (Ust'-Ishim) and each of MA-1, La Braña, or Yamnaya. Himalayan individuals cluster together with other East Asian populations and show equal distance from Ust'-Ishim and the other ancient genomes, probably because Ust'-Ishim belongs to a much earlier period of time.

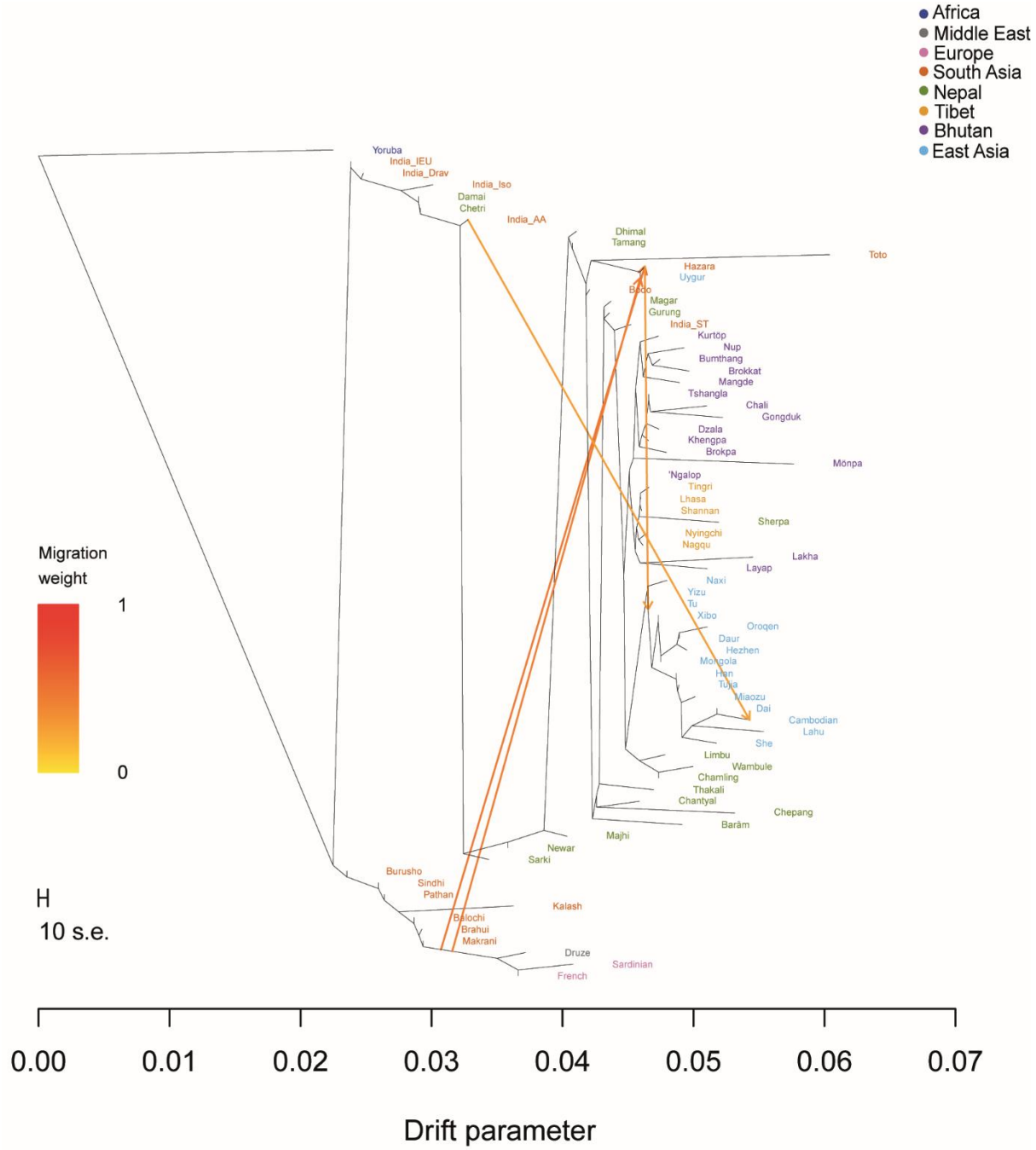


Figure 2. 10 TreeMix results from the worldwide dataset. The tree displays phylogenetic relationships and migration edges between the Himalayan populations and other worldwide populations. The x-axis represents the amount of genetic drift and migration edges are represented by arrows. The tree shows long branches for Toto, Mönpa and Chepang in agreement with their strong genetic drift patterns. Migration edges involving populations from South Asia, East Asia and Nepal are detectable. Dhimal and Bodo are characterised by migration events from South Asia (India and Pakistan).

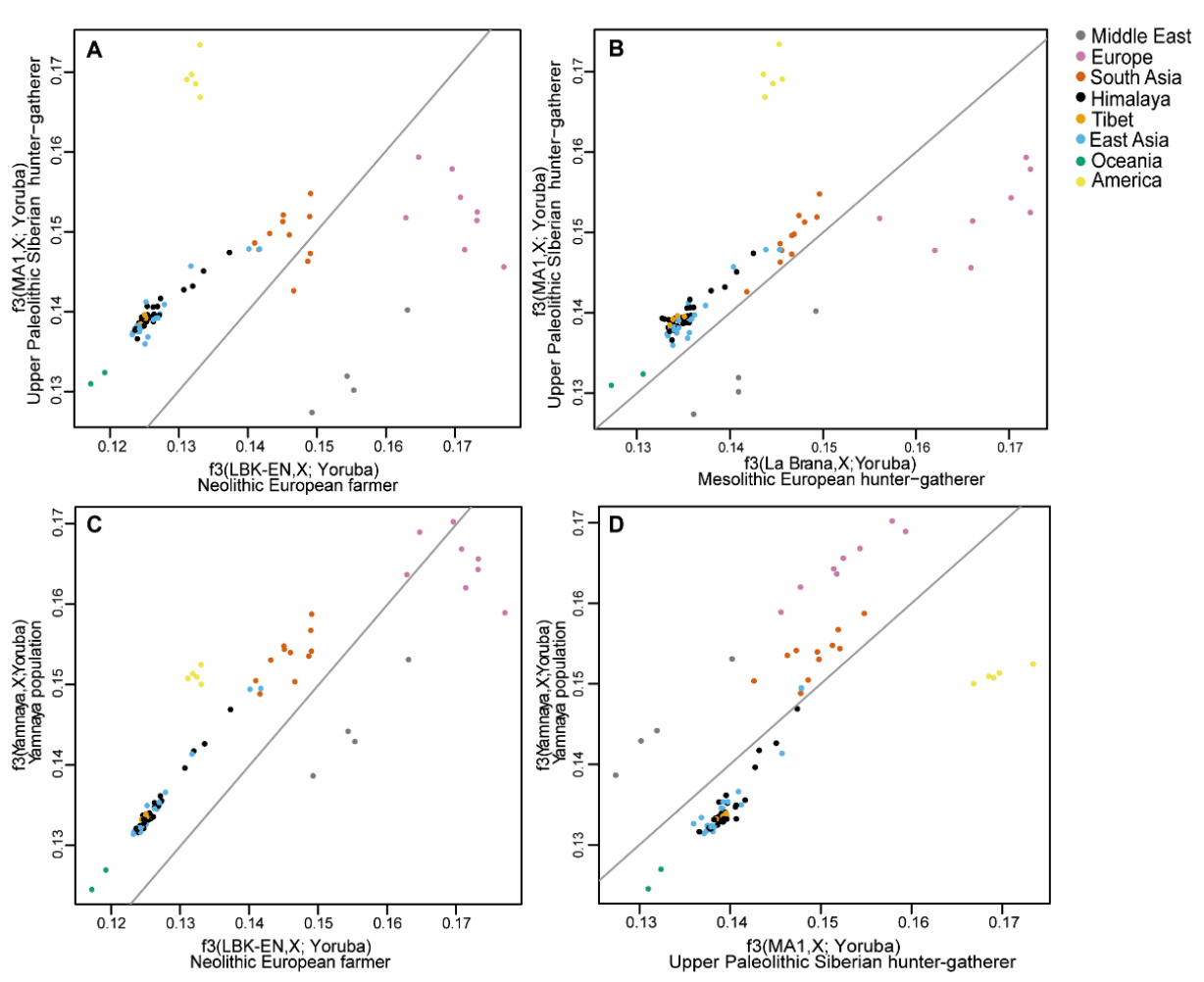


Figure 2. 11 Relative genetic similarity of the Himalayan region and other populations to four ancient DNA samples. A.-D. Each plot shows a comparison between two ancient samples, and equal similarity is represented by the grey line. Each dot represents a present-day population. Thus, section A shows that the Himalayan region populations are more similar to the Upper Palaeolithic Siberian hunter-gatherer than to the Neolithic European farmer.

I also investigated genetic affinity between modern Himalayan populations and five ancient Himalayans (3,150-1,250 years old) from Nepal. The ancient individuals cluster together with modern Himalayan populations in a worldwide PCA (Figure 2.12), and the *f3*-outgroup statistics show modern high altitude populations have the closest affinity with these ancient Himalayans, suggesting that these ancient individuals could represent a proxy for the first populations residing in the region.

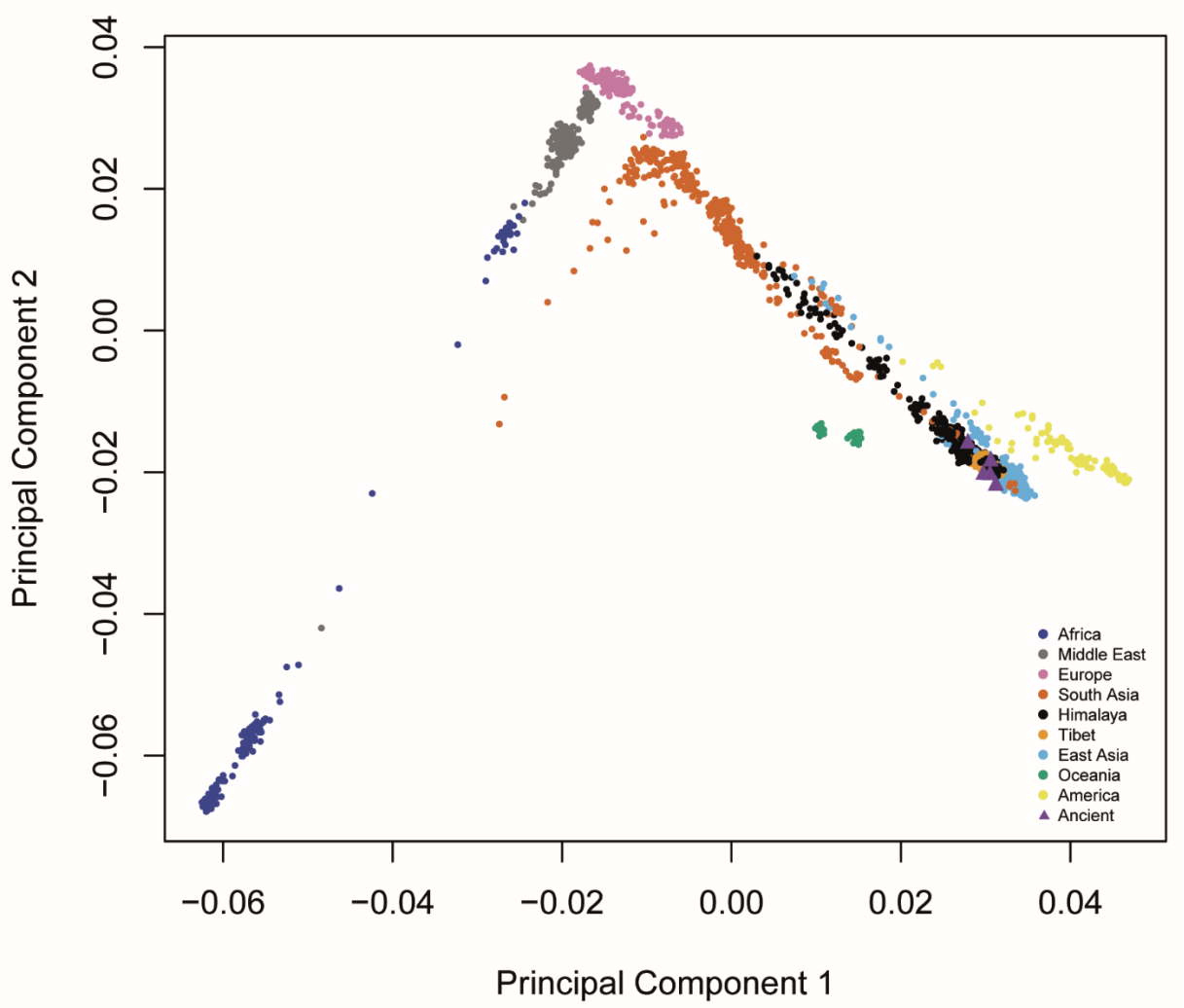


Figure 2. 12 Principal component analysis of the world dataset with ancient Himalayans projected onto the plot. Ancient Himalayan samples (purple) cluster together with modern Himalayan populations.

Finally, I explored the genetic affinity of Himalayan samples with the archaic genomes of Denisovans and Neanderthals (35), and found that they show a similar sharing pattern with Denisovans and Neanderthals as the other South and East Asian populations (Figure 2.13). Individuals belonging to four Nepalese (Dumi, Sunwar, Sherpa, Sarki, Nachiring), one Cambodian, and three Chinese populations show the highest Denisovan sharing (after populations from Australia and Papua New Guinea), but these values are not significantly greater than other South and East Asian populations.

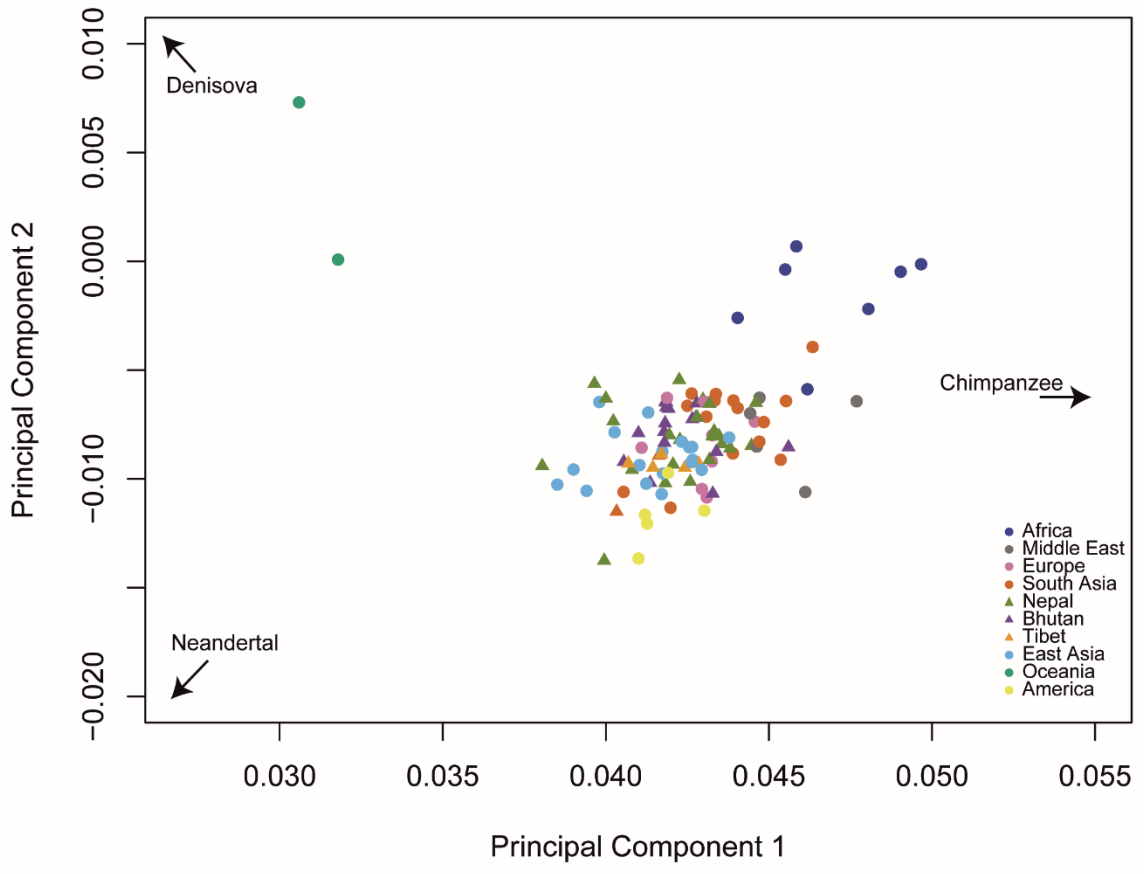


Figure 2. 13 Principal component analysis of modern human populations, Denisova and Neanderthal. The plot shows the projection of modern human samples onto principal components calculated using Denisova, Neanderthal and Chimpanzee. Himalayan populations show a similar pattern of genetic sharing with Denisova to the other South-East Asian populations.

2.3.3 Signatures of adaptation in the Himalayan region

Due to the harsh environment in which many individuals reside, I searched for variants under positive selection within Himalayan populations living at high altitudes, using four approaches: 1) Genome-wide Spearman's correlation between derived allele frequency and altitude; 2) EMMAX, a genome-wide statistical test for association between SNP frequency and altitude that accounts for population substructure (195); 3) the Population Branch Statistic (PBS) which identifies SNPs with unusually high F_{ST} values between high- and low altitude samples, compared with an outgroup population (99); and 4) BayEnv v2, a Bayesian framework for specifically testing association between allele frequency and environmental variables, such as altitude (183, 184).

Genome-wide Spearman's correlations pinpointed 75 derived alleles with frequencies that correlated significantly with altitude (Spearman's *rho* >0.72) (Figure 2.14A). The EMMAX analysis showed that 99.98% of the variance was explained by the kinship matrix, and identified 56 variants where the observed allele frequency nevertheless diverged significantly from the expected frequency (Figure 2.14C). The PBS analysis highlighted 117 variants under possible selection for the derived allele, including ones in regions such as *EPAS1* and Disrupted in Schizophrenia 1 *(DISC1)* previously identified by Tibetan exome sequencing (99) (Figure 2.14E).

Twelve candidate variants lying in three different genomic regions overlap between these first three approaches (Figure 2.14G). Ten of them lie on chromosome 2 in a ~330-kb genomic region that includes *EPAS1*, of which two are of potential functional significance. These are rs1868092, downstream of *EPAS1* in a promoter-flanking region which has previously been associated with high altitude adaptation and shown to be a single-tissue eQTL in whole blood (196, 197), and rs982414, an intronic variant ~231 kb downstream of *EPAS1*, which has been associated with hemoglobin concentration in Tibetans (103). Furthermore, rs12986653, a variant in ATPase H+ Transporting V1 Subunit E2 (*ATP6V1E2*) which falls in a CTCF binding site, shows single-tissue eQTLs associated with the *ATP6V1E2,* CXXC repeat containing interactor of PDZ3 domain *(CRIPT)* and Transmembrane protein 247 *(TMEM247)* genes (198) and has a high CADD

score of 20.6. The second region overlapping between the three methods is on chromosome 1 and includes the eleventh candidate SNP, rs6679627, an intronic variant in the Tripartite Motif Containing 67 (*TRIM67*) near Egl-9 family hypoxia inducible factor 1 (*EGLN1)*, which has previously been associated with high altitude adaptation in Tibetans (126, 199). The third overlapping region is on chromosome 5 and includes a SNP, rs4702062, which shows a strong EMMAX signal together with strong positive correlation with high altitude for the ancestral allele (between 80% and 97% in high altitude populations, compared with a maximum frequency ≤64% in 1000 Genomes Project populations) and has a CADD score of 12.9 (Figure 2.14F). This variant is also in high linkage disequilibrium LD (r^2 = 0.87) with another nearby variant, rs844335, that was picked up by PBS because of its high derived allele frequency, between 80% and 97% in high altitude populations, compared with a maximum frequency ≤61% in 1000 Genomes Project populations. rs4702062 lies in an intergenic region upstream of the ANKH inorganic pyrophosphate transport regulator (*ANKH*) gene on chromosome 5, while rs844335 lies within an open chromatin region nearby, and is also in LD (r^2 = 0.73) with a third variant, rs1550825, that lies in a transcription factor binding site. The *ANKH* gene codes for a transporter that regulates the passage of inorganic phosphate through the cell and contains two hypoxia-responsive elements (HREs) in proximity to its promoter region, and thus its expression is regulated by hypoxic factors (HIFs) (200).

Combining the p-values from the first three methods provides a concise way to merge their findings, although not a measure of the type-1 error rate because the tests are not completely independent. This approach identified 398 variants with Bonferroni-adjusted p-value <0.01 (Figure 2.14G). The fourth method, BayEnv v2, could not be included in this combined p-value analysis as it used an LD-pruned subset of the SNPs. The strongest signals of selection from this last analysis, with multiple significant SNPs in each, included the three regions surrounding *EPAS1, EGLN1* and *ANKH* discussed above, and also a region near the major histocompatibility complex. The *EPAS1, EGLN1* and *HLA-DQB1* regions were also reported as associated with high altitude adaptation in a previous genome-wide association study between Tibetans and Han Chinese using a linear mixed model approach comparable to EMMAX (201). Multiple significant SNPs lying in these regions present single-tissue eQTLs and high CADD scores. An additional six regions with two or more significant SNPs stood out in the combined p-value analysis, surrounding the

RP11-384F7.2, Zinc finger protein 532 *(ZNF532)*, Collagen type IV alpha 4 chain *(COL4A4)*, Solute carrier family 52 member 3 *(SLC52A3)*, Megakaryoblastic leukemia (translocation) 1 *(MKL1)* and Growth factor receptor bound protein 2 (*GRB2)* genes (Figure 2.14G). The results from BayEnv v2 were then used for further validation of the candidate genes highlighted above. It pinpointed 503 variants falling into the category "Decisive" (Bayes Factor (BF) >100, $\log_{10}$ (BF) >2). Eight of the top ten candidate regions discussed above overlapped with the "Decisive" ones: *EGLN1*, *EPAS1*, *COL4A4*, *RP11-384F7.2*, *ANKH*, *HLA-DQB1/HLA-DPB1*, *ZNF532* and *SLC52A3* while the *MKL1* and *GBR2* regions overlapped strong ($10<BF<100$, $1<\log_{10}(BF)<2$) and substantial ($3.2< BF<10$, $0.5 < \log_{10}(BF)<1$) candidates, respectively (Table 2.3).

Candidate gene	Cluster of selected SNPs: GRCh37 coordinates	Number of SNPs in cluster	Top SNP: Combined p-value	Top SNP:	Top SNP frequency high altitude populations	Top SNP frequency East Asian populations	Allele under selection**	eQTLs	Comments
EPAS1	2:46468276-46852033	26	1.83E-27	rs4953359	68%	14%	D	9	Known high altitude selection signal
EGLN1	1:231204794-231897303	21	3.45E-14	rs6655954	58%*	27%*	A	10	Known high altitude selection signal
HLA-DQB1/ HLA-DPB1	6:32582075-33175824	15	1.66E-11	rs10484569	77%	39%	D	6	Known high altitude selection signal
ANKH	5:14908578-149285033	5	2.63E-11	rs4702062	84%*	64%*	A	-	Novel
RP11-384F7.2 AC068633.1	3:117427214-118549344	5	6.68E-09	rs1081896	79%*	64%*	A	2	Novel
ZNF532	18:56562356-56648324	3	8.37E-09	rs3826597	18%	8%	D	-	Novel
COL4A4	2:227770592-227922321	8	1.19E-07	rs3769641	46%	19%	D	1	Novel
SLC52A3	20:744415-745963	2	1.22E-06	rs3746804	63%	25%	D	1	Novel
MKL1	22:40827319-40905072	3	3.97E-06	rs17001997	47%	24%	D	3	Novel
GRB2	17:73326965-73374945	4	3.66E-05	rs4789182	94%	83%	D	4	Novel

NOTE. * Frequency of ancestral allele ** A = Ancestral, D = Derived. Bold candidate genes: Decisive ($\log_{10}$Bayes Factor >2) candidates from BayEnv v2

Table 2. 3 Genomic regions showing the strongest signals of positive selection in the Himalayan populations

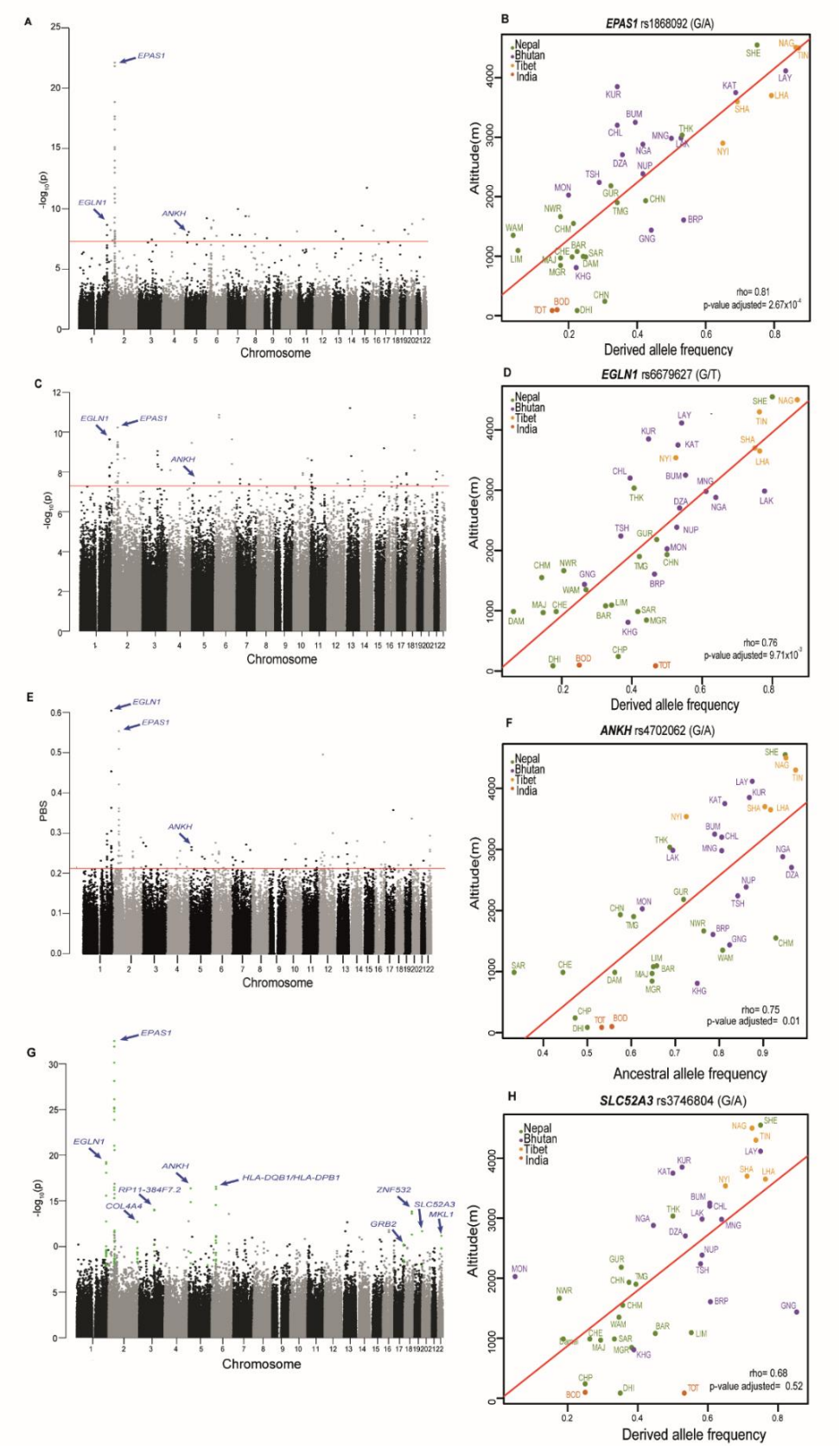


Figure 2. 14 Signals of positive selection (adaptation) in the Himalayan populations. A, C, E, G. Manhattan plots showing a measure of confidence in selection (vertical axis) plotted against genomic coordinate

(horizontal axis). Each dot represents a SNP. A. Spearman's correlation between derived allele frequency and altitude. C. EMMAX. E. Population Branch Statistics. G. Fisher's combined p-value from these three tests. B, D, F, H. Plots of allele frequency against altitude for four selection candidates. Each dot represents a Himalayan region population.

We highlight further features of these candidate regions. The *SLC52A3* region includes a missense variant (Pro267Leu, rs3746804) with derived allele frequency >70% in most high altitude populations compared with a maximum frequency ≤35% in the 1000 Genomes Project populations, and a synonymous variant (rs3746807) with overall high derived allele frequency in Himalayan populations (42-100%) compared with a maximum frequency ≤24% in 1000 Genomes Project populations (Figure 2.14, Table 2.3). rs3746804 shows single-tissue eQTLs for *SLC52A3* in lung and skin, and has a CADD score of 13.3. The *COL4A4* region comprises eight SNPs: the top one, rs3769641, lies in a splicing regulatory region within *COL4A4,* and its derived allele frequency is positively correlated with altitude (Spearman's *rho*= 0.70). This region also contains a missense variant (rs3752895) that shows single-tissue eQTLs in brain tissue for the Rhomboid domain containing 1 (*RHBDD1)* gene and a synonymous variant (rs2228557). These two variants show high CADD scores of 17.2 and 16.7, respectively. The *GRB2* region on chromosome 17 shows four intronic SNPs and has previously been associated with hypoxia-induced oxidative stress level at the intestinal mucosal barrier in Tibetans compared with Han Chinese (202) (Figure 2.14G). Two of the four variants in *GRB2*, rs4542691 and rs4789182, show single-tissue eQTLs. The *MKL1* region on chromosome 22 carries three intronic SNPs, and has previously been associated with the regulation of the cellular response to chronic hypoxia in the vasculature of rats. All three variants in *MKL1*, rs2294352, rs6001931 and rs17001997, show single-tissue eQTLs in muscle-skeletal tissue (Table 2.3) (203).

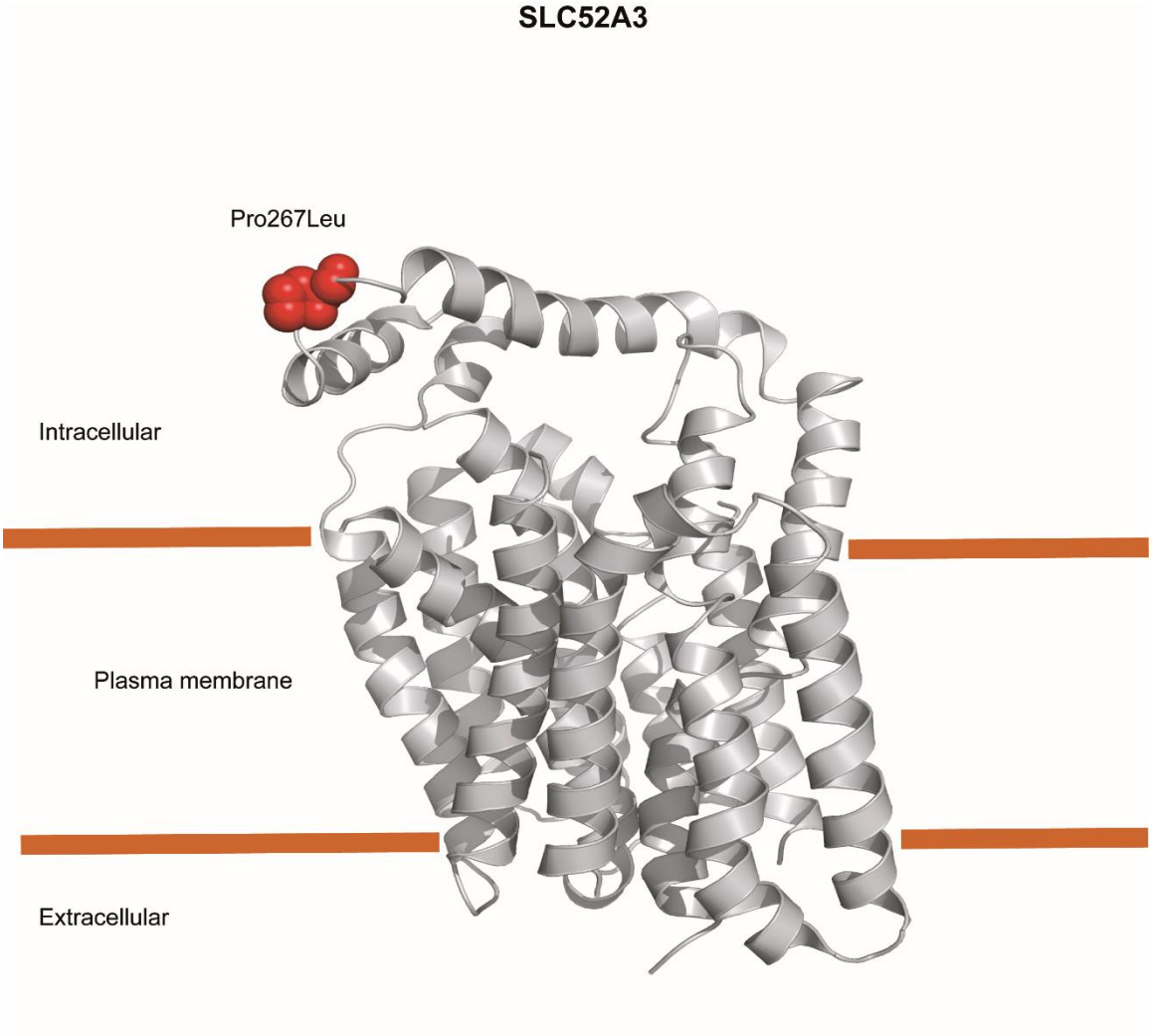


Figure 2. 15 SLC52A3 protein homology model. The plot shows the homology model generated for the SLC52A3 protein and its cellular transmembrane location. The missense variant, rs3746804, is indicated in red (Pro267Leu).

We also examined the allele frequencies of the top SNPs in our ten candidate regions in the five ancient Himalayan genomes, and compared them with the allele frequencies in present-day Himalayans. Six variants in the *EPAS1* region and eleven in the *EGLN1* region show high derived allele frequencies in ancient Himalayans (≥ 0.60). The missense variant rs3746804 in the *SLC52A3* locus also shows a high derived allele frequency of 0.67 in the ancient Himalayans. Variants in *COL4A4, ANKH*, *RP11-384F7.2/AC068633.1* and *HLA-DBP1/DBP2* show derived allele frequencies in the ancient Himalayans of 0.56 -1.00, while two variants, rs4542691 and rs4789182, in the *GRB2* locus show a derived allele frequency of 100% in the ancient samples. Finally, rs3826597 in *ZNF532* region show a derived allele frequency of 0.95 in the ancient Himalayans.

None of the top selection candidate regions, apart from *EPAS1* (50, 104), show signatures of adaptive introgression from archaic Denisovans or Neanderthals according to published introgression maps (204).

We also generated a protein homology model for SLC52A3, and investigated the position of the missense variant, rs3746804. The SLC52A3 structure resembles that of a glucose transporter and rs3746804 is predicted to lie in an exposed intracellular region which could act as an interaction surface for the intracellular environment (205) (Fig 2.15). We finally generated protein-protein interaction networks for our top ten protein candidates. *EPAS1, EGLN1, COL4A4* and *GRB2* were predicted to be part of the same network. Prostaglandin I2 synthase (*PTGIS)* and Vitamin D receptor (*VDR)*, suggested previously by Hu *et al.* to be under selection for high altitude adaptation, are also predicted to be in the same protein-protein interaction network (50).

2.4 Discussion

This study represents the most comprehensive survey thus far of genetic variation in the Himalayan region, aiming to elucidate the genetic ancestry of these populations, including their demographic histories, and the genetic adaptations they have undergone in order to survive in the varied and challenging environments present in the region.

2.4.1 Population structure and demography

In the broadest sense, all the Himalayan populations share ancestry with their geographical neighbours in South and East Asia, reflecting the common pattern of the distribution of human genetic diversity dominated by geography (Figure 2.2). Within this framework, I nevertheless detect an ancestral component that is abundant in most Himalayans, but rare elsewhere (Figure 2.2C), pointing to shared ancestry for these populations, a conclusion reinforced by their similar patterns of shared genetic drift with non-Himalayan ancient samples (Figure 2.12). At finer resolution, I see evidence for both substructure reflecting geography within the Himalayan region, and extreme drift leading to single populations forming outliers in the PCA (Figure 2.3A,B) or specific components in ADMIXTURE analysis (Figure 2.3C). The most striking example is provided by the Toto from North India, an isolated tribal group with the lowest genetic diversity among all the Himalayan populations examined here, indicated by the smallest long-term Ne (Figure 2.5A), and a reported census size of 321 in 1951 (206), although their numbers have subsequently increased. Despite this extreme substructure, shared common ancestry among the high altitude populations (Figure 2.3C and Figure 2.7) can be detected, and the Nepalese in general are distinguished from the Bhutanese and Tibetans (Figure 2.3C) and they also cluster separately (Figure 2.7). In a worldwide context, they share an ancestral component with South Asians (Figure 2.2C). On the other hand, the Tibetans do not show detectable population substructure, probably due to a much more recent split in comparison with the other populations (Figures 2.32 C and 2.5). The genetic similarity between the high altitude populations, including Tibetans, Sherpa and Bhutanese, is also supported by their clustering together on the phylogenetic tree, the PCA generated from the co-ancestry matrix generated by fineSTRUCTURE

(Figure 2.7A), the lack of statistical significance for most of the *D*-statistics tests (Yoruba, Han; high altitude Himalayan 1, high altitude Himalayan 2), and the absence of correlation between the increased genetic affinity to lowland East Asians and the spatial location of the Himalayan populations (Figure 2.8). Together, these results suggest the presence of a single ancestral population carrying advantageous variants for high altitude adaptation that separated from lowland East Asians, and then spread and diverged into different populations across the Himalayan region. Genetic drift and admixture with other Himalayan, South and East Asian populations can explain the widespread distribution of the selected *EPAS1* haplotype at lower frequencies in populations at lower altitudes (127), and the altitude clines in the other selection candidates (Figure 2.14). Our findings suggest a recent split (only a proxy for population differentiation given the limitations of the method applied) between Tibetans, Sherpa and, possibly, other high altitude populations, rather than the Tibetans being a mixture of Sherpa and Han Chinese (113, 117). Whole-genome sequences from multiple high altitude populations will provide better estimates of such divergence times and a more detailed demographic history of the region.

Himalayan populations show signatures of recent admixture events, mainly with South and East Asian populations as well as within the Himalayan region itself. Newar and Lhasa show the oldest signature of admixture, dated to between 2,000 and 1,000 years ago. Majhi and Dhimal display signatures of admixture within the last 1,000 years. Chetri and Bodo show the most recent admixture events, between 500 and 200 years ago (Figure 2.9). The comparison between the genetic tree and the linguistic association of each Himalayan population highlights the agreement between genetic and linguistic sub-divisions, in particular in the Bhutanese and Tibetan populations. Nepalese populations show more variability, with genetic sub-clusters of populations belonging to different linguistic affiliations (Figure 2.7B). Modern high altitude Himalayans show genetic affinity with ancient genomes from the same region (Figure 2.12), providing additional support for the idea of an ancient high altitude population that spread across the Himalayan region and subsequently diverged into several of the present-day populations. Furthermore, Himalayan populations show a similar pattern of allele sharing with Denisovans as other South-East Asian populations (Figure 2.13). Overall, geographical isolation, genetic drift, admixture with neighbouring populations and linguistic

subdivision played important roles in shaping the genetic variability we see in the Himalayan region today.

2.4.2 High altitude adaptation

The harsh environment at high altitude due to increased ultraviolet radiation, hypobaria and hypoxemia is inescapable, so it is expected to have triggered physiological and genetic adaptations including modifications in the cellular responses of the humans who settled there. Genomic scans for positive selection in Tibetans have previously implicated several genes as candidates for high altitude adaptation, especially an extended *EPAS1* haplotype (99, 119, 207-209) that arose by introgression from Denisovans (104), and is widespread in the region (127). Positive selection scans can easily be confounded by population structure, and although simple correlations of SNP frequency with altitude replicated several of the candidates reported in previous studies (Figure 2.14), including those near *EPAS1*, *DISC1* and *ATP6V1E2* which are highly differentiated between lowland Han and Tibetans (99, 119), additional analyses better suited to sub-structured populations only confirmed a subset of these. The signal on chromosome 2 is particularly strong and includes a ~330-kb region encompassing *EPAS1, ATP6V1E2* and *PIGF/CRIPT* (Figure 2.14). An expected signal of selection from *EGLN1* was observed via nearby variants in *TRIM67* and *TSNAX-DISC1* (Figure 2.14) (210, 211). A novel signal of selection was found in the region upstream of *ANKH* on chromosome 5 (Figure 2.14). This region shows extended LD, but the variant driving the selection could not be identified by our analysis. Nevertheless, *ANKH* is itself a strong candidate because it is involved in the regulation of the transportation of inorganic phosphate and its expression is regulated by HIF-2α (*EPAS1*) and HIF-1α (200, 212). ANKH is essential for maintaining cellular function and bone mineralization, and its concentration plays a central role in several metabolic pathways (213).

In order to maximize the power to identify the additional selection candidates, I calculated combined p-values for three different statistics applied to our dataset, and then further validated these candidate genomic regions using a fourth statistic, BayEnv2. (Figure 2.15). Some of these additional variants may play important roles in the hypoxic environment, contributing to physiological responses to hypoxia. *COL4A4* encodes one of

the subunits of collagen type IV, which is an essential component of basement membranes, and plays an important role in angiogenesis. Hypoxia exposure triggers vasoconstriction which requires structural remodelling of arterial vessels, especially in lung, and collagen metabolism is required for this process (214) (215). *GRB2* is involved in the regulation of reactive oxygen species (ROS) production in hypoxic environments and it has been shown that, in Tibetans, downregulation of its expression reduces ROS damage and improves glucose and fat metabolism in intestinal tissues (202). *MKL1* encodes a myocardin-related transcription factor and is involved in smooth muscle cell differentiation (216). Down-regulation of *MKL1* reduces the pulmonary arterial pressure in response to chronic hypoxia and regulates vascular remodelling in rats (203). *SLC52A3* encodes a transporter of riboflavin, a vitamin that modulates fatty acid and amino acid metabolism and reduces cellular oxidative stress (217). Riboflavin supplementation of the diets of mice improves their energetic metabolism under acute hypoxia. Thus, increased riboflavin could be effective in counteracting the alteration of human metabolism in hypoxic conditions (218). SLC52A3 is a transmembrane protein and the homology-based protein model we generated resembles the structure of a glucose transporter; our top candidate variant, rs3746804 (Pro267Leu), lies in the intracellular environment in a possible interaction region of the protein surface (Figure 2.15). This selection signal seems to be specific for Himalayan populations and could be related to the diet and environment, where efficient intake of riboflavin at high altitude would be advantageous (219). Two out of three additional candidates for high altitude adaptation (*PTGIS* and *VDR*) suggested by Hu *et al.* are predicted to be in the same protein-protein interaction pathway as some of our candidates, *COL4A4* and *GRB2*, and linked with other genes (*EPAS1, EGLN1, HIF-1α, VHL*) involved in the hypoxic (50). *ANKH* has also been reported as a candidate for high altitude adaptation in Tibetan pigs (220, 221).

Thus, of the top ten selected candidate regions (seven novel) highlighted by our work, four are members of the most relevant protein-protein interaction network and three others have known functions relevant to high altitude adaptation: findings that are very unlikely due to chance. Furthermore, despite the strong ascertainment bias of the SNPs included on SNP-chips, variants lying in our top ten candidate regions are associated with single-tissue eQTLs and present high CADD scores, suggesting their possible importance in gene regulation and expression. The presence of high derived

allele frequencies of variants in *EGLN1*, *EPAS1*, *SLC52A3* and *GRB2* loci in ancient Himalayans also supports our hypothesis that these candidates may be under selection and important for high altitude adaptation. According to available introgression maps (204), none of the top selected candidate regions, apart from the well-known *EPAS1* intronic region (50, 104), show signatures of adaptive introgression from Denisovans or Neanderthals. In all cases, high-coverage whole-genome sequences and comparisons with other species that have adapted to similar environments should help to identify or confirm the key causal variants and suggest strategies for functional follow-up.

In conclusion, the current analyses have established the broad features of Himalayan genetic variation: a South or East Asian substrate influenced by local differentiation and mixing in ways that are now understood in outline, including extreme genetic drift in several populations. This work has provided a comprehensive dataset from the region for the community to use in future studies. In addition, there is evidence for early strong genetic adaptation to high altitude living followed by spread of the adapted population. Future functional investigations will allow these phenomena to be understood in more detail.

3. Fine-scale demographic history of Himalayan populations using whole-genome sequencing data

3.1 Introduction

In the previous chapter, I presented the benefits of, and insights from, SNP-genotype data for the study of Himalayan population structure and their relationship with worldwide populations and ancient genomes. However, SNP genotype arrays have limitations compared with whole genome sequencing, which can discover SNPs in an unbiased way, as well as discovering other types of variants, such as CNVs, SVs and INDELs. Sequence data can also directly identify causal variants rather than tag ones affecting gene regulation and expression and those responsible for specific phenotypes, including diseases and positively selected traits.

To extend the study based on SNP genotypes in the previous chapter, this chapter is focused on analyses performed using a dataset of eighty-seven high coverage genomes (30x coverage) generated with the Illumina HiSeq X Ten platform at the Wellcome Sanger Institute. The aim is to refine the demographic history and explore the possibility of identifying functional variants for known as well as new signals of positive selection in the Himalayan populations. First, the different types of genetic variation, including SNPs, INDELs and CNVs for autosomal and uniparentally regions of the genome are characterized. Then, refined population structure and recent admixture with neighbouring populations are analysed, using both common and rare variants. A more complete picture of population demographic history can be generated using SMC++, MSMC and other methods which require whole genome sequencing data. Next, Neanderthal and Denisovan archaic introgression in the Himalayan individuals is explored using the Introgression-detection method (222). Finally, narrowing down the signals of positive selection to possible causal driver SNPs is explored, both for the genomic regions identified in the previous chapter, and for new variants positively selected in high altitude populations.

3.2 Materials and Methods

3.2.1 Samples and datasets

Eighty-seven male samples were sequenced from 16 Himalayan populations, including four from each of eight populations in Nepal, six in Bhutan, one in North India and 27 from Tibetans in China. Thirteen of these populations were included in the SNP-genotype data described in the previous chapter, plus three new populations we retrieved samples for that are Ghale and Tharu from Nepal or Lhokpu from Bhutan. Seven of these populations live at high altitude of 2500 meters or more above sea level (135). One population, Tharu, lives in the malaria-endemic Terai tropical forest in the foothills of the Himalayas. It has been reported that Tharu individuals have a lower prevalence of malaria than other ethnic groups in the same region, and this could possibly be explained by the particularly high frequency of α thalassemia in the Tharu population (46, 223).

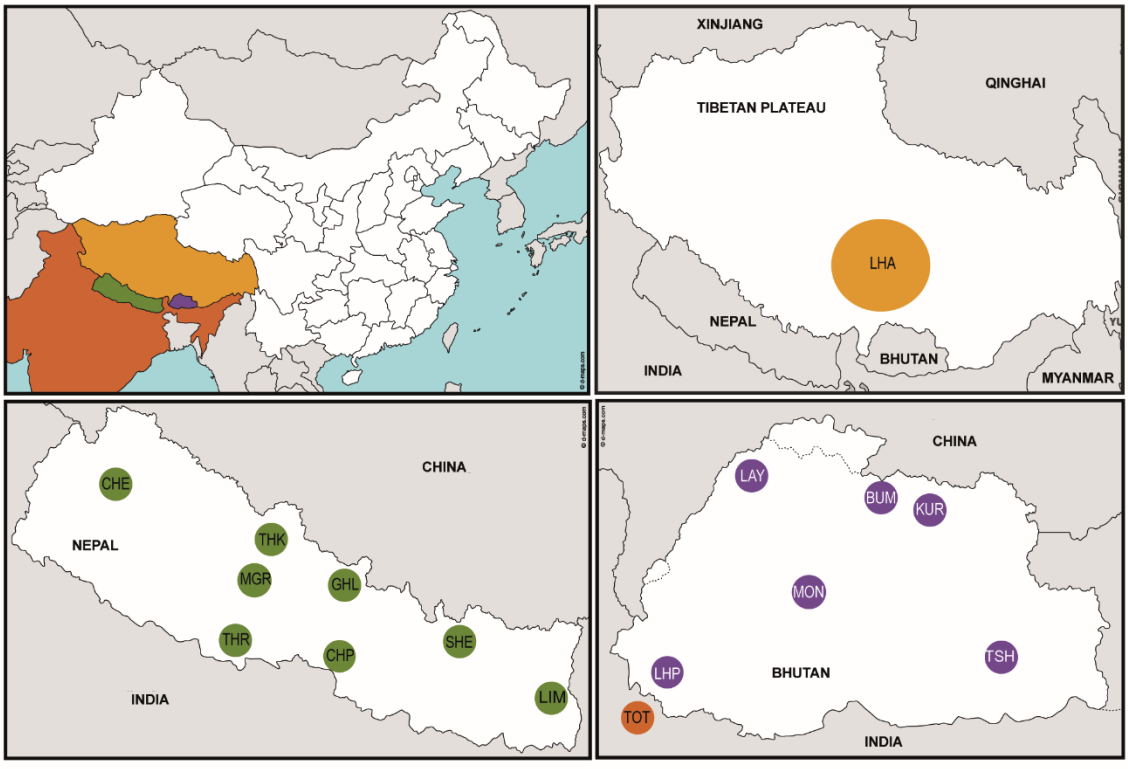


Figure 3. 1 Population samples included in this project. A. Map of South and East Asia, highlighting the four regions where the populations live, with a different colour assigned to each. B. Samples from the Tibetan Plateau. C. Samples from Nepal. D. Samples from Bhutan and India. The circle areas are proportional to the sample sizes. The three letter labels for each population are the same as in the previous chapter; the new populations, Ghale and Tharu from Nepal, and Lhokpu from Bhutan are labelled GHL, THR and LHP.

3.2.2 Variant calling and filtering

3.2.2.1 SNPs and INDELs

3.2.2.1.1 Autosomes

The raw sequence reads were mapped to the human reference assembly GRCh38 (GRCh38_full_analysis_set_plus_decoy_hla.fa) using BWA mem (224), and base quality score recalibration, indel realignment and duplicate removal were performed using the Genome Analysis Toolkit (GATK) by Dr. Shane McCarthy. One Genomic VCF (GVCF) per sample was generated using GATK HaplotypeCaller (225) and then merged into a joint file using the GATK GenotypeGVCFs tool, containing accurate genotype information, genotype likelihoods and annotations for the whole Himalaya dataset. Finally, a set of standard hard filtering parameters for SNPs and INDELs, according to the GATK best practices guidelines, was applied to the joint call set (226, 227). Specifically, the following filtering parameters were used for SNPs (Fig. 3.2):

1. FisherStrand (FS): This is the Phred-scaled p-value estimated using Fisher's Exact Test to identify strand bias in the reads, and values > 60.0 are discarded as they indicate false positive calls.
2. RMSMappingQuality (MQ): This is the root mean square of the mapping quality of the reads across all samples and its value should be < 40.0.
3. MappingQualityRankSumTest (MQRankSum): This is applied to heterozygous calls and it represents the u-based z-approximation of the Mann-Whitney Rank Sum Test for mapping quality (reads with reference bases vs. those with the alternate allele). This value should be < 12.5.
4. QualByDepth (QD): This is the variant confidence divided by the unfiltered depth of non-reference samples and the values have to be < 2.0.
5. ReadPosRankSumTest (ReadPosRankSum): This is applied to heterozygous calls and it is the u-based z-approximation of the Mann-Whitney Rank Sum Test for the distance from the end of the read for reads with alternate allele. The presence of an alternate allele near the ends of reads is indicative of an error. Values < -8.0 were discarded.

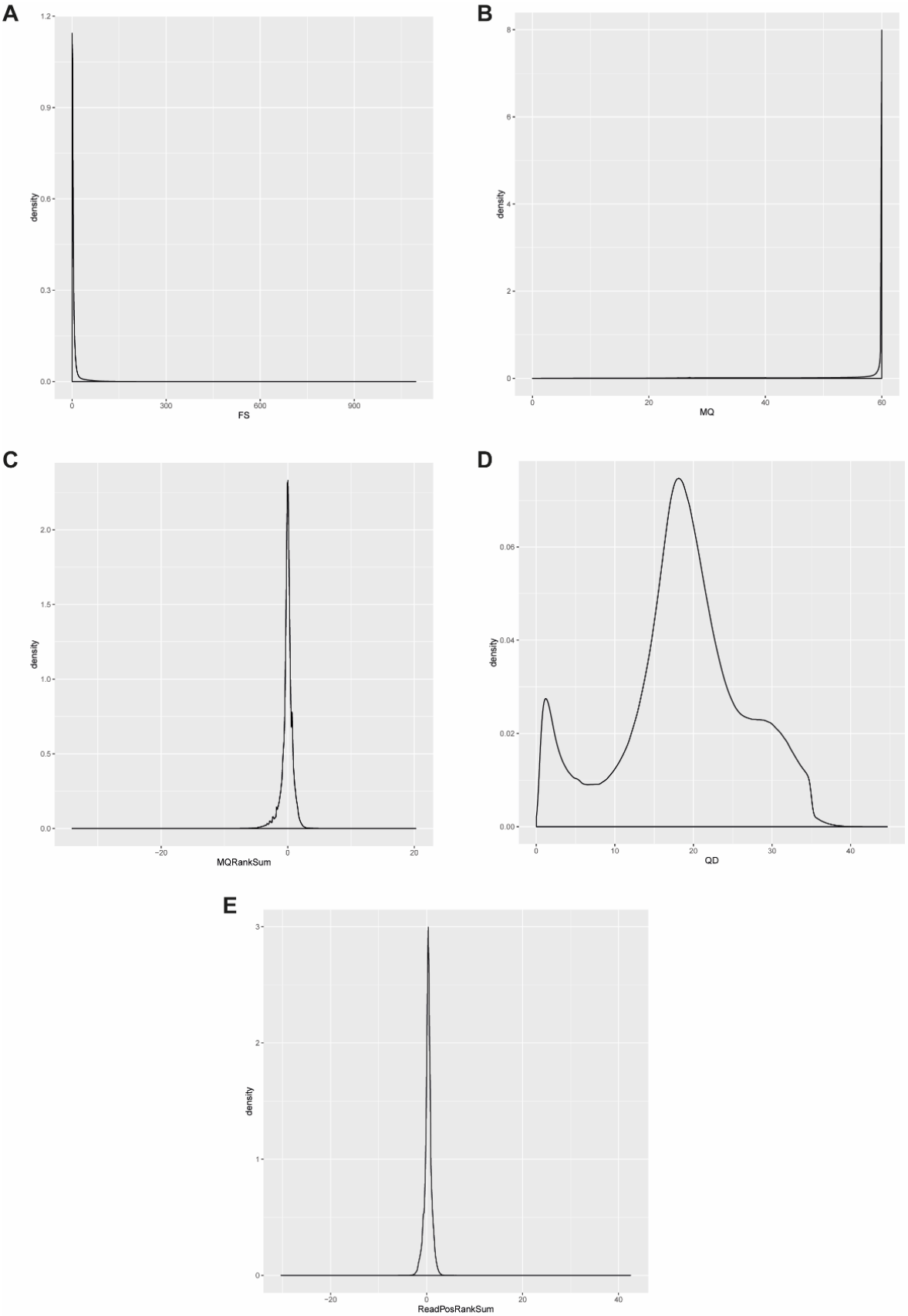


Figure 3. 2 The distribution of SNP parameters after filtering the Himalayan dataset. A. FS, B. MQ, C. MQRankSum, D. QD, E. ReadPosRankSum.

Only three of the above parameters were applied for INDEL filtering (Fig. 3.3):

1. FisherStrand (FS) > 200.0
2. QualByDepth (QD) < 2.0
3. ReadPosRankSumTest (ReadPosRankSum) < -20.0

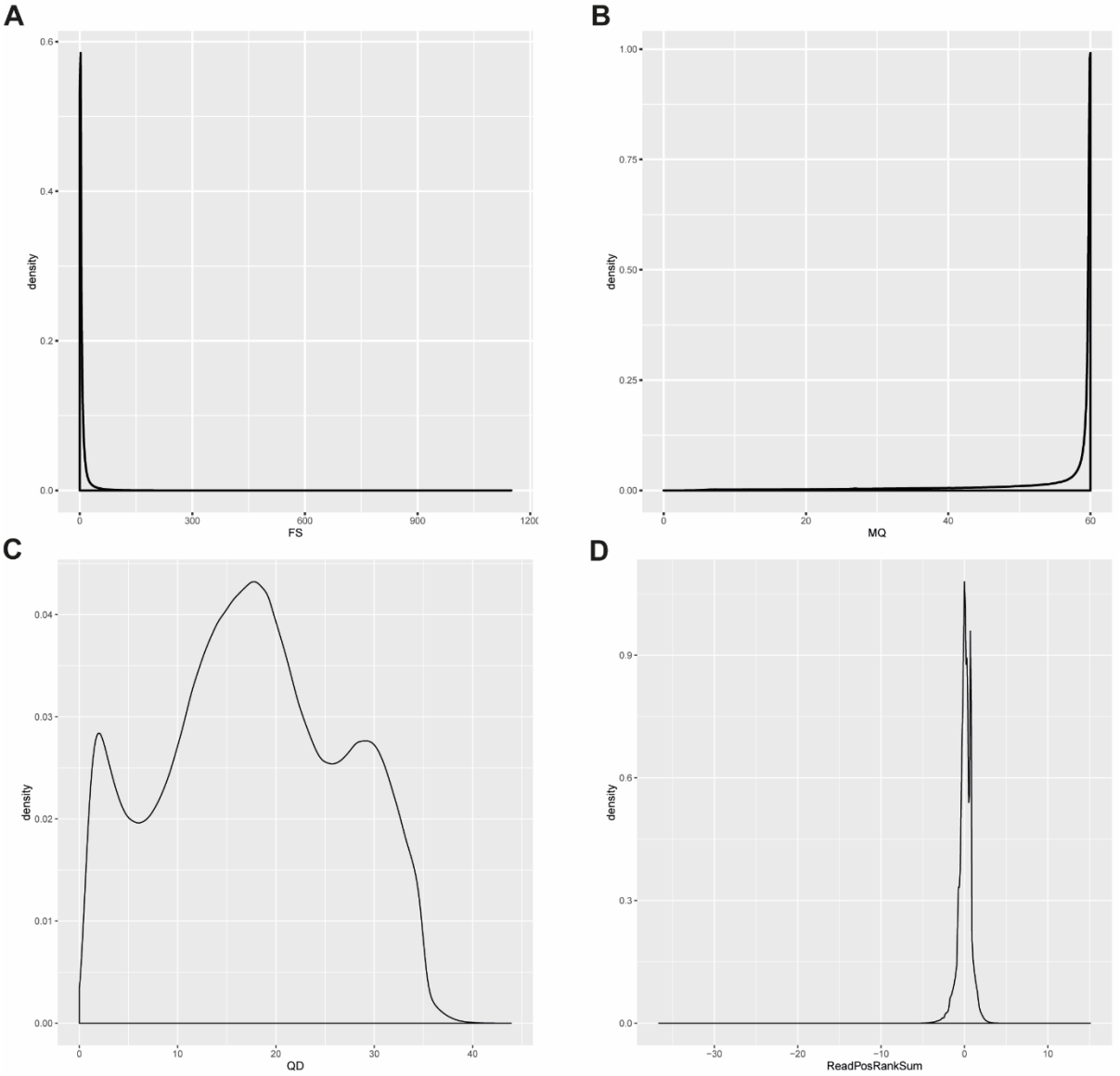


Figure 3. 3 The distribution of INDEL parameters after filtering the Himalayan dataset. A. FS, B. MQ, C. QD, D. ReadPosRankSum.

An additional filter was applied to the SNP dataset: depth (DP) of one third of, or three times greater than the average coverage and genotype quality (GQ) < 30 were also applied using bcftools 1.6 (228). All population-genetic analyses use only sites in the strict masked region of the genome from the 1000 Genomes Project (177, 229): around 78% of the reference sequence. This step has been performed using the GATK

VariantFiltration and SelectVariants tools (226, 227). In total, the call set contains 9,909,760 biallelic SNPs, 28,319 multiallelic SNPs and 1,043,093 INDELs. Finally, for downstream analysis, multiallelic variants, all the sites with missingness greater than 0.2% or Hardy-Weinberg equilibrium violations (hwe > 0.00001) were removed using vcftools 0.1.14 (230), bcftools 1.6 and PLINK 1.92 (155). This resulted in a dataset of 9,901,395 SNPs and 1,021,861 INDELs. Out of these, 9,636,923 SNPs and 991,411 INDELs were autosomal. The functional impacts of SNPs and INDELs were predicted using the Ensembl Variant Effect Predictor (VEP) (54, 189).

3.2.2.1.2 Autosomal SNP LiftOver and filtering

LiftOver is the process of converting genomic coordinates and annotations between assemblies and it is necessary to map all datasets to the same reference build. The GRCh37 assembly was released in 2009, whereas GRCh38 was released at the end of 2013 and the latter assembly has major updates compared to GRCh37 with substantial improvements and better representation of the human genome. However, many of the population genetic and genome-wide association studies still use the GRCh37 assembly, and all archaic sequences (Neanderthals and Denisovans (4, 231)) and other ancient genomes published so far have been mapped onto GRCh37. Some efforts have been made to encourage researchers to use GRCh38 by releasing a LiftOver version of the 1000 Genomes Project phase 3 VCFs and the re-mapped BAMs of the same dataset, for example (229).

The call sets based on GRCh38 were used where possible for this project, but a LiftOver to GRCh37 dataset was also prepared using Picard 2.6.0 LiftoverVcf tool (http://broadinstitute.github.io/picard/) with default parameters for some analyses. Some technical issues were discovered during the merging process and later analyses: around 0.2% (~18,900 SNPs out of ~10 million) of genomic locations shared between the 1000 Genomes Project and the Himalayan dataset showed reference mismatches between GRCh37 and GRCh38, some of which are due to strand flips and others represent genuine difference between the references. For example, the reference for certain positions has been fixed from GRCh37 to GRCh38. These positions were called homozygous for the alternative allele (ALT/ALT) for all the 1000 Genomes Project

individuals in GRCh37, although the ALT allele is the ancestral allelic state. This has been swapped in GRCh38, hence all 1000 Genome Project samples are now called homozygous reference (REF/REF) for these positions. Most of the LiftOver tools available convert genomic coordinates within assemblies, however they are not aware of changes in the actual reference sequence. This can lead to large technical biases in the data as it can create artificial genetic variability. Thus, having taken care of all of these technical issues, I filtered the LiftOver call set by removing all these problematic position before proceeding with the downstream analyses.

3.2.2.1.3 Y chromosome

The call set for the Y chromosome was restricted to the commonly used ~10.3Mb region accessible by NGS (232). First, all of the samples were confirmed to be males by analysing the average read depth across the accessible regions on both X and Y chromosomes using GATK DepthOfCoverage, and found to be similar (Table 3.1).

In order to improve consistency across the different datasets and reduce the number of technical artefacts, three different sets of joint calls from BAM files of Himalayan samples and two reference datasets, the 1000 Genomes Project Phase 3 (1000GP) (66, 232) and the Simons Genome Diversity Project (SGDP) (23) were generated. First, all of the reads from the Y chromosome in the Himalayan BAMs in GRCh38 were extracted using SAMtools 1.6 (233) and the Picard 2.6.0 SamToFastq tool, then were remapped to GRCh37(hs37d5.fa) using BWA 0.7.15 (224). The BAMs mapped GRCh38 for the 1000GP were available from 2017. Hence, three sets of joint calls were generated from the three Y chromosome datasets below:

1. 1000 Genomes Project (1243 males, Y mean coverage ~ 4.7x) and Himalayan samples in assembly GRCh38
2. 1000 Genomes Project (1243 males, Y mean coverage ~ 4.7x) and Himalayan samples in assembly GRCh37
3. Simon Genome Diversity Project (184 males, Y mean coverage ~ 20x) and Himalayan samples in assembly GRCh37

The all-sites calling was performed for all male samples using bcftools 1.6 mpileup and call flags for the ~10.3Mb callable regions, setting the ploidy argument to 1, using –m (multiallelic-caller) and setting the minimum base quality (-Q) and the minimum mapping quality (-q) to 20. Further filters applied are missingness less than 30% across the samples and the same read depth filter as for autosomal SNPs for the high coverage datasets (SGDP + Himalaya). This resulted in a dataset of 36,691 SNPs for the joint call set of SGDP and Himalayan samples and 73,493 SNPs for the call set of 1000GP and Himalayan ones.

Sample	Population	Country	X	Y	mtDNA
BUM1919	Bumthang	Bhutan	18.21	18.69	1250.59
BUM1922	Bumthang	Bhutan	18.04	18.64	846.74
BUM1932	Bumthang	Bhutan	19.04	19.43	1512.02
BUM1933	Bumthang	Bhutan	17.32	17.53	802.45
KUR1554	Kurtöp	Bhutan	19.42	19.86	914.57
KUR1568	Kurtöp	Bhutan	19.4	19.75	1129.61
KUR1576	Kurtöp	Bhutan	19.91	19.87	1186.23
KUR1579	Kurtöp	Bhutan	20.38	20.61	1425.07
LAY1973	Layap	Bhutan	16.66	16.96	390.75
LAY1987	Layap	Bhutan	16.5	16.82	1565.87
LAY1989	Layap	Bhutan	16.14	16.48	1573.06
LAY1993	Layap	Bhutan	17.55	17.82	1681.35
LHP1013	Lhokpu	Bhutan	17.89	18.44	569.51
LHP1027	Lhokpu	Bhutan	16.13	16.44	605.89
LHP1034	Lhokpu	Bhutan	18.61	18.75	750.06
LHP1052	Lhokpu	Bhutan	18.89	19.5	681.26
MON1226	Mönpa	Bhutan	18.55	17.85	717.45
MON1228	Mönpa	Bhutan	17.51	17.81	873.78
MON1306	Mönpa	Bhutan	19.08	19.57	1123.45
MON1310	Mönpa	Bhutan	18.31	18.61	946.15
TSH0963	Tshangla	Bhutan	17.76	18.32	1464.57
TSH0966	Tshangla	Bhutan	17.81	18.22	1494.11
TSH0992	Tshangla	Bhutan	18.43	18.63	781.27
TSH1137	Tshangla	Bhutan	18.17	18.23	1098.75
TOT1094	Toto	India	18.95	19.41	659.27
TOT1124	Toto	India	18.4	18.76	1271.01
TOT1125	Toto	India	19.32	19.73	991.7
TOT1126	Toto	India	17.42	17.78	872.5
CHE0774	Chepang	Nepal	16.23	16.43	651.82
CHP0782	Chepang	Nepal	17.05	17.15	1517.9
CHP0785	Chepang	Nepal	16.03	16.32	1323.53
CHP0826	Chepang	Nepal	16.15	16.4	1056.94

CHE0020	Chetri	Nepal	15.71	15.63	3302.36
CHE0043	Chetri	Nepal	16.54	15.42	4418.68
CHE0047	Chetri	Nepal	17.65	17.82	4263.63
CHE0133	Chetri	Nepal	17.6	18.01	3011.62
GHL0693	Ghale	Nepal	17.89	18.22	1186.29
GHL0696	Ghale	Nepal	18.93	19.27	1158.02
GHL0935	Ghale	Nepal	19.26	19.72	782.08
GHL0937	Ghale	Nepal	16.3	16.78	704.95
LIM0052	Limbu	Nepal	16.05	15.05	1961.5
LIM0053	Limbu	Nepal	15	14.03	2469.55
LIM0063	Limbu	Nepal	14.38	13.61	3466.27
LIM0071	Limbu	Nepal	17.04	16.41	2214.74
MGR0389	Magar	Nepal	14.08	14.08	2511.06
MGR0482	Magar	Nepal	17.91	18.61	974.87
MGR0488	Magar	Nepal	17.79	18.65	1142.74
MGR0511	Magar	Nepal	18.04	18.48	893.7
SHE0298	Sherpa	Nepal	16.53	16.85	3997.17
SHE0635	Sherpa	Nepal	18.69	19.07	1184.35
SHE0838	Sherpa	Nepal	18.49	19.01	1312.03
SHE0841	Sherpa	Nepal	14.44	14.37	1439.25
THK0424	Thakali	Nepal	18.54	19.04	1269.93
THK0427	Thakali	Nepal	18.47	17.9	635.28
THK0437	Thakali	Nepal	18.22	18.64	933.74
THK0439	Thakali	Nepal	19.08	19.74	954.2
THR0641	Tharu	Nepal	17.82	18.26	829.59
THR0788	Tharu	Nepal	17.02	17.31	1115.18
THR0795	Tharu	Nepal	20.29	20.68	629.63
THR0816	Tharu	Nepal	18.82	19.26	471.09
Tib1	Lhasa	Tibet	22.29	23.18	1887.49
Tib10	Lhasa	Tibet	22.19	22.39	2003.88
Tib13	Lhasa	Tibet	20.26	20.28	1791.99
Tib14	Lhasa	Tibet	19.49	19.49	1877.85
Tib15	Lhasa	Tibet	21.68	21.99	1661.93
Tib16	Lhasa	Tibet	14.35	14.42	1744.9
Tib18	Lhasa	Tibet	18.8	19.03	2655.86
Tib21	Lhasa	Tibet	13.99	14.03	2004.54
Tib22	Lhasa	Tibet	16.85	16.98	2022.63
Tib24	Lhasa	Tibet	21.08	20.96	2287.29
Tib25	Lhasa	Tibet	21.09	21.29	3690.96
Tib26	Lhasa	Tibet	19.86	20.07	1954.41
Tib28	Lhasa	Tibet	22.41	22.42	1392.1
Tib29	Lhasa	Tibet	18.43	18.7	1799.37
Tib3	Lhasa	Tibet	22.73	22.89	3180.15
Tib30	Lhasa	Tibet	17.89	18.06	1659.34
Tib31	Lhasa	Tibet	26.23	26.32	2610.96
Tib32	Lhasa	Tibet	19.89	19.86	2634.09

Tib33	Lhasa	Tibet	19.26	19.31	1861.72
Tib34	Lhasa	Tibet	24.35	24.46	2283.43
Tib35	Lhasa	Tibet	20.65	20.91	1590.91
Tib4	Lhasa	Tibet	21.12	21.27	2287.49
Tib5	Lhasa	Tibet	20.57	20.78	3407.48
Tib6	Lhasa	Tibet	20.57	22.36	3034.15
Tib7	Lhasa	Tibet	18.75	19.09	2587.84
Tib8	Lhasa	Tibet	17.48	17.47	2280.84
Tib9	Lhasa	Tibet	19.9	20.01	3705.23

Table 3. 1 Read depth estimated from BAM files. The table reports average read depth for each of the Himalayan samples of chromosome Y, X and mtDNA.

3.2.2.1.4 mtDNA

All the reads that mapped to the mitochondrial DNA (mtDNA) were extracted from the whole-genome Himalayan BAM files. The read depth was calculated (Table 3.1) and all mitochondrial positions were called similarly to the Y chromosome using bcftools mpileup, and then merged with the 1000 Genomes Project Phase 3 mtDNA calls (2,534 individuals) with bcftools merge as suggested by the HaploGrep 2.0 programme (234). Mitochondrial DNA calls from the 1000 Genomes Project have been curated by removing nuclear mitochondrial DNA variants (NUMTs) and resolving heteroplasmies (66). All the positions showing mitochondrial heteroplasmy in the Himalayan individuals were removed for downstream analyses. This resulted in a dataset of 3,911 SNPs.

3.2.2.3 CNVs

A CNV call set from the Himalayan samples was generated by senior bioinformatician in our team, Yuan Chen, using GenomeSTRiP 2.0 (http://software.broadinstitute.org/software/genomestrip/genome-strip-20). The software discovers and genotypes CNVs using sequencing data (235). GenomeSTRiP was run following the standard pipeline, which has three steps: SVPreprocess, CNVDiscovery, GenerateHaploidCNVGenotypes. In total, 9,322 CNVs were called in the Himalayan dataset, 20 of which were filtered out using a genotype quality cutoff (GQ) > 20 with vcftools 0.1.14 (--minGQ 20), or for violating Hardy-Weinberg equilibrium (hwe > 0.00001), with PLINK 2.0 (https://www.cog-genomics.org/plink/2.0/). A final dataset of

9,302 CNVs was considered for further analyses. Out of these, 9,072 located on autosomes and 230 found on the sex chromosomes, X and Y. The dataset contains 4,497 deletions (delCNV) and 1,654 duplications (dupCNV) and 3,151 multi-allelic variants (mCNV). The functional impacts of CNVs were predicted using the Ensembl Variant Effect Predictor (VEP) (54, 189).

3.2.3 QC

3.2.3.1 SNPs

3.2.3.1.1 Concordance with SNP genotypes

Twenty-six out of the 87 sequenced samples have been also genotyped and analysed in the previous chapter. The SNP genotypes from these 26 samples were converted to GRCh38 and compared with the genotypes called from the whole genome sequencing data for the same sample. Shared genomic positions were extracted and the genotypes compared within the two datasets with a custom script. The genotype concordance was calculated counting the number of concordant genotypes (ref/ref, ref/alt and alt/alt genotypes) and divided it by the total number of genotypes (Table 3.2) (236) .

Individual	Country	SNP-chip ID	Sequencing ID	Concordance
BUM1919	Bhutan	HPG5196910	ERS724833	0.998
BUM1922	Bhutan	HPG5196913	ERS724834	0.998
BUM1932	Bhutan	HPG5197099	ERS724835	0.997
BUM1933	Bhutan	HPG5197100	ERS724848	0.998
LAY1973	Bhutan	HPG5197102	ERS724857	0.997
LAY1987	Bhutan	HPG5197200	ERS724858	0.996
TSH0963	Bhutan	HPG5197075	ERS724873	0.998
TSH0966	Bhutan	HPG5197078	ERS724874	0.998
TSH1137	Bhutan	HPG5197117	ERS724876	0.998
CHE0774	Nepal	HPG5197047	ERS724853	0.993
CHP0782	Nepal	HPG5197050	ERS724854	0.995
CHP0785	Nepal	HPG5197053	ERS724855	0.996
CHP0826	Nepal	HPG5197061	ERS724856	0.995
CHE0020	Nepal	HPG5196735	ERS724849	0.997
CHE0043	Nepal	HPG5196746	ERS724850	0.996
CHE0047	Nepal	HPG5196749	ERS724852	0.997
CHE0133	Nepal	HPG5196819	ERS724851	0.998
LIM0052	Nepal	HPG5196753	ERS724861	0.998
LIM0063	Nepal	HPG5196764	ERS724863	0.998
LIM0071	Nepal	HPG5196772	ERS724864	0.998
MGR0389	Nepal	HPG5196934	ERS724865	0.997
MGR0482	Nepal	HPG5196978	ERS724866	0.996
MGR0488	Nepal	HPG5196984	ERS724867	0.997
MGR0511	Nepal	HPG5196990	ERS724868	0.985
SHE0635	Nepal	HPG5197013	ERS724869	0.999
SHE0838	Nepal	HPG5197067	ERS724870	0.997

Table 3. 2 Genotype concordance between whole-genome sequencing and SNP array genotypes in 26 Himalayan samples.

3.2.3.1.2 Ti/Tv

The quality of the Himalayan SNP dataset was also assessed using the GATK VariantEval tool and the vcflib package popStats (237). In total, 8,264,072 (82%) SNPs are in the Single Nucleotide Polymorphism database (dbSNP) 146 with the same alternative alleles with consequent concordant rate of 99.14. The heterozygous/non-reference homozygous ratio (Het/Hom Alt ratio), which can be an indicator of potential sample contamination if it is greater than two, was 1.26 for SNPs in the Himalayan dataset compared to 1.24 in dbSNP. The transition/transversion ratio (Ti/Tv ratio) for whole-genome sequences should be around 2.1, and the Himalayan dataset follows this criterion for the known variants, although it is lower when the novel sites are considered. This

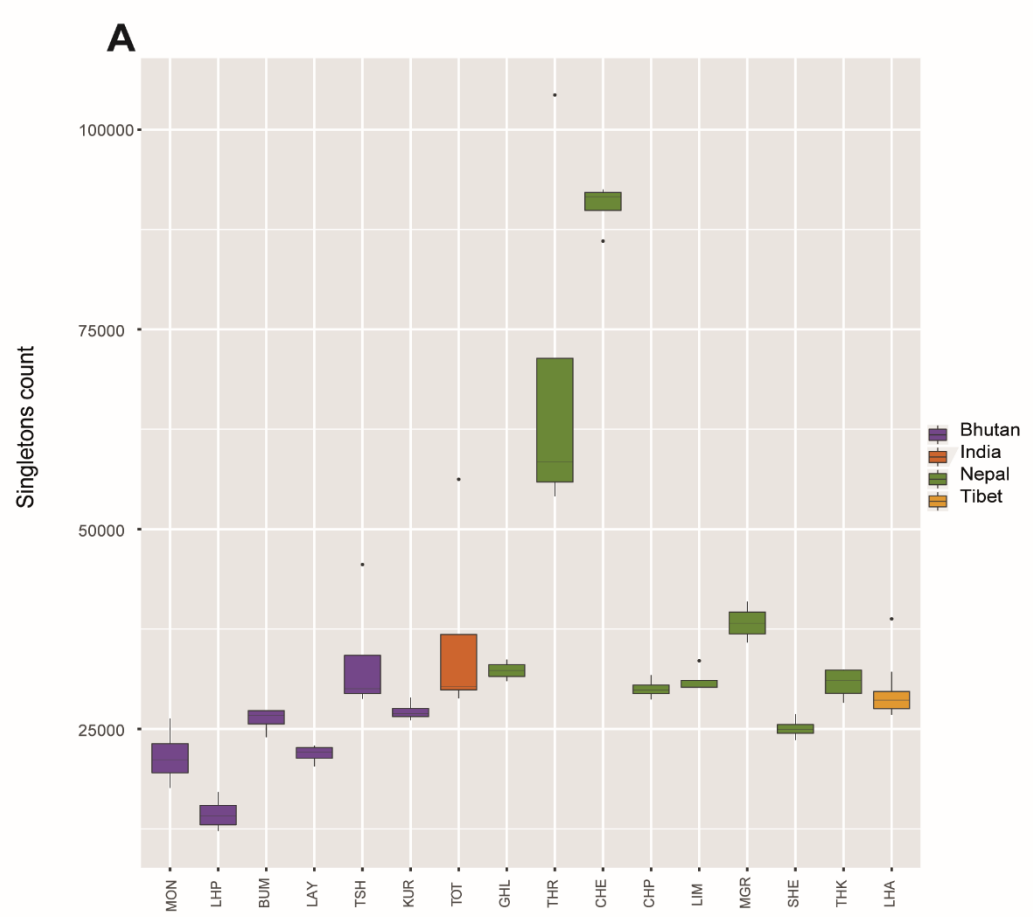
could be due to the occurrence of false positives within the novel sites and extra care is needed when including these variants in downstream analyses (Table 3.3).

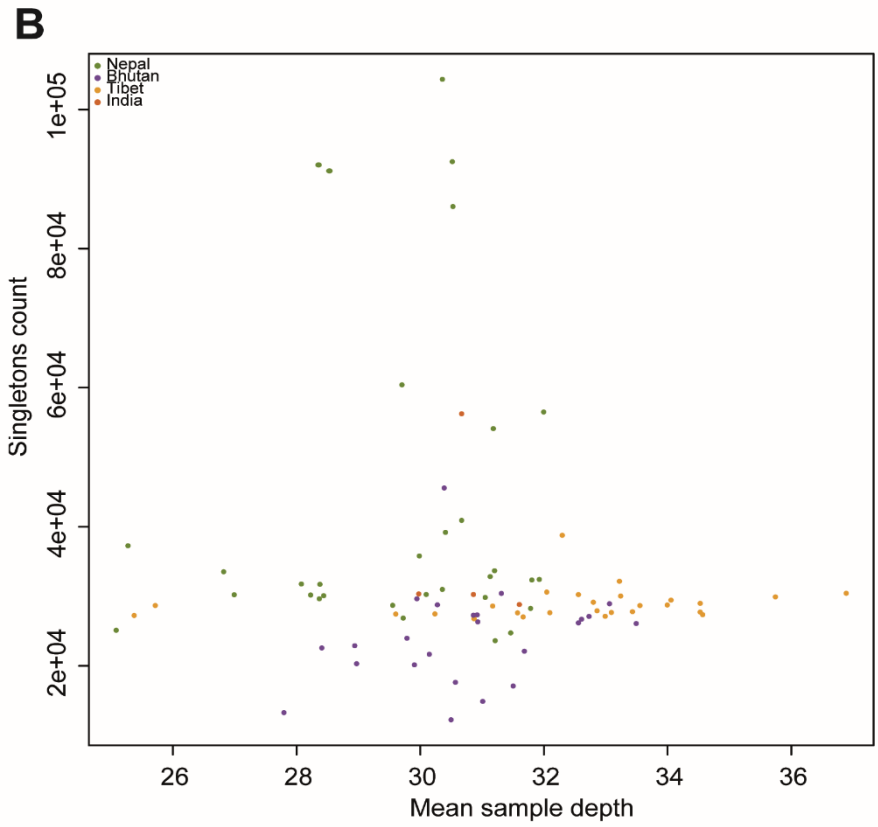
Novelty	nSamples	nSNPs	concordance Rate	SNPNoveltyRate	TiTvRatio
All	87	9938079	99.14	16.12	2.09
Known	87	8335797	99.14	0	2.14
Novel	87	1602282	0	100	1.86

Table 3. 3 SNP call set QC metrics. Al l= all variants in the Himalayan dataset. Novel = variants found in the Himalayan dataset only. Known= variants present in dbSNP 146.

3.2.3.1.3 Singleton counts

Singletons are variants with only one alternative allele in all of the samples, and the number of singletons per sample is often used as a metric to assess dataset quality. Random sequencing errors are often seen as singletons and an elevated number could indicate sequencing errors. The total number of singletons for SNP variants in the Himalayan individuals using vcftools 0.1.14 was 2,781,024. The distribution of singletons per individual in each population was plotted (Figure 3.4A) and the correlation between the number of singletons and the mean coverage of depth per individual was analysed. Chetri and Tharu show the highest number of singletons, and Lhokpu the lowest number, but these are not due to differences of sequencing coverage (Figure 3.4B). Overall, Bhutanese populations are characterised by the lowest number of singletons, whereas Nepalese ones have higher proportions (Figure 3.4A and B). Finally, the genome-wide distribution of singletons was also plotted and it showed that they were uniformly distributed across the genome (Figure 3.4C).


A
Singletons count
100000
75000
50000
25000
MON
LHP
BUM
LAY
TSH
KUR
TOT
GHL
THR
CHE
CHP
LIM
MGR
SHE
THK
LHA
Bhutan
India
Nepal
Tibet


B
Singletons count
Mean sample depth
Nepal
Bhutan
Tibet
India
2e+04
4e+04
6e+04
8e+04
1e+05
26
28
30
32
34
36

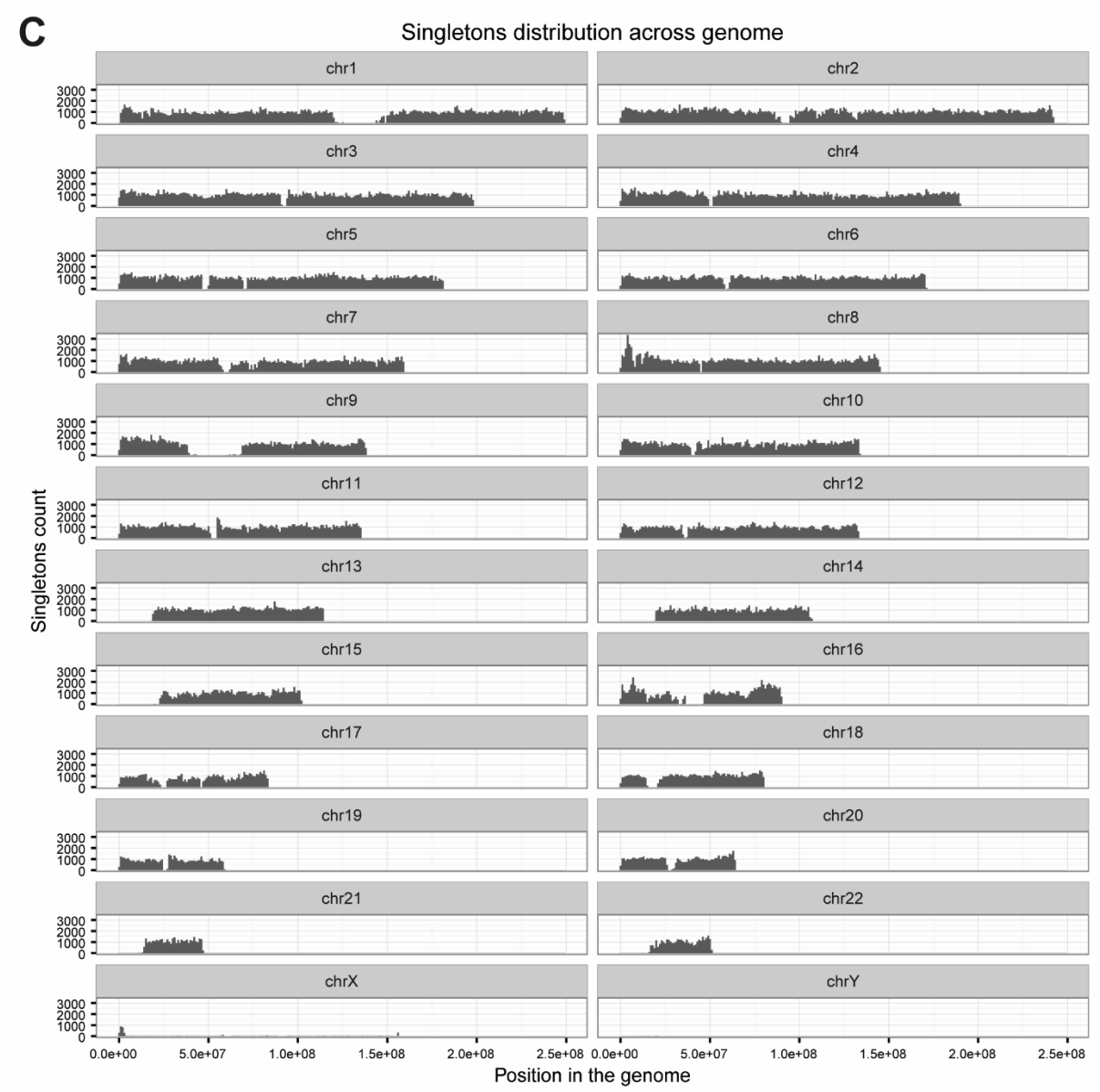


Figure 3. 4 The distribution of singleton SNPs in the Himalayan dataset. A. Total number per individual in each of the Himalayan populations. B. The correlation between the number of singleton SNPs and the average sequence coverage of each individual. C. Genome-wide distribution of all SNP singletons in the Himalayan dataset.

3.2.3.1.4 Site Frequency Spectrum (SFS)

The site frequency spectrum at bialleic sites across all 87 samples using frequency bins of 0.05 was calculated with vcflib popStats (237) (Figure 3.5). The allele frequency is the number of non-reference alleles divided by the total number of alleles. The distribution is as expected, with a large excess of rare variants (<0.05) and small excess of high frequency variants (>0.95).

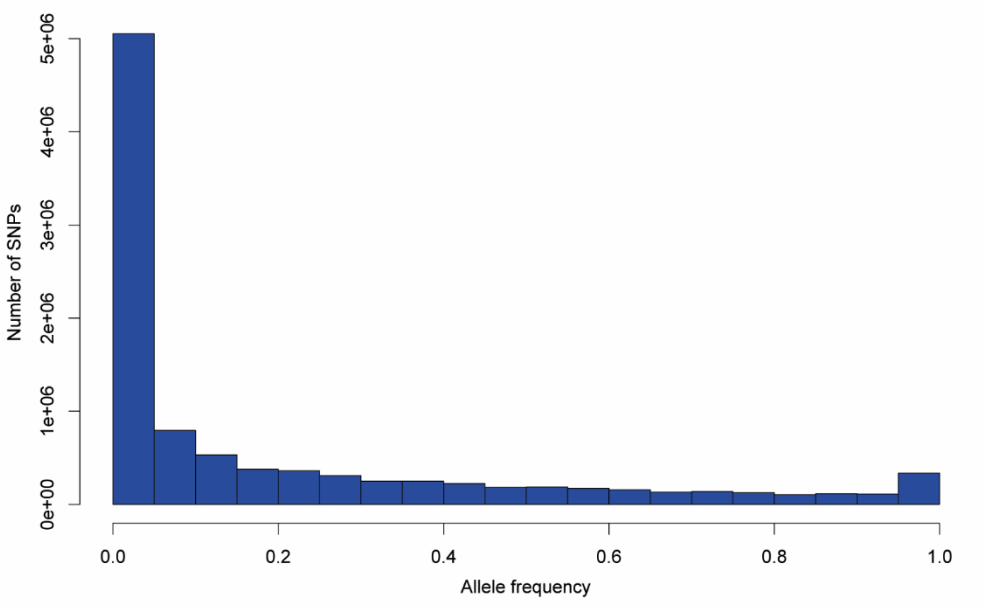


Figure 3. 5 Site frequency spectrum (SFS) of the Himalayan SNP dataset. The x-axis reports the allele frequency bins and the y-axis shows the proportion of the variants.

3.2.3.1.5 PCA and ADMIXTURE analyses

Principal component analyses (PCA) of the Himalayan dataset using both common (MAF> 0.01) and rare (MAF< 0.01) SNPs were carried out using PLINK 1.92 on an LD pruned file. The number of common variants used was 1,223,478, whereas the number of rare variants was 1,319,138. A PCA with the dataset including Himalayans and 1000 Genomes Project individuals (worldwide dataset) was also computed using EIGENSOFT 6.0 (157, 159). For this, I computed the eigenvectors with global diversity, and then the Himalayan individuals were projected onto it. ADMIXTURE v1.2 (74) analyses were run on the Himalayan SNP dataset using the cross validation (CV) error to identify the best K value.

3.2.3.2 INDELs

The quality of the INDEL calls was assessed similarly to the SNP dataset, and 206,659 (22%) novel INDELs in the Himalayan dataset were identified that were absent in dbSNP 146. The Het/Alt Hom ratio was 1.37 for INDELs in the Himalayan dataset compared to 1.34 in dbSNP.

Novelty	nSamples	nINDELs	INDELNoveltyRate
All	87	1043093	18.59
Known	87	836434	0
Novel	87	206659	100

Table 3. 4 INDEL call set QC metrics. All = all variants in the Himalayan dataset. Novel = variants found in the Himalayan dataset only. Known = variants present in dbSNP146.

The frequency distributions of the different INDEL lengths for novel and known INDELs in the Himalayan dataset were calculated, showing an overall slightly higher proportion of novel deletions compared to known ones. In contrast, there are slightly more known duplications than novel ones. Interestingly, at the extreme tails of the length distribution on both sides there are only novel INDELs, perhaps reflecting improvements in read length and INDEL calling compared with the dbSNP average (Figure 3.6).

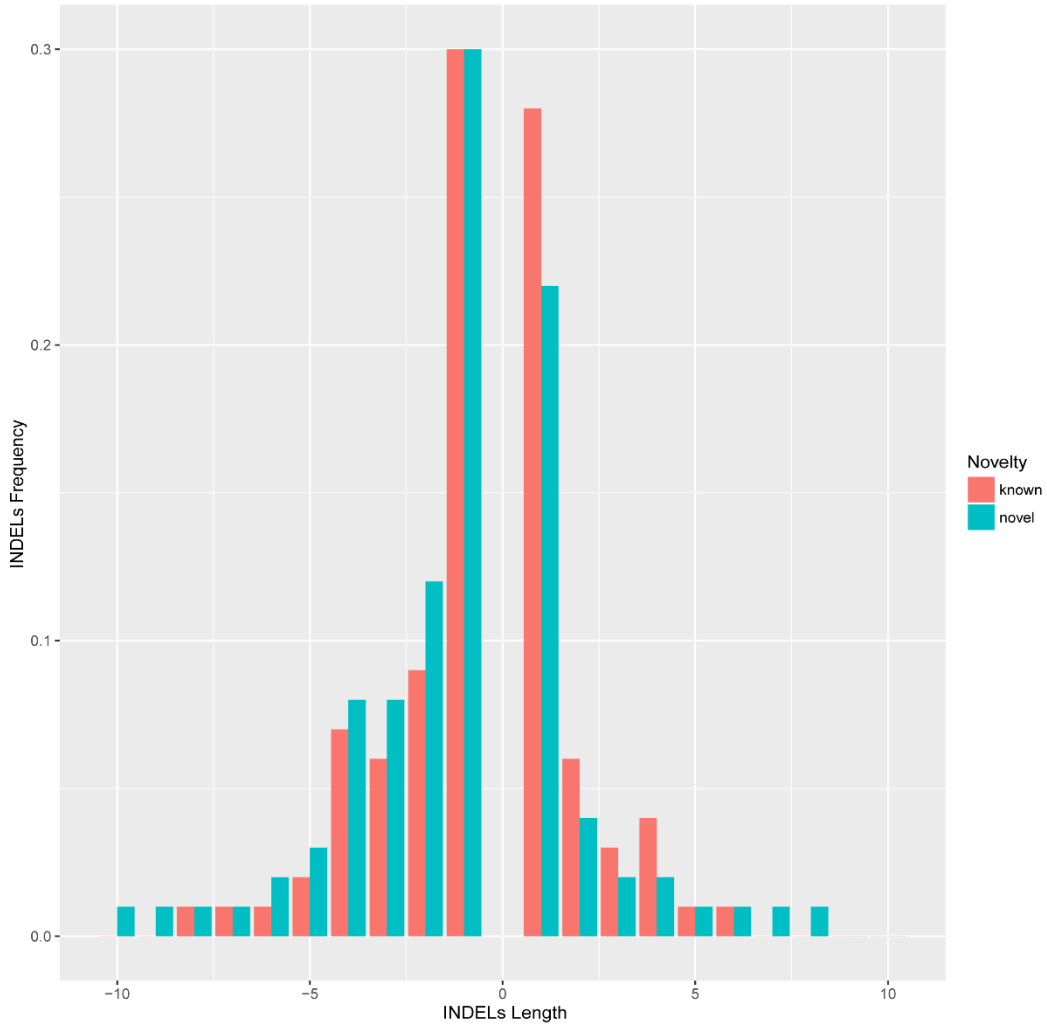


Figure 3. 6 INDEL frequency distribution in the Himalayan dataset. The plot reports the novel and known INDELs in the Himalayan dataset compared to dbSNP 146. The x-axis shows the INDEL length in bins and the y-axis reports the INDEL frequency in the Himalayan populations.

The total number of INDELs is 262,472. Similarly to SNPs, Chetri and Tharu show the highest number of singletons, and Lhokpu the lowest number; singletons are uniformly distributed across the genome.

The whole-genome site-frequency distribution (SFS) of INDELs resembles the SNPs SFS, with a large excess of rare variants and slight excess of high-frequency variants (Figure 3.7; compare with Figure 3.5).

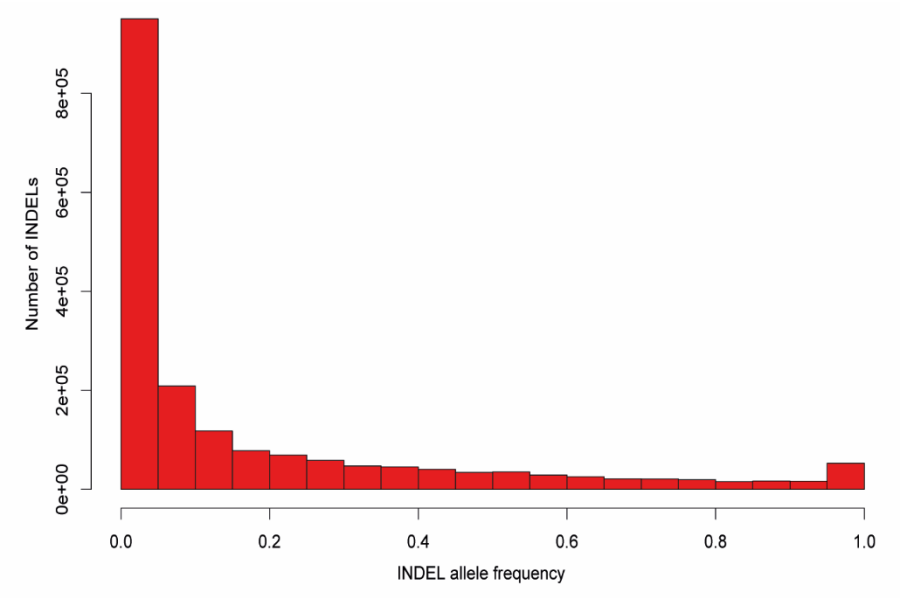


Figure 3. 7 Observed folded site frequency spectrum (SFS) distribution of the Himalayan INDEL dataset. The x-axis reports the minor allele frequency and the y-axis the proportion of variants.

A PCA using LD-pruned INDELs was carried out using PLINK 1.92.

3.2.3.3 CNVs

PCA was performed on both deletions and duplications using PLINK 2.0, and the results were compared. CNV discovery is a difficult process and there is concern that very rare and singleton variants can be enriched for artefacts. In previous studies, it has been reported that deletions identified by GenomeSTRiP showed an estimated FDR of 2.9% and this value is overall higher for rare variants (238, 239). Thus, I compared common Himalayan CNV intervals (MAF > 0.01) with the ones present in the 1000 Genomes Project individuals (GRCh38 LiftOver) with bedtools 2.22.0 (intersect argument) (240) to test if Himalayans carried novel common CNVs. Bedtools intersect searches for overlaps between two sets of genomic regions. I set the minimum overlap required at 0.3 as described in the SGDP data publication (241) (-f flag) and the fraction of overlap required as reciprocal for the two sets of intervals (-r flag). Thus if two CNVs share at least 0.3 of their length, they are considered to belong to the same CNVR. Finally, I set a "left outer join" (-loj flag) to report both overlaps and no overlaps in the output.

3.2.4 Datasets

3.2.4.1 Phasing and pruning

The SNP call set on GRCh38 for Himalayan populations was phased with Eagle v2.4 (242, 243) using the LiftOver 1000 Genomes Project data to GRCh38 as a reference panel. The new version incorporated a new pipeline developed by Giulio Genovese which includes extra processing steps applied to the 1000 Genomes Project LiftOver data to remove duplicated and multi-allelic variants.

I have also generated two pruned datasets by filtering out SNPs and INDELs in high LD ($r^2 > 0.5$) using PLINK 1.92, resulting in a dataset of 2,669,936 SNPs and 462,049 INDELs.

3.2.4.2 Datasets used for the analyses

Several combination datasets were generated for different analyses because the comparison datasets were mapped to different human reference genomes (Table 3.5). Overall, the data mapped to GRCh38 were used for most of the analyses. A merged SNP-only dataset of Himalayan samples plus the LiftOver version of the 1000 Genomes Project Phase 3 individuals (GRCh38), which contained 2,591 individuals and 9,636,923 SNPs, was used for most population demographic history analyses. The merged datasets of the LiftOver to GRCh37 and filtered Himalaya SNP dataset described earlier with the 1000 Genomes Project samples or SGDP data were generated for the analyses of the Y chromosome, archaic ingression as well as for Fine*MAV* for positive selection (244) .

Dataset	Reference mapped to	Analysis
Himalaya (SNPs, INDELs, CNVs)	GRCh38	Population structure and demography
Himalaya + 1000GP low coverage data*	GRCh38	Population structure and demography, Y chromosome and MtDNA phylogenies
Himalaya + 1000GP high coverage data	GRCh38	Effective population size, population split time (MSMC2)
Himalaya* + 10000 GP low coverage data	GRCh37	Fine*MAV,* Y chromosome phylogeny
Himalaya* + SDGP	GRCh37	Archaic introgression, Y chromosome phylogeny

Table 3. 5 Different datasets used for the different analyses. The table reports the reference genome of the samples and the analyses performed with each dataset. *Data that have been lifted over and not directly mapped onto the reference genome reported in the table. This does not apply to the Y chromosome.

3.2.5 Genetic variation analyses

The genome-wide observed heterozygosity rate per individual was calculated using PLINK 1.92 (--het argument). For each sample, PLINK computes method-of-moments F coefficients from observed and expected autosomal homozygous genotype counts:

F = (Observed hom. count - Expected count) / (Total observations - Expected count)

Thus heterozygosity rate was derived from PLINK's output using the formula:

$$Het = (N(NM) - O(Hom))/N(NM)$$

where N(NM) is the number of non-missing genotypes and O(Hom) denotes the number of observed homozygous genotypes per individual (156). A pruned dataset (LD, $r^2 > 0.5$) with only common variants (MAF >0.01) was used to detect Runs of Homozygosity (ROH) with PLINK 1.92. To maximize the detection of autozygous segments in the Himalayan populations (169), the parameter settings were: the minimum number of SNPs to call an ROH was 100, the heterozygote allowance was zero, the missing SNP allowance was less than 5, and the window threshold to call an ROH was 0.05. Identity-by-descent (IBD)

tracts for each unphased individual within a single population were estimated using the IBDSeq programme (245).

Pairwise population Fixation Index (F_{ST}) values were calculated for the Himalayan and worldwide datasets using EIGENSOFT 6.0 with 1000 Genomes Project admixed American populations removed.

Phylogenetic relationships and gene flow between Himalayans and other worldwide populations were studied using TreeMix 1.12 (77), again with the sequencing data. Genetic affinity to other worldwide populations, in particular other South and East Asian individuals, was tested by computing outgroup *f3* statistics with the phylogeny *f3*(X, Y; YRI), where X is each of the Himalayan populations and Y are the 1000 Genomes Project populations except admixed Americans and the Yoruba (YRI) used as an outgroup, using the qp3Pop function in the ADMIXTOOLS v5.0 package (82). Outgroup *f3* statistics was computed separately for 6,106,953 common SNPs (MAF > 0.01) and 3,823,094 rare SNPs (MAF < 0.01) using the dataset of Himalayans and 1000 Genome Project individuals. Local ancestry was inferred using phased data with PCAdmix 1.0 (79). Phased VCFs were first converted to Beagle (246) format with a custom script as an input for PCAdmix. The South and East Asian populations in the 1000 Genomes Project were used as parental populations and Himalayan populations modelled as a mixture of them.

3.2.6 Demographic analyses

The fine-scale demographic history of the Himalayan populations was investigated using several statistical approaches.

3.2.6.1 Rare variant sharing

Rare variants are informative about recent population demographic events and fine population structure. *f2* variants are the rarest shared variants, with only two alternative alleles in the samples considered (177). Three different sharing patterns were analysed using custom scripts; the admixed American populations in the 1000 Genomes Project were excluded from these analyses:

1. Singletons discovered in Himalayan populations shared with variants at any frequency in the 1000 Genomes Project.
2. *f2* variants in the merged 1000 Genomes Project and Himalayan dataset. Singletons discovered in Himalayan populations shared with singletons discovered in the 1000 Genomes Project dataset.

3.2.6.2 Effective population size and population split times

Effective population size (Ne) and population split time were investigated using the multiple sequentially Markovian coalescent approach (MSMC2) (84) and SMC++ (247) . Single sample consensus calls and mask files with the callable region for each individual were generated from both Himalayan and the high coverage 1000 Genomes Project (remapped to GRCh38) BAMs using a combination of samtools, bcftools and the bamCaller.py as explained in the manual of MSMC2 and MSMC tools. The phased VCFs were used as input for MSMC2, which was run as explained in the MSMC tutorial with default parameters for both Ne and split time estimation (time segment patterning: 1*2+25*1+1*2+1*3). A generation time of 30 years and a genome-wide average mutation rate of $1.25e^{-8}$ per base pair per generation were used (84). SMC++ for Ne estimation was run with the default parameters using the VCF dataset including Himalayan and 1000 Genomes Project individuals, varying the pair of distinguished lineages for forming

composite likelihoods as reported in the SMC++ manual. A mask file to filter out large uncalled regions of the genome was used.

3.2.6.3 Y chromosome and mtDNA phylogeny analyses

The haplogroups for Y chromosomes were defined using yHaplo (248) . A fasta format of Y SNPs was generated using vcf-to-tab in vcftools and custom scripts, and used as an input to construct a phylogenetic tree with RAxML (version 8.2) using GTRCAT as the substitution model(249). The phylogenetic tree from RAxML was then manually rooted and visualised using FigTree v.1.4.3 (http://tree.bio.ed.ac.uk/software/figtree/). The ages of internal nodes were estimated based on the SNP branch length (250) using the mutation rate 0.76×10^{-9}/site/year (95% C.I. $0.67–0.86 \times 10^{-9}$/site/year) (153).

mtDNA haplogroups were defined using HaploGrep 2.0 . The mtDNA phylogenetic tree was constructed and the ages of the internal nodes estimated in the same way as for the Y chromosome, but using a mutation rate of 2.53×10^{-8}/site/year (95% highest posterior density: $1.76–3.23 \times 10^{-8}$) (153). The possible functional consequences of mtDNA heteroplasmic variants were predicted by running the Variant Effect Predictor (VEP).

3.2.6.4 Archaic introgression

Introgression maps for the Himalayan individuals were generated by Dr. Laurits Skov using a method he has developed and published (222). The method does not use the genome of an archaic sample, but instead searches for genomic regions with a high SNP density in the test sample of SNPs not seen in a reference population assumed to be unadmixed. A dataset including modern Himalayan and other worldwide individuals from the SGDP mapped to GRCh37 was used for this analysis, with non-African populations as the test and Africans as the unadmixed reference. High coverage archaic genomes from Denisovans (4) and the Altai (231) and Vindija (251) Neanderthals were used to assign introgressed segments to these species. Only segments with mean probability ≥ 0.8 (corresponding to 5% FDR) were retained. Himalayan-specific introgressed segments were identified by comparing Himalayan segments with those in

other populations using bedtools 2.22.0 (intersect argument) with the minimum overlap at 0.5 (-f flag) and the fraction of overlap required to be reciprocal for the two sets of intervals (-r flag) and a “left outer join” (-loj flag) to report both overlaps and no overlaps in the output. These introgressed regions then were annotated using BEDOPS v2.4.35 (252) from regulatory annotations in Ensembl GRCh37 Release 93 (54) and only those falling within a gene or regulatory features (distance =0) were retained.

3.2.7 Fine mapping of positive selection

We took the advantage of the whole genome sequencing data which provide a full catalogue of genetic variation, both coding and non-coding, free from ascertainment bias to further refine the signals of selection in the Himalayan populations using three approaches:

- Genome-wide single locus estimates of F_{ST} (253) between low and high altitude (> 2500 meters above sea level) populations for SNPs, INDELs and CNVs respectively with PLINK 1.92. Variants showing F_{ST} values > 0.3 were considered significant. Single locus F_{ST} for the *EPAS1* region between the Tibetans sequenced here and Chinese Han (CHB) sequences from the 1000 Genomes Project was calculated as a positive control for highly differentiated SNPs. Variants with F_{ST} values > 0.8 where considered highly differentiated between Han Chinese and Tibetans. LD blocks within the *EPAS1* region were estimated using PLINK 1.92 via Haploview with default parameters (58, 181).
- The Fine-Mapping of Adaptive Variation (Fine*MAV*) algorithm (244). Fine*MAV* is specifically designed to localise a signal of population-specific selection to a single most likely variant, and thus to differentiate between selection-driving and passenger variants for functional follow-up studies. A Fine*MAV* score is calculated for the derived allele of each SNP by combining its Derived Allele Purity (*DAP*), Derived Allele Frequency *(DAF*) and functional prediction (the *CADD* PHRED-scaled C-score)(190). Thus it assigns high scores for derived alleles that are common, population-specific and have a strong predicted functional effect. Firstly, the population genetic component of Fine*MAV* (*DAPxDAF*) is calculated, which summarises population differentiation. Then *DAPxDAF* is combined with the measure of functionality (*CADD*) as implemented in Fine*MAV* to further prioritise highly differentiated variants and elucidate the most likely functional targets of selection. *DAP*, *DAF* and Fine*MAV* values were calculated for derived and ambiguous alleles (annotated accordingly to Ensembl Compara) (54) using a custom script (SNPs only; INDELs were omitted).

The Fine*MAV* analyses were performed in the following contexts:

1. Himalayan high altitude populations (> 2500 meters above sea level) vs low altitude populations to explore differentiation between different Himalayan groups; n=2, x=4.96.
2. Comparison between Himalayan high altitude and AFR, EAS, EUR, SAS (the1000 Genomes Project data) to explore high altitude specific signals; n=5, x=2.98.
3. Comparison between Himalayan low altitude and AFR, EAS, EUR, SAS (the1000 Genomes Project data) to explore low altitude specific signal; n=5, x=2.71

The top 100 outlier SNPs in the whole-genome Fine*MAV* distribution were reported and some of the strongest hits commented. Fine*MAV* analyses were run by Dr. Michal Szpak.

- Hypoxia responsive element analysis. The *EPAS1*-encoded protein, HIF-2α, binds to a specific sequence motif called a hypoxia responsive element (HRE). To test whether any variant in the high altitude individuals was lying in a HRE and putatively affecting HIF-2α binding, possible binding motif regions for HIF2-α from the HOCOMOCO v11 Human TF collection (254, 255) were predicted using PWMScan (256) from PWMTools with default parameters (Ambrosini G., PWMTools, http://ccg.vital-it.ch/pwmtools) for the human reference genome (GRCh38). The Position Weight Matrix Logo for HIF-2α was generated by PWMScan only for the regions of interest: the regions with ENCODE Project data (GRCh38) (257) for DNA accessibility (DNase-seq) which are associated with transcription factor (TF) binding and histone modifications in aorta and lung tissues. Then the highly differentiated variants (F_{ST} >0.8-between Tibetans and Han Chinese) within these regions were identified.

3.3 Results

3.3.1 A large high-quality sequencing dataset of Himalayan populations

3.3.2.1 SNPs and INDELs

The dataset of 87 individuals from 16 populations contained 9,938,079SNPs, 1,602,282 novel. After filtering, there are 1,043,093 INDELs in total in the Himalayan dataset, 206,659 of which are novel. Deletions make up 54.2%, duplications 40.2%, and 5.6% are classified as sequence alterations. As expected, the majority of INDELs fall within intronic and intergenic regions. However, 6,302 INDELs lie within coding sequences with different functional consequences (Figure 3.8).

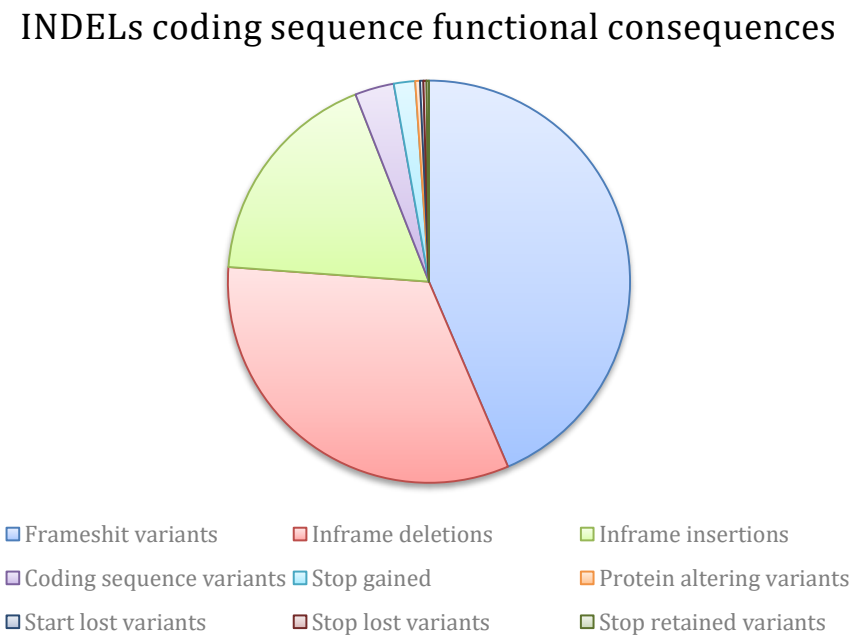
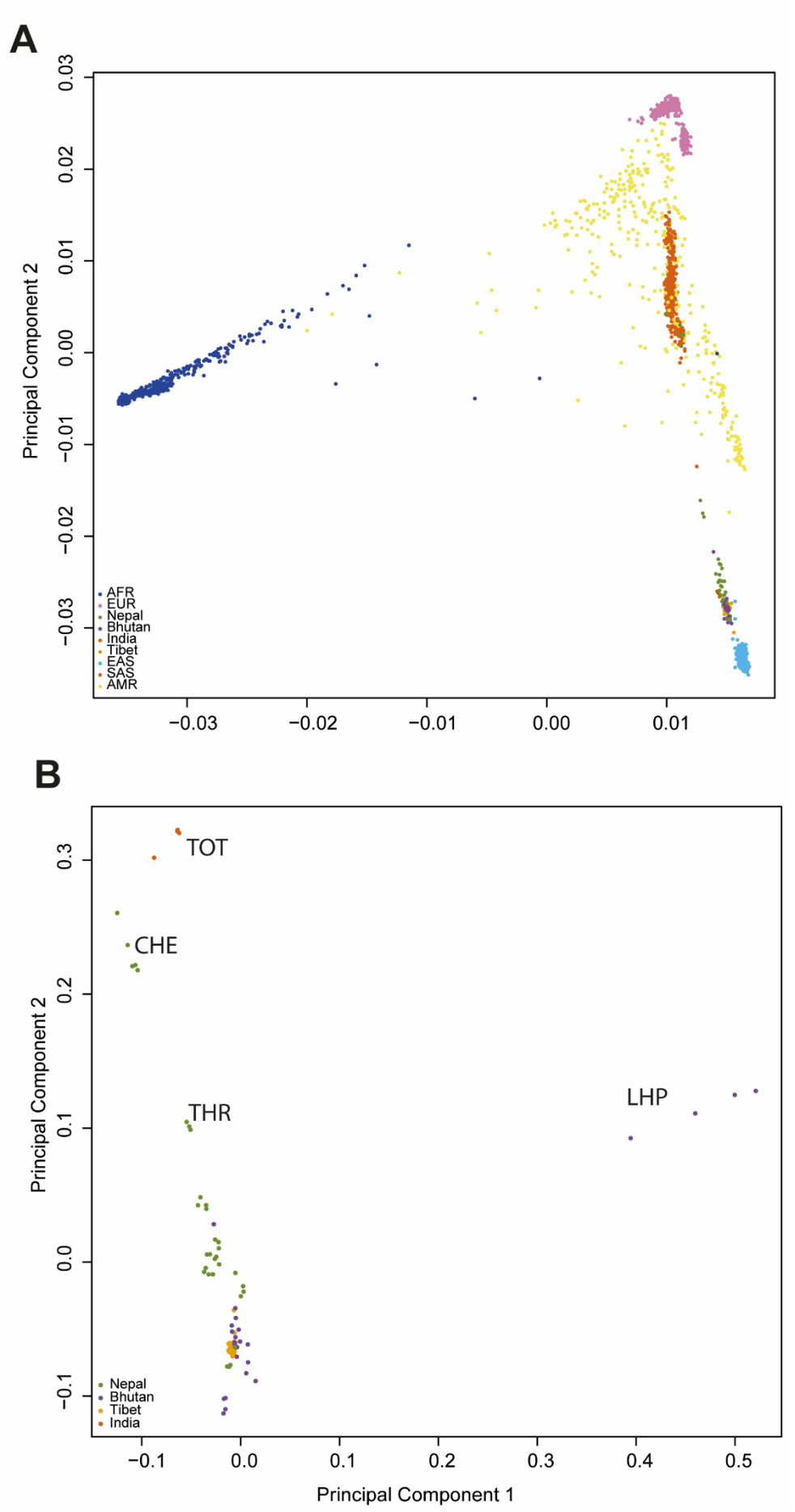


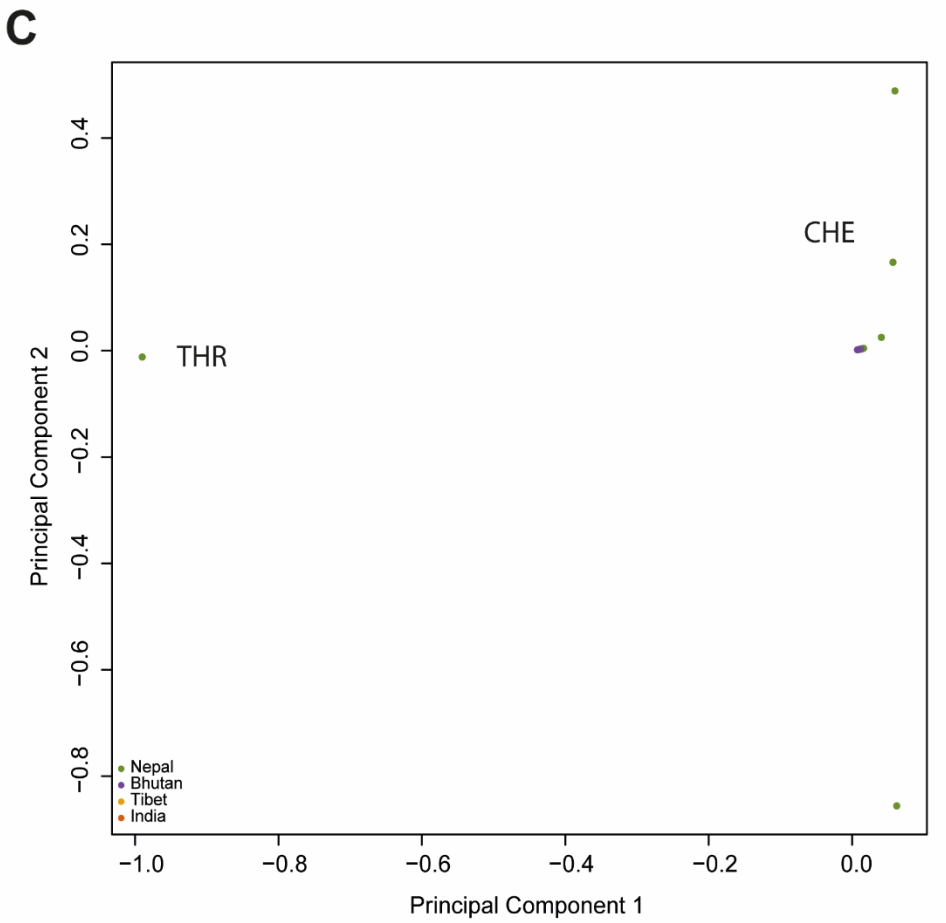
Figure 3. 8 Functional consequences of INDELs in the coding sequence. The pie chart reports the proportion of variants falling in different functional categories retrieved using VEP.

The total of low frequency SNP variants (MAF < 0.01) was 2,961,006 SNPs and 267,175 INDELs. The majority of these were extremely rare: 2,892,999 singletons SNPs and 262,472 INDELs, respectively. No correlation between number of singletons and depth of coverage was identified in any of the samples, and singletons were randomly distributed across the genome.

PCA using the Himalayan samples together with other worldwide populations from the 1000 Genomes Project (Figure 3.9A) showed similar structures to those with SNP genotype data from the last chapter. Himalayan individuals lie on a cline within other South and East Asian populations, with Chetri and Tharu clustering with South Asians and the rest of the Himalayan populations genetically closer to East Asians. PCA of the Himalayan samples alone using only common variants displays a clear population separation, with the first two components of the PCA showing strong sub- clustering, with the Chetri plus one Tharu individual and Lhokpu populations forming outliers (Figure 3.9B). The PCA using rare variants shows that the first two principal components separate the majority of the Himalayan populations from the four Chetri and one Tharu individuals that are extreme outliers (Figure 3.9C) due to their excess of singletons compared to other Himalayan populations. PCA computed using INDELs replicated the SNP results with Chetri, Tharu, Toto and Lhokpu forming outliers and the rest of the populations forming sub-clusters (Figure 3.9D).

The ADMIXTURE analysis using the SNPs from the Himalayan dataset displays patterns consistent with the PCA, with different proportions of ancestral components between Nepal, Bhutan, North India and Tibet. Using the best number of ancestral components according to the cross-validation error (K=2), the Bhutanese, Tibetan and Sherpa populations are mainly represented by the red component, while other Nepalese populations show a mixture of green and red components. Outlier populations (Lhokpu, Toto, Chetri and Tharu) are characterised by the green component. The increase of one additional component (K=3) leads to the separation of Lhokpu and Mönpa characterised by the green component, and Toto, Chetri and Tharu by an additional yellow component (Figure 3.10).


A
Principal Component 2
0.03
0.02
0.01
0.00
−0.01
−0.02
−0.03
−0.03
−0.02
−0.01
0.00
0.01
AFR
EUR
Nepal
Bhutan
India
Tibet
EAS
SAS
AMR
B
Principal Component 2
0.3
0.2
0.1
0.0
−0.1
TOT
CHE
THR
LHP
Nepal
Bhutan
Tibet
India
−0.1
0.0
0.1
0.2
0.3
0.4
0.5
Principal Component 1


C
CHE
THR
Nepal
Bhutan
Tibet
India
Principal Component 2
Principal Component 1
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
−1.0
−0.8
−0.6
−0.4
−0.2
0.0

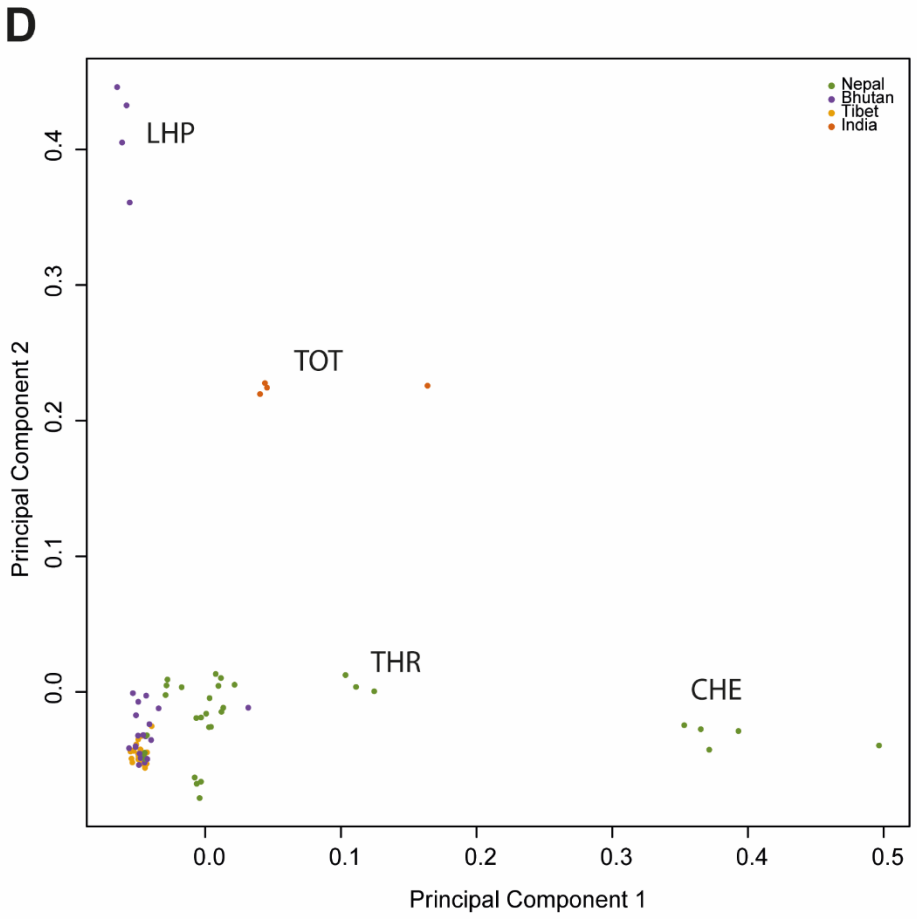

D
LHP
TOT
THR
CHE
Nepal
Bhutan
Tibet
India
Principal Component 2
Principal Component 1
0.4
0.3
0.2
0.1
0.0
0.0
0.1
0.2
0.3
0.4
0.5

Figure 3. 9 Principal component analyses (PCA). A. PCA of the world dataset. Each dot represents a sample, coded by region as indicated. The Himalayan region samples lie between other East Asian and South Asian samples. B. PCA of the Himalayan populations alone for common SNPs variants. Chetri, Lhokpu and Toto act as outliers. C. PCA of the Himalayan populations alone for rare SNPs variants. Chetri, and one Tharu individuals act as extreme outliers. The rest of the samples are indistinguishable. D. PCA of the Himalayan populations alone for INDELs. The pattern of variation is similar to the PCA on common SNPs.

Figure 3. 10 ADMIXTURE (K value of 2 and 3) analysis of the Himalayan samples. Purple= Bhutan, orange= India, Nepal=green, Gold= Tibet

3.3.2.2 CNVs

CNVs were investigated in the Himalayan populations, and compared with other worldwide populations from the 1000 Genomes Project. After CNV discovery and filtering, 9,302 CNVs were analysed: 9,072 located on autosomes and 230 on the X and Y. The dataset contains 4,497 deletions (delCNV), 1,654 duplications (dupCNV) and 3,151 multi-allelic variants (mCNV). There are 6,820 common variants (MAF> 0.01). PCA on both the all variants and common variants datasets for duplications, deletions and mCNV display similar patterns, although duplications detect more detailed substructure. On the first principal component, Limbu and one Chetri individual, showing the highest proportion of CNVs, are separated from the rest of the populations for both deletions and duplications. In the PCA applied to duplications, Limbu outliers lie separated from the Chetri individual, and the second principal component splits the remaining samples into two groups. Similarly to the PCA performed on SNPs and INDELs, Tibetan, Bhutanese and Sherpa from Nepal mostly cluster together, whereas other Nepalese samples tend to form a cline and sub-groups (Figure 3.11).

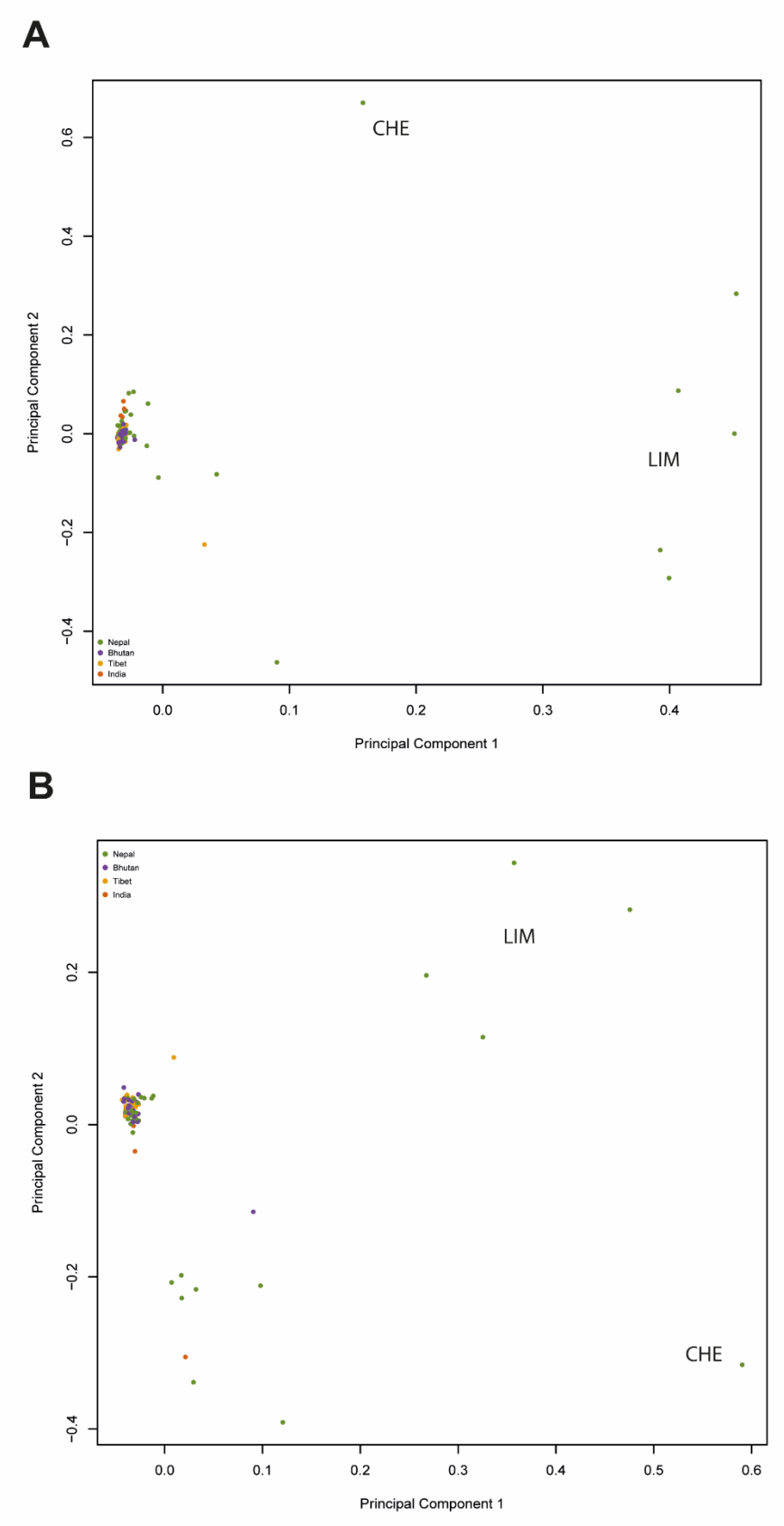


Figure 3. 11 PCA on CNVs. The plots show the PCA on deletions (A) and duplications (B). Limbu and one Chetri individuals are seen as outliers.

GenomeSTRiP produces a high quality CNV call set, yet some further filtering and QC are still needed before further population-genetic analyses, as illustrated by the PCA plots. Although this is beyond the scope of my thesis, some known interesting CNVs in these populations were rediscovered in my dataset. One is a 3.4 kb deletion upstream of *TMEM247* near *EPAS1* in Tibetans (258). High altitude populations show the highest frequency of this deletion, with Tibetans (overall frequency 0.83) showing 19 homozygotes and seven heterozygotes. The deletion is present in the other high altitude populations (Ghale, Kurtöp, Layap, Thakali and Sherpa) with an allele frequency ranging from 0.5 to 0.625. The second is the alpha-thalassemia deletion in the Tharu (7) and Ghale. All the four Tharu individuals in the dataset carry heterozygous $-\alpha^{3.7}$ deletions, thus with an allele frequency of 0.5. Surprisingly, three out of four Ghale and one Tshangla individuals also carry the deletion, with one Ghale individual homozygous for it and the other three heterozygous. All the other samples in the dataset are homozygous for the reference allele.

3.3.2 Recent demographic history – genetic structure

3.3.2.1 Genetic similarity and gene flow in the Himalayas

The genetic structure and relationships between Himalayan populations and other worldwide populations were investigated using several metrics. Firstly, the aim was to explore the genetic relationship of the three new populations in this dataset, Tharu, Ghale and Lhokpu to the other Himalayans, and more generally to compare the patterns to those from the genotype data analysed in the previous chapter. Secondly, I was interested in understanding why the Tharu, Lhokpu and Chetri appeared as extreme outliers in the PCA.

The phylogenetic tree reconstructed using TreeMix on the worldwide dataset displays long branches for Lhokpu, Toto, Mönpa and Chepang, in agreement with the genetic drift patterns. It also shows that Chetri and Tharu form separate branches from the rest of the Himalayan populations, and that the populations from higher-altitude, such as Lhasa, Layap and Sherpa cluster together, in agreement with the SNP-genotype analysis (Figure 3.12). This genetic pattern was also confirmed by the pairwise population F_{ST} analysis, which shows Lhokpu and Toto having the highest F_{ST} values in comparison with the other populations, and the Tibetan and Bhutanese populations showing more genetic affinity. These analyses thus confirmed the results from the SNP-genotype data.

The position of Chetri and Tharu as outliers in PCAs might be due to their different genetic affinity with other South and East Asian populations, or to gene flow events into these individuals. To distinguish between these two hypotheses, the Himalayan populations were compared with 1000 Genomes Project individuals using: 1) Pairwise F_{ST}; 2) TreeMix migration edges; 3) *f3-outgroup* statistics; and 4) local ancestry analyses using PCAdmix.

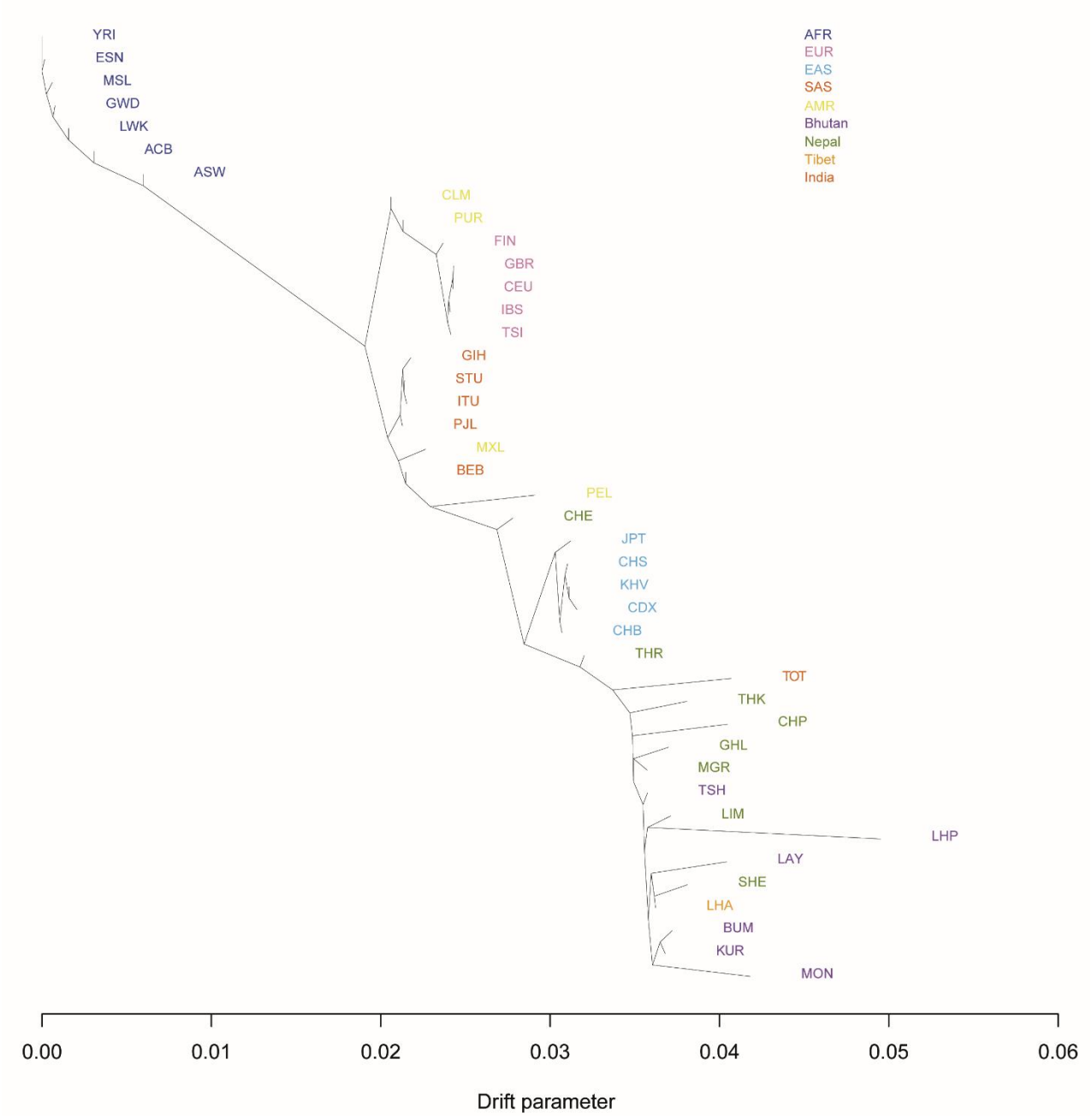


Figure 3. 12 TreeMix results from the worldwide dataset. The tree displays a phylogeny of the Himalayan and other worldwide populations. The x-axis represents the amount of genetic drift. The tree shows long branches for Toto, Mönpa, Lhokpu and Chepang, in agreement with their strong genetic drift patterns.

The pairwise F_{ST} shows five clusters with high genetic affinity that reflect geographical distributions: the first one including African populations, the second one of Europeans, the third of East Asians, the fourth one of South Asians and the last one including 6 Himalayan populations. Interestingly, this last one includes populations from all the three geographical areas (Nepal, Bhutan and the Tibetan plateau). Subsequently, genetic affinity is also found between South Asians and Europeans, and within Himalayan

populations. Most of the Himalayan populations show higher genetic affinity to East Asians. On the other hand, Chetri and Tharu have higher genetic affinity with South Asians. Chetri and Tharu show also genetic similarity to Europeans. Lhokpu, followed by Toto and Chepang show higher genetic differentiation from the rest of the Himalayan populations sequenced here (Figure 3.13).

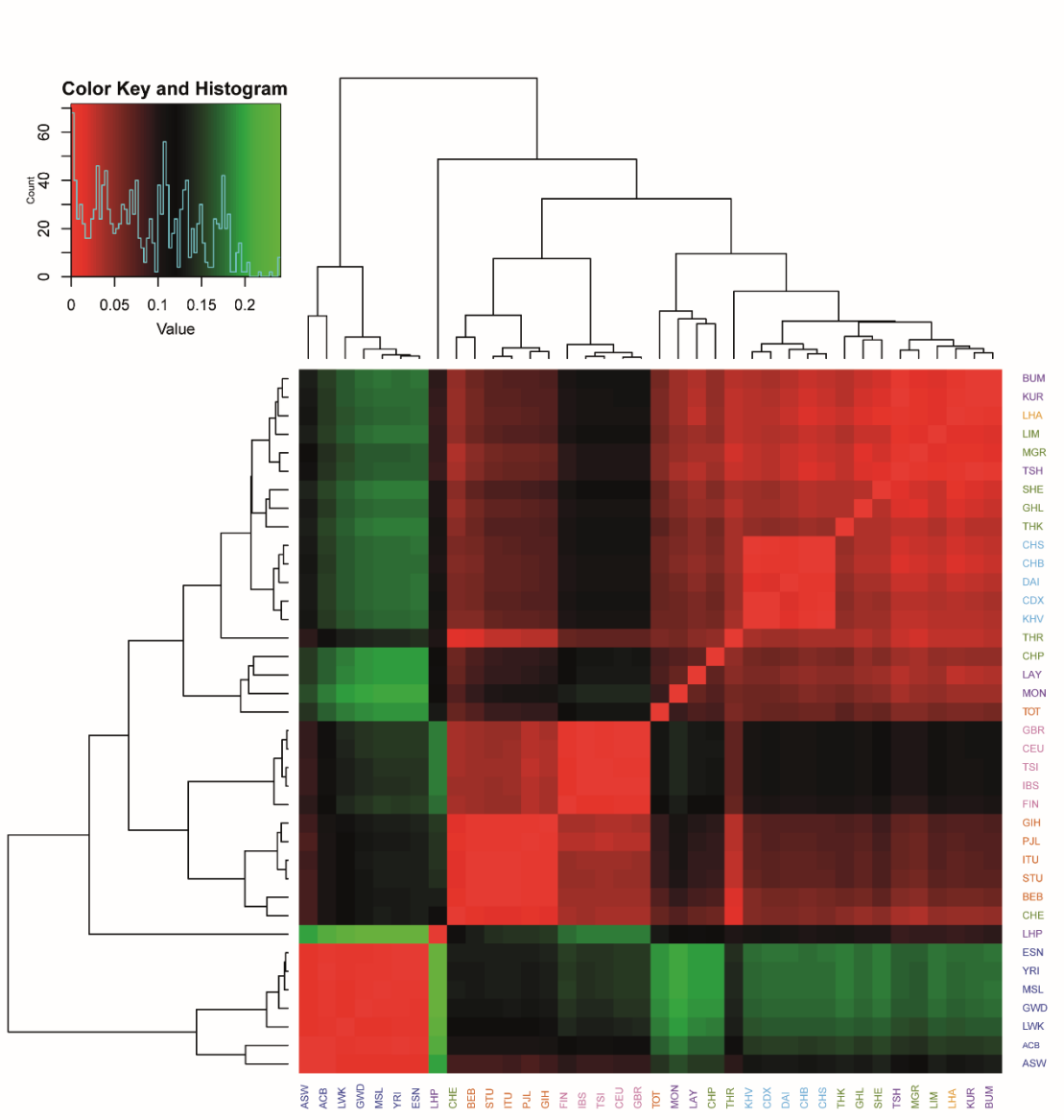


Figure 3. 13 Heatmap of pairwise F_{ST} values for the worldwide dataset. The amount of genetic divergence is indicated by the colour scheme at the top left corner.

TreeMix analysis indicates possible gene flow into Tharu and Chetri from other South Asian, European and the Nepalese Limbu populations. Migration edges from European individuals could be due to European admixture into South Asian individuals or shared ancestral components (194). Finally, it is possible to detect gene flow from Chetri and Tharu ancestries into Toto individuals (Figure 3.14).

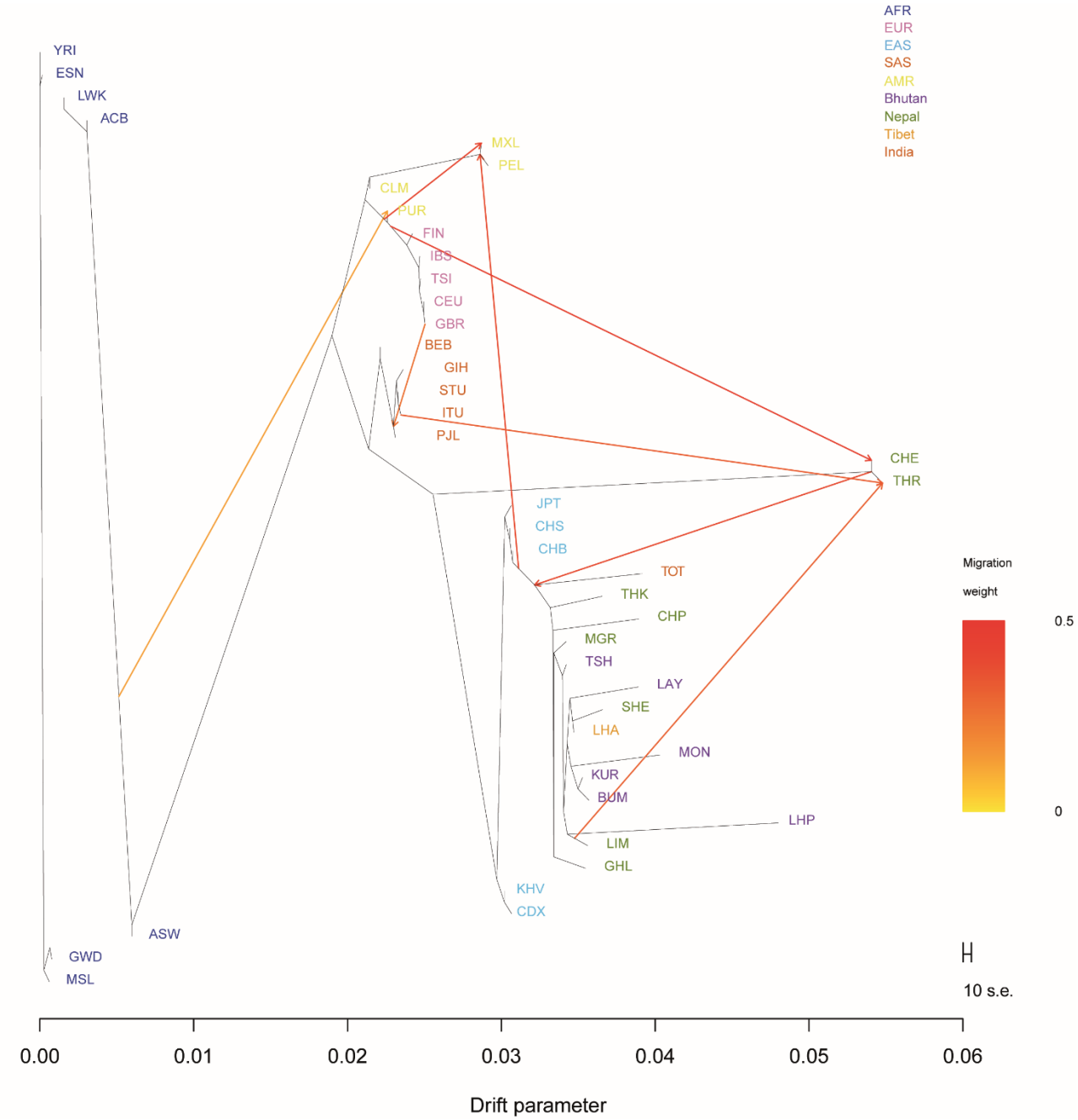


Figure 3. 14 Migration edges for the worldwide dataset from the Treemix analysis. The colour intensity of the arrows indicates the migration weight.

The *f3-outgroup* statistics based on the phylogeny *f3*(X, Y; YRI) for both rare and common variants shows that the Chetri and Tharu share less genetic affinity with East Asian individuals (common variants: 0.155 <*f3*< 0.172, rare variants: 0.027 <*f3*< 0.037) compared to other Himalayan populations (common variants: *f3*>0.189, rare variants: *f3*>0.070), with Tibetans, Bhutanese, Limbu and Sherpa from Nepal showing the highest genetic affinity to East Asian individuals (Figures 3.15 and 3.16). On the other hand, Chetri and Tharu display more genetic similarity to South Asians and Chetri show an excess of genetic sharing with European individuals. Interestingly, for rare variants, other Nepalese populations, Toto from India and Tshangla from Bhutan show more genetic similarity to South Asians compared to other populations. Moreover, rare variants also give more detailed sharing with East Asians: Himalayan populations display more genetic sharing with CHB compared to CDX and KHV (Figure 3.15). Finally, all Himalayan populations show a similar pattern of genetic drift when Bengali population (BEB) is used as a proxy for South Asia (Figure 3.15 and 3.16). This could be due to the potential admixture of Bengali individuals with East Asians (259, 260).

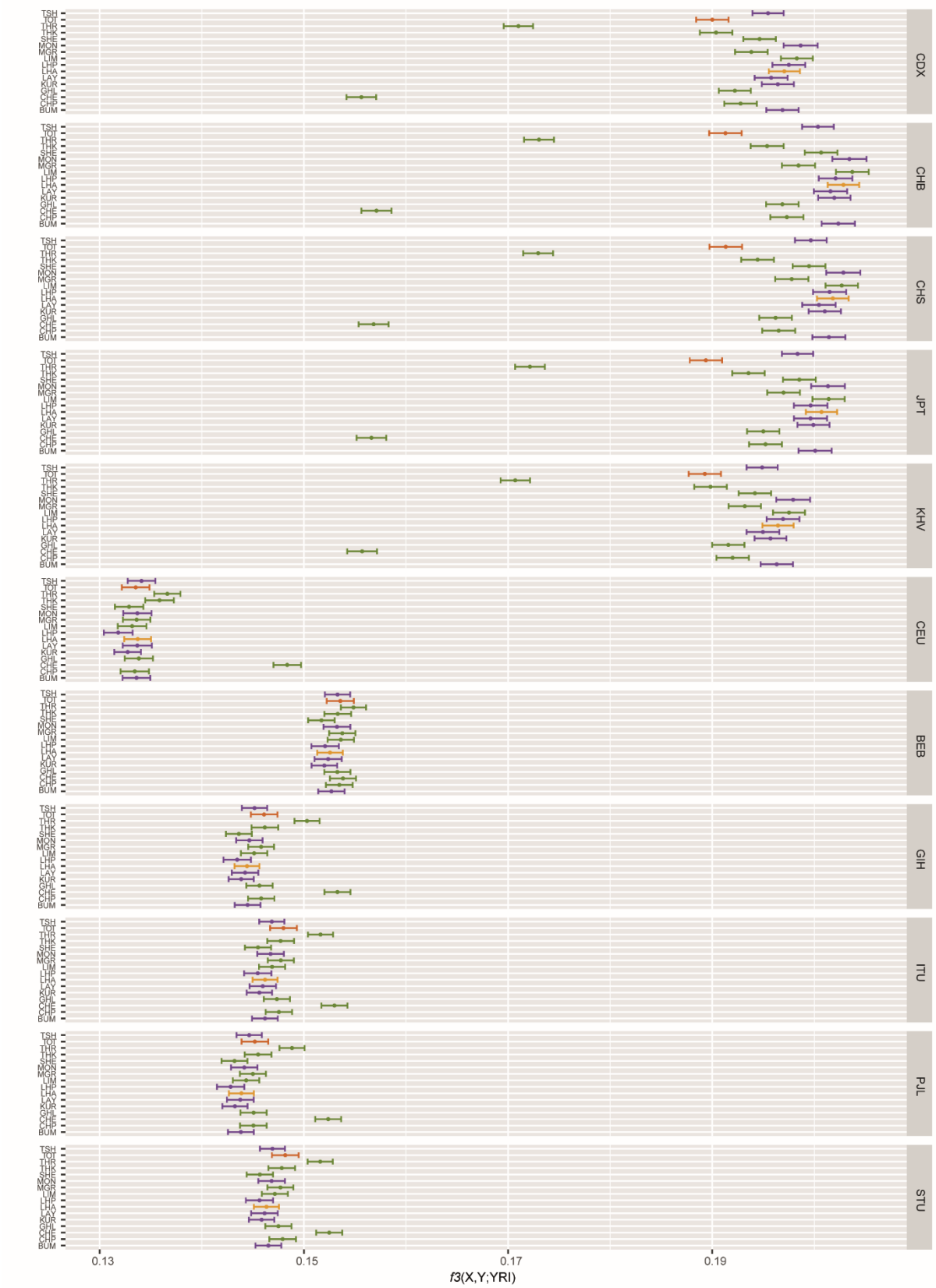


Figure 3. 15 Common variant outgroup *f3* statistics for the Himalayan samples. The plot shows the amount of genetic affinity (x-axis) of each Himalayan population (y-axis) to each of the worldwide populations (individual panels).

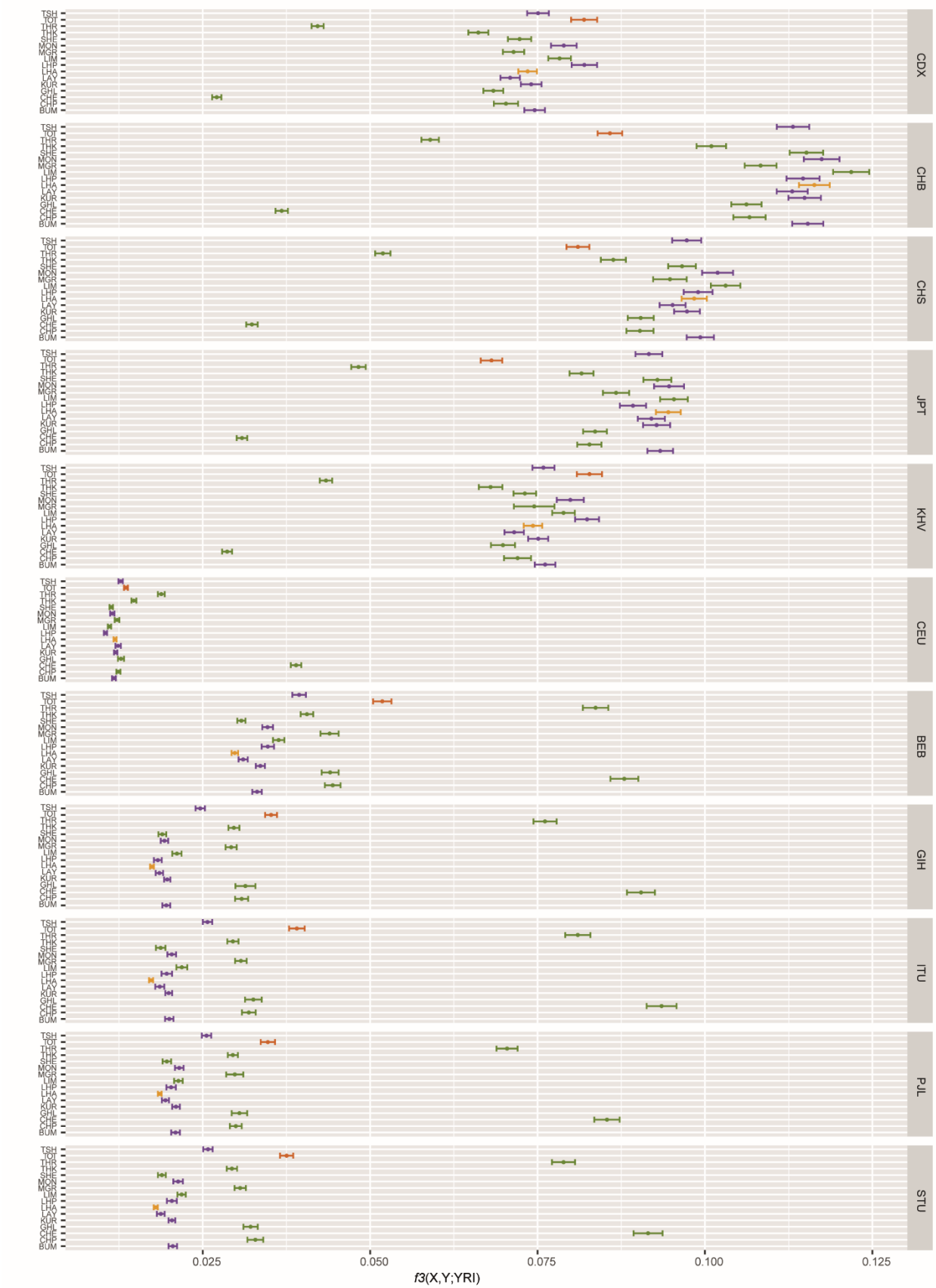


Figure 3. 16 Rare variant outgroup *f3* statistics for the Himalayan samples. The plot shows the amount of genetic affinity (x-axis) of each Himalayan population (y-axis) to each of the worldwide populations (individual panels).

This pattern of genetic affinity to South and East Asians is confirmed when these populations are used as sources in PCAdmix analysis. Tharu and Chetri individuals show the highest proportion of South Asian ancestry haplotypes, whereas Tibetans, display the lowest values. Bhutanese populations and the Nepalese Limbu and Sherpa have a low proportion of shared South Asian ancestry comparable to the Tibetans (Figure 3.17).

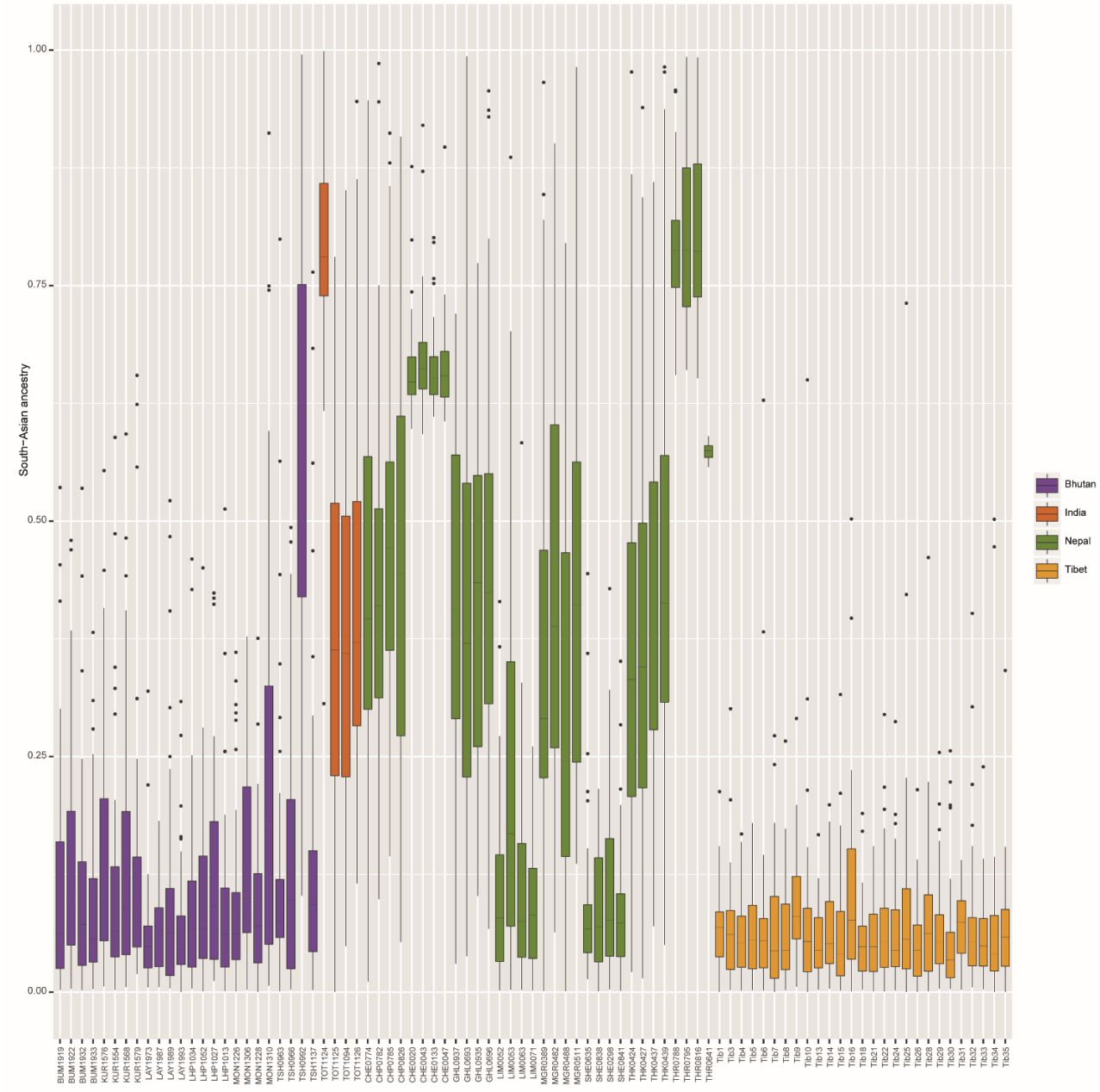


Figure 3. 17 Local ancestry PCAdmix analysis. The plot shows the genome-wide proportion of South Asian ancestry haplotypes for each Himalayan individual.

3.3.2.2 Himalayan genetic variation

Heterozygosity rate, runs of homozygosity (ROHs) and identity-by-descent (IBD) are standard parameters to study population demography. Heterozygosity is indicative of the genetic variation in the population and very low levels of heterozygosity are indicative of reduced genetic variation, for example due to events such as population bottlenecks and genetic drift that have severe effects on population sizes. In contrast, high heterozygosity suggests high levels of genetic variability, and gene flow may have played an important role by introducing new alleles. Himalayan populations show different degrees of heterozygosity, with Lhokpu, Monpa and Toto having the lowest values whereas Chetri and Tharu have the highest (Figure 3.18).

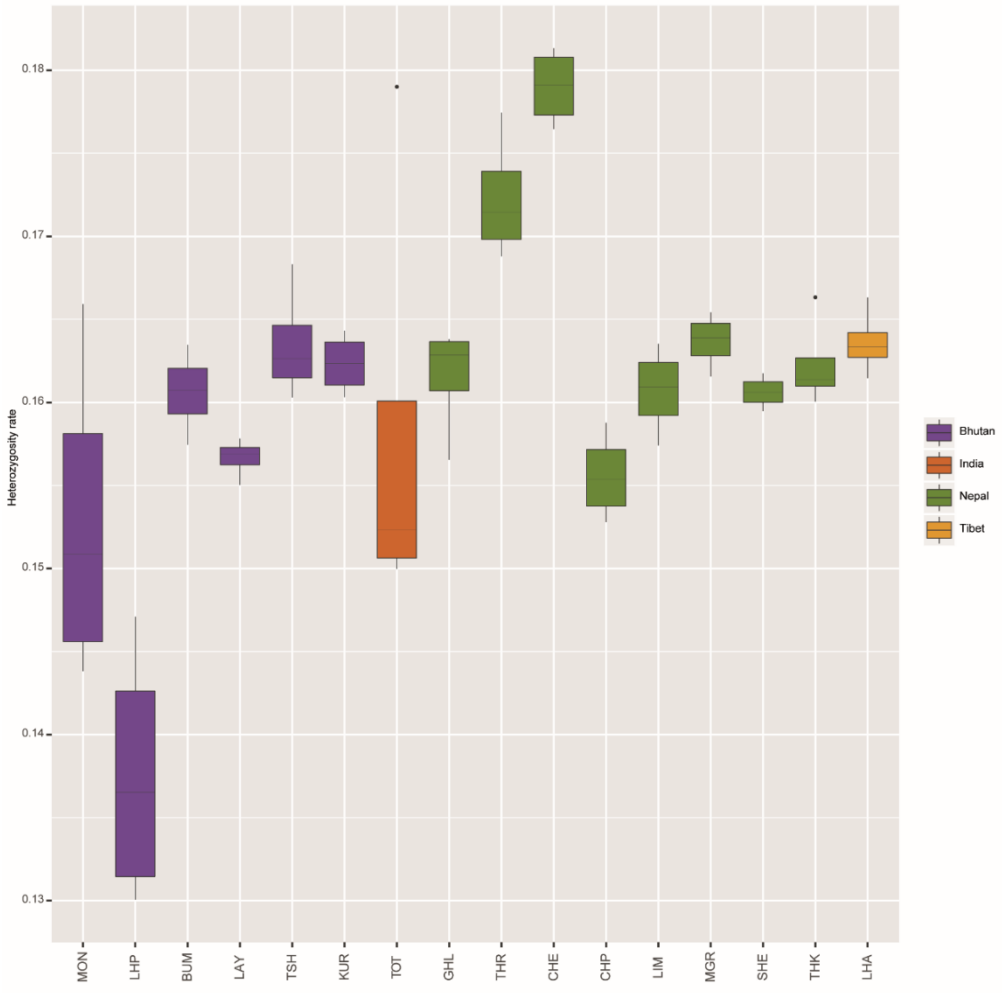


Figure 3. 18 Heterozygosity rate in the Himalayan populations. The plot shows the genome-wide heterozygosity rate for each population.

Compared to other South and East Asians populations, Himalayan individuals show values of heterozygosity similar to their neighbours and in agreement with previous estimates (261, 262). In particular, most Nepalese, Bhutanese and Tibetan populations show similar levels of heterozygosity to other South Asians, whereas Lhokpu, Mönpa, Toto and Chepang show values similar to other East Asians, who have an overall lower heterozygosity rate than other 1000 Genome Project populations (heterozygosity ratio ≈1.4) (261)). Chetri and Tharu show the highest heterozygosity rates even compared to other South-East Asian samples. This could be due to the excess of rare heterozygous variants in these populations (Figure 3.19).

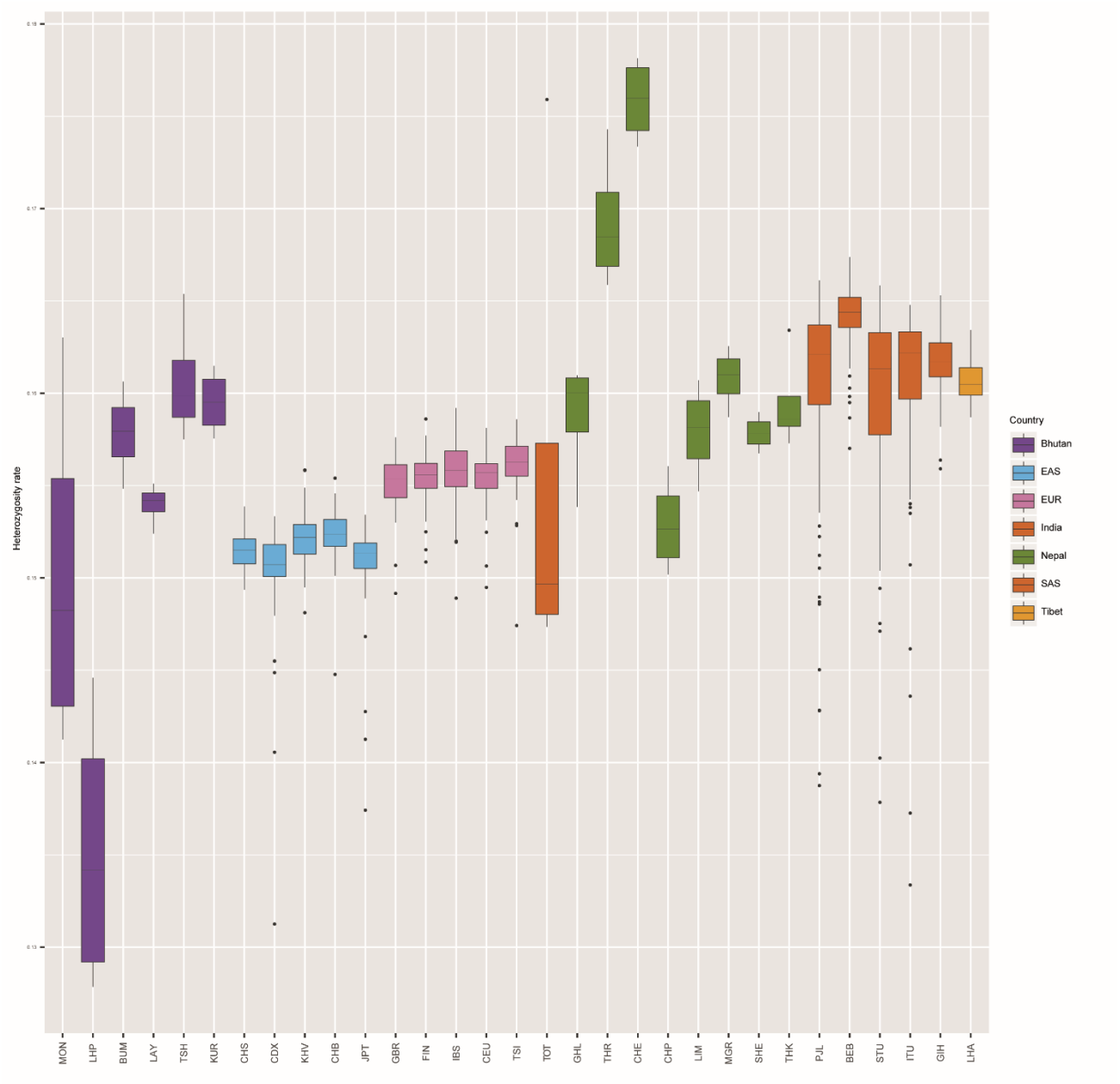


Figure 3. 19 Heterozygosity rate in the Himalayan populations compared to other worldwide populations.

The number and length of ROHs and IBD tracts are strong predictors of population isolation, genetic drift and inbreeding. Similarly to the results in the previous chapter, ROHs are widely distributed across human populations but, generally, isolated populations show higher numbers of ROHs whereas large and admixed populations have fewer and shorter ones (263). Similarly, IBD is a useful measurement of relatedness between individuals and thus used to estimate population bottlenecks and substructure (264). In agreement with the analyses performed in the previous chapter, Bhutanese populations overall show the highest proportion of ROHs whereas Tibetans have the

lowest. Lhokpu, Kurtöp and Thakali are characterised by the highest number and longest ROHs (Figure 2.6A).

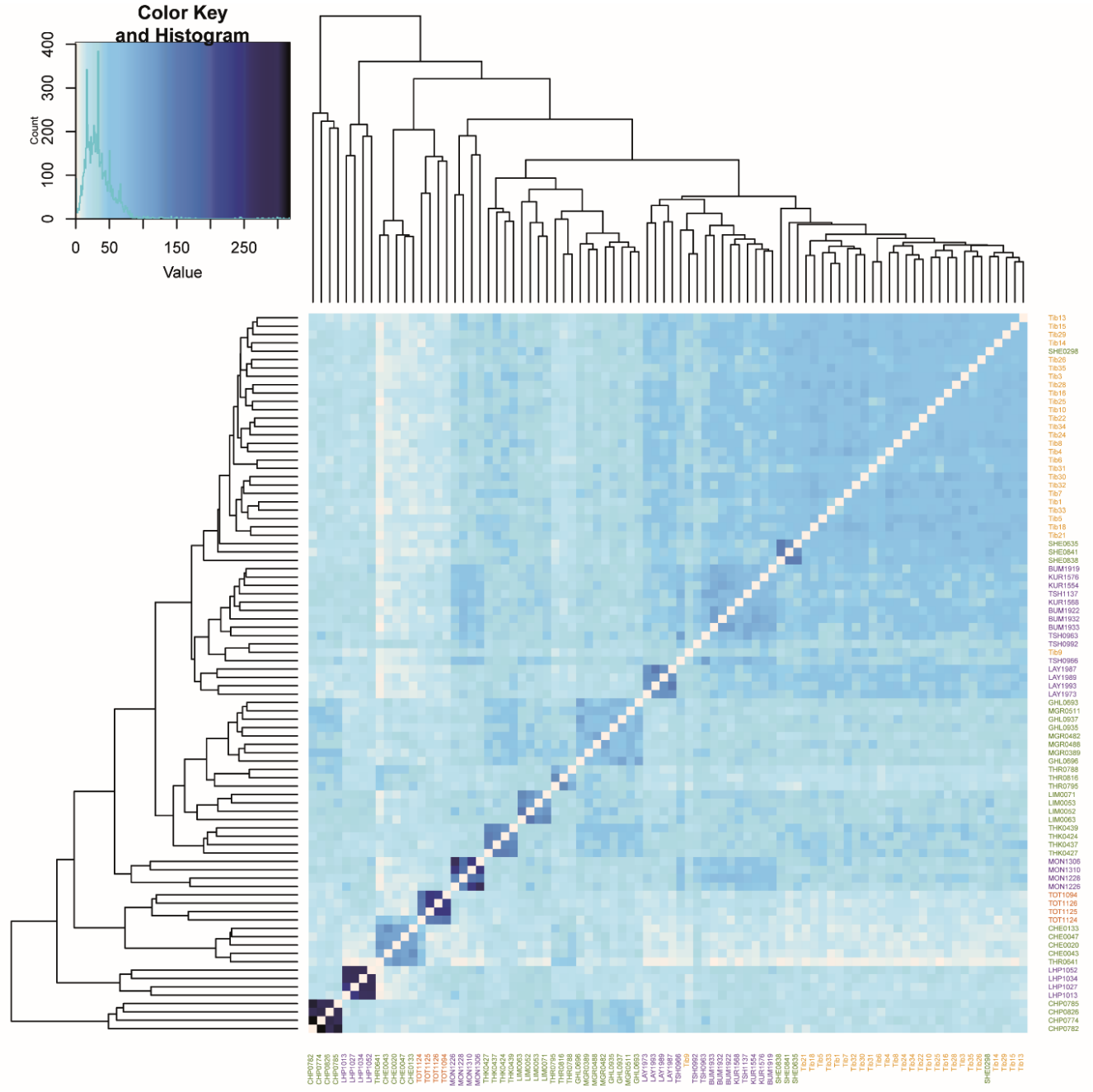


Figure 3. 20 Heatmap of shared IBD tracts. The plot shows the number of shared IBD tracts between different Himalayan individuals. The number of tracts is shown by the colour scheme in the top left corner.

Lhokpu, Mönpa, Layap and Toto also show the highest number and longest IBD tracts, while the Nepalese Magar, Chetri and Tharu show the lowest. The pairwise matrix of shared IBD tracts in the Himalayan populations replicates many of the patterns of

genetic similarity found with previous analyses. Populations showing signatures of genetic drift and isolation, Lhokpu, Chepang, Mönpa and Toto, share the highest number of IBD segments within individuals of the same population (Figure 3.20). Tibetan, Sherpa and other Bhutanese populations form a cluster in agreement with the PCA, ADMIXTURE and F_{ST} analyses. Another smaller cluster is formed by individuals from the Nepalese Ghale and Magar.

3.3.2.3 Rare variant sharing

The singletons (Figure 3.4) found in the Himalayan dataset tend to be shared with the East and South Asian populations in the 1000 Genome Project: Tibetans from Lhasa share the most with East Asians, and Tharu and Chetri with South Asians (Figure 3.21). More detailed information is gained by looking at *f2* variants in all the individuals from the Himalayas and the 1000 Genomes Project. *f2* variants tend to be shared more between individuals from the same population or geographical group, or between recently admixed groups (177) (Figure 3.22). However, within the Himalayan populations, sub-clusters of *f2* variant sharing were observed. Himalayan populations share *f2* variants at three levels: within each population, between different Himalayan populations and with other South and East Asians. Magar, Ghale, Thakali and Chepang form a clear cluster, while Tibetans, Tharu, Chetri, Toto, Lhokpu and Mönpa form a second cluster. An additional cluster was formed by Tshangla, Bumthang Limbu and Kurtöp. This is consistent with the recent shared ancestry of the Himalayan populations within themselves and with other neighbouring populations. Interestingly, *f2* variant sub-clusters do not follow geographical proximity; populations that are physically distant group together (Figure 3.23). Finally, the sharing of the subset of *f2* variants with one alternative allele found in Himalayan populations and one in the 1000 Genomes Project populations showed that, overall, all Himalayan populations share singletons mostly with South and East Asian individuals. Tibetans from Lhasa have the highest number of singletons shared with East Asians, in particular with Chinese Han (firstly CHB and secondly CHS), in agreement with the known recent shared demographic history of Tibetans and Chinese Han individuals (50). It is important to note that the overall higher number of shared singletons in Tibetans compared to other Himalayan populations may be due to the larger sample size (27 vs 4). Chetri and Tharu display higher singleton

sharing with South Asians compared to other Himalayans and, overall, all Himalayan populations share more singletons with the Bengali populations, confirming the pattern of the outgroup *f3* statistics. Intriguingly, Chetri and secondly Tharu show higher variant sharing with African populations. Chetri and Tibetan individuals show a slightly increased sharing with Europeans (Figure 3.18).

Figure 3. 21 Heatmap of singleton sharing between Himalayan and 1000 Genomes Project individuals. The plot shows sharing of variants that are singletons in both the Himalayan and 1000 Genomes Project datasets.

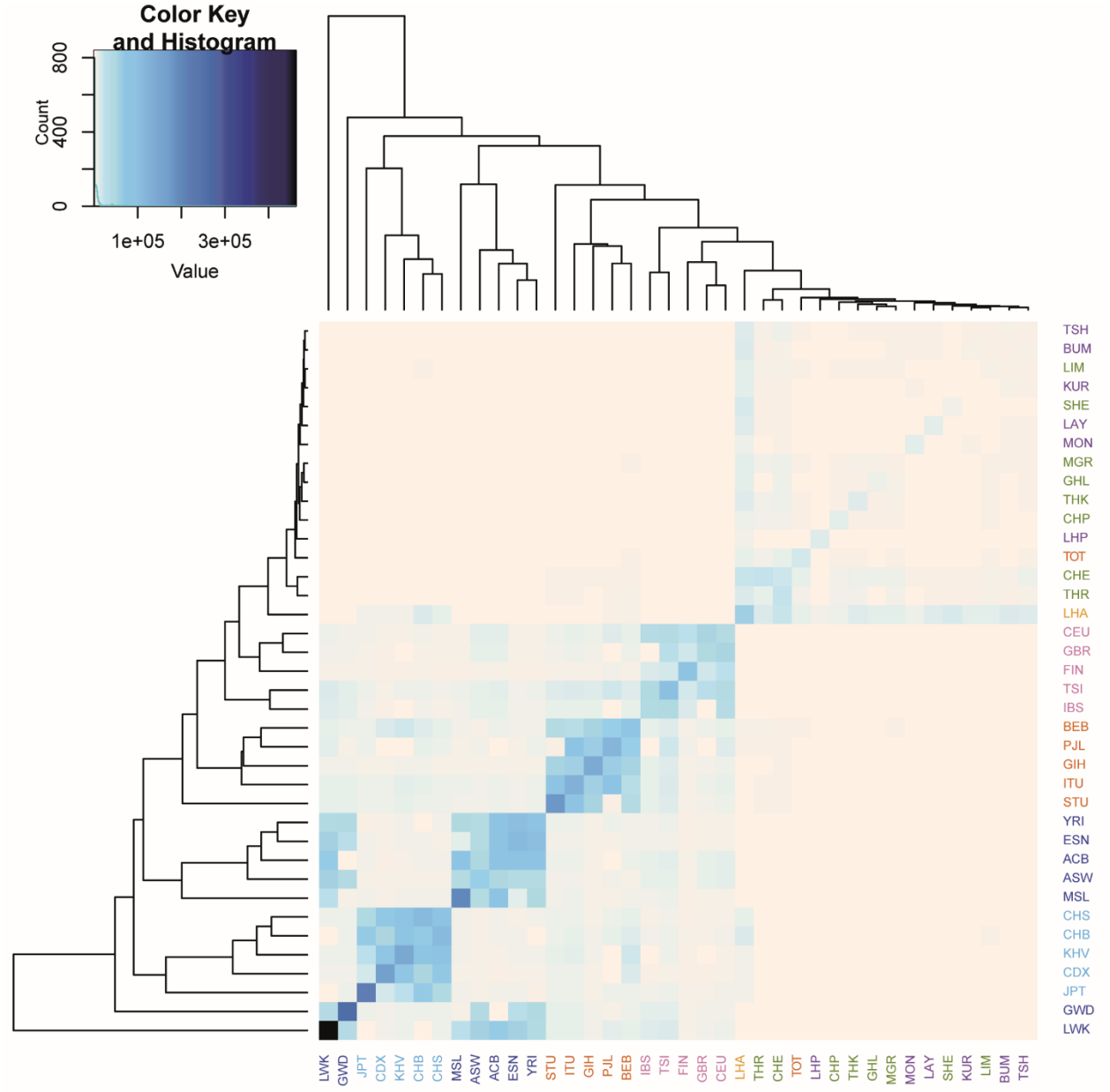


Figure 3. 22 Heatmap of *f2* variant sharing in the worldwide dataset. The highest *f2* variant sharing is within populations of the same geographical region.

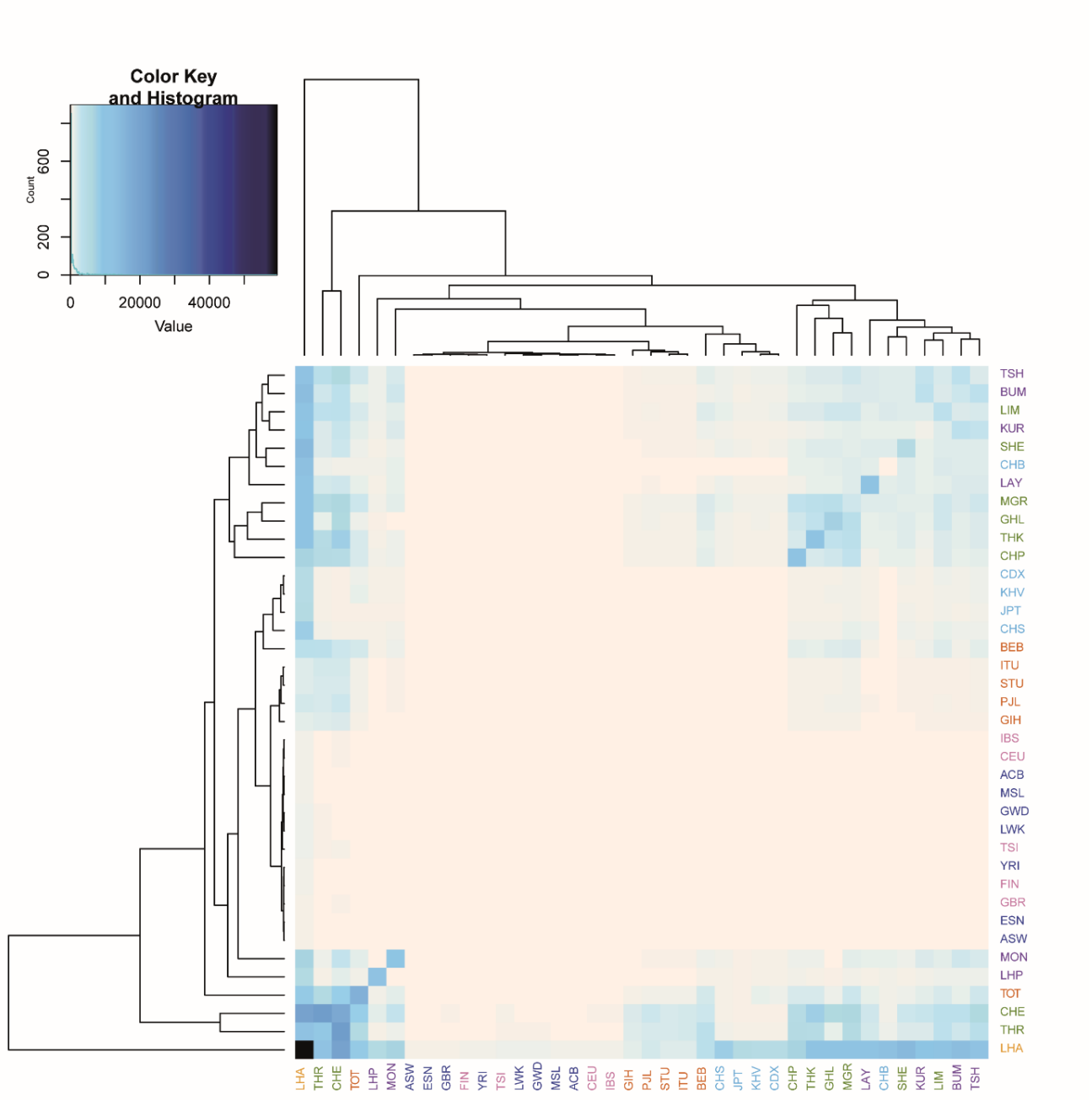


Figure 3. 23 Zoom showing the Himalayan samples from the heatmap of *f2* variants in Figure 3.22. Himalayan populations share *f2* variants within themselves and with other East and South Asian populations.

3.3.2.4 Recent population expansion

3.3.2.4.1 Coalescent based demographic inference

The effective population size (Ne) history of Himalayan populations was investigated using three methods: MSMC2 and SMC++. MSMC2 estimates effective population size in older times, whereas in more recent times (<~1e+04 years) it loses resolution because there are few very recent coalescences. SMC++ is somewhat more accurate in estimating Ne in recent times as it incorporates allele frequencies from a larger number of individuals, but it requires a considerable sample size to attain this potentiality. MSMC2 and SMC++ gave similar results for Ne values of the Himalayan populations, and in a similar time frame, with some showing an increase of Ne in more recent times but others maintaining low Ne. In MSMC2 runs, most of the populations show a Ne curve similar to other East Asian individuals, with the exceptions of Chetri and Tharu who display higher Ne than other Himalayan populations, and similar Ne curves to other South Asians (Figure 3.24).

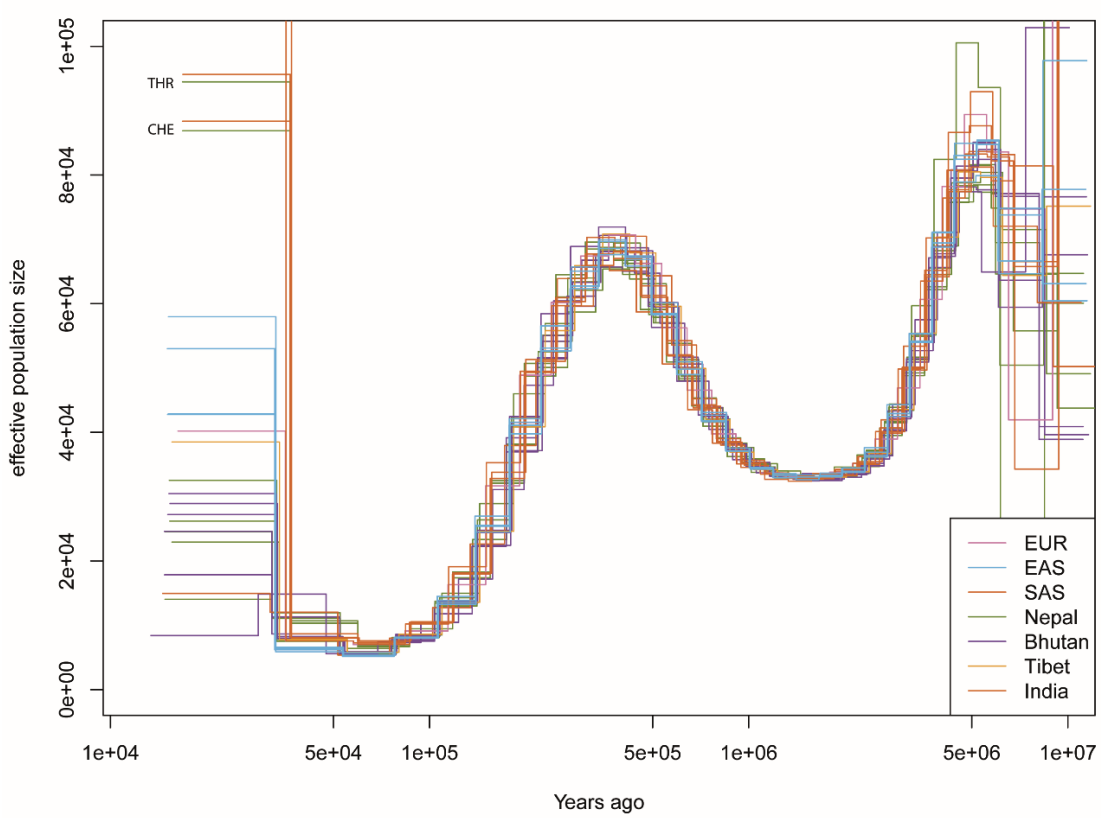


Figure 3. 24 MSMC2 Ne plot. The plot shows the Ne for the different Himalayan individuals for each population compared to other European, South and East Asians. The x-axis shows the time in years, and the y-axis the Ne values.

Within the Himalayan samples, Chepang, Lhokpu, Mönpa and Toto show the lowest values of recent Ne, in agreement with a scenario of population isolation and genetic drift, which also supports the results from the heterozygosity rate, ROHs and IBD analyses. On the other hand, Chetri, Tharu and Lhasa show the highest recent Ne values within the Himalayas, in agreement with their recent population growth (Figure 3.25).

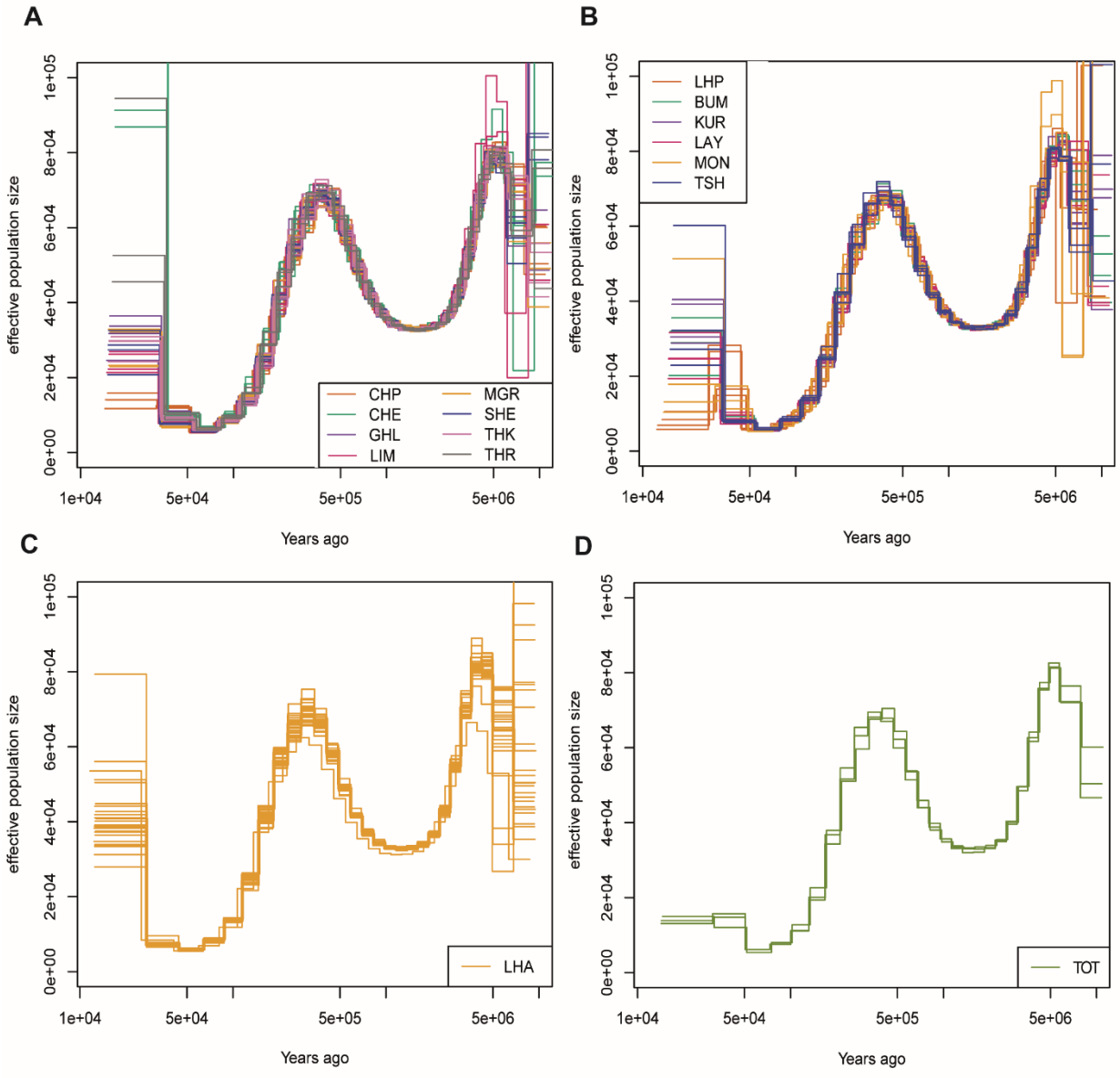


Figure 3. 25 MSMC2 Ne plot. The plots show the Ne for the different Himalayan individuals coloured by country. The x-axis shows the time in years and the y-axis the Ne values. A. Nepal, B. Bhutan, C. Tibet, D. India.

Overall, SMC++ runs display comparable results to MSMC2, with generally low recent Ne values across Himalayan populations. Chetri, Tharu and Tibetans from Lhasa show the highest Ne compared to other Himalayans. The higher Ne and different curve shape in Tibetans may be due to the bigger sample size. Lhokpu show the lowest Ne together with Mönpa and Toto (Figure 3.26). It is important to note that the low sample sizes for most Himalayan populations prevent reliable estimates in recent times and the drop in Ne in recent times in some curves may be affected by this technical limitation.

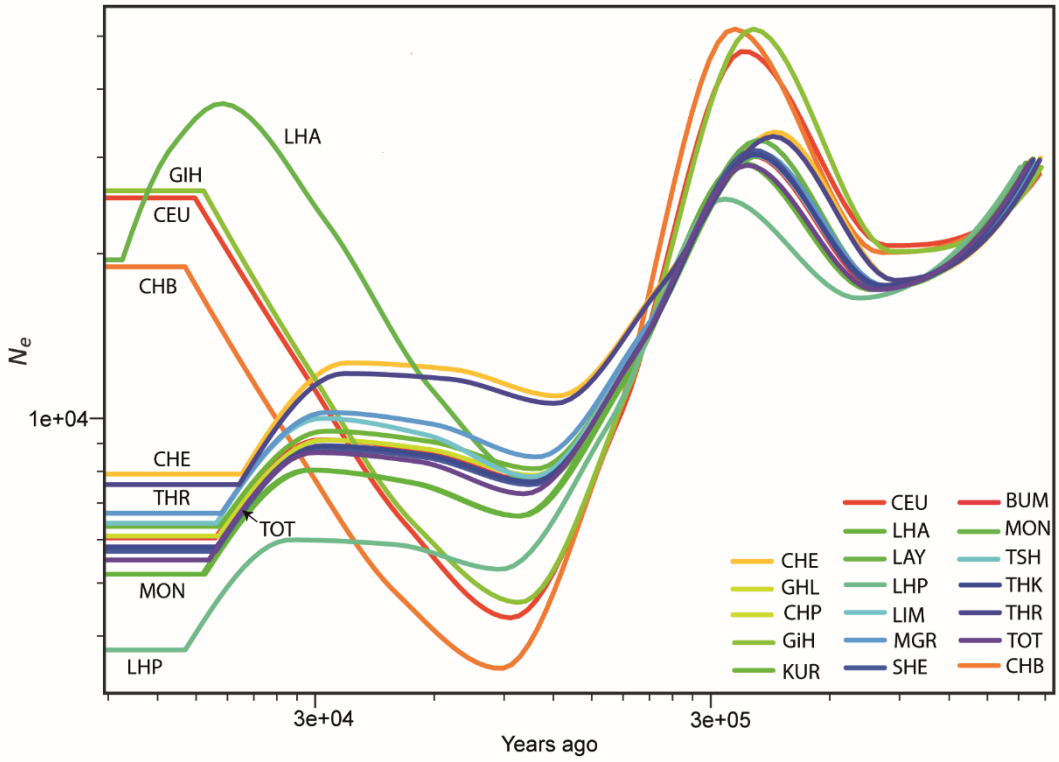


Figure 3. 26 SMC++ Ne plot. The plots show the Ne for the different Himalayan individuals, together with CEU, CHB and GHI. The x-axis shows the time in years and the y-axis the Ne values.

3.3.2.4.2 Y chromosome and mtDNA demographic inference

The Y chromosomes and mitochondrial DNA (mtDNA) of 87 Himalayan individuals were studied and compared with other worldwide populations to understand the patterns of population structure and migration as reflected by uniparental makers. Both Y chromosome and mtDNA haplogroups of the Himalayan individuals show a mixture of South and East Asian lineages. The most common haplogroups for the Y chromosome are D1 and O2. The O2 lineage, common in South-East Asian populations

(265, 266), is the most widespread in the Himalayan region (Figure 3.27). Except for Layap, at least one individual from each population belongs to the O2 lineage. Within O2, O2a1 and the O2a2b1 form two separate clades (Figure 3.27).

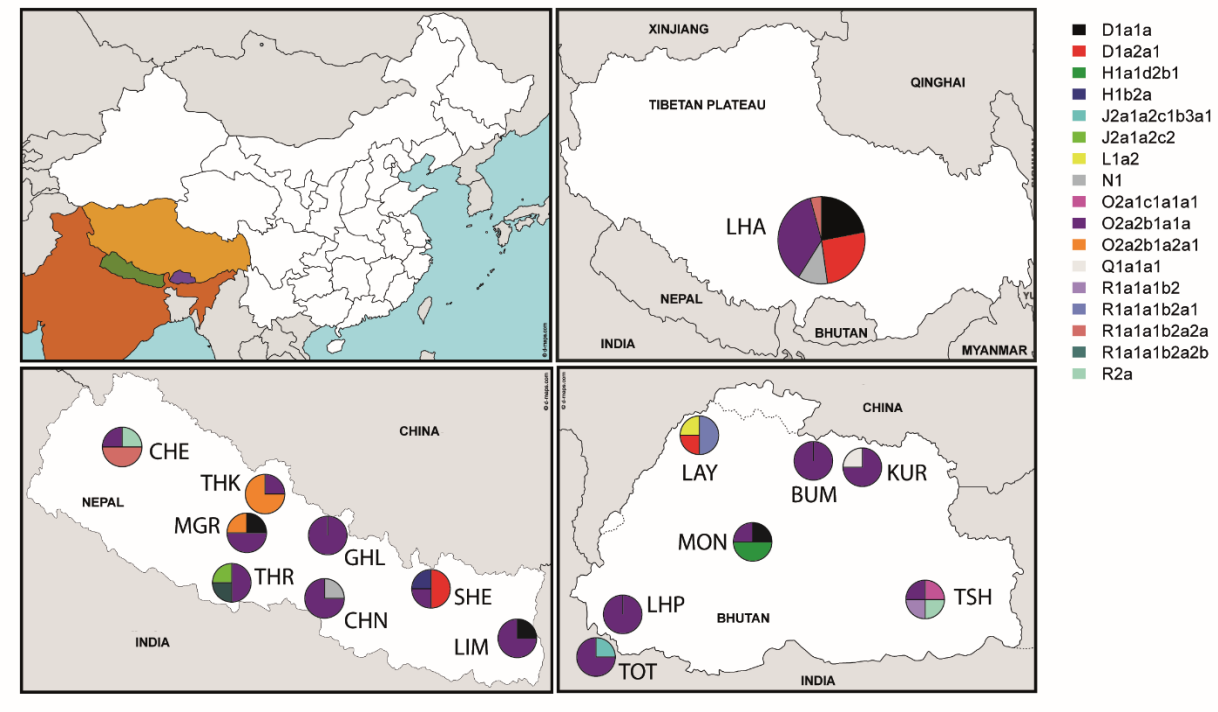


Figure 3. 27 Map of Y chromosome haplogroups. The circle areas are proportional to the sample size (Bhutanese, Nepalese and Indian = 4, Lhasa=27).

The phylogenetic tree of the Himalayan Y chromosomes plotted together with the SGDP samples showed the expected structure, with the Himalayan individuals closely clustered with either East or South Asians (Figure 3.28). The most striking feature is the star-like expansion of O2a2b1 lineages. Its TMRCA is around 1000 years ago, suggesting an enormous male population expansion at this time (Table 3.6). In addition, smaller star-like expansions are seen in D1a2a1 and R1a1a1b2. The D lineage, that is common in Tibetan and Japanese and Andamanese populations (267), is present in Tibetan, Sherpa and Layap individuals with the haplogroup D1a2a, whereas the D1a1a is shared between Tibetans, one Mönpa and one Limbu individuals. It is interesting to note that the D1a2a1 cluster contains all high altitude populations (Tibetans, Sherpa and Layap), resembling the pattern seen before in this chapter from autosomal data (Figure 3.28). The split time for the D lineages are very recent with a TMRCA around 600 years ago (Table 3.6). These times are consistent with one another and with the recent split time between the high

altitude populations (see the section below). Although the O2 and D haplogroups, that indicate an East Asian genetic substrate, are the most prevalent in the samples, South and Central Asian haplogroups are also present: the R1a1* haplogroup in Chetri, Layap and Tharu individuals showing recent split times although older than O2a2b1 and D1a2 lineages (TMRCA ~5800 years ago) (Table 3.6) and the R2a* in Chetri and Tshangla that are mostly prevalent in the Indian sub-continent (268).

Haplogroup	TMRCA (years ago)	95% C.I. TMRCA (years ago)
O2a2b1	~ 1,100	~980-1,250
R1a1a1b2	~5,800	~5,100-6,600
D1a2a1	~580	~500-640

Table 3. 6 Y chromosome split times within Himalayan individuals. The table reports the TMRCA and their 95% C.I. for most prevalent lineages in the Himalayas.

The phylogenetic tree reconstructed using the 1000 Genomes Project samples, both GRCh37 and GRCh38, show a very similar phylogeny to the one reconstructed with the Simon Diversity Project individuals. (APPENDIX A).

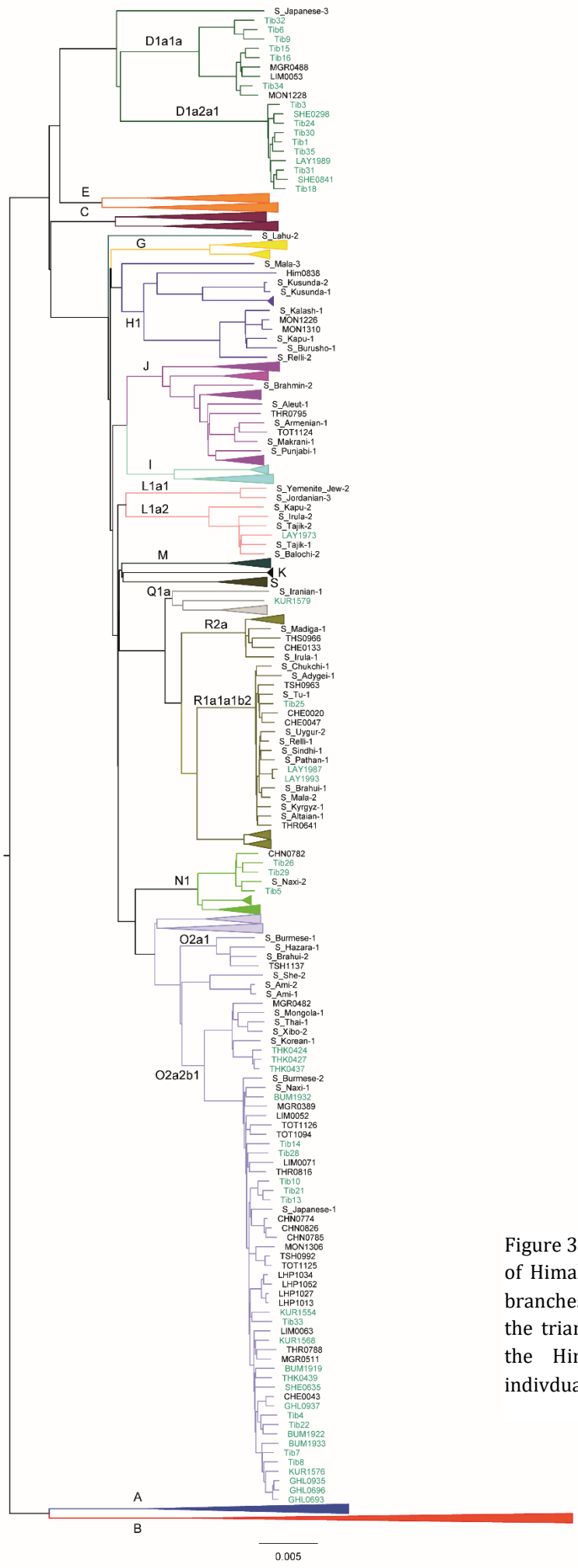


Figure 3. 28 Phylogenetic tree of the Y chromosomes of Himalayan and worldwide (SGDP) samples. The branches are coloured to indicate the haplogroup and the triangles represent SGDP lineages absent from the Himalayan individuals. The high altitude indivduals are coloured in green.

The mtDNA analysis shows substantial haplogroup variability, with a mixture of East and South Asian lineages (Figure 3.29). The most common lineages are M and the D. However, a clear separation between haplogroup distributions is noticeable. The D haplogroups, prevalent in North and Eastern Asia (269), are mostly seen in Nepal and the region neighbouring it, where Lhokpu and Toto live. The M lineage is widespread in South and East Asia with M2, M3, M4 and M5 mostly present in South Asia (270) and also found in lowland Himalayan populations. In contrast, high altitude populations from Tibet, Bhutan and the Sherpa show mainly M9a that previously has been associated with hypoxic adaptation in Tibetans (271). Sherpa also carry lineages within the A and C haplogroups which have previously been reported to be Sherpa-specific (117).

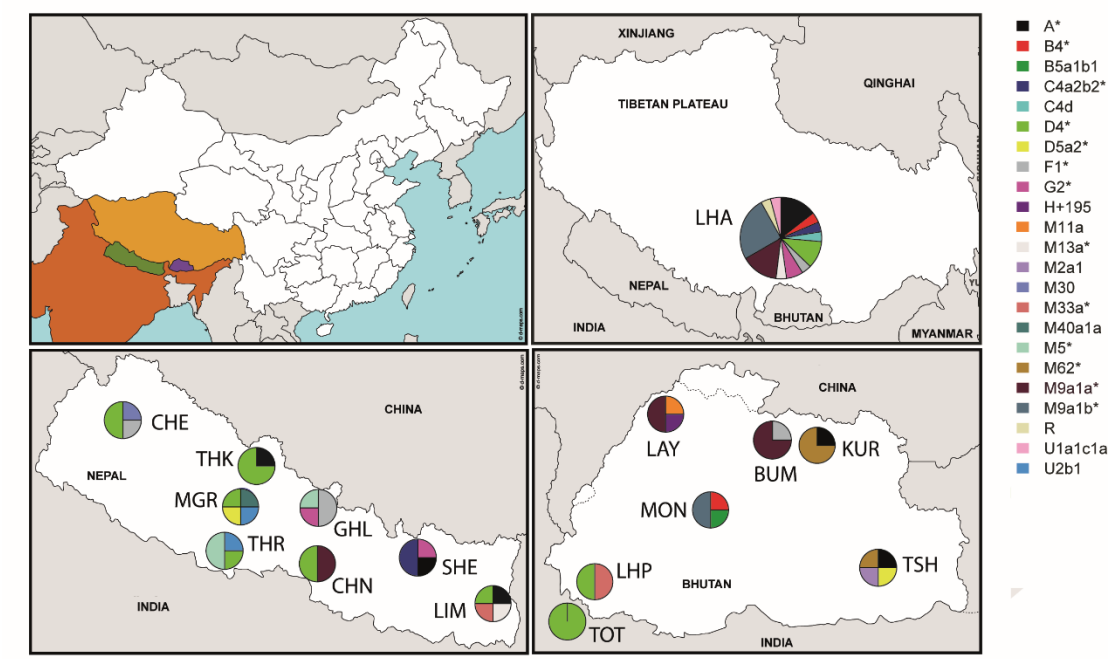


Figure 3. 29 Map of mtDNA haplogroups. The circle areas are proportional to the sample size (Bhutanese, Nepalese and Indian = 4, Lhasa=27).

In contrast to the Y chromosome, Himalayan mtDNAs do not form large clades on the phylogenetic tree, but tend to be dispersed with other worldwide individuals, in particular from South and East Asia. A main cluster of Himalayans is formed by individuals of the haplogroup M9a1a (Figure 3.30).

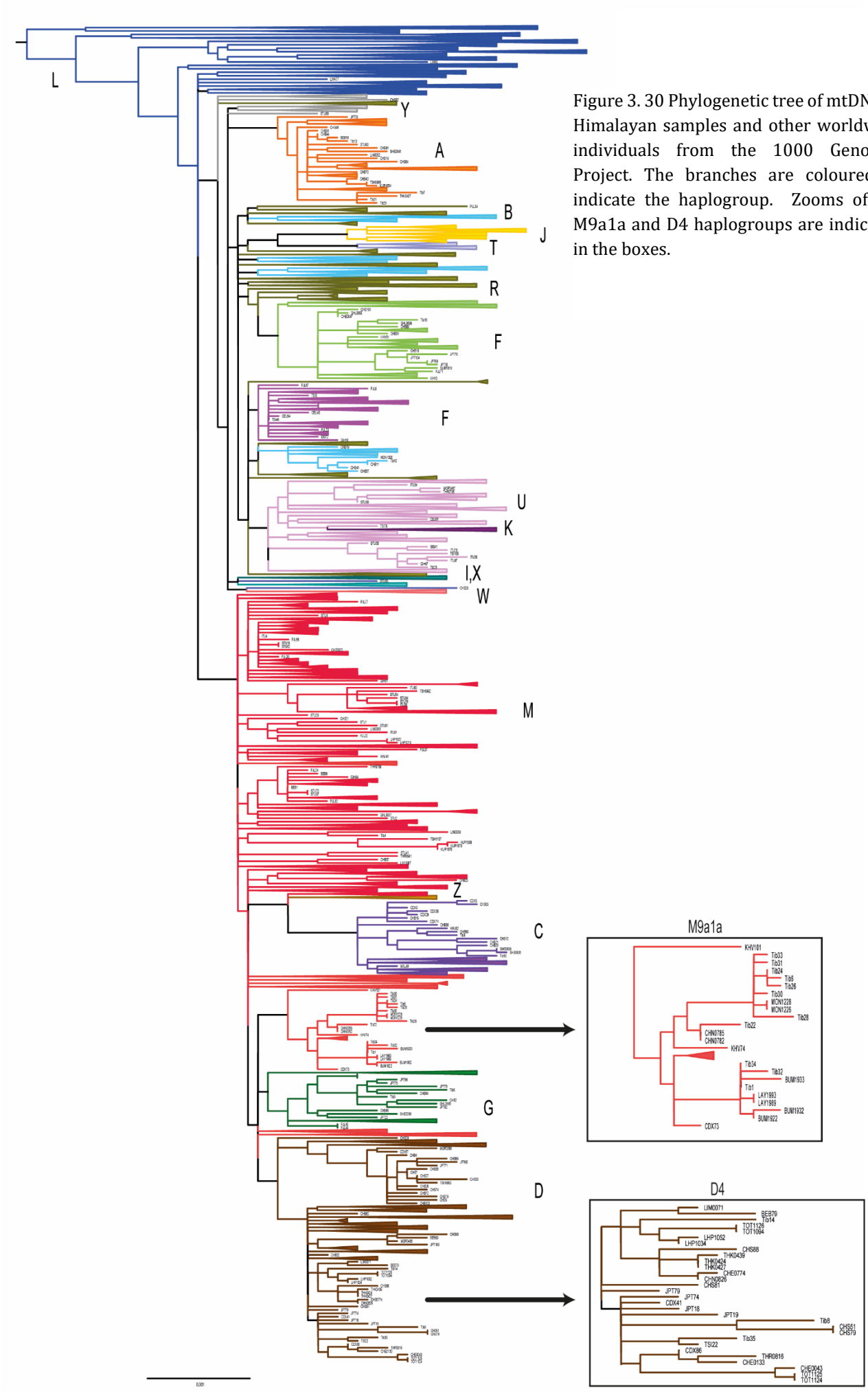


Figure 3. 30 Phylogenetic tree of mtDNA of Himalayan samples and other worldwide individuals from the 1000 Genomes Project. The branches are coloured to indicate the haplogroup. Zooms of the M9a1a and D4 haplogroups are indicated in the boxes.

Some of the terminal branches in the different haplogroups, in particular M, include two or more Himalayan individuals with identical sequences, and thus indicate very recent divergence. The oldest pairwise split time is dated around 4,700 years (95% C.I. 3,700-6,800 years) between one Magar and one Tharu individuals belonging to the haplogroup U. I found 25 heteroplasmies that, according to VEP, are all intergenic and without any identifiable functional consequences.

The analyses of the Y chromosome and mtDNA in the Himalayans show that both are characterised by South and East Asian lineages. The most striking feature is that Y chromosome haplogroups D, R1a and especially O2 are widespread and common in the region and show signatures of a strong recent male-specific expansion. (Figure 3.28 and 3.30, Table 3.6).

3.3.3 Long term demographic history

3.3.3.1 Population split times from other Asians

Population split times were estimated using MSMC2, taking the 50% cross-coalescence rate as the point estimate for the split time. All of the Himalayan populations split from Africans around 50,000-60,000 years ago (Figure 3.31A), as expected. Chetri shows less than full separation from the Africans at the most recent times, which could be due to the gene flow from Africans, as also seen in ALDER results in the previous chapter (Figure 2.10). This is also true for the split time with Europeans which is around 25,000-35,000 years ago for all the Himalayan populations, but incomplete for Chetri (Figure 3.31B). All of the Himalayan populations split from South Asians around 10,000-20,000 years ago, with Chetri, Tharu, and Toto splitting most recently at around 10,000-11,000 years ago and also showing recent gene flow from South Asians. Ghale also display recent gene flow from South Asian populations (Figure 3.31C). All of the Himalayan populations have recent split times from East Asians at 10,000 years ago or less (Figure 3.31D); the Tibetans from Lhasa and other high altitude populations split from the Chinese Han around 9,000 years ago.

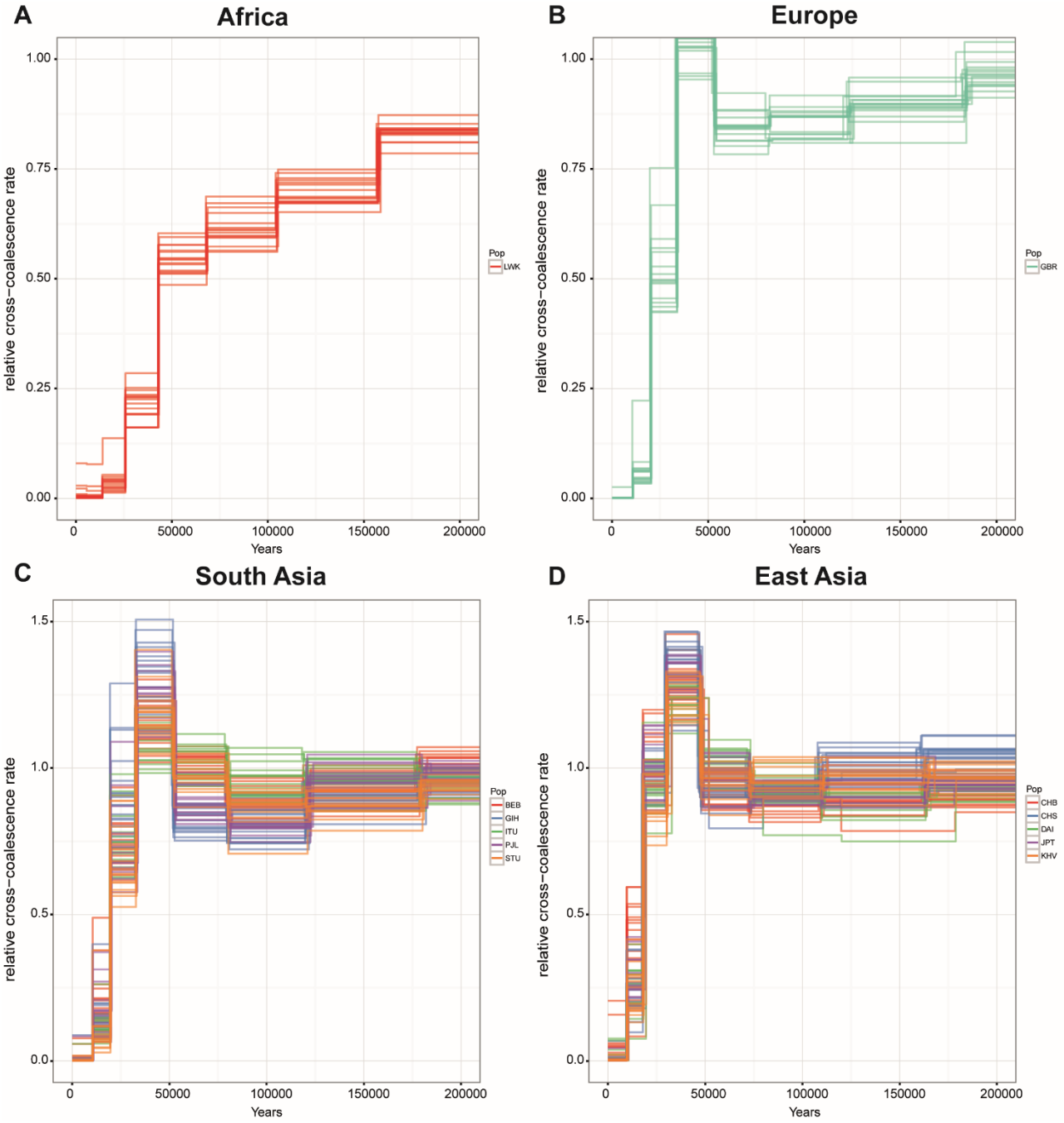


Figure 3. 31 MSMC2 analysis of split times. The x-axis shows the time and the y-axis the cross-coalescence rate. A rate of 0.5 is taken as indicative of population separation. The curves represent the split time for each Himalayan population with the continental populations indicated in the legend. A. Africa, B. Europe, C. South Asia, D. East Asia

Himalayan populations separated from each other between 3,000 and 10,000 years ago, and many of them are not clean splits (Figures 3.32, 3.33, 3.34 and 3.35). High altitude populations show very similar and recent split times, and lack of full separation for many of them, emphasising their genetic similarity and the idea of a single ancestral high altitude population that split from East Asians and then diverged into the current communities (Figure 3.32A). When comparing populations from Tibet and Bhutan, all with a high proportion of East Asian ancestry, Tharu, Chetri and Toto split from Tibetans at an earlier time (Figure 3.32B). Although these estimates cannot be interpreted too literally because splits are complicated by gene flow, admixture and possible technical biases due to phasing, they nevertheless provide important insights into their population demography.

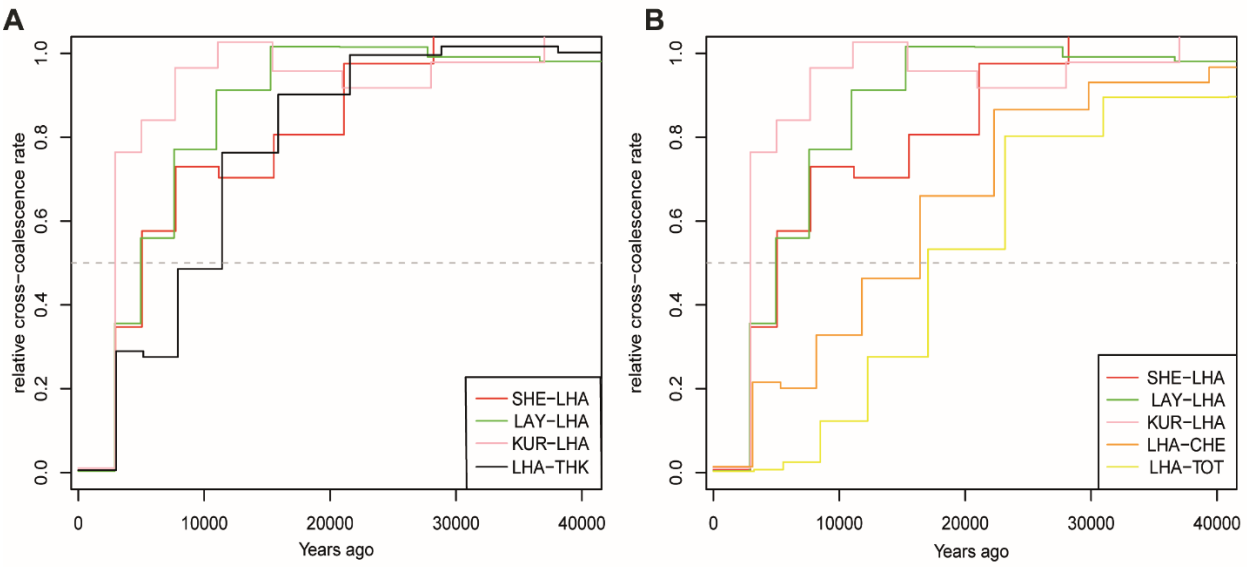


Figure 3. 32 MSMC2 analysis of split times between Himalayan populations. A. High altitude Himalayan populations split times from Tibetans (Lhasa). B. Split times between Chetri, Tharu and Toto from Tibetans.

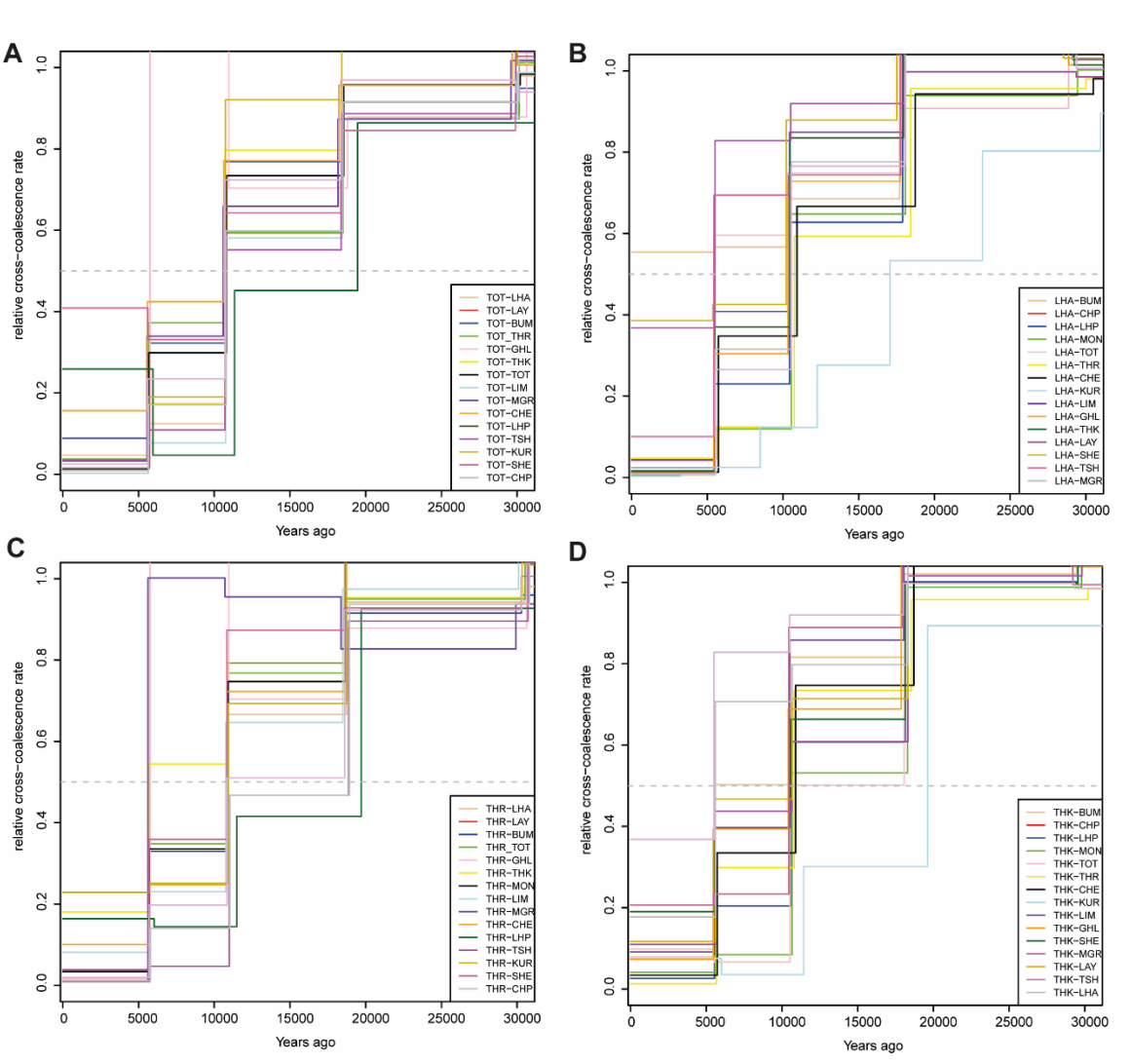


Figure 3. 33 MSMC2 analysis of split times between Himalayan populations. A. Himalayan populations split times from Toto. B. Himalayan populations split times from Tibetans (Lhasa). C. Split from Tharu D. Split from Thakali.

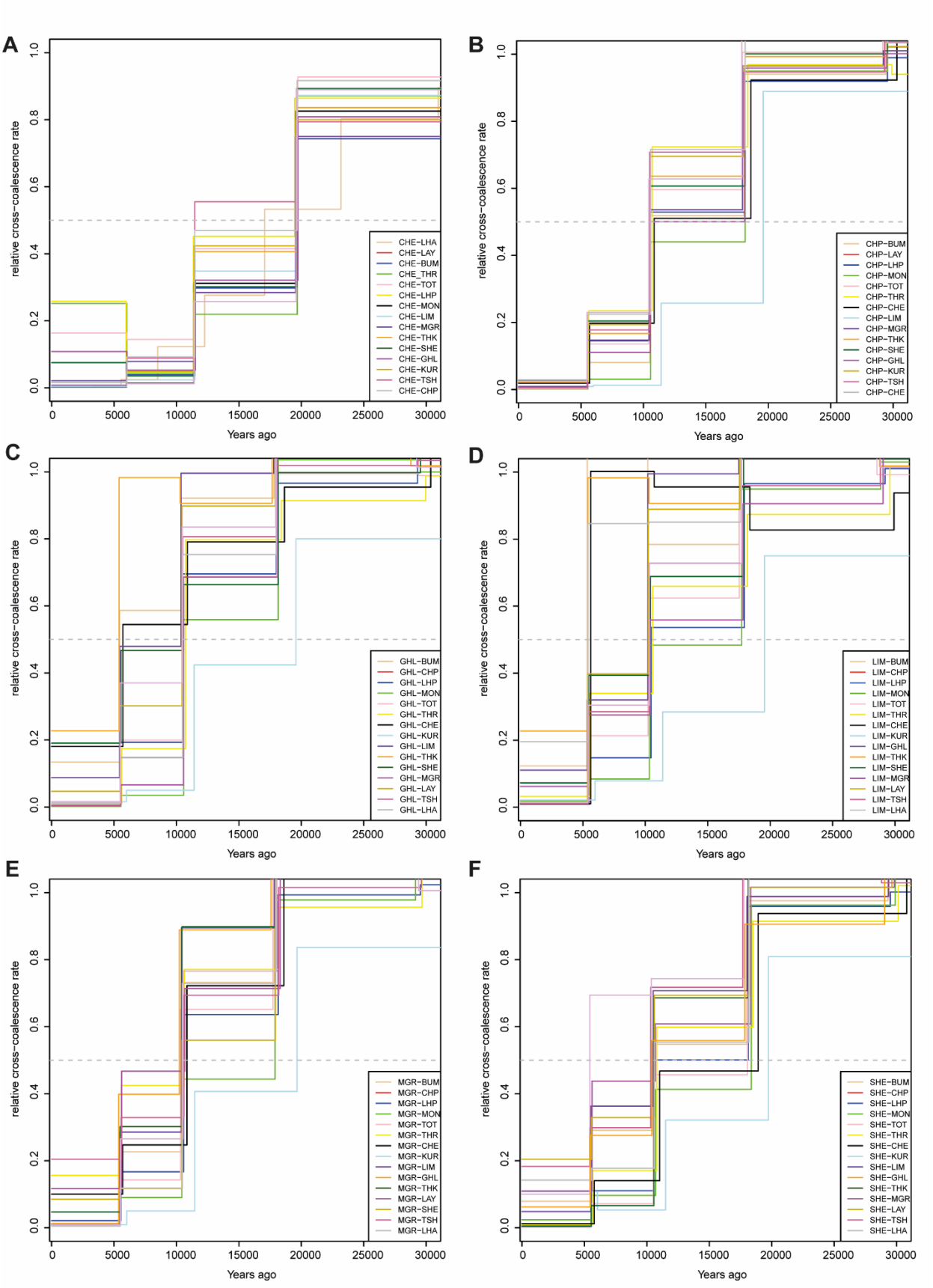


Figure 3. 34 MSMC2 analysis of split times between Himalayan populations. A. Split from Chetri B. Chepang C. Ghale D. Limbu E. Magar F. Sherpa

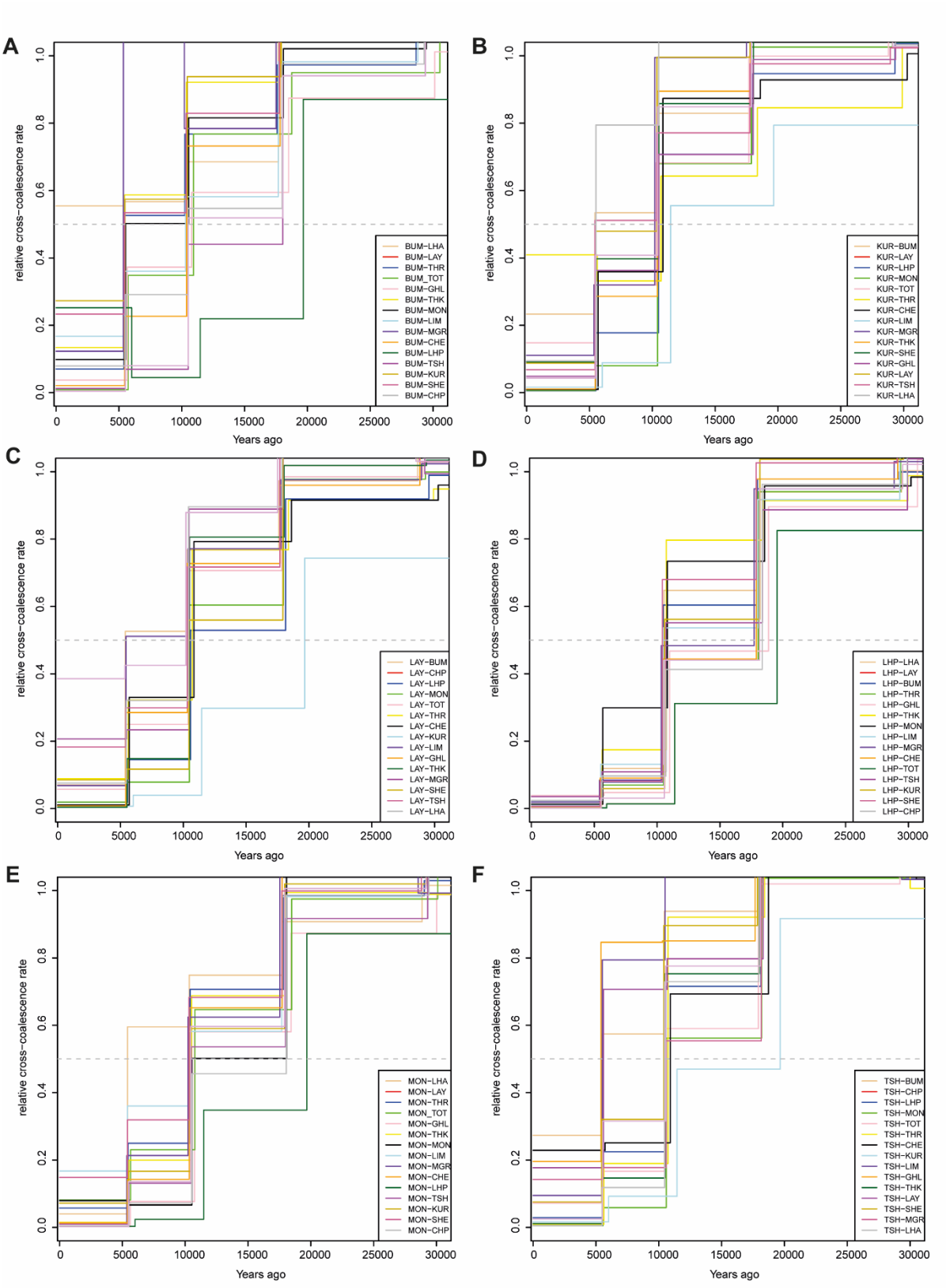


Figure 3. 35 MSMC2 analysis of split times between Himalayan populations. A. Split from Bumthang B. Kurtöp C. Layap D. Lhokpu E. Mönpa F. Tshangla

3.3.3.2 Archaic introgression landscape in the Himalayas

Denisovan and Neanderthal introgressed genetic material in present-day Himalayans was investigated using a novel method that does not require an archaic reference and is thus more powerful in finding intregressed segments, but allows subsequent mapping to these archaic genomes (222). Compared with other worldwide individuals from the SGDP, the genome-wide distribution of Neanderthal segments in the Himalayans is similar, as expected from a model of shared Neanderthal ancestry in all non-African populations due to a single admixture event soon after the exit from Africa (contributing approximately 2%) (272) (Figure 3.36). Each Himalayan individual carries around 200-260 Neanderthal regions (Vindija + Altai Neanderthals) with a mean probability ≥ 0.8. The number and distribution of introgressed segments from the Vindija and Altai Neanderthals were very similar across all the individuals tested. Each Himalayan individual carries around 50-60 additional Denisova segments, a lot less than in Melanesians, as expected (Figure 3.36). The genome-wide distribution of Denisovan introgressed segments in Himalayans is similar to those in other East Asian populations and more than in South Asians. A total of 2,635 Denisova introgressed segments were inferred in the Himalayan individuals, 479 of which are not shared with other SGDP worldwide individuals (minimum overlap within segments set at 0.5 to call a shared segment). Overlapping the introgressed segments with known gene and regulatory annotations from Ensembl revealed 239 to explore for the possible functional impact.

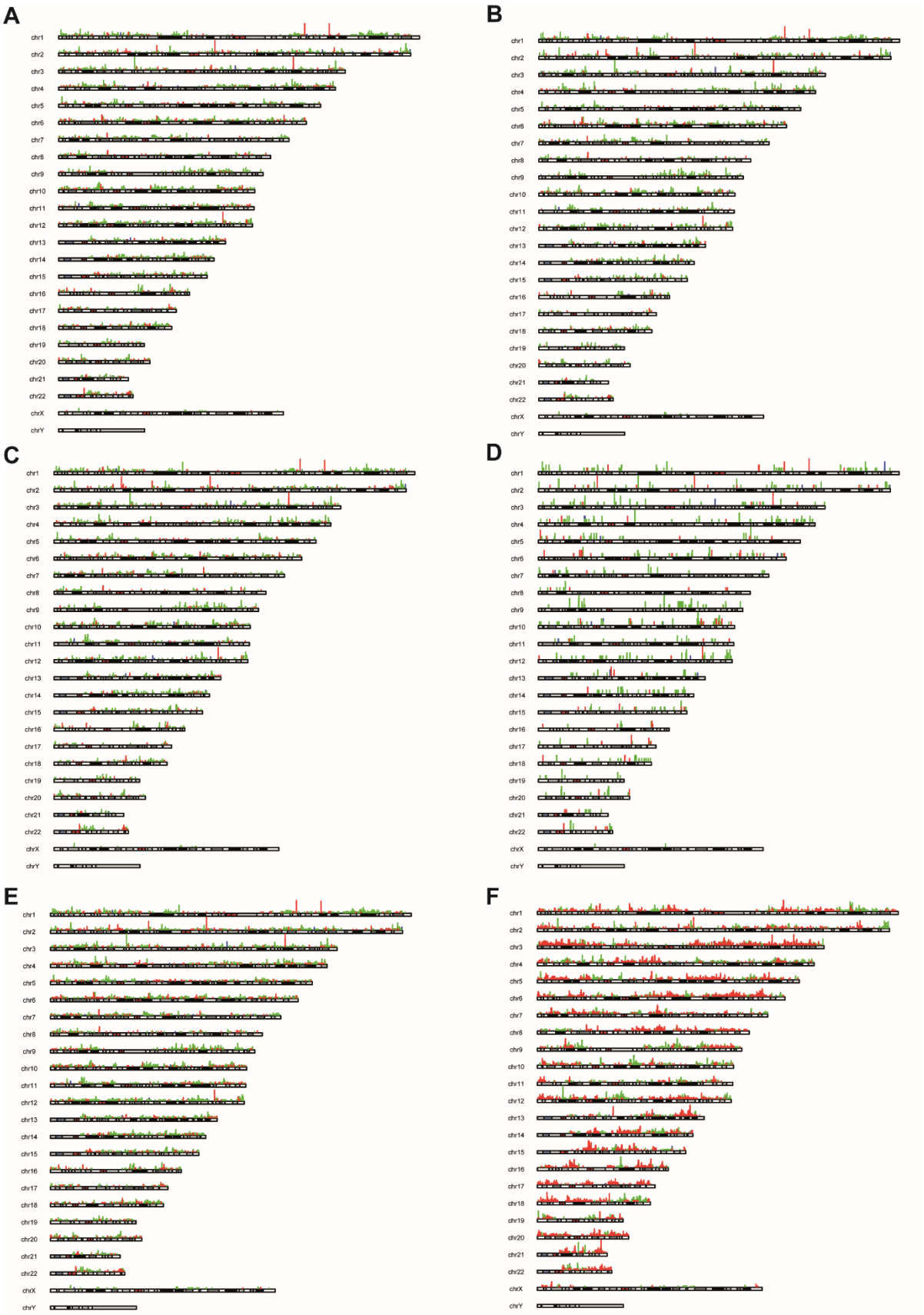


Figure 3. 36 Comparison of archaic introgressed segments in different populations. The green and blue segments are Neanderthal (most of the blue are covered by green ones as shared between Altai and

Vindija), and the red segments Denisovan. The height of the coloured bar is proportional to the number of individuals carrying the segment. A. Nepal, B. Bhutan, C. Tibet, D. India, E. East Asia, F. Melanesia

Because the introgressed regions potentially contain a mix of modern human, Neanderthal and Denisovan variants, I specifically extracted SNPs predicted to be Denisovan-like by the method. Within these 239 regions, 1,060 SNPs are Denisovan specific. I ran VEP on these variants to predict specific function and most likely gene association for each of them. Six variants were predicted to be missense and nine synonymous, 164 fall within regulatory regions and three are splice region variants (APPENDIX B). Possible evolutionary implications of these variants are described later in this chapter.

EPAS1 has been widely studied in Tibetans because they carry an extended introgressed Denisovan haplotype associated with high altitude adaptation (104). Thus, I have explored this region in more detail, comparing Himalayans with other worldwide populations from the SGDP dataset. The introgressed *EPAS1* region is widespread across the Himalayan populations (127) but mostly absent from other worldwide populations. In particular, all Melaneasian individuals examined lack the introgressed *EPAS1* region, supporting the idea of two pulses of Denisovan introgression, one involving both Melanesian and East Asian and a second affecting only East Asians (36). Nevertheless, the absence of the *EPAS1* introgressed region from Melanesians could also be the result of negative selection in this region. Within Himalayans, the region extends over more than ~350 kb including sequences both upstream and downstream of *EPAS1* (Figure 3.38). Although the Denisovan segments differ between individuals, *EPAS1* and the downstream region are present in the majority of Himalayans and include six genes with a complicated LD pattern (Figure 3.37).

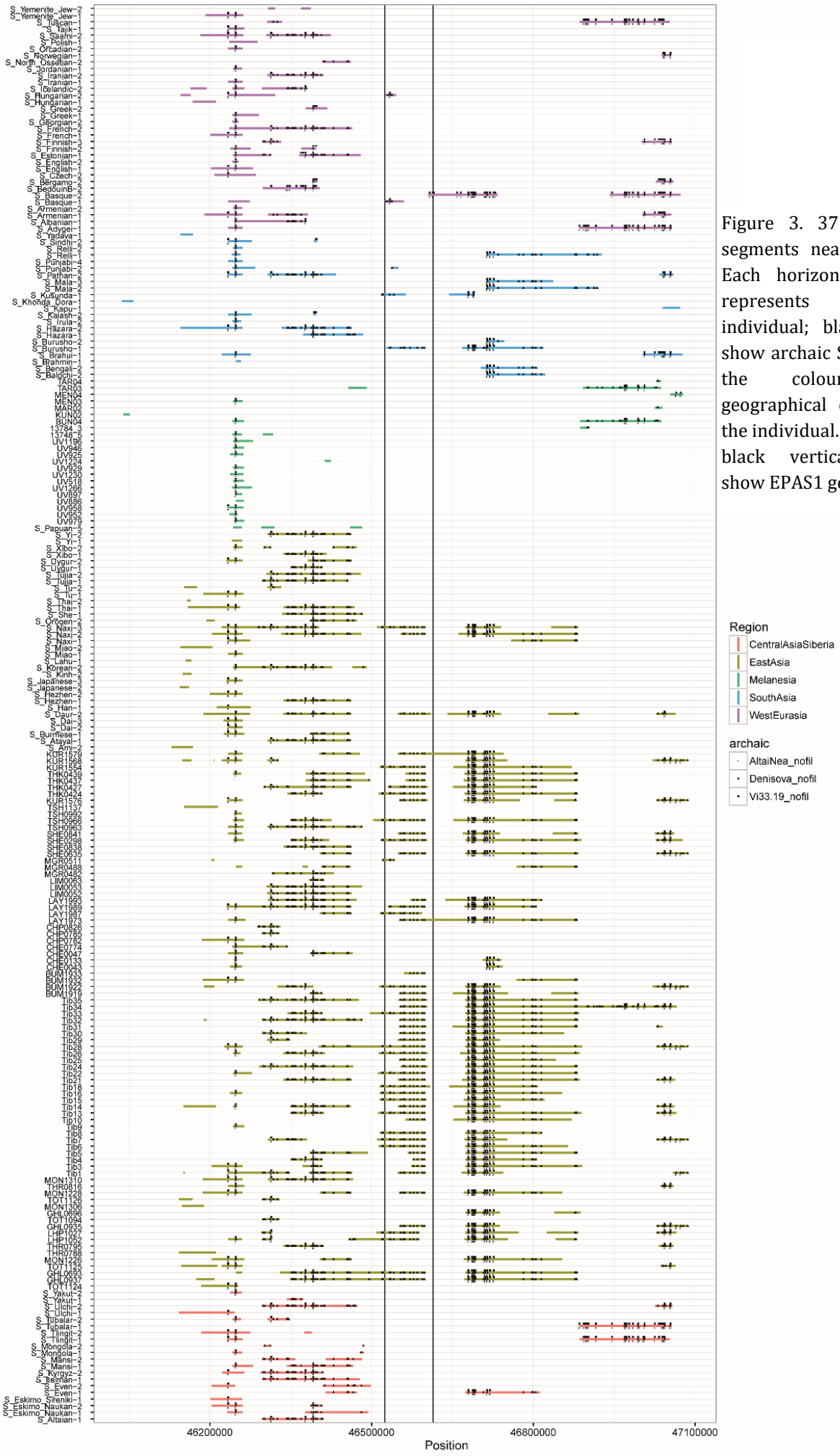


Figure 3. 37 Archaic segments near *EPAS1*. Each horizontal lines represents an individual; black dots show archaic SNPs and the colour the geographical origin of the individual. The two black vertical lines show EPAS1 gene.

3.3.4 Fine-scale positive selection signals

Positive selection in the Himalayan populations was investigated using three major approaches to look for new candidate signals and refine the ones reported in the previous chapter. Genome-wide single-locus F_{ST} (60) between low and high altitude Himalayan populations for SNPs and INDELs replicated two regions reported in the last chapter, *EPAS1* on chromosome 2 and *HLA-DQ* on chromosome 6 with F_{ST} > 0. 3, and also discovered a new region on chromosome 7 overlapping the gene ArfGAP With Dual PH Domains 1 (*ADAP1*) with F_{ST} between 0.25 and 0.32 (Figures 3.38A and B). *ADAP1* is involved in the B Cell Receptor signalling and Arf6 signalling pathways. Single locus F_{ST} of deletions and duplication only replicated the known Tibetan-specific deletion around *EPAS1* with F_{ST} of 0.45 between high and low altitude populations, and found one duplication on chromosome 17 (chr17: 46262318-46270696) with derived allele frequency of 0.23 in the low altitude populations and absent in high altitude populations (Figure 3.38C). This duplication overlaps with a promoter region and it was also found at frequency of 0.48 in South Asians in the 1000 Genomes Project (chr17:46264046-46289131), so the presence for this duplication in Chetri, Toto, Chepang and Tharu could be explained by their South Asian genetic background and admixture.

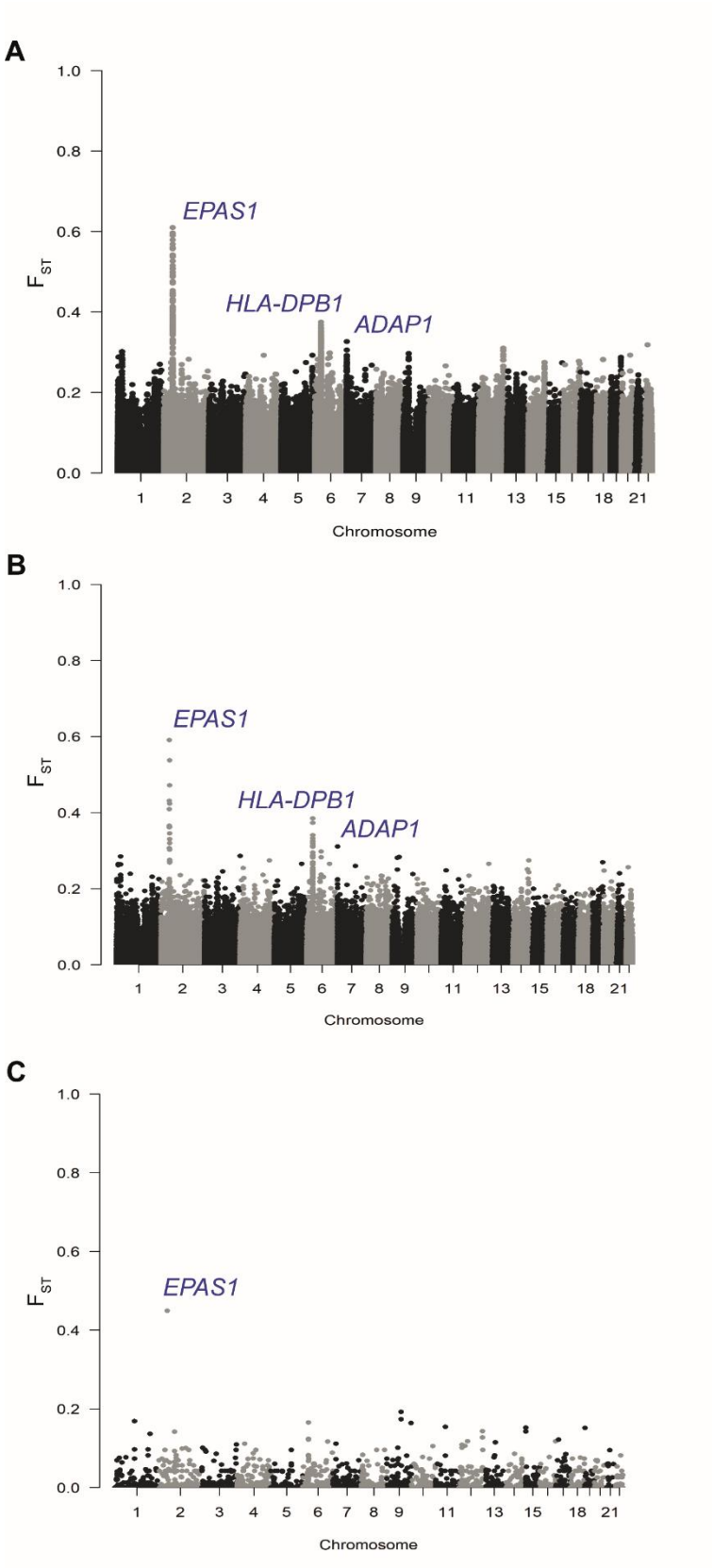


Figure 3. 38 Manhattan plots of genome-wide F_{ST} between low and high altitude Himalayan populations. The x-axis shows the genomic positions and the y-axis the F_{ST} values. Each dot is a variant. A. SNPs B. INDELs C. CNVs (duplications and deletions).

Single locus F_{ST} values for the *EPAS1* region between Tibetans and Chinese Han are very high: In a region of more than 300 kb containing 1,558 SNPs, 95 showed $F_{ST} \geq 0.8$, and of which 455 are derived variants in the Denisova genome (Figure 3.39). Variants with the highest F_{ST} values lie within *EPAS1* ($F_{ST} \geq 0.9$), followed by the region downstream of *TMEM247* and *ATP6V1E2* ($F_{ST} \geq 0.8$). It is difficult to identify the positively selected target variant(s) from these numbers alone, although one of them, rs150877473 (chr2:46360880-46360880, F_{ST} of 0.91), is predicted to fall into a splice site with a derived allele frequency of 0.85 in Tibetans and 0.01 in Han Chinese(132). The derived allele (G) has also been associated with Primary familial polycythemia (273).

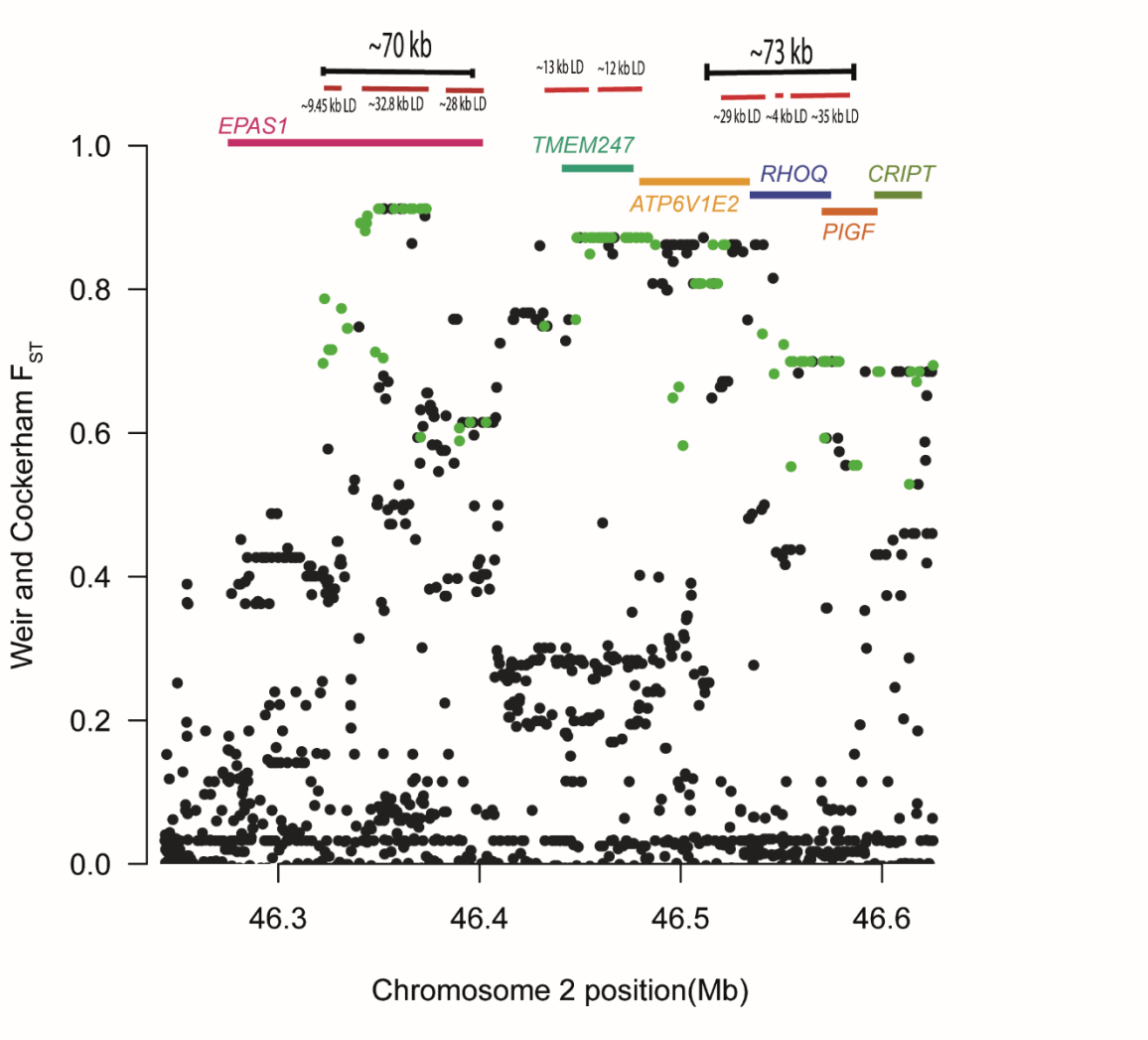


Figure 3. 39 Manhattan plot of single-locus F_{ST} values around *EPAS1* between Tibetans and Han Chinese. The x-axis shows the genomic positions and y-axis the F_{ST} value. Each dot is a variant. The green dots are derived variants in the Denisova genome. The black and red lines represent LD blocks within the region.

In the attempt to narrow down casual variants, the Fine*MAV* algorithm was applied to compare low versus high altitude Himalayan populations, and low and high altitude Himalayan populations, to other worldwide populations (1000 Genomes Project). Fine*MAV* picked up a strong peak underlying selection in *EPAS*1 in high altitude populations compared to low altitude ones, with the strongest hit being an intronic variant rs141366568 (Figure 3.40A). Out of 100 top outliers, almost half (47) fall in or nearby *EPAS1*. This is similar to a pattern of multiple selection candidates observed for other adaptively introgressed genomic regions in Europeans (244). The second strongest peak in high altitude populations was observed on chromosome 6 in high altitude populations and included a missense variant (rs11551421) in *HLA-DPB1* as the top scoring candidate, followed by variation in Collagen Type XI Alpha 2 Pseudogene 1 (*COL11A2P1*) and Orofacial Cleft 1 Candidate 1 (*OFCC1*) (Figure 3.40A, APPENDIX C) (Fine*MAV* score ≥ 3) (Figure 3.41A). On the other hand, no convincing low altitude specific adaptation signals were observed. The strongest hits included a missense variant on chromosome 2 in the Tenascin N (*TNN*) gene (rs4894028) previously associated with skeletal and cardiac myopathy (274, 275) and two synonymous variants on chromosome 11 in the Neuron Navigator 2 (*NAV2*) gene (rs1867115), and on chromosome 5 in the DEAD-Box Helicase 41 (*DDX41*) gene (rs148853192) respectively (APPENDIX C) Figure 3.40B). However, the derived allele frequency of these variants in low altitude populations is around 015-0.20, as seen in other South Asian populations (~0.10-0.25); therefore the observed signals illustrate differentiation between high and low altitude populations, rather than strong positive selection specific to lowlanders.

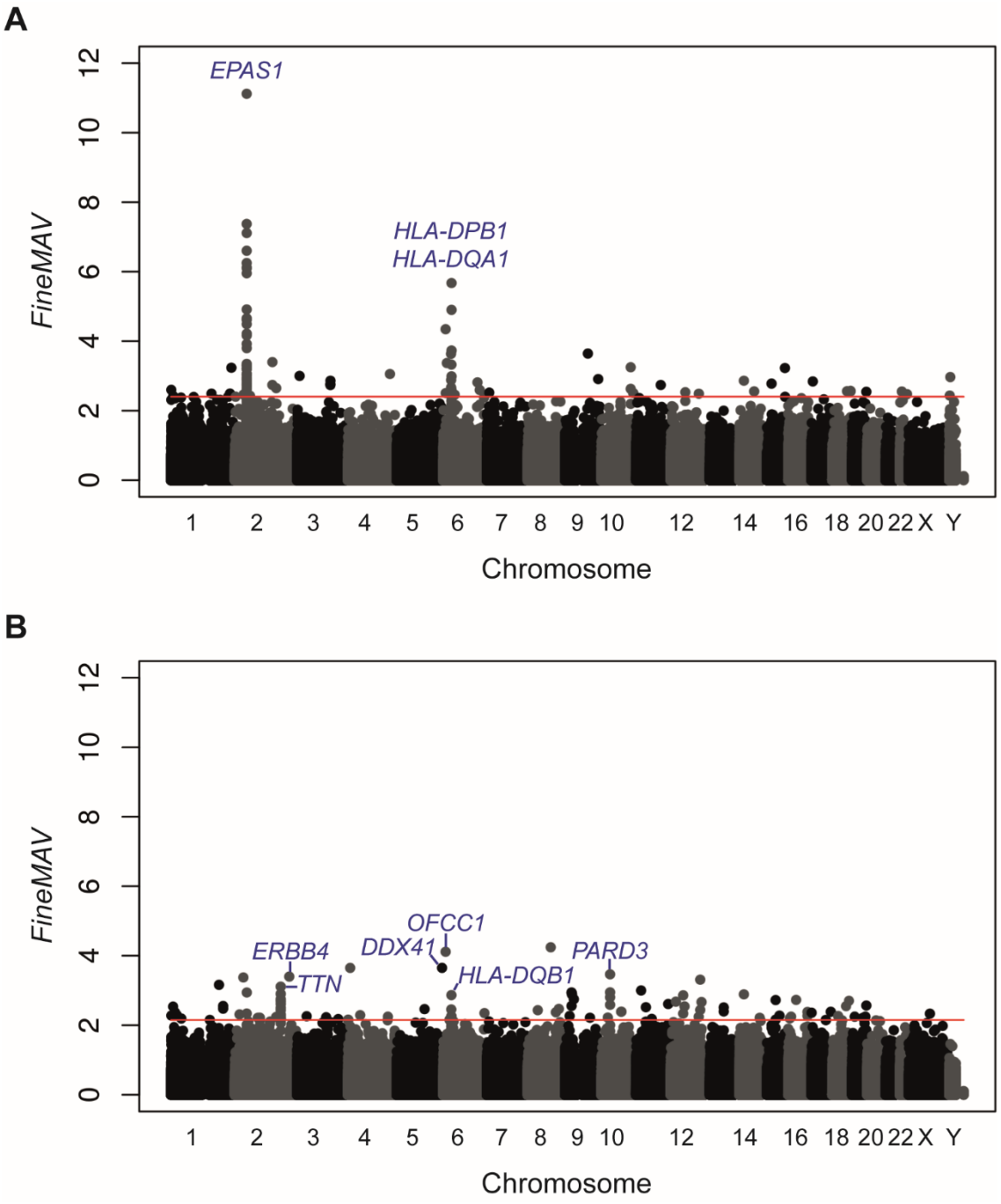


Figure 3. 40 Manhattan plot of genome-wide Fine*MAV* scores in Himalayans. Fine*MAV* scores calculated for genome-wide SNPs in: A. High altitude populations (run against low altitude populations); B. Low altitude populations (run against high altitude populations). Each dot in the Manhattan plots represents a single SNP plotted according to coordinates in GRCh37. The threshold (dashed lines) was set to include the top 100 variants.

Subsequently, high and low altitude Himalayans were compared to worldwide populations from 1000 Genomes Project phase 3 (AFR, EAS, EUR and SAS) to detect high and low latitude specific signals, respectively (Figure 3.41 A and B). Apart from a strong *EPAS1* signal in highlanders, we also detected a peak on chromosome 22 underlying the Von Willebrand Factor Pseudogene 1 (*VWFP1*), the TPTE Pseudogene 1 (*TPTEP1*) and the XK Related 3 (*XKR3*) genes found in both high and low altitude populations. This signal is shared by both high and lowlanders, therefore was not picked up in the previous analyses, and appears to be specific to Himalayan region (found at ~0.7 frequency in Himalaya but rather rare elsewhere). However, the function of those genes remains enigmatic. Other Himalayan-specific signals shared across altitudes include an intronic variant in the Myomesin 2 (*MYOM2*). *MYOM2* is expressed in heart and skeletal muscles and interconnects the major structure of sarcomeres (276). Similarly, a signal underlying the Serine and Arginine Rich Splicing Factor 10 (*SRSF10*) on chromosome 1 was picked up, a gene that plays a crucial role in cell survival under stress conditions by inhibiting the splicing machinery (277). Variants in genes linked to pain perception, thermal sensitivity and olfaction were also picked up (*CACNA2D3, TXNDC15, TAAR5*). Two of the top scoring candidates were described as disease associated: missense rs140598 in Fibrillin 1 (*FBN1*) associated with cardiovascular abnormality and Marfan syndrome (278), and splice donor rs143909348 in Serine Protease 1 (*PRSS1*) associated with pancreatitis (279). One of the top scoring variants (rs608415) in lowlanders is intronic and lies within the Killer Cell Immunoglobulin Like Receptor, Three Ig Domains and Long Cytoplasmic Tail 1 (*KIR3DL1*) gene, involved in the immune response. Fine*MAV* has also picked-up well know East Asian signals of selection which are shared with Himalayans, namely causal SNPs in *ABCC11*, *EDAR*, *FUT2*, *PCDH15*, *PRSS53* and *ZAN* (244). However, the strength of the signal on those SNPs is reduced by allele sharing with the EAS used in this analysis (Figures 3.41 A and B, APPENDIX C).

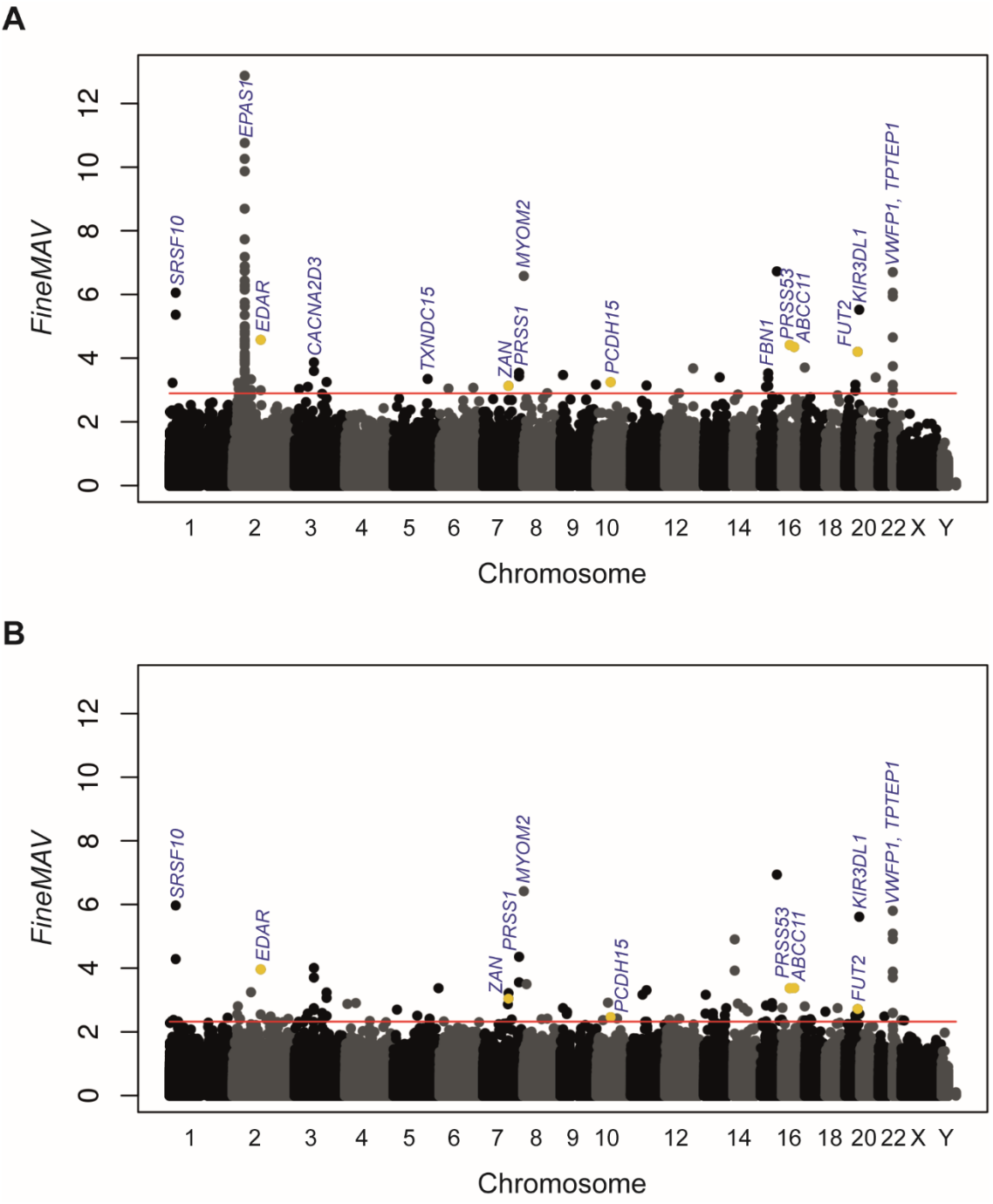


Figure 3. 41 Manhattan plot of genome-wide Fine*MAV* scores in Himalayans. Fine*MAV* scores calculated for genome-wide SNPs in: A. High altitude populations (run against 1000GP populations); B. Low altitude populations (run against 1000GP populations). Each dot in the Manhattan plots represents a single SNP plotted according to coordinates in GRCh37. Yellow dots represent known candidates of positive selection in East and South Asians. The threshold (dashed lines) was set to include the top 100 variants.

Finally, I investigated whether any of the highly differentiated variants between Tibetans and Chinese Han ($F_{ST} \geq 0.8$) fell within an HIF-2α (*EPAS1*) binding motif, hypoxia responsive element (HRE). The HIF-2α core binding motif is ACGTG with a 2 bp flanking region at each side making the extended motif (Figure 3.42). PWMScan predicted 13,198 genome-wide HIF-2α binding sites on both DNA strands with a p-value threshold of 1.000e-05 and a cut-off percentage of 99.59%. A total of 1,254 SNPs found in Tibetan samples fell within these regions, of which 75 overlapped with broad peaks and 45 with narrow peak intervals in aorta tissue, while 81 lay in broad peak regions and 41 in narrow peaks in lung tissue. However, only one SNP (chr10: 1056621: C/G), falling in a broad peak in aorta tissue, is highly differentiated between Tibetans and Chinese Han. It is fixed in Tibetans and absent in Chinese Han (F_{ST} =1) and it overlaps with a promoter region according to Ensembl Regulatory features. The predicted HIF-2α binding motif is 9 bp on chromosome 10:1056617-1056626 (zero-based coordinate system: first base is 0) falling in the WD Repeat Domain 37 (*WDR37*) gene and overlapping with a promoter region.

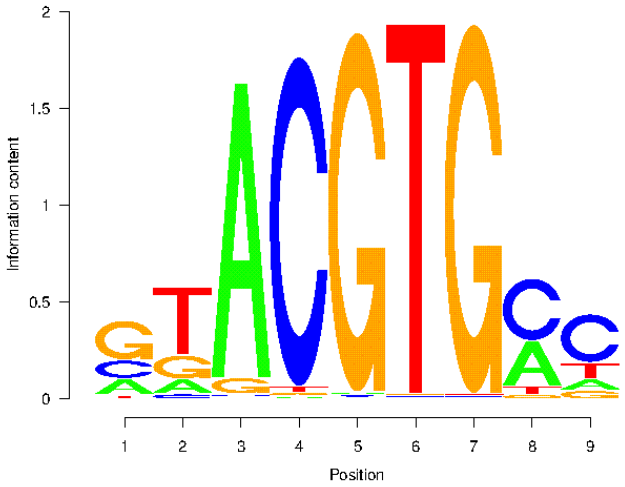


Figure 3. 42 HIF-2α Position Weight Matrix Logo. The height of each letter represents its specific weight or importance within the motif.

WDR37 is member of a family of genes involved in gene regulation, signal transduction and cell cycle progression. The highly differentiated variant is at position 4, within the binding core motif (Figure 3.42) and a change from C->G is predicted to strongly affect the binding and potentially disrupt the binding site. However, this variant

is present also in most of the other Himalayan individuals and is not highly differentiated between low and high altitude populations.

3.4 Discussion

This study further refines the Himalayan population demographic history using both autosomal and uniparental markers of SNPs, INDELs and CNVs. It also attempts to refine the signals of positive selection to pinpoint the potential functional variants.

In the broadest sense, all the Himalayan populations share proportions of their ancestry with other South and East Asian individuals (Figure 3.9A). Nevertheless, within Himalayans, population structure is present with Chetri, Tharu, Lhokpu and Toto forming outliers in the PCA for both SNPs and INDELs, and in the ADMIXTURE analysis. Tibetan, Bhutanese and Sherpa samples cluster together replicating the pattern seen in the previous chapter (Figures 3.9 B, C, D and 3.10). On a finer scale, Himalayan populations are described by different proportions of South and East Asian genetic background and gene flow from them (Figures 3.15, 3.16 and 3.17). Chetri and Tharu show more genetic affinity with South Asians compared to other Himalayan populations and this can explain their outlier status in the PCA (Figure 3.9B, C, D). The genetic affinity to South Asians is also supported by the high rare variant sharing (*f2* and singletons) between Chetri and Tharu with South Asian individuals (Figures 3.21, 3.22 and 3.23). Overall, Himalayan populations share rare variants within themselves and with neighbouring populations suggesting that the connection between these populations extends into very recent times (Figures 3.21, 3.22 and 3.23) (280). However, Himalayan populations showed very diverse demographic histories. Lhokpu and Toto, followed by Chepang and Mönpa, showed extreme genetic drift, replicating the conclusions of the previous chapter, characterised by long branches in TreeMix analysis, lower heterozygosity rates (Figure 3.12) and an increased number of ROHs and shared IBD segments (Figure 3.20), as well as very low effective population sizes (Figures 3.25 and 3.26). In contrast, Chetri and Tharu show higher heterozygosity than other populations and carry a significantly higher proportion of singletons (Figures 3.4, 3.18 and 3.19), as well as a recent increase of effective population size detectable by MSMC2 and SMC++ (Figures 3.25, 3.26), which can be explained by very recent gene flow with South Asians. Tibetans are the closest population to East Asians (Figure 3.21), in particular Han Chinese (CHB), as expected (Figure 3.31) (50, 99).

Himalayan populations show recent split from one another in the last 5,000-9,000 years. Some of the populations are not completely separated, and show signatures of gene flow (Figures 3.32, 3.33. 3.34 and 3.35). In particular, high altitude population show very recent split times (between 3,000 and 5,000 years ago) and several are not completely separated (Figure 3.32, 3.33, 3.34, 3.35). This supports the hypothesis that high altitude populations derive from a single ancestral population that separated from lowland East Asians in the last 10,000 years, and then spread and diverged into different populations across the Himalayan region in recent times. On the other hand, Chetri and Tharu split first from other Himalayan populations in agreement with their lower genetic affinity to other Himalayans and higher affinity to South Asians (Figure 3.32B, 3.33, 3.34, 3.35).

The presence of South and East Asian admixture and the recent split times within Himalayans is also supported by the Y chromosome and mtDNA analyses (Figures 3.28 and 3.30), and many lineage split times within the last 5,000 years (Table 3.6). The widespread distribution of O2, R1a and D Y chromosome haplogroups, the strong clustering of these on the tree and the recent split times within them suggest rapid spreading within the region driven by males (Figures 3.27, 3.28 and Table 3.6).

Himalayan populations show a distribution of archaic introgressed tracts similar to other South-East Asian populations. However, they carry 239 segments not present in other populations (SGDP), some of which overlap with genes and contain Denisovan-specific SNPs (Figure 3.36). Within the regions overlapping genes, a missense variant (rs35211496) has been found in *TNFRSF11A* on chromosome 18 and a missense and synonymous variants (rs76335535, rs139705553) in *CYP2F1* on chromosome 19. *TNFRSF11A* codes for a receptor that is essential for osteoclast and lymph node development, and been associated with bone pathologies such as arthritis (281). *CYP2F1* codes for a member of the cytochrome P450 superfamily of enzymes and catalyses reactions involved in drug metabolism. These genes are involved in immune response (282) and xenobiotic biotransformation (283) respectively, and introgressed segments in these genes could have played an important role in the adaptation to new environments and against pathogens (284-286). The *EPAS1* introgressed region is widespread across Himalayan individuals and spans over 300 kb downstream and partially upstream of *EPAS1* (Figures 3.37 and 3.39). *EPAS1* itself and the downstream

region are enriched of Denisova-specific variants (Figures 3.37 and 3.39). Melanesian individuals do not carry the *EPAS1* introgressed region, supporting the idea of two pulses of Denisovan admixture in Melanesian and East Asians (36).

The high altitude signals of selection in *EPAS1* and *HLA-DP/DQ* from the previous chapter were replicated here, and an additional novel region identified (Figure 3.38, Figure 3.40A) on chromosome 7 near *ADAP1* using both SNPs and INDELs. *ADAP1* is regulated by *ERB44*, which is a direct regulator of HIF-1α and is necessary for the transcription of known HIF-1α target genes. *ADAP1* is involved in intracellular signalling and microtubule regulation (287, 288). No low altitude specific signal was detected. The extensive shared ancestry and gene flow with surrounding populations have decreased differentiation and, thus, population-specific derived allele frequency (Figure 3.40B). Interestingly, positive selection signals shared between low and high altitude populations (Himalayan-specific) were identified. *The KIR3DL1* gene is involved in viral immunity and it has been linked with immune response against malaria in individuals living in the Solomon Islands (Melanesia) (289) (Figure 3.41 A and B). This gene represents one of the highest Fine*MAV* scores in lowlanders and it could be associated with resistance against malaria in populations living at the foothills of Himalayas, in the Terai tropical forest. Although many highly differentiated variants have been discovered in the *EPAS1* locus, sequence variation data alone were not sufficient to pinpoint the functional variants due in part to the complicated LD patterns in the region. Thus, *in vitro* functional exploration of the *EPAS1* region is needed to understand the molecular mechanism of high altitude adaptation and narrow down the important functional variants within the *EPAS1* locus.

4. Functional investigation of *EPAS1* in high altitude adaptation

4.1 Introduction

The first genome-wide surveys of high altitude adaptation in Tibetans were carried out in 2010. In one study, using a cohort of Tibetans, the authors found a cluster of 31 *EPAS1* SNPs in high linkage disequilibrium that correlated significantly with lower haemoglobin concentration in these individuals (145). In a second study, using 50 exome sequences from individuals of Tibetan ancestry, the authors reported that variants in *EPAS1* exhibited the strongest signal for high altitude adaptation followed by those in other genes associated with the response to hypoxia (99). Several SNPs in and around *EPAS1* on chromosome 2 had a striking allele frequency difference between highland Tibetans and lowland dwelling Han Chinese samples (99). Subsequently, a more detailed analysis of the *EPAS1* region by sequencing the whole region of 40 Tibetans and 40 Han Chinese samples found that the region had an unusual haplotype structure that could be accounted for introgression of DNA from an archaic human species, the Denisovans, into modern humans (104). The authors reported that the core Denisovan haplotype in *EPAS1,* which was characterized by five derived alleles of closely linked SNPs (AGGAA) lay within a ~32.7 kb introgressed region. Several SNPs in this region were highly differentiated and included those with the highest F_{ST} value between Tibetans and Han Chinese, so this region was therefore considered the best candidate for contributing to the high altitude adaptation phenotype in Tibetans (104).

High coverage whole genome sequencing of 87 Himalayan samples has, as discussed in the previous chapter, now shown that the total length of the introgressed region spans a ~300 kb genomic region that includes six genes and has an extended LD pattern that makes it difficult to narrow down putative functional variants (119). The highest F_{ST} values between Tibetans and Han Chinese are within the previously reported ~32.7 kb region; within this region, the five SNPs of the Denisovan core and a candidate regulatory variant (rs370299814) upstream it are within the highest differentiated between Tibetans and Han Chinese and, thus, they are good functional candidates for being positively selected (127). The availability of a testable phenotype described in the first chapter, such as the low red blood cell count and low haemoglobin levels in Tibetans,

and these genetic candidates make it possible to investigate the underlying biological function of high altitude adaptation in Himalayans at a molecular level. A first attempt at understanding the role of *EPAS1* in high altitude adaptation indicated that the gene was somehow downregulated in Tibetans which was in agreement with the low red blood cell count in this population (209). However, the experimental conditions used in this study could confound results related to hypoxic adaptation, as they compared expression profiles of lowlanders sampled at only low altitude with Tibetans samples only at high altitude without having reciprocal controls. Moreover, its population samples were not properly genetically classified prior to inclusion, with no genotype data available. So, I have attempted to address these issues by conducting *in-silico* analyses and *in vitro* experiments to study the expression levels of *EPAS1* at different levels of oxygen in the human lymphoblastoid cell lines (LCLs) with and without the introgressed haplotype from the 1000 Genomes Project.

The aims in the chapter, were first to develop a hypoxic protocol that could be used to test if the commercially available LCLs could be used as a cellular model for studying *EPAS1* gene expression. If so, the differentiation of gene expression in the cell lines with and without the introgressed haplotype was assessed using both quantitative PCR (qPCR) and RNA-sequencing (RNA-Seq).

4.2 Material and methods

4.2.1 Samples

Twelve human LCLs and one human lung epithelium cell line (A549) were used in this study (290). LCLs grow in suspension and tend to form clusters of cells in culture, while the A549 cell line is derived from adenocarcinomic human alveolar basal epithelial cells and the cells form a monolayer adherent to the culture flask (291). The LCLs were from the 1000 Genomes Project populations (177) and obtained from the NHGRI Repository at the Coriell Institute for Medical Research. These included ten cell lines belonging to the Chinese Dai population from Xishuangbanna (CDX), China and one control sample of European ancestry (CEU) and an additional control sample of Mexican ancestry (MXL) which was only used for the RNA-seq experiment (Table 4.1). Four of the LCLs are heterozygous for the derived alleles of the five core SNPs in the intogressed haplotype (AGGAA) within the *EPAS1* region. Within these, three out of four are heterozygous for rs370299814. The A549 cell line is commercially available and vials were provided by another laboratory (Team 147, Paul Kellam's lab) within the Wellcome Sanger Institute.

		EPAS1						*EGLN1*		*ATP6V1E2*
		Core SNPs in the introgressed haplotype ("AGGAA")								
		rs115321619	**rs73926263**	**rs73926264**	**rs73926265**	**rs55981512**	**rs370299814**	**rs186996510**	**rs12097901**	**rs12986653**
Pop	**Cell line**	**G/A**	**A/G**	**A/G**	**G/A**	**G/A**	**G/A**	**G/C**	**C/G**	**G/A**
CDX	HG00844	G/G	A/A	A/A	G/G	G/G	G/G	G/G	C/G	G/G
CDX	HG01031	G/G	A/A	A/A	G/G	G/G	G/G	G/C	G/G	G/G
CDX	HG00881	G/G	A/A	A/A	G/G	G/G	G/G	G/G	C/C	G/G
CDX	HG00978	G/A	A/G	A/G	G/A	G/A	G/A	G/G	G/G	G/A
CDX	HG02390	G/G	A/A	A/A	G/G	G/G	G/G	G/G	C/G	G/A
CDX	HG02396	G/A	A/G	A/G	G/A	G/A	G/G	G/G	C/G	G/G
CDX	HG02397	G/A	A/G	A/G	G/A	G/A	G/A	G/G	C/G	G/A
CDX	HG02190	G/G	A/A	A/A	G/G	G/G	G/G	G/G	G/G	G/G
CDX	HG02187	G/A	A/G	A/G	G/A	G/A	G/A	G/G	C/G	G/A
CDX	HG02153	G/G	A/A	A/A	G/G	G/G	G/G	G/G	C/G	G/G
CEU	GM12878	G/G	A/A	A/A	G/G	G/G	G/G	G/G	C/C	G/G
MXL	GM19661	G/G	A/A	A/A	G/G	G/G	G/G	G/G	C/C	G/G

Table 4. 1 *EPAS1* genotypes of the LCLs used in this study. The table contains genotypes for the 12 LCLs analysed as anc/der alleles. Three of them carry the derived alleles for both the 5 core SNPs and the putative functional one (rs370299814) while the fourth one only carries the derived alleles of the 5 core SNPs. All other eight LCLs are ancestral homozygous for these 6 SNPs.

4.2.2 Methods and protocols used for the analyses

4.2.2.1 In silico analysis of 1000 Genomes Project cell line sequences

The ~32.7 kb *EPAS1* region was compared between 1000 Genomes Project individuals and four high coverage Tibetan whole genome sequences (Tib8, Tib14, Tib29, Tib34 which have mapped to the same reference sequence), discussed in the previous chapter (4). To retrieve haplotype information, the genomic region was phased with SHAPEIT using the 1000 Genomes Project data as a reference panel (177, 292). Comparison between sequences were performed using a combination of custom scripts and vcftools 0.1.14 (230). The haplotype network, which represents the relationships between the different haplotypes in the dataset, was generated using the haploNet function implemented in the pegas R package (293, 294).

The genotypes for interesting variants previously associated with high altitude adaptation were also retrieved from the 1000 Genome Project samples. Those include two variants in *EGLN1* (126) and one in *ATP6V1E2* from the Himalayan SNP-chip data (Table 4.1). Likewise, the genotypes for the 4 bp deletion located in the second intron of *EPAS1* were inspected (50). This small INDEL was reported to have a derived allele frequency of 0.629 in Tibetans and 0.08 in Han Chinese. It overlaps the binding sites of three transcription factors: the polymerase subunit *POLR2A*, the HIF co-activator *EP300*, and *GATA2* which controls erythroid differentiation. It also lies within an activating H3K27Ac mark in seven cell lines (50).

4.2.2.2 *In silico* analysis of publicly-available RNA-seq data

The expression profile of *EPAS1* in different tissues was retrieved from the GTEx Portal (198) and ENCODE human RNA-seq data (257, 295). For ENCODE data, gene quantification for all RNA-seq (polyA RNA-seq) experiments mapped to build GRCh38 were downloaded. The *EPAS1* expression across all tissues and cell lines available was compared with each other as well as with the expression of the *HIF-1α*, *HIF-3α* and *EGLN1*.

In addition, Tibetan placenta RNA-seq data from a previous study were analysed (209). In this study, 63 term placental samples from highland Tibetans were compared with 14 specimens from Han Chinese and tested for *EPAS1* differential expression. According to their results, both Tibetans and Han Chinese show relatively similar low levels of *EPAS1* expression. After downloading paired-end reads (fastq files), two independent analyses were performed:

1) Variant calling from RNA-seq data. The GATK pipeline for variant calling using RNA-seq was followed (225, 227). Aligned reads generated with STAR 2.5.2b (296) were processed with Picard 2.6.0 for adding read group information, sorting, marking duplicates and indexing. The GATK tool specifically developed for RNA-seq data, SplitNCigarReads, was used to split reads into exon fragments and hard-clip any sequences overhanging into intronic regions. Finally, variant calling was performed using GATK HaplotypeCaller with the –dontUseSoftClippedBases argument, which is a new functionality taking into account the information about intron-exon split regions from the previous step (225, 227). The called positions were merged with Tibetan and Han Chinese genomic sequences from the previous chapter. Principal component analysis (PCA) was performed calculating the eigenvectors from the whole-genome sequencing samples and projecting the RNA-seq individuals onto the plot to check the ancestry of these samples.
2) Expression quantification and differential expression. Genomic indexing and mapping were generated with STAR 2.5.2b (296) using the GRCh38 human reference genome and Ensembl gene annotation files version 88 (54). Gene and transcript assembly and quantification was performed with the Cufflinks pipeline (297). Cufflinks reports expression levels in Fragments Per Kilobase of transcript per Million mapped reads (FPKM), a normalised estimation of gene expression. FPKM are estimated from the number of reads mapped to each particular gene sequence taking into account the sequencing depth and gene length. A FPKM method was applied to the data to see if it was possible to retrieve similar results to the original paper using HTseq for transcript quantification which counts the number of reads mapping to each specific genomic feature from aligned sequencing reads and a file of genomic features (298).

4.2.2.3 *In silico* analysis of *EPAS1* evolutionary conserved regions

Evolutionarily conserved regions (ECRs) in *EPAS1* were compared across different species using the ECR Browser from NCBI Dcode.org Comparative Genomics Development with the default parameters (ECR length=100 bp and ECR similarity=70%)(299). SNPs within these regions were overlapped with the F_{ST} results between Tibetans and Han Chinese discussed in the previous chapter. A list SNPs showing highly differentiation (F_{ST} values ≥ 0.5) was retrieved.

4.2.2.4 Cell culture

Cells were routinely cultured in RPMI 1640 medium with L-glutamine (Thermo Fisher Scientific, Cat No. 11875093), penicillin and streptomycin (Thermo Fisher Scientific, Cat No. 10378-016) and 15% foetal bovine serum (FBS) (Thermo Fisher Scientific, Cat No. 10500-056) at 37^0C in a 5% carbon dioxide 95% air mixture. LCLs were passaged or sub-cultured at confluency (1-1.5x10^6 cells/ml). Similarly, A549 were passaged when confluence was reached. Cells were washed with the Dulbecco's phosphate-buffered saline solution (DPBS, Thermo Fisher Scientific, Cat No. 14190144), detached from the flask surface using phenol red Trypsin-EDTA (0.25%) (Thermo Fisher Scientific, Cat No. 25200-056), split in a 1:5 ratio and sub-cultured. All cell lines were tested for mycoplasma contamination and resulted negative.

Eleven LCLs (ten CDX and one CEU) were karyotyped by the Cytogenetic Core Facility at the Wellcome Sanger Institute to check potential chromosomal abnormalities (300). Ten randomly selected metaphases from each cell line were karyotyped by M-FISH and DAPI-banding. The nomenclature of chromosomal aberrations used is in agreement with the International System for Human Cytogenetic Nomenclature (301).

4.2.2.5 Hypoxic protocol

A hypoxic protocol designed to be performed in a standard cell culture incubator (Panasonic MCO-19MUV-PE) was developed (APPENDIX D). Hypoxic conditions (1% O_2) were reached by controlling the nitrogen levels in the incubator (APPENDIX D). Three key points for the success of the protocols are: 1) The use of conditioned medium for culturing LCLs; 2) Overnight medium O_2 degasification; and 3) Use of six-well plates with a small culture volume. Medium degasification and the culturing volume are very important for O_2 diffusion in the medium. O_2 concentration in the gas phase it is different from the one experienced by cells at the bottom of a culture flask or plate. To have efficient gas diffusion in the medium in both normoxic and hypoxic conditions, I used six-well plates with a total volume of ~4 ml to reduce diffusion distance and have better oxygen diffusion in the liquid phase via overnight medium incubation to degasify hypoxic medium and have a stable O_2 supply in normoxia.

LCLs require a quite high cell density to grow properly, and have different lengths of the lag phase. The latter is the period of the cell growth cycle where cells adapt to the culture conditions and its length depends upon the growth phase of the cell line at the time of subculture and the seeding density. Conditioned medium is exhausted medium harvested from cultured cells mixed with fresh medium. It contains growth factors, metabolites and extracellular matrix proteins secreted into the medium by the cultured cells and it considerably helps cell growth. So I used conditioned medium and seeded plates to have around 1.5-2x10-^{6} cells at each time point of the experiment.

A549 cells were used as positive control as they are widely used to test hypoxic conditions (302, 303). It is known that in A549 the HIFs (*HIF-1α* and *EPAS1*) are induced by acute hypoxia at 4 h and hypoxic factors are constitutively expressed in lung tissues (198, 304-306). Different time points (4h, 16h, 24h, 48h, 72h) were used to check for *EPAS1* expression in LCLs and whether *EPAS1* expression was induced similarly to A549, or whether it needed chronic exposure to hypoxic conditions (APPENDIX D).

4.2.2.6 Gene expression profiling methods

Expression profiling is the measure of the gene activity at a specific time and can assay the expression of a few or thousands of genes at once. The quantitative polymerase chain reaction (qPCR), also known as real-time PCR, monitors the amplification of a targeted DNA molecule during the PCR. Quantitative reverse transcription PCR (RT-qPCR) is used to quantify gene expression and the starting material is RNA. RNA is first retro-transcribed into its complementary DNA (cDNA) by reverse transcriptase. Then, the cDNA is used as the template for the qPCR reaction. It can be one-step or two-step, with the retro-transcription and quantification being performed in one or two separate reactions. Gene expression is detected by using fluorescent molecules such as SYBR green, a fluorescent double-stranded DNA (dsDNA)-binding dye. As the PCR reaction progresses, at each cycle of amplification SYBR Green dye binds to dsDNA as it polymerizes, resulting in an increase in the level of fluorescence at the end of each extension step. The quantity of dsDNA product in the reaction is proportional to the amount of fluorescence (307, 308). This approach gives a very accurate estimation of gene expression for specific target genes, but it is not feasible for a genome-wide screen of expression levels. High throughput measurements, such as RNA-seq, can estimate the presence and quantity of thousands of genes in a biological sample at a given time, providing a detailed view of gene expression, alternative splicing, and allele-specific expression. After RNA extraction and QC steps, barcoded sample libraries are prepared. These pools of samples are then sequenced using one of the high throughput platforms (309).

4.2.2.7.1 Quantitative reverse transcription PCR (RT-qPCR)

Before conducting RNA-Seq, I carried out preliminary work to establish whether *EPAS1* could be detected in LCLs and whether its expression was modified in LCLs with the introgressed *EPAS1* haplotype. A two-step RT-qPCR was performed on the samples challenged with hypoxia to assay *EPAS1* and other genes (*EGLN1*, *HIF-1α*) involved in the hypoxic pathway. Expression of target genes was compared between normoxic and hypoxic samples and between samples with and without the introgressed haplotype. To calibrate target gene expression, housekeeping gene (*TUBB*) expression was assayed (310, 311). RNA extraction was carried out using the QIAGEN RNeasy Mini Kit (QIAGEN,

Cat No. 74104) with the RNase-Free DNase Set (QIAGEN, Cat No. 79254). RNA was quantified using Nanodrop N80. Subsequently, 500 ng of RNA per sample was reverse transcribed into cDNA and used for RT-qPCR (APPENDIX D).

Primer sequences for qPCR were designed with NCBI Primer-BLAST from the NCBI Reference Sequence (RefSeq) for each gene (312) (Table 4.2). I set the minimum primer melting temperature (Tm) at 60°C, the optimum at 62°C, the max at 64°C and the max Tm difference at 1°C. The amplicon size (PCR product) was set between 70-150 bp (optimal length around 100 bp). To avoid amplification of contaminating genomic DNA, the primers were designed to span an exon/exon junction. The GC content was around 40-60% to ensure maximum product stability and each pair of primers had low self-complementarity to decrease the possibility of primer-dimer formation (313-315).

Gene	Primer pair	Primer sequences	
		Forward	Reverse
EPAS1	01	TTCCTGCGAACACACAAGCT	TCGGCTTCGGACTCGTTTT
	02	CATGCGCTAGACTCCGAGAAC	TACCTGACCCTTGGTGCACAA
EGLN1	01	CAAAGTTAATTTCTATGCCTGGAAGA	CAGGCCCTGCCAGACTTCTA
HIF-1α	01	ATGTGACCATGAGGAAATGAGAGA	TTTTGTTCTTTACCCTTTTTCACAAG
TUBB	01	TGTATTTGGTCAGTCTGGGGCAG	GCAGGCAGTCACAGCTCTCT
ATP6V1E2	02	AGCCATCCTCTCGGCCTCTA	AGAGTTCAGGCCTCCCTTTTGG
TMEM247	01	ACCTTCCCCAAGATGGTGCC	GGACTCTGCCTCCAGATAAGCC
CRIPT	01	GAACCGTCGTGGGGAAGGAT	GCTTTCTTCCACCACTTTCTGTG
PIGF	01	TAGCTTTGTAGGAGCATGGCTTG	GAGATGGGCCATACCTGCCA

Table 4. 2 Primer sequences used for RT-qPCR analyses.

Gene expression was quantified by qPCR with KAPA SYBR FAST (Kapa Biosystems, Cat No. KK4604) using an ABI StepOne Plus thermocycler. SYBR Green, together with the set of target gene primers (Table 4.2), binds to the DNA and emits a green light detectable by the qPCR machine (APPENDIX D). Target genes were standardised to the housekeeping gene *TUBB* and the gene expression level was calculated as the standardised gene expression value:

$$2^{-\Delta\Delta CT} = [(C_T \text{ gene of interest} - C_T \text{ internal control})_{\text{hypoxia}} - (C_T \text{ gene of interest} - C_T \text{ internal control})_{\text{normoxia}}] \quad (307)$$

where C_T is the cycle threshold. C_T is inversely related to the amount of amplicon in the reaction. Thus, the lower the C_T, the greater is the gene expression (greater amount of amplicon). Differential expression was also assessed as $\Delta C_T = C_{T\,(\text{normoxia})} - C_{T\,(\text{hypoxia})}$ for each gene in the PCR reaction (307).

4.2.2.7.2 RNA-seq

RNA-seq analysis was carried out on samples from the experiment performed with the hypoxic protocol previously described. A total of 12 lymphoblastoid cell lines and a A549 cell line were cultured under normoxic (state conditions) and hypoxic conditions over 72 hours and samples collected at five different time points (4h, 16h, 24h, 48h and 72h). Two technical replicates for all samples were used. RNA-seq was performed by the Wellcome Sanger Institute core facility. Briefly, lysed cells in 350µl of QIAGEN buffer RLT (QIAGEN, Cat No. 79216) were provided to the facility where automated RNA extraction was performed for all 260 samples. Of these, 257 RNA samples passed automated QC step (Qubit 4 Fluorometer). Sequencing libraries were constructed for these 257 samples via the New England BioLabs (NEB) RNAseq-AUTO library preparation procedure using 100 ng of total RNA in 50 µl water with a RIN (RNA Integrity number) of >6. The 257 samples were sequenced using paired end (PE) 75 bp reads over 4 lanes using an Illumina HiSeq 4000 platform. The raw reads (fastq files) were mapped to human reference genome GRCh38.p10 using STAR_2.5.3a (296), incorporating GENECODE v27 human gene annotations (gtf file) (316). Sample QC was performed with FastQC 0.11.5 and multi-sample aggregation with MultiQC v1.3 (317) and QTLtools 1.1 bamstat (318). Transcript quantification was performed using featureCounts 1.6.2 (319). The Deseq2 1.6.3 R package (320) was used for data exploration following its tutorial (https://www.bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.html), to test sample-to-sample distances using PCA and whether it was possible to detect differential expression in the dataset between hypoxia and normoxia and between introgressed and non introgressed samples (Figure 4.1).

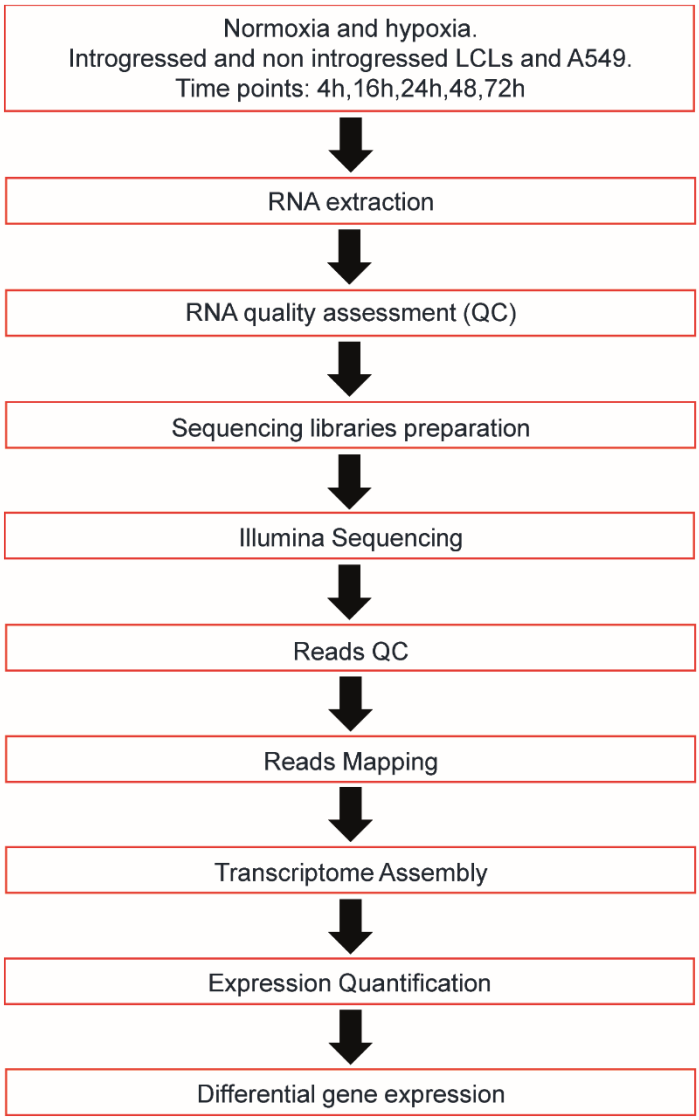


Figure 4. 1 Workflow of RNA sequencing experiment.

4.3 Results

4.3.1 The *EPAS1* introgressed haplotype in the 1000 Genomes Project data

The *EPAS1* ~ 32.7 kb introgressed region (104) was compared between samples from the 1000 Genomes Project, Tibetan and Densiovan (GRCh37, chr2:46567917-46600661; GRCh38, chr2: 46340778-46373522). Four Tibetan high coverage whole genome sequences generated and analysed in the previous chapter were also mapped to GRCh37 (Tib8, Tib14, Tib29, Tib34) who have high sequence similarity to the Denisovans. The individual showing the smallest number of sequence differences with the Denisovan is Tib29 (8 single nucleotide differences). Out of all the 1000 Genome Project individuals, 10 carry the full Densivan-core (AGGAA) haplotype (Table 4.3), wheareas 29 individuals carry a partial haplotype (3-4 of the core alleles). Of these 29 samples, 28 are of African ancestry (6 YRI, 2 MSL, 5 LWK, 4 GWD, 6 ESN, 3 ASW, 2ACB) and one of Han Chinese (CHB) ancestry. Three of four inviduals carrying the full Denisova-core haplotype, also have the putative functional variant A at rs370299814 (Table 4.3).

A haplotype network was generated from phased data of all the 39 individuals for the ~32.7 kb introgressed region (Figure 4.2). The Tibetans show both haplotypes closest to the Denisovan. The CDX individual HG00978 displays the closest haplotype to the Tibetan, followed by other South-East Asian samples carrying the full Denisovan-core (Figure 4.2). African samples show many single nucleotide differences in the entire ~32.7 kb region. Partial core Denisovan haplotypes in Africans can be explained by the presence of these alleles in the modern human gene pool before the introgression event (inherited from the modern human/Denisovan common ancestor). In contrast, the hypothesis of incomplete lineage sorting (ILS) between modern humans and Denisovans for the full *EPAS1* region seems unlikely (286). A more plausable explanation is an introgression event followed by strong positive selection on the haplotype that led to an extended haplotype with derived allele frequency > 0.8 in Tibetans and with very strong LD.

Other intersting variants in the region around *EPAS1* were also screened in the 1000 Genomes Project samples. The variant in *ATP6V1E2*, rs12986653, highlighted in the positive selection scans discussed in the second chapter, was present in all four introgressed cell lines used in the functional analyses (Table 4.1). This variant is an eQTL in different tissues including lung and arterial tissues (198, 305). Furthermore, the 4 bp deletion in the second *EPAS1* intron has an allele frequency of 0.85 in Tibetans, and between 0.38 and 0.63 in other high altitude populations. It shows low frequency (0-0.25) in low altitude Himalayans. All four introgressed CDX cell lines were heterozygous for this deletion. A cell line, HG02390, that does not carry the introgressed haplotype is heterozygous for this deletion. The deletion was also heterozygous in the other 1000 Genomes Project individuals carrying the full Denisovan core (HG00625, NA18566, NA18643, HG01847, HG02026, HG03814, HG03931) (Table 4.3). All the other individuals in 1000 Genomes Project were homyzygous for the ancestral allele. Two *EGLN1* variants, rs186996510 and rs12097901, with high derived allele frequency in Tibetans have been suggested to promote increased HIF factor degradation under hypoxic conditions (126). Surprisingly, these variants are present in both introgressed and non-introgressed samples. The derived allele of rs186996510 is found in only one heterozygous individual (HG01031) among those used for functional analysis. On the other hand, the derived allele of rs12097901 is present in all CDX cell lines in the heterozygous or homozygous state (Table 4.1).

		EPAS1					
		Core Denisovan haplotype ("AGGAA")					
		rs115321619	**rs73926263**	**rs73926264**	**rs73926265**	**rs55981512**	**rs370299814**
Pop	**Cell line**	**G/A**	**A/G**	**A/G**	**G/A**	**G/A**	**G/A**
CDX	HG00978	G/A	A/G	A/G	G/A	G/A	G/A
CDX	HG02396	G/A	A/G	A/G	G/A	G/A	G/G
CDX	HG02397	G/A	A/G	A/G	G/A	G/A	G/A
CDX	HG02187	G/A	A/G	A/G	G/A	G/A	G/A
CHB	NA18643	G/A	A/G	A/G	G/A	G/A	G/A
CHS	HG00625	G/A	A/G	A/G	G/A	G/A	G/G
KHV	HG01847	G/A	A/G	A/G	G/A	G/A	G/A
KHV	HG02026	G/A	A/G	A/G	G/A	G/A	G/A
BEB	HG03814	G/A	A/G	A/G	G/A	G/A	G/A
BEB	HG03931	G/A	A/G	A/G	G/A	G/A	G/A

Table 4. 3 The 1000 Genomes Project samples carrying the *EPAS1* Denisovan-core introgressed haplotype. The table reports individuals with the Denisovan haplotype. Eight out of ten carry also the putative functional variant rs370299814.

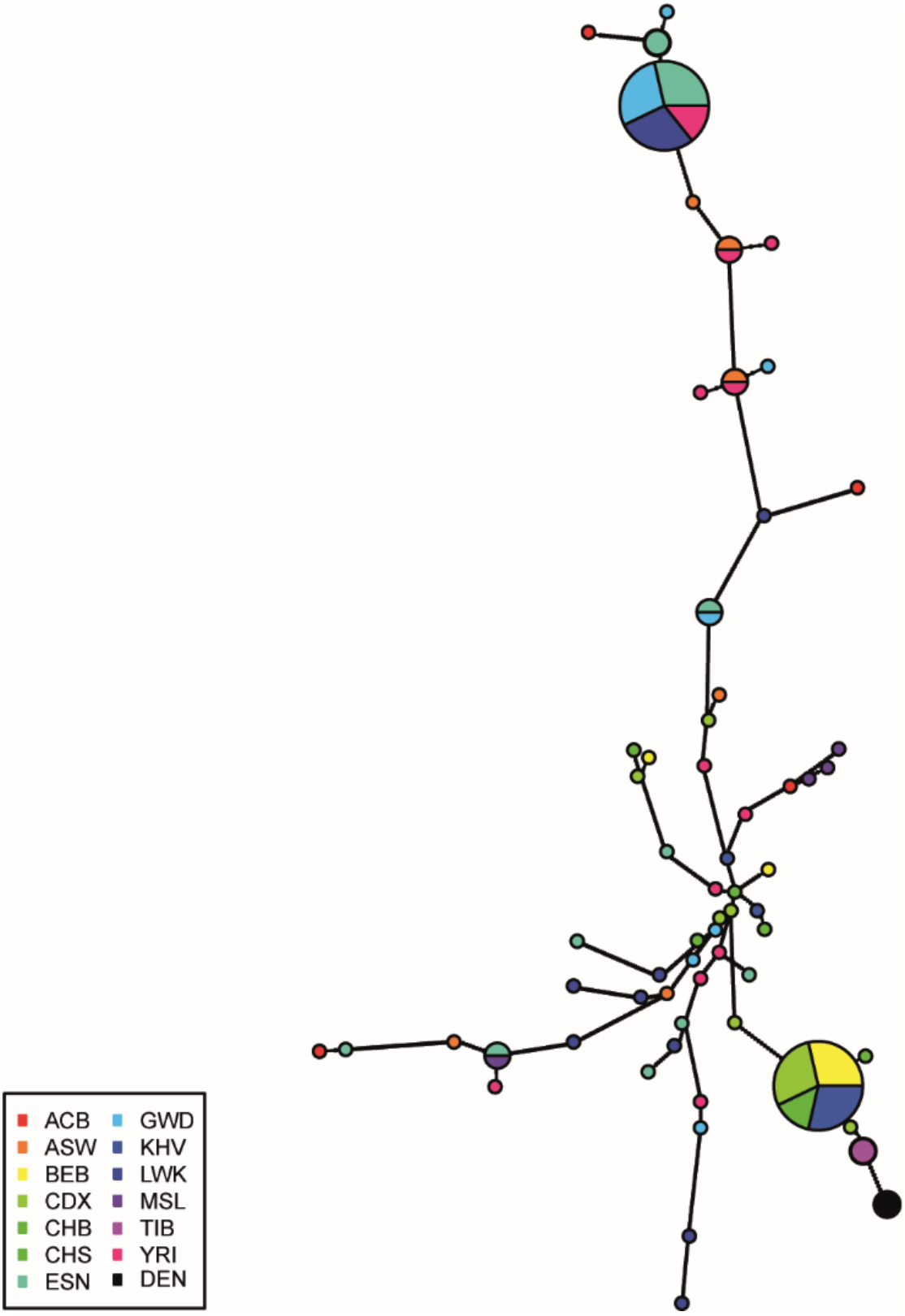


Figure 4. 2 Haplotype network of the ~32.7 kb *EPAS1* introgressed region. The Tibetan individuals show the closest haplotype to the Denisovan sample. African haplotypes show large number of differences from the Denisovan haplotype.

4.3.2 Re-analysis of *EPAS1* expression from publicly available data

EPAS1 expression was investigated in human tissues and cell lines available in the GTex Portal and ENCODE database (257). According to GTE, *EPAS1* is highly expressed in lung tissue followed by aortic and adipose tissues (Figure 4.3). Regarding ENCODE data, tissue-specific gene quatifications for *EPAS1* were compared in order to detect where the gene is most highly expressed. Aorta, lung, cardiac tissues, endothelial cells and placenta show the highest expression. LCL of GM12878 shows low constitutive levels of *EPAS1* expression, at a level similar to bronchial fibroblast and certain nervous system tissues (Figure 4.4).

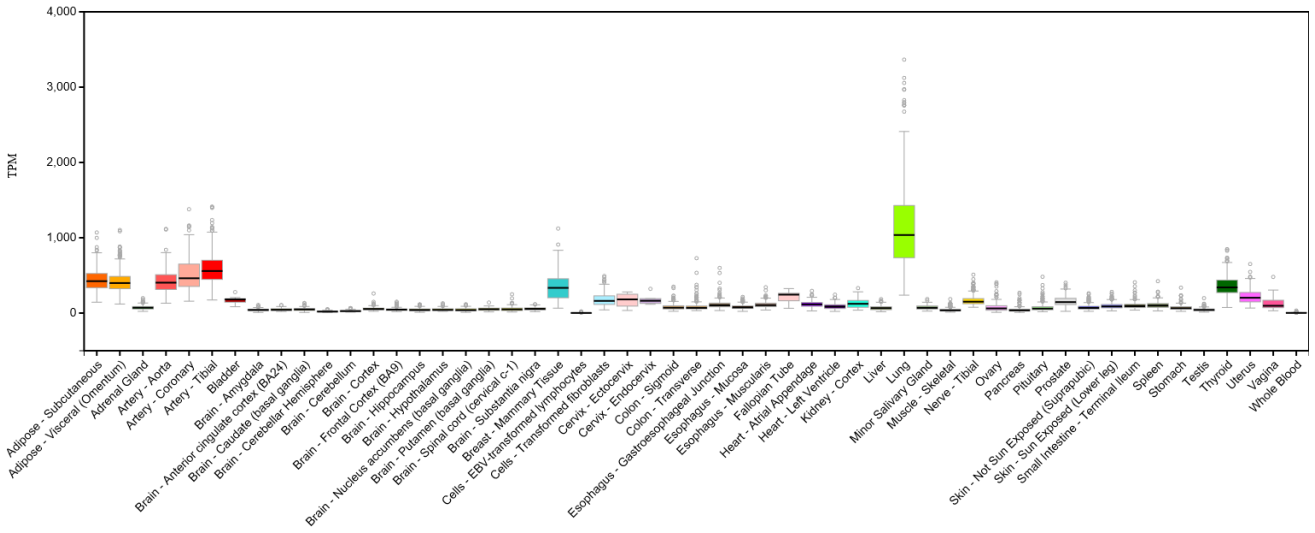


Figure 4. 3 *EPAS1* expression across different tissues in the GTEx Portal. On the x-axis are indicated the different tissues and on the y-axis are indicated the levels of expression in FTPM (Transcript Per Million).

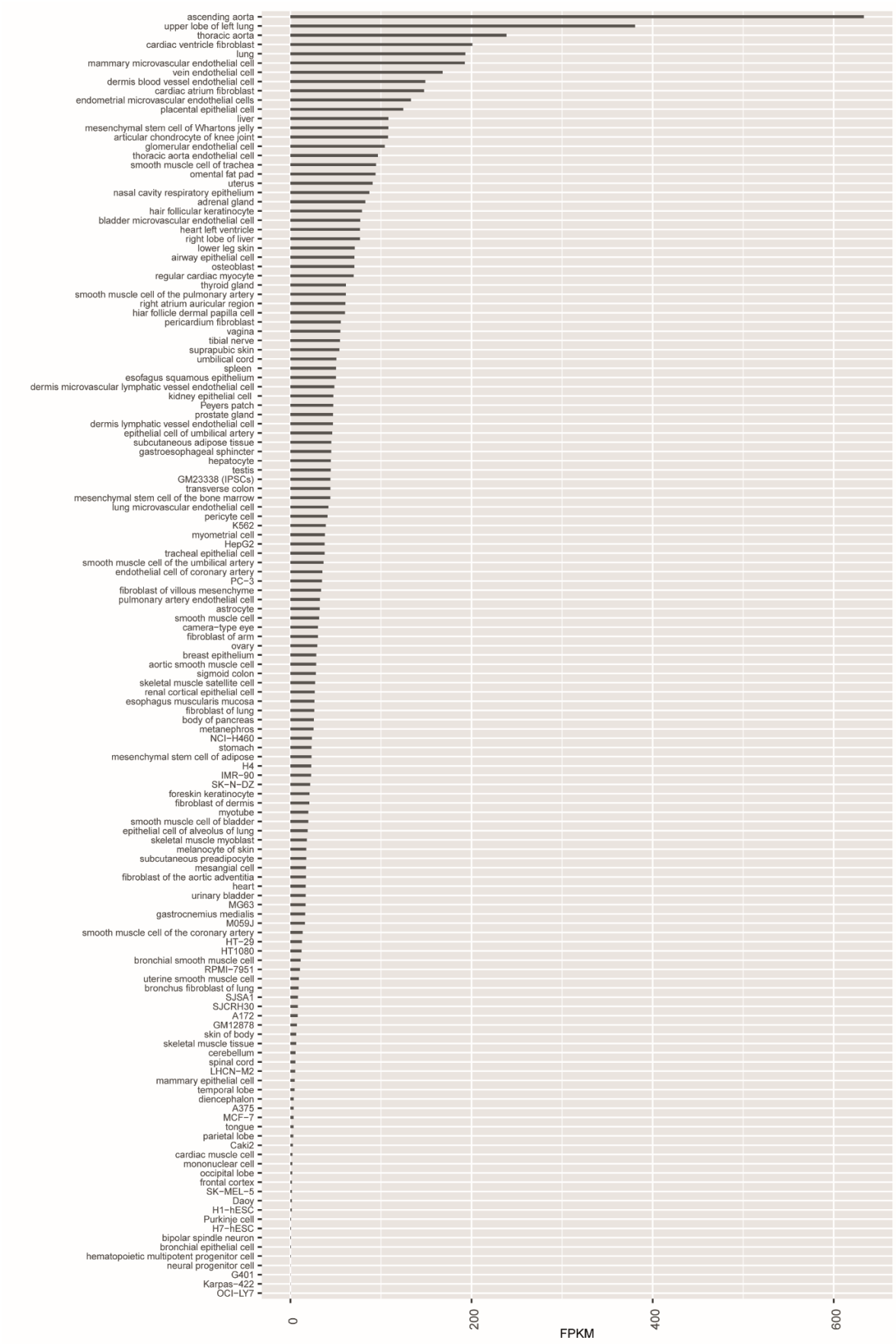


Figure 4. 4 *EPAS1* expression across different tissues in the ENCODE database. On the y-axis are indicated the different tissues and on the x-axis are indicated the levels of expression in FPKM.

EPAS1 expression was compared with other hypoxic factors to see whether their expression patterns were similar across the different tissues. *HIF-1α* shows considerable expression across different tissues, with myoblasts from skeletal musculature (LHCN-M2) and fibroblasts showing the highest values. On the other hand, *EPAS1* expression is more localised in specific tissues compared to *HIF-1α*. Overall, *HIF-3α* has lower expression across multiple tissues and is absent from many of them. *HIF-1β* and *EGLN1* show broad expression patterns across tissues although at very low levels (Figure 4.5). *EPAS1* and *HIF-1α* share high sequence homology and can both dimerize with *HIF-1β* in hypoxia and modulate gene expression. However, the two hypoxic factors regulate distinct sets of target genes (142). *HIF-1α* is broadly expressed and probably activates a general hypoxic response across tissues, whereas *EPAS1* seems to have an action in specific tissues (141). There is evidence of temporal organisation of *HIF-1α* and *EPAS1* hypoxic responses. It seems that the activation of target genes in acute hypoxia is primarly by *HIF-1α* (few hours after hypoxia exposure), whereas in response to chronic hypoxia *EPAS1* seems to exert a major role. This would explain the broader tissue expression of *HIF-1α* compared to *EPAS1* (141) (Figure 4.5).

EPAS1 expression was also investigated in publicly-available Tibetan placental villus parenchymal RNA-seq data (209). In the original paper, 63 Tibetan samples and 14 Han Chinese were sequenced and 72 samples of the 77 were availalble for download. However, they do not provide any genotype calls for the Tibetan and Han samples and only report the number of Tibetan individuals carrying the adaptive haplotype (Adaptive haplotype homozygous= 34, Adaptive haplotype heterozygous= 26, Wild-type haplotype homzygous=3). Thus, I performed variant calling from the RNA-seq data according to the GATK best practice and merged these genotypes with the Tibetans and Han Chinese individuals (CHB from the 1000 Genomes project) described in the previous chapter. PCA on these combined samples was performed, projecting the RNA-seq samples onto eigenvectors estimated from the Tibetan and Chinese Han sequences to check the ancestry of the RNA-seq samples (Figure 4.6A). PC1 separates Tibetans from Han Chinese and all RNA-seq samples lie together with Tibetans. The second PC divides Tibetans into subsclusters, with the RNA-seq samples mostly grouping with them. None of the RNA-seq samples cluster with Han Chinese (Figure 4.6A).

Figure 4. 5 Comparison of the hypoxic factors expression levels in different human tissues. The y-axis reports the tissues in the ENCODE database, and the x-axis the five major hypoxic genes. Expression levels (FPKM) are represented by the different colours as indicated in the legend.

After that, I tested *EPAS1* differential expression between individuals in the two main clusters of RNA-seq samples from the transcript quantifications performed with Cufflinks (297). In agreement with the original paper, *EPAS1* does not show differential expression within the samples (Figure 4.6B). However, it is important to note that only Tibetans seem to have been sampled for RNA-seq analysis (at least among the 72 available), so no comparison between Tibetans and Han Chinese was possible.

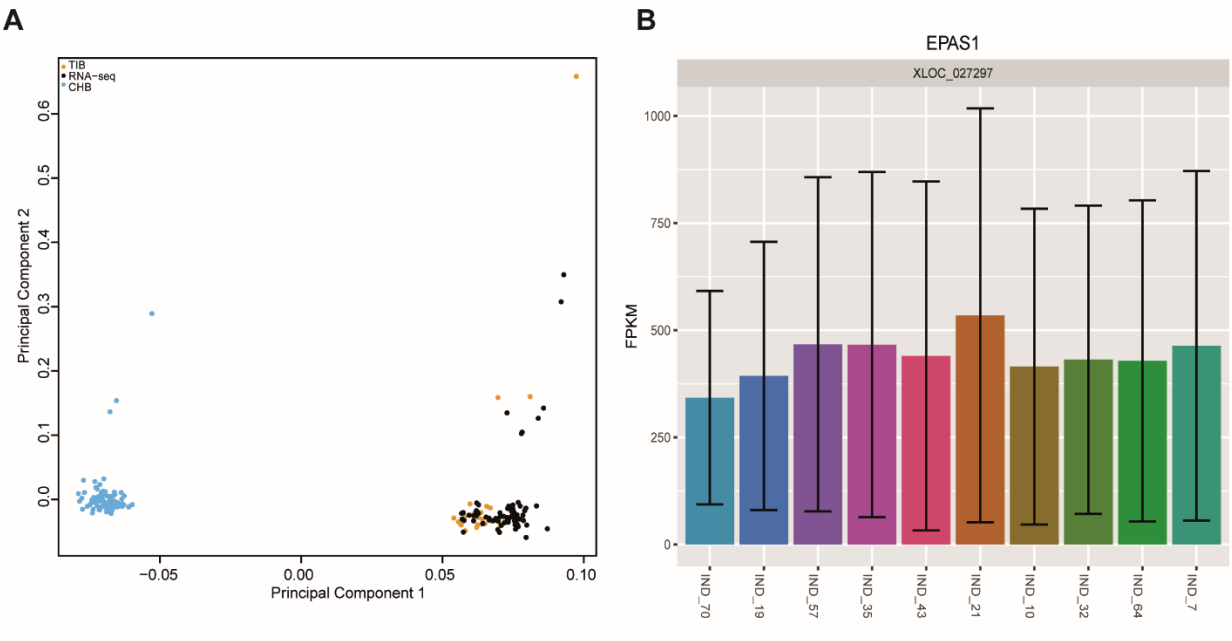


Figure 4. 6 Analysis of publicly-available Tibetan placental RNA-seq data. A. PCA of RNA-seq samples from (209) and Tibetan and Han Chinese whole-genome sequencing data. B. *EPAS1* differential expression across some of the RNA-seq samples. All individuals show similar levels of *EPAS1*.

4.3.3 Sequence conservation patterns in the *EPAS1* locus

Evolutionarily conserved regions (ECRs) were investigated in the *EPAS1* locus using the NBCI ECRs Browser. Overall, the entire *EPAS1* region shows a high degree of similarity to other primates, in particular Chimpanzee (*panTro3*) with > 98% sequence identity. Genomic similarity decreases with the increase of phylogenetic distance. ECRs are found both at the protein-coding (exon) level and in introns, UTRs, simple repetitive elements and transposons, and intergenic regions (Figure 4.7). Compared with Chimpazee, three large ECRs spanning the entire *EPAS1* locus with identity > 98.1% have been identified (ECR1 = chr2:46293668-46296898; ECR2 = chr2:46296963-46381755; ECR3 = chr2:46381818-46386697). The ~32.7 kb introgressed region falls within ECR2. Within these ECRs, 737 SNPs have been found. In particular, 20 SNPs lie in ECR1, 670 in ECR2 and 47 in ECR3. Of these, 54 SNPs show high F_{ST} values ($F_{ST} \geq 0.5$) between Tibetans and Han Chinese. All variants are categorised as non-coding by VEP (189). One, rs7557402, falls within a splice region and has previously been associated with polycythemia and familial erythrocytosis (Illumina Clinical Services Laboratory). This variant has a derived allele frequency (DAF) for the G allele of 0.83 in Tibetans and 0.18 in Han Chinese.

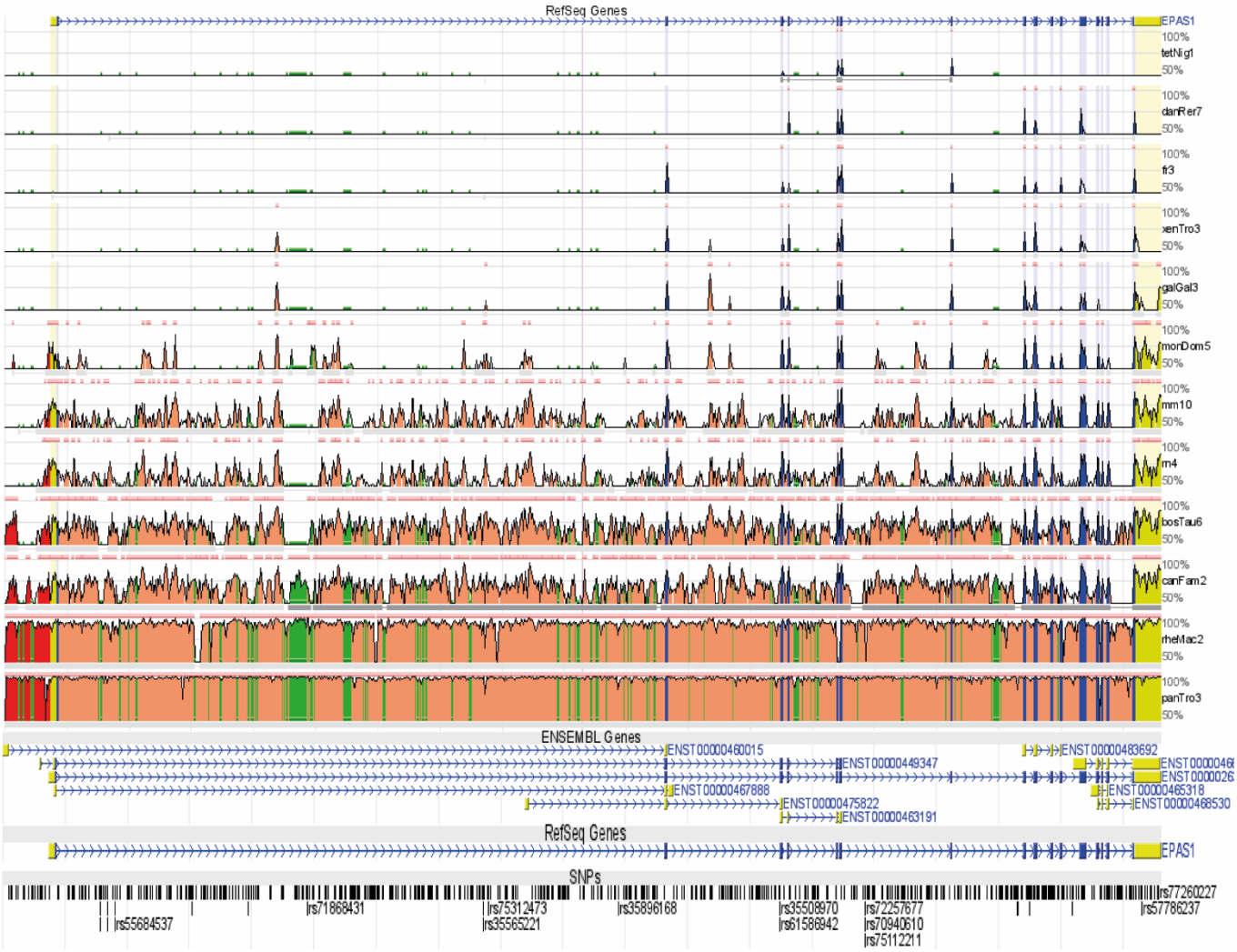


Figure 4. 7 *EPAS1* evolutionarily conserved regions (ECRs). The plot shows the proportion of sequence identity between humans and other species (tetNig=tetraodon, danRer=zebrafish, fr=fugu, xenTro=frog, galGal=chicken, monDom=opussum, mm=mouse, rn=rat, bosTau=cow, canFam=dog, rheMac=rhesus, panTro=chimpanzee). Orange= intronic regions, blue= protein-coding exons, green= transposons and single repeats, yellow= UTRs, red= intergenic regions. *EPAS1* transcripts and SNPs within the locus are also indicated.

4.3.4 Cell line karyotyping

The karyotypes of the eleven LCLs were determined. This has been done to characterize any additional chromosomal abnormalities, particularly involving chromosome 2, that could confound the results. Seven cell lines do not show any chromosomal aberration. The other four cell lines display different chromosomal aberrations, although none involving chromosome 2 (Table 4.4).

Cell line	Karyotype [number of cells]	% Cells
HG00844	46,XY[10]	100%
HG00881	46,XY[10]	100%
HG02187	46,XX[4]	40%
	46,XX,del(9),der(18)t(9;18), der(19)t(10;19)[6]	60%
HG02190	46,XX[9]	90%
	46,XX,del(7)[1]	10%
HG02397	46,XY[10]	100%
GM12878	46,XX[9]	90%
	46,XX,t(3;6)[1]	10%
HG02396	46,XY [8]	80%
	46,XY,dup(4)[1]	10%
	47,XY,t(2;3),+14[1]	10%
HG01310	46,XY [10]	100%
HG02390	46,XY[10]	100%
HG02153	46,XX[10]	100%
HG00978	46,XX[10]	100%

Table 4. 4 Cell line karyotypes. Del = deletion, dup = duplication, t = translocation, der = a derived chromosome, + = additional copy of chromosome.

4.3.6 *EPAS1* expression profiles

The expression of *EPAS1* and other genes involved in the hypoxic molecular pathway was assessed in LCLs with or without the introgressed haplotype. The cell lines were grown under normoxic (21% O_2) and hypoxic (1% O_2) conditions for different time points up to 72 hours (4h, 16h, 24h, 48h, 72h) and gene expression quantified by qPCR and RNA-seq.

4.3.6.1 Hypoxic protocol optimisation

Pilot studies using three LCLs with introgressed haplotype and three without together with the A549 cell line were set up to check the effectivness of the hypoxic protocol described previously and identify appropriate time points for assessment of changes in *EPAS1* and other genes (*EGLN1* and *HIF-1α*) expression under hypoxia in the different cell lines. Previous studies showed that two hours of hypoxic exposure was deemed to be a too short time to obtain reliable induction of hypoxia in LCLs and A549 cell culture. Hence, 4h was selected as the first time point for acute hypoxic exposure as in previous studies (304, 306). A second 8h time point was assessed, but did not show substantial differences (ΔC_T values) in gene expression from the 4 h one. *EPAS1* differential expression in LCLs was also detected at 24h and 48h of hypoxic exposure. Hence, the 4h and 24h time points were used to optimise the protocol. Subsequently, I set up an intermediate time point at 16h to investigate the induction of *EPAS1* expression. I also added a 72h time point to explore putative change in *EPAS1* expression under chronic hypoxic exposure. This led to a final 5 time point experiment (4h, 16h, 24h, 48h, 72h) that was performed and RNA from these experiments were sent for RNA-seq.

4.3.6.2 RT-qPCR

RT-qPCR was initially carried out to establish whether *EPAS1* was expressed in human LCLs and, if expressed, whether there were differences between the LCLs with and without the introgressed haplotype. The total RNA was manually extracted, retrotrascribed and qPCR of *EPAS1* and other genes (*EGLN1* and *HIF-1α*) involved in the hypoxic response was performed at time point of 4 and 24 hours.

The shape of the amplification curves passing the C_T threshold is similar across samples and two time points (Figure 4.8), indicating similar PCR efficiency (307). The qPCR results show detectable tested and control gene expression, indicating that the LCLs can be a suitable *in-vitro* model for understanding the role of *EPAS1* in hypoxia.

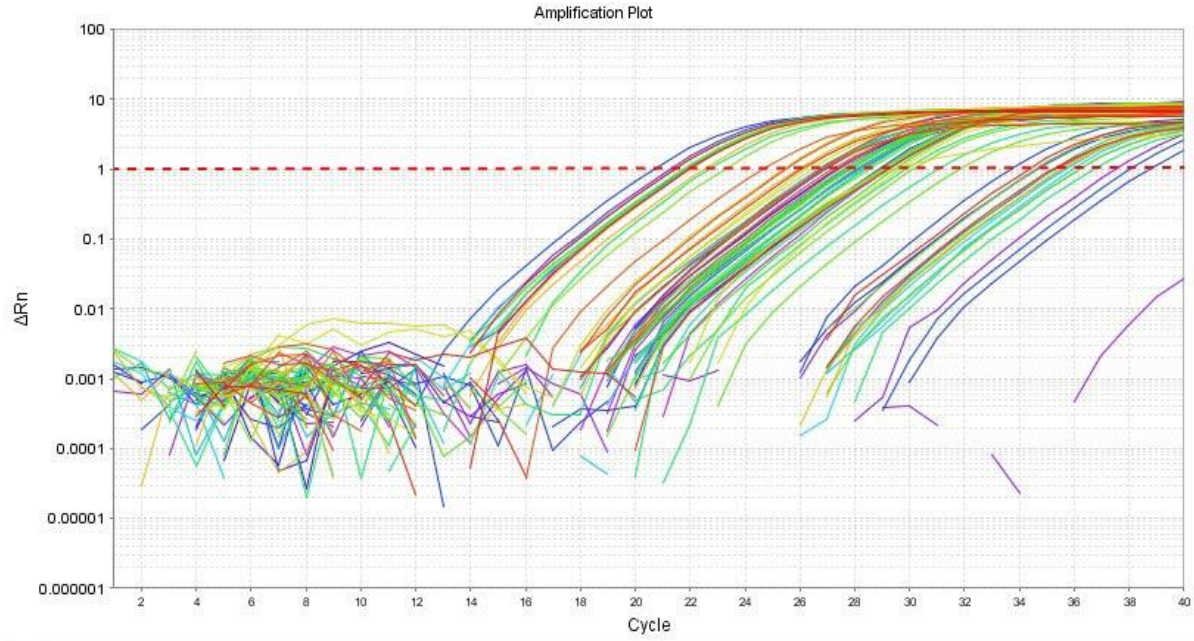


Figure 4. 8 qPCR amplification plot of target genes. The amplification plot shows the variation of log (ΔRn) with PCR cycle number. Rn is the reporter signal normalized to the fluorescence signal of the reference dye (ROX). ΔRn is Rn minus the baseline. All tested genes show an expression profile that passes the C_T threshold (dashed line). Each colour represents the row of the input plate. There no association between sample conditions and row.

Primer melting curves were inspected for single peaks of amplification around the primer melting temperature, indicating the likely amplification of a single product. Primer pairs showing multiple peaks were excluded to avoid biases in the interpretation of expression quantifications (Figure 4.9). In fact, the number of peaks is commonly used to determine the purity of qPCR products (315). *EPAS1* expression was compared with *HIF-1α*, *EGLN1* and *TUBB* in LCLs and A549. The housekeeping gene *TUBB* does not show a change in expression across conditions and, thus, can be used as standard control. Slight variation in the cycle threshold (C_T) can reflect pipetting errors, due to the manual preparation of the qPCR plate and the limited precision of the pipette (Figure 4.9). Target genes were standardised to the housekeeping gene *TUBB* and gene expression profiling was calculated as

standardised gene expression ($2^{-\Delta\Delta CT}$ = [(C_T gene of interest – C_T internal control)$_{hypoxia}$ – (C_T gene of interest – C_T internal control)$_{normoxia}$]) (307). Differential expression was also assessed as $\Delta C_T = C_{T(normxia)} - C_{T(hypoxia)}$. Each cycle change corresponds to a two-fold change in expression (307).

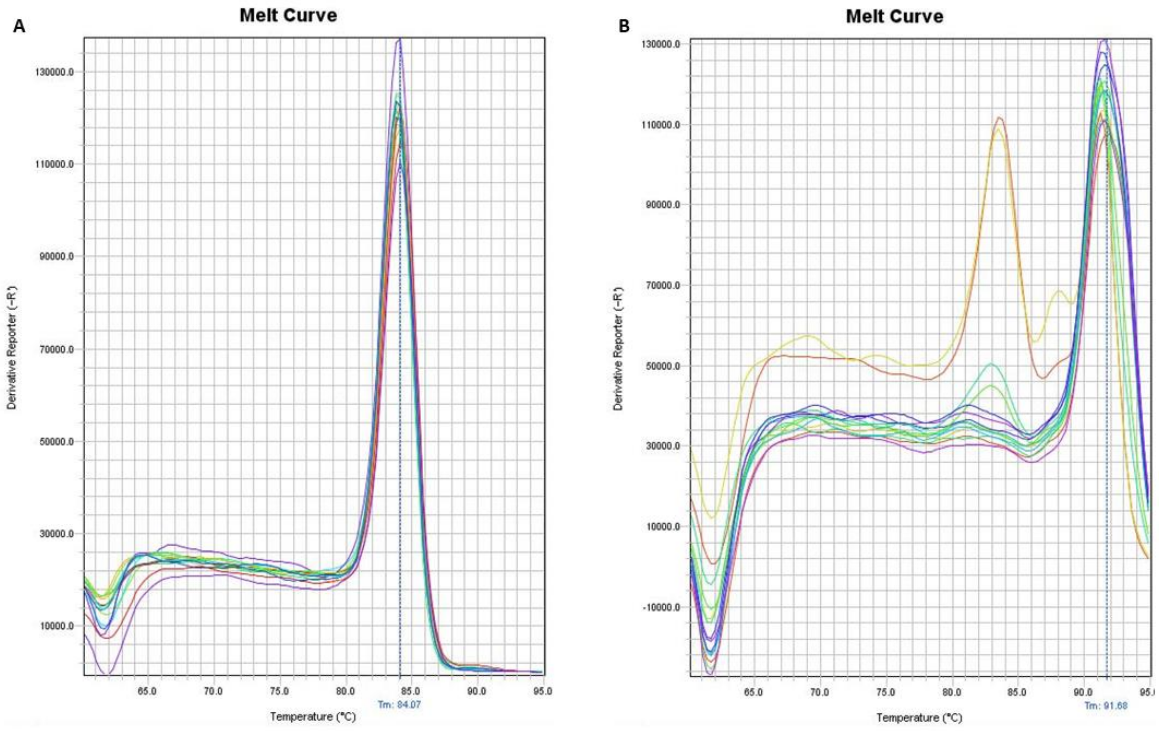


Figure 4. 9 Primer melting temperature curves. The plots show example of melting curves for each primer pair across samples. Each colour represents the row of the input plate. There is no association between sample conditions and row. The x-axis reports the temperature whereas the y-axis shows the rate of the fluorescence variation. The plots report melting curves of two sets of *EPAS1* primers. A. Example of a good single peak trace. B. Example of a bad melting curve with multiple peaks.

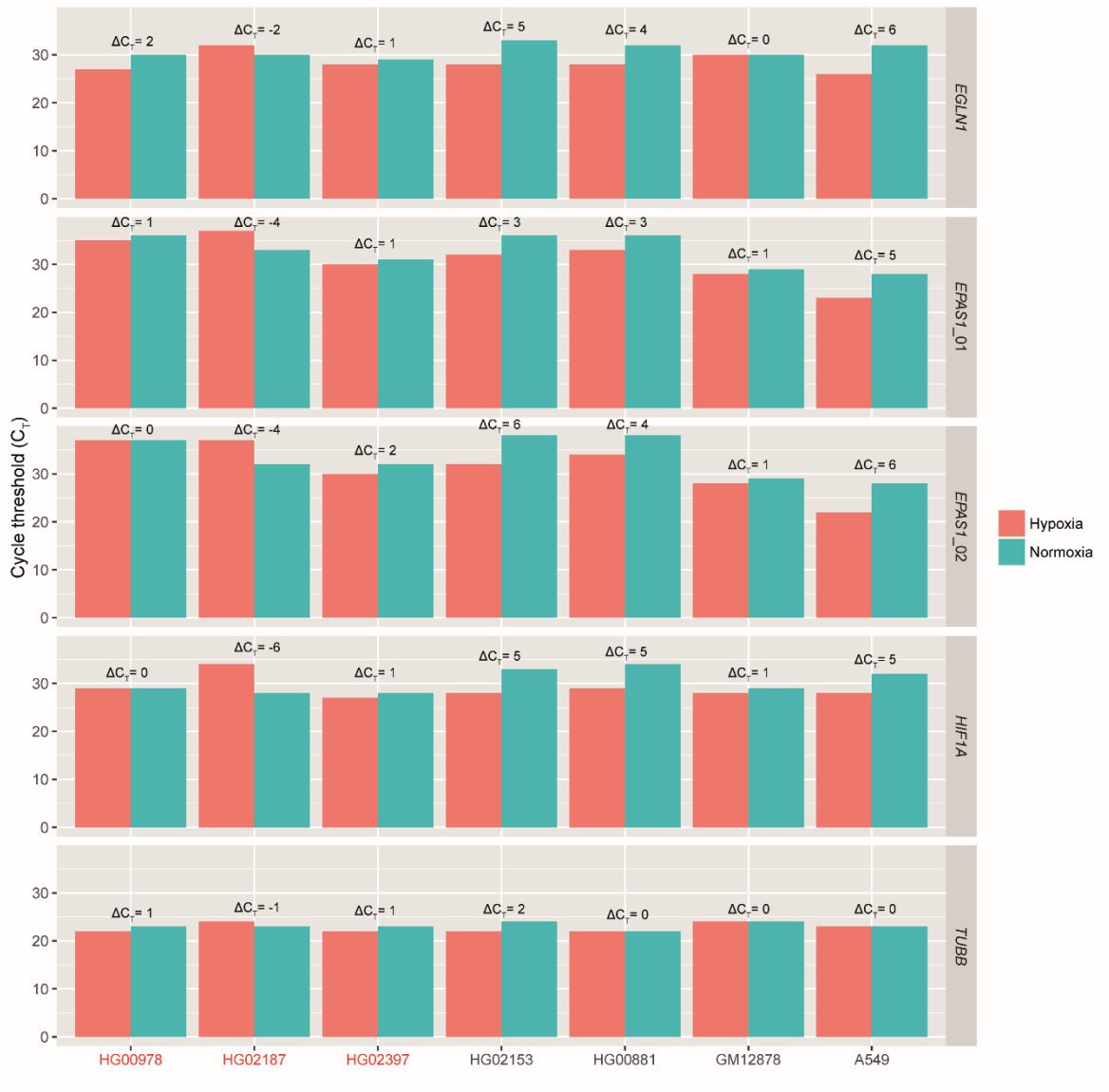


Figure 4. 10 C_T differences between normoxia and hypoxia at 4 hours. The plot shows the C_T for each gene across introgressed and non introgressed cell lines. The lower the C_T, the higher the expression. ΔC_T values are also reported for each sample between normoxia and hypoxia. LCLs with the introgressed core Denisovan haplotype are indicated in red.

At 4 hours of hypoxic exposure, *EPAS1* is highly expressed in A549 cell lines (*EPAS1*_01 $\Delta CT_{(normoxia-hypoxia)} = 5$; *EPAS1*_02 $\Delta CT_{(normoxia-hypoxia)} = 6$), but it shows very low levels in LCLs. Furthermore, no differential expression between the LCLs with and without the introgressed haplotype is detectable (Figure 4.10 and 4.11). This confirms that the peak of hypoxic factors expression in A549 is around 4 hours.

Changes in expression between hypoxia and normoxia (ΔCT=4-5) was observed for *HIF-1α* and *EGLN1.* Both genes showed higher expression (ΔC_T) in hypoxia in three non introgressed samples (HG02153, HG00881, A549) whereas all

introgressed and GM12878 do not show increased *EGLN1* and *HIF-1α* expression in hypoxia compared to normoxia.

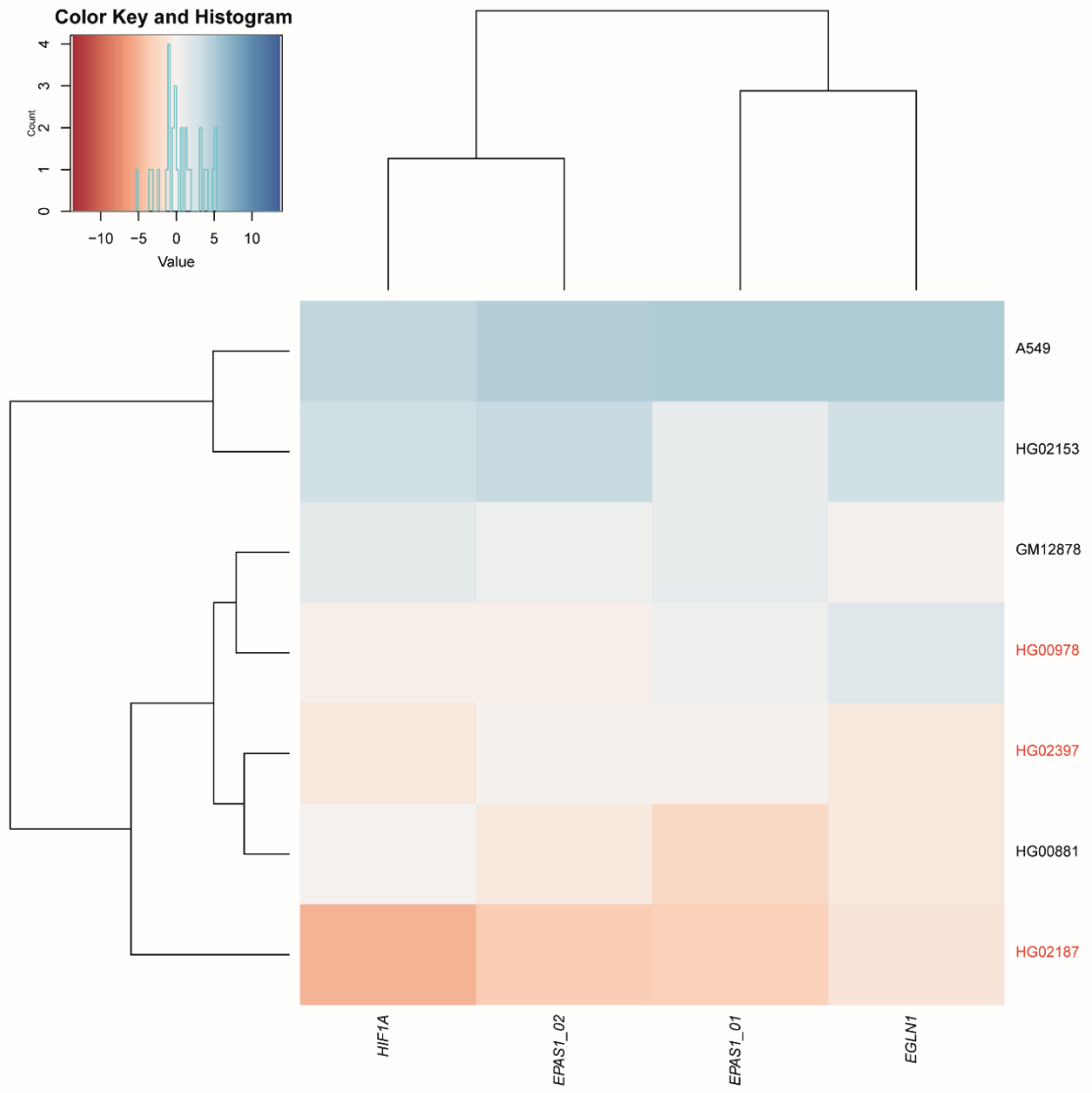


Figure 4. 11 *EPAS1* expression at 4 hours hypoxic exposure. The plot reports the standardised gene expression for the target genes. *EPAS1* shows high expression in the A549 cell line but not in LCLs. No significant differences between the LCLs with and without the introgressed haplotype are detectable.

EPAS1 shows an increase in expression at 24h of hypoxic exposure, when the differential expression between the LCLs with and without the introgressed haplotype was observed. This was further confirmed in different biological samples (three LCLs with introgressed haplotype), technical replicates as well as by two different pairs of primers for *EPAS1* (*EPAS1*_01 and *EPAS1*_02). Specifically, in the LCLs without the introgressed haplotype, *EPAS1* expression increases under hypoxic conditions, whereas it does not in the ones heterozygous for the introgressed haplotype. Another pilot experiment showed that introgressed cell lines carriyng the rs370299814 (HG00978, HG02187, HG02397) and the one without it (HG02396) show the same steady *EPAS1* expression between normoxia and hypoxia. Thus rs370299814 might not be the functional variant associated with this phenotype. At 24 hours, A549 does not show high *EPAS1* expression or any differential expression between normoxia and hypoxia (Figure 4.12 and 4.13). Small changes in expression between hypoxia and normoxia (ΔCT=1-3) was observed for *HIF-1α* and *EGLN1.*

Figure 4. 12 C_T differences between normoxia and hypoxia at 24 hours. The plot shows the C_T for each gene in LCLs with and without the introgressed haplotype. The lower the C_T, the higher the gene expression. ΔC_T values between normoxia and hypoxia are also reported. The LCLs with the introgressed haplotype are in red.

This could mean that in this cell line at 24h, the downstream hypoxia response has already been triggered. Small changes in expression between hypoxia and normoxia (ΔCT=1-3) was observed for *HIF-1α* and *EGLN1* but it was lower than at the 4 hours time points and there was no differential expression between the LCLs with and without the introgressed haplotype. (Table 4.1, Figure 4.13).

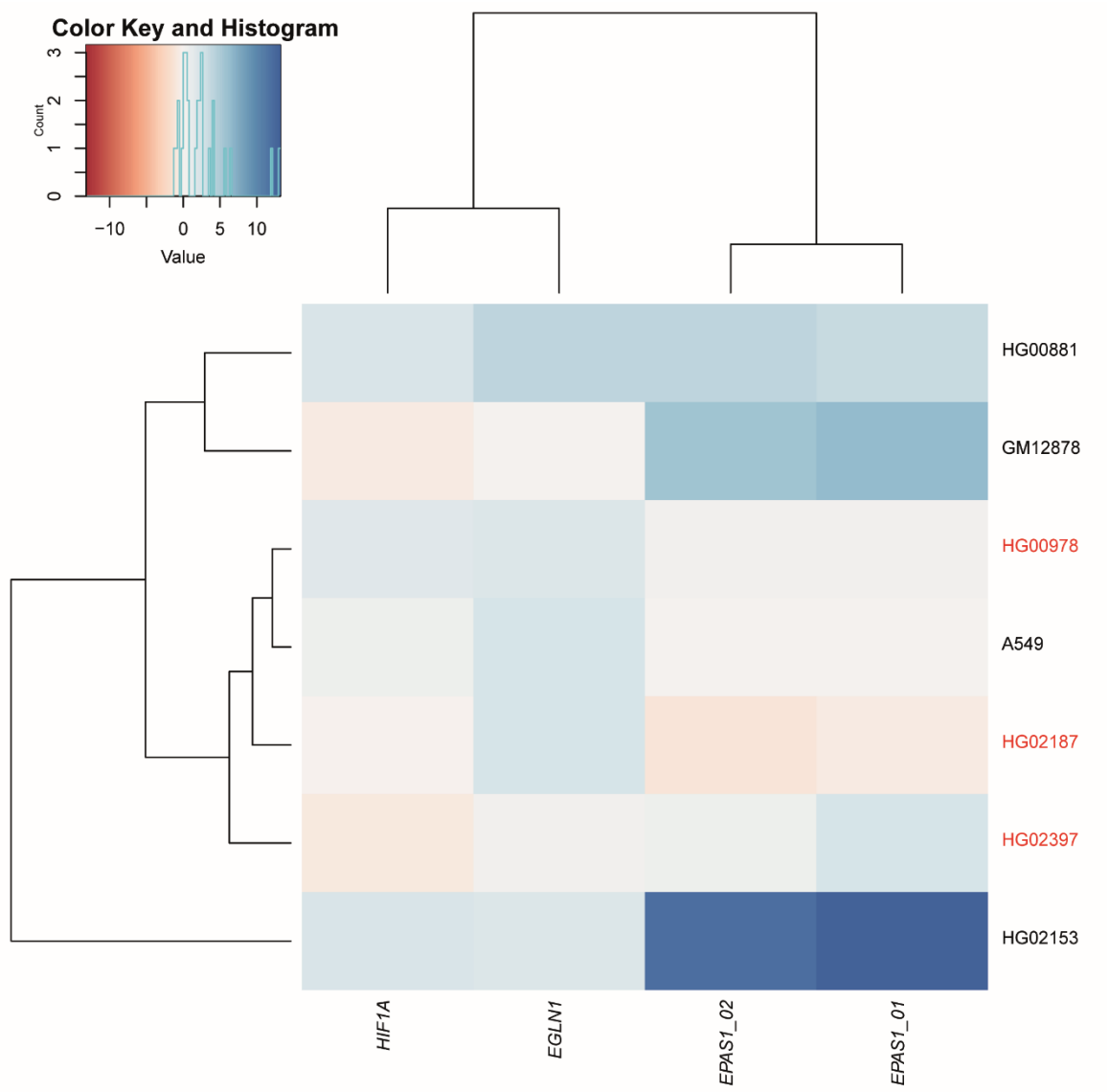


Figure 4. 13 *EPAS1* expression at 24 hours hypoxic exposure. The plot reports the standardised gene expression for the target genes. *EPAS1* shows differential expression between introgressed and non-introgressed cell lines. Cell lines in red are introgressed cell lines.

The expression of all the genes lying in the ~300 kb introgressed region was assessed (*EPAS1*, *TMEM247*, *ATP6V1E2*, *PIGF* and *CRIPT*). *EPAS1* was the only one showing a significant change of expression between normoxia and hypoxia and a differential expression between LCLs with and without the introgressed haplotypes. However, this conclusion needs to be interpreted cautiously because of the extremely low expression of some of the genes in the LCLs, where in some of the samples the expression level remained undetermined (C_T too high), and the results are not discussed further or presented here because an improved approach is being pursued, as described in the next section.

4.3.6.3 RNA-seq

The analysis of the RNA-seq data is ongoing and is currently in its preliminary phase. The quality of the data was evaluated and basic analyses to assess the success of the experiment were carried out.

4.3.6.3.1 Data QC

The quality of RNA-seq data was assessed using FastQC and samtools flagstat. The samples were sequenced over two runs (run one with one lane, run two with 3 lanes), and QC was performed per lane to identify and exclude possible technical biases. Per sample FastQC results were aggregated using Multi QC v1.3. The total number of mapped reads per sample ranged between ~5 and 7 million (Figures 4.14 and 4.15). In all samples, most of the reads were uniquely mapped and most onto exons (Figures 4.14 and 4.15 A,B). Most of the samples have around 50% duplicated reads, which may be due to the high abundance of a small number of genes rather than an indication of PCR over-amplification or low library complexity. Per base sequence mean quality score (the range of quality values across all bases at each position)) is above 30 for all samples. All reads show good average quality score (per sequence quality score > 27); no reads have poor quality. None of the samples show adapter contamination. Finally, the N content across all bases was very low across all samples.

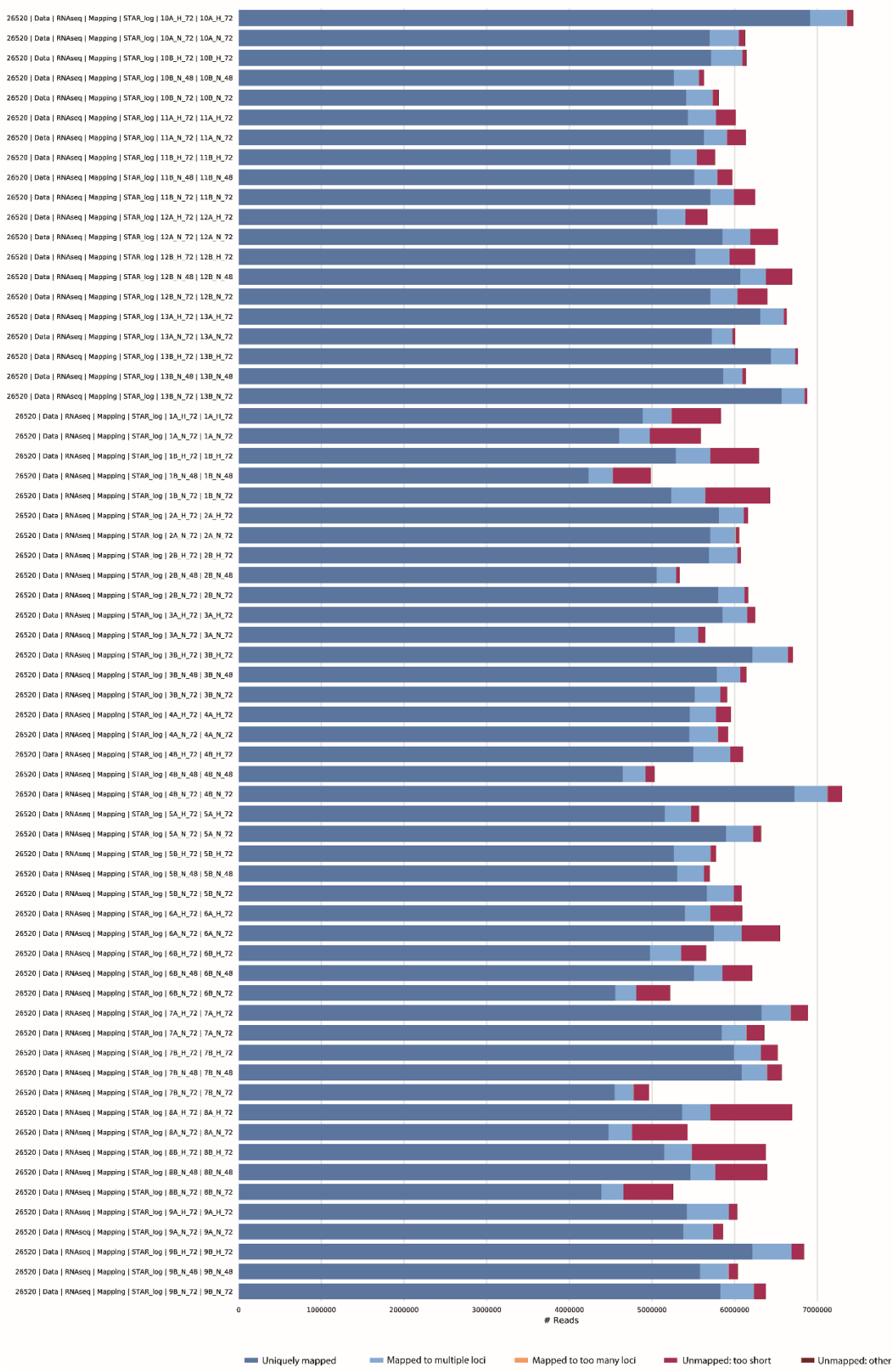


Figure 4. 14 Total number of reads sequenced in the first run (26520). The x-axis shows the total number of reads, the y-axis the sample sequenced (first element = run ID, last element= name of the sample representing each cell line (1-13), the replicate (A,B), the condition (N, H) and the time point (4,16,24,48,72).

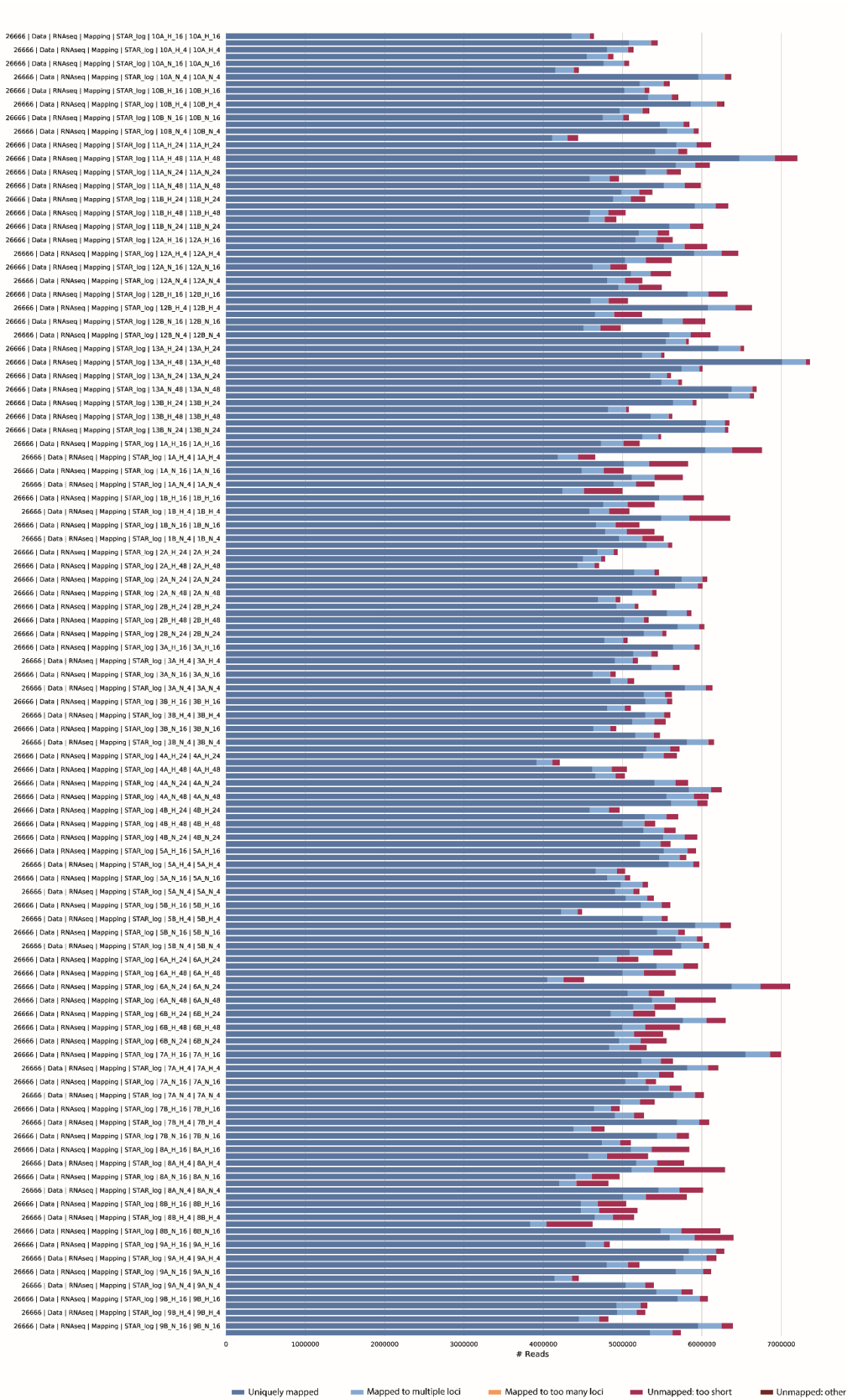

26666 | Data | RNAseq | Mapping | STAR_log | 10A_H_16 | 10A_H_16
26666 | Data | RNAseq | Mapping | STAR_log | 10A_H_4 | 10A_H_4
26666 | Data | RNAseq | Mapping | STAR_log | 10A_N_16 | 10A_N_16
26666 | Data | RNAseq | Mapping | STAR_log | 10A_N_4 | 10A_N_4
26666 | Data | RNAseq | Mapping | STAR_log | 10B_H_16 | 10B_H_16
26666 | Data | RNAseq | Mapping | STAR_log | 10B_H_4 | 10B_H_4
26666 | Data | RNAseq | Mapping | STAR_log | 10B_N_16 | 10B_N_16
26666 | Data | RNAseq | Mapping | STAR_log | 10B_N_4 | 10B_N_4
26666 | Data | RNAseq | Mapping | STAR_log | 11A_H_24 | 11A_H_24
26666 | Data | RNAseq | Mapping | STAR_log | 11A_H_48 | 11A_H_48
26666 | Data | RNAseq | Mapping | STAR_log | 11A_N_24 | 11A_N_24
26666 | Data | RNAseq | Mapping | STAR_log | 11A_N_48 | 11A_N_48
26666 | Data | RNAseq | Mapping | STAR_log | 11B_H_24 | 11B_H_24
26666 | Data | RNAseq | Mapping | STAR_log | 11B_H_48 | 11B_H_48
26666 | Data | RNAseq | Mapping | STAR_log | 11B_N_24 | 11B_N_24
26666 | Data | RNAseq | Mapping | STAR_log | 12A_H_16 | 12A_H_16
26666 | Data | RNAseq | Mapping | STAR_log | 12A_H_4 | 12A_H_4
26666 | Data | RNAseq | Mapping | STAR_log | 12A_N_16 | 12A_N_16
26666 | Data | RNAseq | Mapping | STAR_log | 12A_N_4 | 12A_N_4
26666 | Data | RNAseq | Mapping | STAR_log | 12B_H_16 | 12B_H_16
26666 | Data | RNAseq | Mapping | STAR_log | 12B_H_4 | 12B_H_4
26666 | Data | RNAseq | Mapping | STAR_log | 12B_N_16 | 12B_N_16
26666 | Data | RNAseq | Mapping | STAR_log | 12B_N_4 | 12B_N_4
26666 | Data | RNAseq | Mapping | STAR_log | 13A_H_24 | 13A_H_24
26666 | Data | RNAseq | Mapping | STAR_log | 13A_H_48 | 13A_H_48
26666 | Data | RNAseq | Mapping | STAR_log | 13A_N_24 | 13A_N_24
26666 | Data | RNAseq | Mapping | STAR_log | 13A_N_48 | 13A_N_48
26666 | Data | RNAseq | Mapping | STAR_log | 13B_H_24 | 13B_H_24
26666 | Data | RNAseq | Mapping | STAR_log | 13B_H_48 | 13B_H_48
26666 | Data | RNAseq | Mapping | STAR_log | 13B_N_24 | 13B_N_24
26666 | Data | RNAseq | Mapping | STAR_log | 1A_H_16 | 1A_H_16
26666 | Data | RNAseq | Mapping | STAR_log | 1A_H_4 | 1A_H_4
26666 | Data | RNAseq | Mapping | STAR_log | 1A_N_16 | 1A_N_16
26666 | Data | RNAseq | Mapping | STAR_log | 1A_N_4 | 1A_N_4
26666 | Data | RNAseq | Mapping | STAR_log | 1B_H_16 | 1B_H_16
26666 | Data | RNAseq | Mapping | STAR_log | 1B_H_4 | 1B_H_4
26666 | Data | RNAseq | Mapping | STAR_log | 1B_N_16 | 1B_N_16
26666 | Data | RNAseq | Mapping | STAR_log | 1B_N_4 | 1B_N_4
26666 | Data | RNAseq | Mapping | STAR_log | 2A_H_24 | 2A_H_24
26666 | Data | RNAseq | Mapping | STAR_log | 2A_H_48 | 2A_H_48
26666 | Data | RNAseq | Mapping | STAR_log | 2A_N_24 | 2A_N_24
26666 | Data | RNAseq | Mapping | STAR_log | 2A_N_48 | 2A_N_48
26666 | Data | RNAseq | Mapping | STAR_log | 2B_H_24 | 2B_H_24
26666 | Data | RNAseq | Mapping | STAR_log | 2B_H_48 | 2B_H_48
26666 | Data | RNAseq | Mapping | STAR_log | 2B_N_24 | 2B_N_24
26666 | Data | RNAseq | Mapping | STAR_log | 3A_H_16 | 3A_H_16
26666 | Data | RNAseq | Mapping | STAR_log | 3A_H_4 | 3A_H_4
26666 | Data | RNAseq | Mapping | STAR_log | 3A_N_16 | 3A_N_16
26666 | Data | RNAseq | Mapping | STAR_log | 3A_N_4 | 3A_N_4
26666 | Data | RNAseq | Mapping | STAR_log | 3B_H_16 | 3B_H_16
26666 | Data | RNAseq | Mapping | STAR_log | 3B_H_4 | 3B_H_4
26666 | Data | RNAseq | Mapping | STAR_log | 3B_N_16 | 3B_N_16
26666 | Data | RNAseq | Mapping | STAR_log | 3B_N_4 | 3B_N_4
26666 | Data | RNAseq | Mapping | STAR_log | 4A_H_24 | 4A_H_24
26666 | Data | RNAseq | Mapping | STAR_log | 4A_H_48 | 4A_H_48
26666 | Data | RNAseq | Mapping | STAR_log | 4A_N_24 | 4A_N_24
26666 | Data | RNAseq | Mapping | STAR_log | 4A_N_48 | 4A_N_48
26666 | Data | RNAseq | Mapping | STAR_log | 4B_H_24 | 4B_H_24
26666 | Data | RNAseq | Mapping | STAR_log | 4B_H_48 | 4B_H_48
26666 | Data | RNAseq | Mapping | STAR_log | 4B_N_24 | 4B_N_24
26666 | Data | RNAseq | Mapping | STAR_log | 5A_H_16 | 5A_H_16
26666 | Data | RNAseq | Mapping | STAR_log | 5A_H_4 | 5A_H_4
26666 | Data | RNAseq | Mapping | STAR_log | 5A_N_16 | 5A_N_16
26666 | Data | RNAseq | Mapping | STAR_log | 5A_N_4 | 5A_N_4
26666 | Data | RNAseq | Mapping | STAR_log | 5B_H_16 | 5B_H_16
26666 | Data | RNAseq | Mapping | STAR_log | 5B_H_4 | 5B_H_4
26666 | Data | RNAseq | Mapping | STAR_log | 5B_N_16 | 5B_N_16
26666 | Data | RNAseq | Mapping | STAR_log | 5B_N_4 | 5B_N_4
26666 | Data | RNAseq | Mapping | STAR_log | 6A_H_24 | 6A_H_24
26666 | Data | RNAseq | Mapping | STAR_log | 6A_H_48 | 6A_H_48
26666 | Data | RNAseq | Mapping | STAR_log | 6A_N_24 | 6A_N_24
26666 | Data | RNAseq | Mapping | STAR_log | 6A_N_48 | 6A_N_48
26666 | Data | RNAseq | Mapping | STAR_log | 6B_H_24 | 6B_H_24
26666 | Data | RNAseq | Mapping | STAR_log | 6B_H_48 | 6B_H_48
26666 | Data | RNAseq | Mapping | STAR_log | 6B_N_24 | 6B_N_24
26666 | Data | RNAseq | Mapping | STAR_log | 7A_H_16 | 7A_H_16
26666 | Data | RNAseq | Mapping | STAR_log | 7A_H_4 | 7A_H_4
26666 | Data | RNAseq | Mapping | STAR_log | 7A_N_16 | 7A_N_16
26666 | Data | RNAseq | Mapping | STAR_log | 7A_N_4 | 7A_N_4
26666 | Data | RNAseq | Mapping | STAR_log | 7B_H_16 | 7B_H_16
26666 | Data | RNAseq | Mapping | STAR_log | 7B_H_4 | 7B_H_4
26666 | Data | RNAseq | Mapping | STAR_log | 7B_N_16 | 7B_N_16
26666 | Data | RNAseq | Mapping | STAR_log | 8A_H_16 | 8A_H_16
26666 | Data | RNAseq | Mapping | STAR_log | 8A_H_4 | 8A_H_4
26666 | Data | RNAseq | Mapping | STAR_log | 8A_N_16 | 8A_N_16
26666 | Data | RNAseq | Mapping | STAR_log | 8A_N_4 | 8A_N_4
26666 | Data | RNAseq | Mapping | STAR_log | 8B_H_16 | 8B_H_16
26666 | Data | RNAseq | Mapping | STAR_log | 8B_H_4 | 8B_H_4
26666 | Data | RNAseq | Mapping | STAR_log | 8B_N_16 | 8B_N_16
26666 | Data | RNAseq | Mapping | STAR_log | 9A_H_16 | 9A_H_16
26666 | Data | RNAseq | Mapping | STAR_log | 9A_H_4 | 9A_H_4
26666 | Data | RNAseq | Mapping | STAR_log | 9A_N_16 | 9A_N_16
26666 | Data | RNAseq | Mapping | STAR_log | 9A_N_4 | 9A_N_4
26666 | Data | RNAseq | Mapping | STAR_log | 9B_H_16 | 9B_H_16
26666 | Data | RNAseq | Mapping | STAR_log | 9B_H_4 | 9B_H_4
26666 | Data | RNAseq | Mapping | STAR_log | 9B_N_16 | 9B_N_16
0
1000000
2000000
3000000
4000000
5000000
6000000
7000000
Reads
Uniquely mapped
Mapped to multiple loci
Mapped to too many loci
Unmapped: too short
Unmapped: other

Figure 4. 15 Total number of reads sequenced in the second run (26660). The x-axis indicates the total number of reads, the y-axis the sample sequenced (first element = run ID, last element= name of the sample representing each cell line (1-13), the replicate (A,B), the condition (N, H) and the time point (4,16,24,48,72).

No significant differences between the two runs were detected. Both runs show samples with a similar number of uniquely mapped reads with most of the reads mapping on exonic regions (Figure 4.16) and similar QC results.

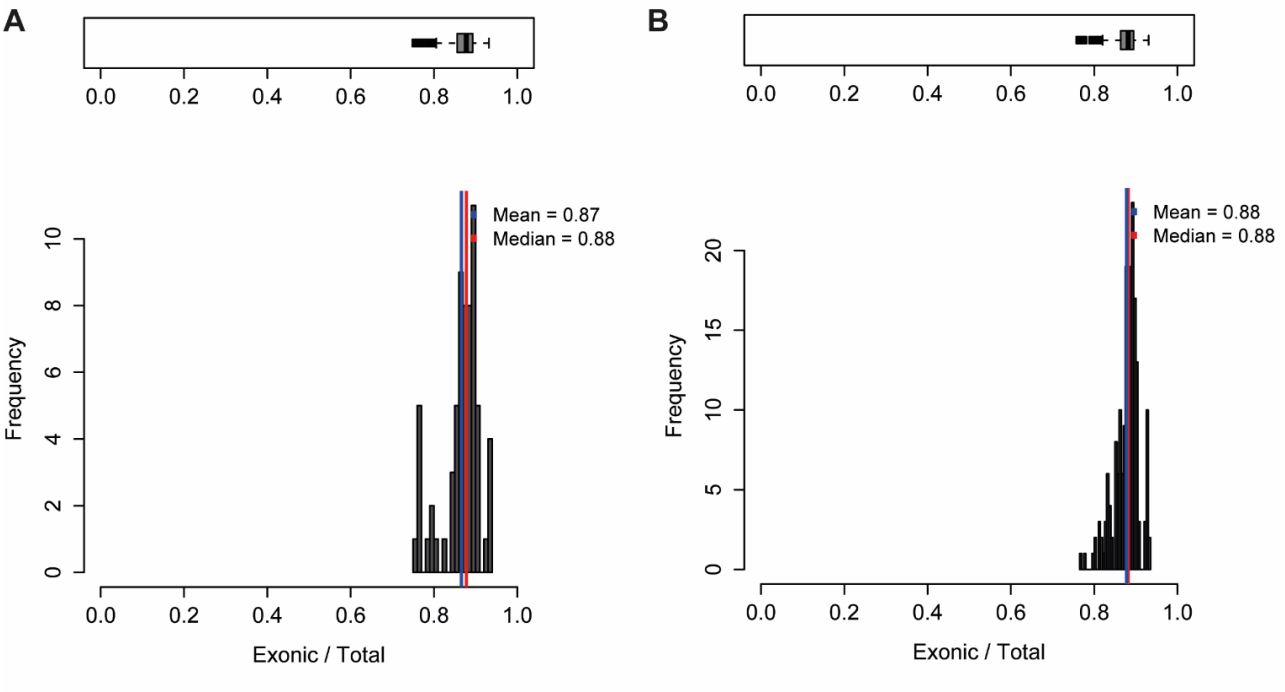


Figure 4. 16 Proportion of exonic reads in the samples. Most of the reads map to exons for both runs. Mean (blue) and median (red) values are indicated on the plots. A Plot for samples in the first run (1 lane). B Samples in the second run (3 lanes).

4.3.6.3.2 Gene expression in hypoxia and normoxia

Expression quantification was performed with featureCounts. The quantification matrix was analysed using Deseq2. PCA using whole-genome expression data shows that the samples are clustering for technical replicates and by the presence of the introgressed haplotype rather than the oxygen concentration exposure (normoxia versus hypoxia) (Figure 4.17 A and B).

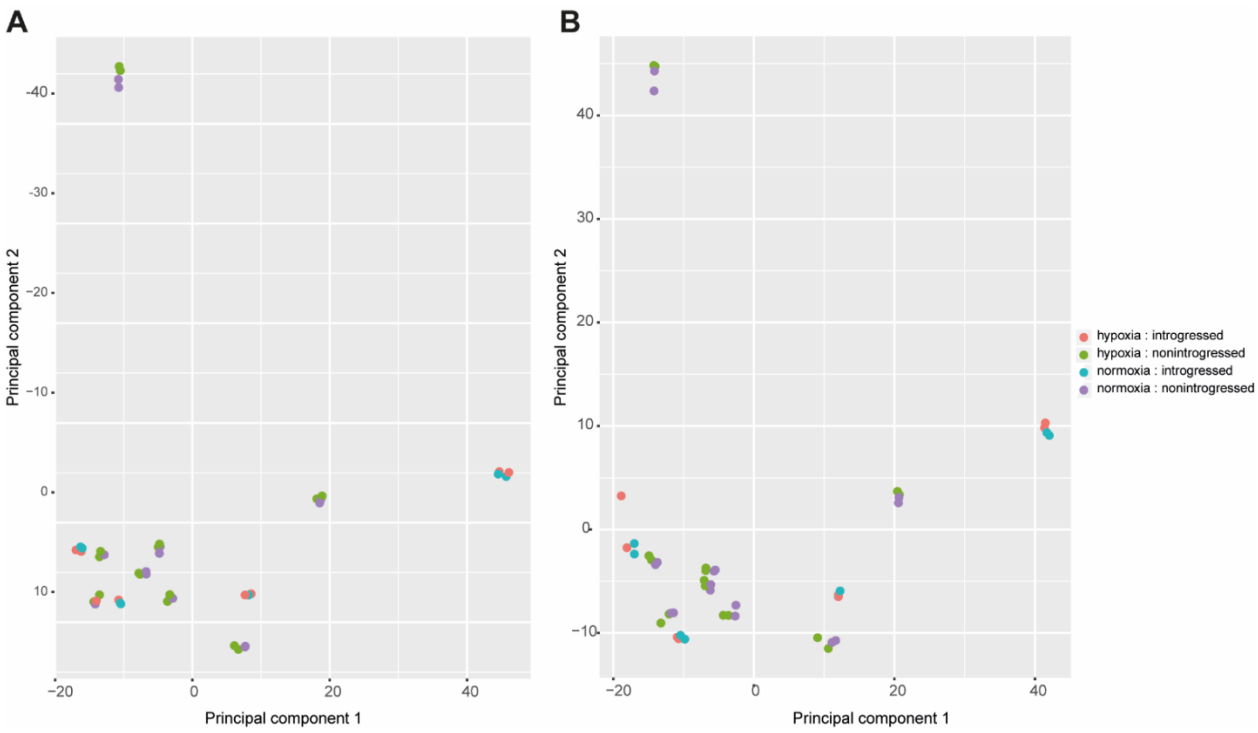


Figure 4. 17 PCA on whole-genome expression in LCL samples. A. 4 hour time point B. 24 hour time point.

To test the effectivness of the experiment, I explored differential expression in genes between hypoxia and normoxia at the 4 and 24 hours time points. Gene expression is known to be regulated by the oxygen levels in the enviroment: in particular, expression of hypoxic factors (HIFs) is boosted by hypoxia and, consequently, the expression of their target genes. Importantly, response to hypoxia depends also on the cell type and the heterogeneity of gene regulation responses reflects the different cellular requirements (321, 322). Therefore, I estimated differential expression (log2 fold change) from normalised read counts between normoxic and hypoxic LCL introgressed and non introgressed samples at 4 and 24

hours to see whether I could detect genes that are differentially expressed. Inclusion of A549 in preliminary analyses would bias differential expression results due to differences in basal gene expression between LCLs and A549. The normalised gene read counts for hypoxic conditions were higher than normoxia, supporting possible higher gene expression in hypoxia and the activation of set of genes (Figure 4.18).

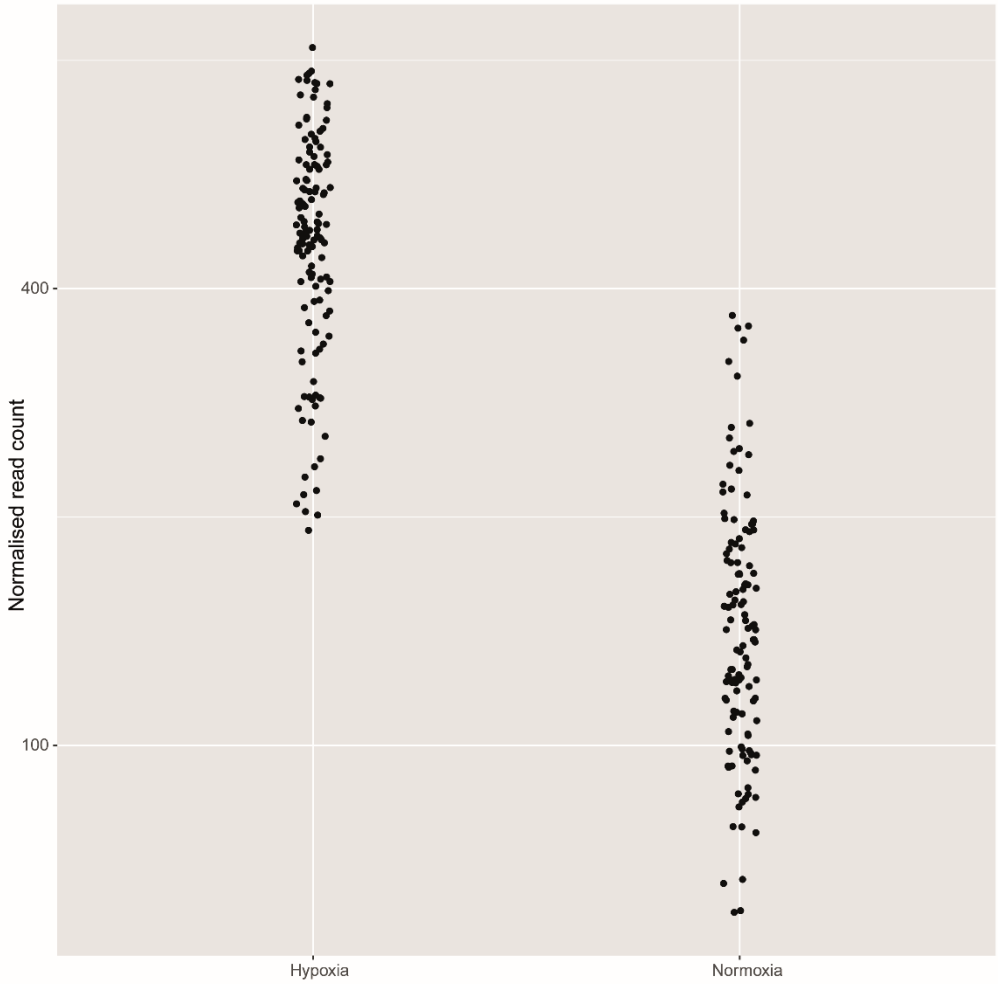


Figure 4. 18 Normalised read count between normoxia and hypoxia. The plot shows the gene read counts for all samples in hypoxia and normoxia. Hypoxic samples show higher read count and thus possible higher expression of a set of genes.

LCLs across the different time points do show differentially expressed genes between normoxia and hypoxia and between introgressed and non introgressed cell lines, confirming that the hypoxic condition in the cells was reached throughout the experiment (Figure 4.16). *EPAS1* does not show statistically significant differential

expression (pvalue adjusted < 0.1) between the samples. However, *EPAS1* is not highly expressed in LCLs compared to other genes and this could affect the detection of its expression using Deseq2 (323) (Figure 4.19 A and B). Different filtering and normalisation methods are needed to test reliable detection of differentially expressed genes.

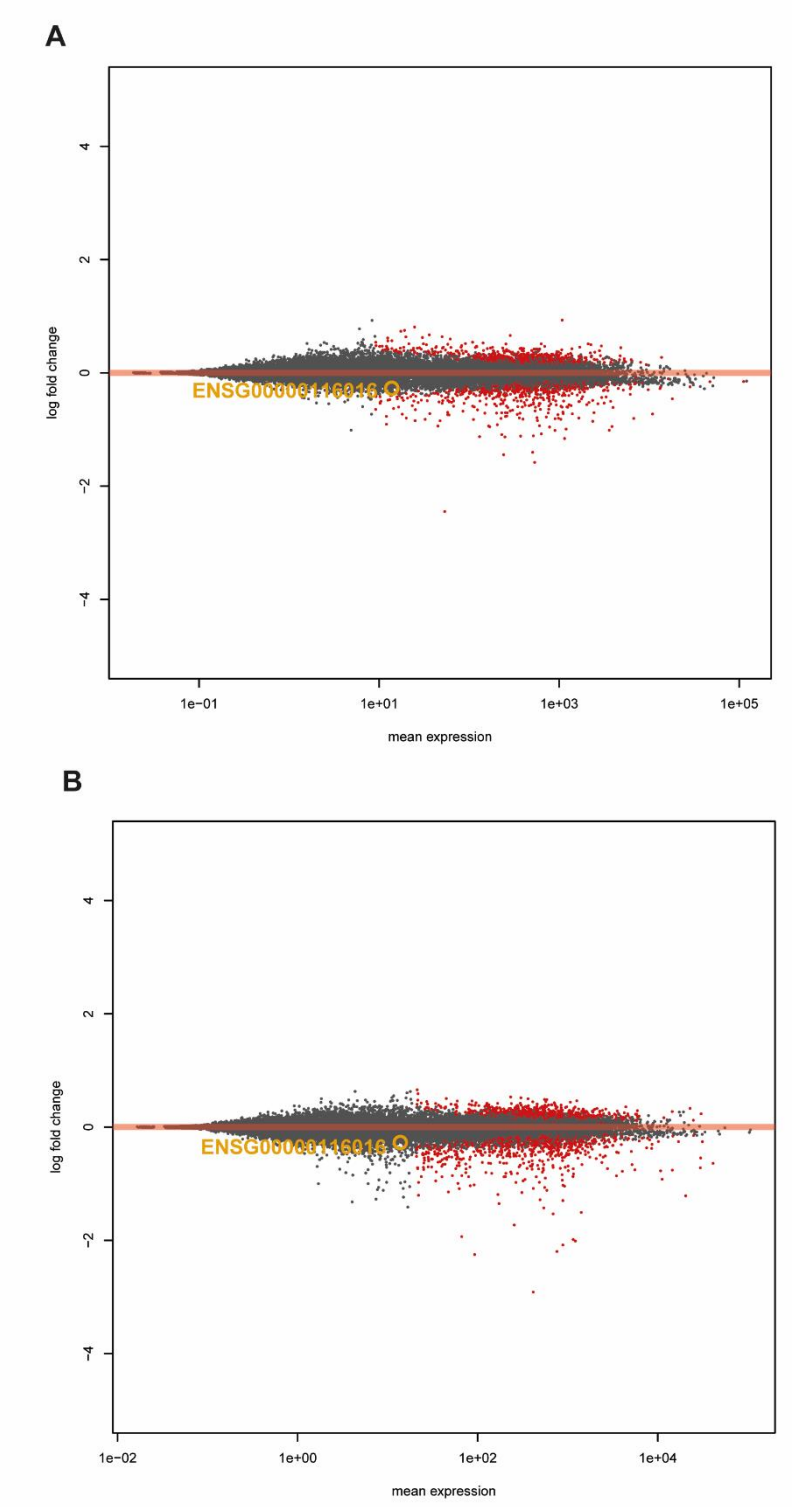


Figure 4. 19 Differential gene expression between normoxia and hypoxia and introgressed and non introgressed samples. The y-axis shows the log2 fold change between the different conditions, the x-axis the average of the read counts normalized by size. Each gene is represented with a dot. Genes with a pvalue adjusted < 0.1 for a significant difference are shown in red. In gold is indicated *EPAS1*. A. 4 hour time point B. 24 hour time point.

4.4 Discussion

In this study, *in silico* and *in vitro* work on understanding the underlying biological function of high altitude adaptation in Himalayans has been reported. Although the Denisovan introgressed region spans over ~300 kb and contains multiple genes and variants, the *EPAS1* ~32.7 kb region with the introgressed core, together with the adjacent SNP rs55981512, have previously been proposed as candidates for involvement in the high altitude phenotype (104, 119). This conclusion is supported by multiple lines of evidence presented in this study. Firstly, as seen in the previous chapters, the highest F_{ST} values between Tibetans and Han Chinese are reported for several variants within this region. Secondly, the genomic region around the *EPAS1* locus shows high evolutionary conservation of DNA sequence across different species with three ECRs (Figure 4.7) that span both intronic and exonic regions, suggesting an important role of *EPAS1* during development, hypoxic response and diseases (132, 142, 324, 325). Lastly, the importance of this locus is supported by the preliminary functional results showing differential expression of *EPAS1* between hypoxia and normoxia and between introgressed and non introgressed samples (Figures 4.12 and 4.13).

We do not know if the Denisovan population was adapted to live at high altitude and if the *EPAS1* haplotype was selected in Denisovans for this function. It could be also that this haplotype carried out a different function in Denisovan physiology, unrelated to high altitude adaptation. Thus, the biological functional consequences of *EPAS1* of the Denisovan haplotype are unknown. Initial attempts at characterizing the *EPAS1* molecular function showed its downregulation in Tibetans and has been tentatively interpreted as agreement with the Tibetan phenotype of low red blood cell count (209). However, the published study had many limitations and the necessity for new functional studies stimulated the development of the hypoxic protocol and experiments described in this chapter. The detection of expression of *EPAS1* in human LCLs makes them a usable model for hypoxic experiments. The comparison of *EPAS1* expression levels in LCLs together with a positive control lung cell line, A549, confirmed that *EPAS1* expression is induced at different time points in different cell lines and, thus, can play a role in response to hypoxia (Figures 4.10-4.13).

Analyses of publicly available RNA-seq data show that *EPAS1* expression is more localised to specific tissues than *HIF-1α*, which seems to have a more generalist function in the hypoxic response (141, 302) (Figure 4.2). In particular, expression of *EPAS1* shows very high levels in endothelial, cardiac and lung tissues (Figures 4.3and 4.4). The differential expression of *EPAS1* in LCLs with or without the introgressed haplotype supports the hypothesis that this gene is critical for the adaptive response to hypoxia (Figures 4.12 and 4.13). However, all cell lines with the introgressed haplotype used in this study show similar haplotypes for the entire ~32.7 kb core region, which makes it difficult to pinpoint the exact functional variant(s) (Figure 4.2). Other genes such as *EGLN1* have also been implicated in the high altitude adaptation of Tibetans. The qPCR results show differences in expression of *EGLN1* and *HIF-1α* at 4 hours of hypoxic exposure although this variability involve only two non introgressed LCLs (HG02153 and HG00881) and A549 (Figures 4.9 and4.10). *HIF-1α* expression at 4 hours can support the idea that this gene plays central role in the acute phase of hypoxic exposure (141, 306). Particularly, *EGLN1* differences in expression seem to not correlate with the presence of rs186996510 and rs12097901 suggesting that these variants do not have large effects on expression levels, consistent with their annotation as missense variants (Table 4.1, Figures 4.10 and 4.11).

RNA-seq data generated by this study will provide a better understanding of the molecular pathway of high altitude adaptation in the Himalayan populations. The preliminary QC and analyses of the data show that their quality is sufficiently good (Figure 4.13, 4.14 and 4.15). They also show that the hypoxic condition in cell culture was reached, and differential expression between hypoxia and normoxia is detectable (Figure 4.16 and 4.17). Further comprehensive analyses of the RNA-seq data will provide some insight of the possible underlying biological function which could hopefully explain some aspects of the mechanism of high altitude adaption of the introgressed haplotype in *EPAS1* gene.

5. General discussion

5.1 Summary

The geographical structure of the Himalayan region has played a strategic role in shaping the genetic, cultural and ethno-linguistic mosaic of South and East Asia. Isolation, genetic drift and natural selection have all potentially moulded the genetics of populations residing in this area. All these factors make the Himalayan populations exceptional candidates for understanding human evolutionary dynamics in a complex area. Although many genetic studies have been published on the Himalayan region, most of them have focused on specific populations such Tibetans and Sherpa, or the surrounding regions, or on the relationship between Tibetans and Han Chinese (50, 99, 117, 119).

In a previous collaborative study, we carried out the first genetic study of a broad number of Himalayan populations using autosomal STRs, and highlighted the high genetic complexity of the region (111). However, a comprehensive survey of the genetic variability and the broad population history of the Himalayan region has been missing. In this dissertation, I describe the work done on the genetic characterisation of 52 autochthonous groups from Nepal, Bhutan, North India, and the Tibetan Plateau in China, and have carried out a preliminary exploration of the biological mechanism underlying the genetic adaptation in these populations. Firstly, I elucidated their genetic structure and population demography using a combination of SNP-genotype data and high coverage whole-genome sequences. All the Himalayan populations share ancestry with their geographical neighbours in South and East Asia. Nevertheless, Himalayans share a common ancestral population that is abundant in most Himalayans, but rare elsewhere, followed by the development of local fine structure which is influenced by both language and geography. An exception, however, is that all of the high altitude populations from Tibet, Bhutan and Sherpa from Nepal showed a high genetic affinity with recent split times from one another during the last 5,000 years, regardless of their geographical and linguistic distribution. This conclusion is in contrast to a previous study which suggested a geographical cline within high altitude populations, with geographically closer populations having more genetic affinity (120) which could be due to the limited

number of populations included and lack of Bhutanese populations. The high genetic similarity of high altitude populations could be explained by archaeological evidences of permanent settlement at extreme high altitudes, above 3000 meters, only from ~3,600 years ago. The agricultural shift from a millet based economy to a barley one seems to have played a role in this transition to higher altitudes. Barley can bear better hypoxic conditions, has a longer growing season and it is more frost-hardy compared to millet (326, 327). This agricultural shift together with sheep husbandry may have allowed people to thrive at extreme high altitudes only in very recent times (327). This would support our results suggesting that high altitude adaptation originated in a single ancestral population and subsequently spread widely across the Himalayas, illustrated by the presence of a shared high altitude genetic component correlating with altitude reported here, and the *EPAS1* adaptive haplotype being widely distributed across the region in an altitude-dependent manner as described previously (127). The archaeological evidences would also explain our results of the very recent split times within the different high altitude populations.

It has been suggested that the Himalayas were a corridor for human migrations from the Tibetan plateau to South Asia in ancient times, or alternatively have represented a natural barrier to gene flow between South and East Asia and remained uninhabited until more recent times (105-107, 328). Archaeological evidences for intermittent human presence in the Tibetan Plateau date back from at least 40,000-30,000 years ago (110, 329). These mostly include human handprints and footprints, animal bones and stone artefacts (327). It is plausible that these remains are traces of early waves of people reaching those altitudes, possibly tracking game. However, it is unlikely these early people would have been the first permanent dwellers of the Himalayan region. The population split times between Himalayan and South and East Asian populations, as well as their genetic affinities, can shed light on the settling and the movements of people in the region. Himalayan populations separated first from South Asians between 20,000 and 10,000 years ago and subsequently from East Asians around 9,000 years ago. Himalayas also show different proportions of South and East Asian recent admixture among themselves. These time estimations suggest that the ancestors of present-day Himalayans probably arrived through waves of migration during the early Neolithic. Previous studies reported that the Tibetans have a high genetic affinity with East Asians and gene flow, in particular

to Han Chinese, with a population split time around 7,000-9,000 years ago (118, 119). Here, I have further shown that other high altitude populations within this region also have a high affinity to East Asians, with a more recent split from East Asian populations in comparison to South Asians. This is again supported by archaeological evidences that report the establishment of the first villages at elevated altitudes (~2,500 m) in the north-eastern Tibetan Plateau, along the Yellow river, from around 5,200 years ago, possibly due to population expansion from neighbouring areas. It is plausible that from these areas populations expanded further up in altitudes and colonised the Himalayan arc. As previously mentioned, permanent settlements at altitudes greater than 3,000 meters were probably allowed thanks to a shift in the agro-pastoral economical system in the last ~3,600 years (327). Although high altitude populations are genetically closer to East Asians, they also show different degrees of South Asian admixture, like other Himalayan populations probably due recent admixture events.

In contrast, some Nepalese populations (for example, Chetri and Tharu) showed a higher proportion of South Asian admixture with an earlier population split time from East Asians compared to the South Asian populations. This could suggest extensive admixture with other South Asians populations in recent times: This hypothesis is supported by the analyses of autosomal markers, Y chromosomes and mtDNA. The patterns of rare variant sharing (singletons and *f2* variant sharing) and admixture analyses such ALDER, highlighted that these gene flows occurred recently, between 2,000 and 200 years ago. The Y and mtDNA lineages in the Himalayans are commonly found in South and East Asia, and are widespread across the region. The Y chromosomes of the Himalayans, however, often form distinct Himalayan-specific clusters with very short branches, suggesting a rapid male expansion across the region around 1,000 years ago. This could suggest a second hypothesis of the colonisation of the Himalayan region: Populations in the South Asian side of the Himalayan foothills settled in the area in older times and then mixed with the people expanding from East Asia in more recent times. This could explain the later population split time from South Asian populations compared to the East Asians. It would also explain the rapid male expansion seen in the Y chromosome, comprising mostly East Asian haplogroups (O2 and D). Demographic model simulations and additional

research on these populations is needed to further understand the peopling of the Himalayan region.

Overall, the genetic evidence supports the hypothesis of the Himalayas being used as a corridor for human migration with bidirectional gene flow. The Himalayas could also possibly have been used by archaic populations as a corridor between South and East Asia. However, further research is needed to understand the area occupied by archaic humans such as Denisovans (330, 331).

The current work thus provides a comprehensive description of the genome-wide genetic variation and fine-scale population structure in the previously underrepresented Himalayan region. The data generated here fill a gap in South and East Asia genetic variability and can be used as a reference panel for the region. Furthermore, the study highlights how the complex geography and culture in the Himalayan region have shaped the evolutionary and population-genetic dynamics. Different populations show different degrees of isolation and genetic differentiation within the same region. For example, Lhokpu, Mönpa and Toto showed the highest genetic drift. Others such as Chetri and Tharu show high population growth and high heterozygosity which can be explained by admixture with South Asians. Socio-cultural structures such as language and the caste system might also have played an active role in shaping the population structure in present day Nepal.

Himalayan populations reside at a broad range of altitudes and in different environments, ranging from tropical forest to some of the world's highest mountain peaks, which make them ideal to study genetic adaptation. Several statistical frameworks were applied to both SNP-genotype data and whole-genome sequences to explore the genetic adaptation at both high and low altitudes. The most convincing example of positive selection for low altitude adaptation was the α-thalassaemia variant in the Tharu population that is reportedly protective against malaria (46, 223). High altitude adaptation in Himalayans has been of great interest in the past decade and many candidate genes have been associated with it in studies of Tibetans and Sherpa. The availability of populations living at a wide range of different altitudes gave additional power to validate known loci and identify novel candidates associated with high altitude adaptation. *EPAS1* selection, its relevance to high altitude adaptation, and introgression from Denisovans have previously been reported (99, 104). However, functional studies of *EPAS1* variants has not been systematically

carried out and it is still unknown which variants are key and what their underlying biological mechanism is. The complicated linkage disequilibrium pattern of the *EPAS1* region, together with the length of over 300 kb of the introgressed region, makes it hard to prioritise functional variants. The previous work and studies reported in this thesis have together highlighted 14 potential drivers (Figure 5.1, Table 5.1).

Variant	Genomic location (GRCh38)	Gene	Present in four introgressed cell lines
rs370299814	chr2: 46334261	*EPAS1*	YES*
rs115321619	**chr2: 46340777**	***EPAS1***	**YES**
rs73926263	**chr2: 46341541**	***EPAS1***	**YES**
rs73926264	**chr2:46341878**	***EPAS1***	**YES**
rs73926265	**chr2:46342631**	***EPAS1***	**YES**
rs55981512	**chr2: 46343203**	***EPAS1***	**YES**
rs150877473	**chr2:46360880**	***EPAS1***	**YES**
rs141366568	**chr2:46366983**	***EPAS1***	**YES***
rs1868092	chr2:46387063	*EPAS1*	YES**
rs12986653	chr2:46490951	*ATP6V1E2*	YES**
rs982414	chr2:46618437	*CRIPT*	NO
4bp deletion	**chr2:46350662-46350664**	***EPAS1***	**YES****
TED (3.4 kb deletion)	chr2:46467138-46470544	*TMEM247*	YES****

NOTE: * derived allele absent in HG02396, ** present in various 1000 Genomes Project samples, *** present in the four introgressed cell lines and HG02390, **** present in various 1000 Genomes Project samples but not in HG02396.

Table 5. 1 Potential drivers of selection in the *EPAS1* region. In bold are the variants falling within the highly differentiated ~32.7 kb region in *EPAS1*.

EPAS1 plays an important role not only in response to physiological hypoxia, but also in development and diseases like cancer. It is essential for correct heart development during embryonic stages (132, 324, 332). Understanding the genetics and physiology of high altitude adaptation is important not only for those interested in human adaptation and evolution, but can be also medically relevant for understanding mountain sickness and chronic obstructive pulmonary disease, and for cancer development mechanisms and angiogenesis. So understanding the molecular

mechanisms and validating variants of *EPAS1* for high altitude adaption is a key step and *in vitro* work on *EPAS1* differential expression between cell lines with and without the introgressed core Denisovan haplotype under hypoxic condition could potentially shed light on this. I developed an optimised protocol to carry out the hypoxic experiments in normal cell culture incubators. The protocol is user-friendly and can be used across different laboratories on any available cell lines. The steady and low expression of *EPAS1* in introgressed cell lines under hypoxic condition is in agreement with the physiological response and phenotype in Himalayans, which is characterized by low red blood cell count that is tightly regulated by erythropoietin (*EPO*). In light of these results, among the potential drivers of selection, I could exclude rs370299814 and rs141366568 because there was not a difference in *EPAS1* expression between cell lines with the introgressed haplotype carrying those alleles (HG00987, HG02187 and HG02397) and the one (HG02396) ancestral for them. Likewise, the TED deletion downstream of *EPAS1* has been detected in multiple individuals of the 1000 Genomes Project, yet is absent in HG02396. I could also exclude rs982414 because this variant was not present in the cell lines used for this study. Furthermore, because change in expression levels between introgressed and non-introgressed cell lines only involved the *EPAS1* gene and not the neighbouring genes (*TMEM247*, *ATP6V1E2*, *CRIPT*, *PIGF*), it is also likely that rs12986653 is not the responsible variant for the assessed phenotype. Finally, the 4bp deletion is carried by all four introgressed cell lines together with HG02390, a cell line that does not carry the derived state of the other variants. Further analyses of RNA-seq and qPCR profiles could help to understand if this deletion could be functional in HG02390 and the other introgressed cell lines. On the other hand, candidates supported for being potential drivers of the adaptive phenotype are the SNPs within the core Denisovan haplotype (rs115321619, rs73926263, rs73926264, rs73926265, rs55981512) and the nearby variant predicted to affect splicing (rs150877473). All four introgressed cell lines used in this project carried all these variants that lie in the ~32.7 kb which includes the most highly differentiated variants between Han and Tibetans ($F_{ST} > 0.85$).

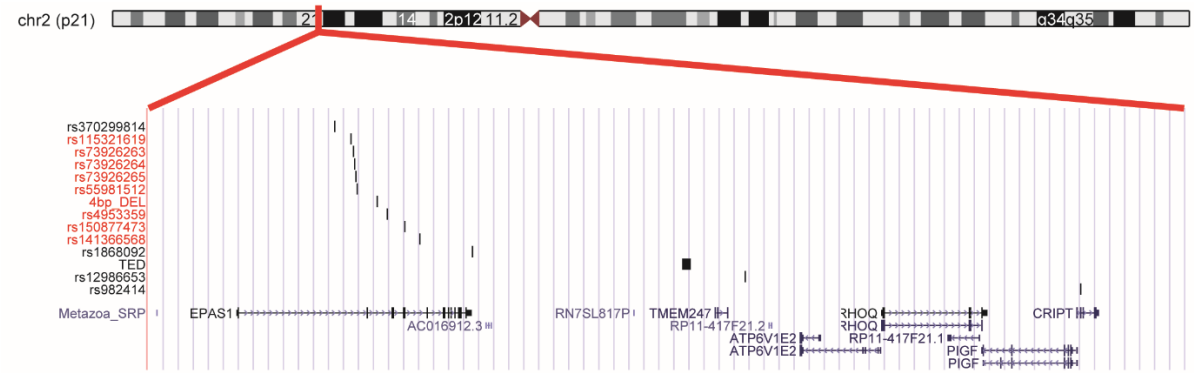


Figure 5. 1 Genomic location of potential drivers of selection in *EPAS1* region. In red are indicated variants falling the highly differentiated ~32.7 kb region in *EPAS1*.

Further studies to understand whether the *EPAS1* differential expression observed between the introgressed and non-introgressed LCLs involves a *cis* or *trans* regulation, and to further narrow down the functional variants are still required. Furthermore, exploring the potential interaction of *EPAS1* with the other candidate genes associated with high altitude adaptation would help to identify the molecular cascade activated in response to hypoxia in Himalayans.

In conclusion, the combination of SNP genotyping with whole-genome sequences is a powerful approach to study population demographic history and genetic adaptation; functional studies are essential to further understand the underlying biological mechanism of the genetic adaption to high altitude living in the Himalayas.

5.2 Next steps

In-depth analyses of the RNA-seq data generated for the introgressed and non-introgressed cell lines cultured under normoxic and hypoxic conditions will be the next steps. These include analysing differential expression of *EPAS1* and other genes involved in the hypoxic pathway and estimating the impact of specific variants in splicing and predicting possible eQTLs, as well as genome-wide changes of expression levels in both acute (4 hours) and chronic hypoxia (72 hours) in LCLs and the A549 cell line. These should allow the understanding of differential response in cell lines with and without the *EPAS1* core Denisovan introgressed haplotype and other variants. Any potentially interesting differentiated genes discovered from RNA-seq analyses can subsequently be validated using qPCR.

The population-genetic studies presented in this thesis have identified several other candidates for positive selection at high altitude. A similar approach of pinpointing likely functional variants in the sequence data and further exploring their function in cell lines can potentially be applied in a systematic way.

5.3 Future research

Himalayan populations have been poorly represented in genome-wide studies of human variation, and even fewer studies of various diseases in populations residing at high altitude have been published. The genetic and physiological changes that enabled adaption to the harsh environment at high altitude, which include increased ultraviolet radiation, hypobaria and hypoxemia, may have altered the susceptibility to diseases and innate defence mechanisms against microbial infection in these populations. It has been reported that people living in high altitude environments have a higher incidence of digestive system disease; the gut wall is an essential barrier for preventing bacterial infection and endotoxin entering human organs (333). Tibetans have evolved a mechanism of protection against intestinal injuries by downregulating *EGFR*, *GRB2*, and *PTPN11* genes, but have a higher prevalence of congenital heart disease in children (334). It has also been reported that young children and infants at high altitudes have higher rates of hospitalization for bronchiolitis and pneumonia due to respiratory syncytial virus (RSV) infection, and it is suggested that physiological adaptation to altitude leads to more severe consequences in RSV infections (335). These, together with other reports suggest that there are differences in predisposition to diseases and in the immune response to pathogens in the high altitude populations (336-339). Both genome-wide association and functional genomic studies are very much needed in these populations to both better understand the consequences of high altitude adaptation and improve health care for these populations. Some of the Himalayan populations have experienced severe bottlenecks and it is important to understand the resulting genetic burden of these events.

A lot more work is needed to further understand the detailed biology and function of *EPAS1* and the underlying mechanism of high altitude adaptation. Since the first reports of *EPAS1* being under strong positive selection in Tibetans, it has been widely accepted that this gene was associated with high altitude adaptation in this population and it was from an introgression event from the Denisovans into modern humans (99, 104, 145). Reduced oxygen levels and hypobaric hypoxia can considerably challenge the normal life and reproduction of different species and thus acting as selective pressure (340). However, *EPAS1* plays a plethora of roles in

different biological pathways, including immunity against pathogens and athletic performance that would have played a key role to escaping predators (341-343). Hypoxia inducible factors play an important role regulating both innate and adaptive host immune response (344). Oxygen deprivation can dramatically affect running and physical resistance performance and hypoxic factors play an important role in regulating hypoxic levels in tissues (345). Currently, we do not know which function *EPAS1* had in the Denisovan population and if the original purpose of the introgressed haplotype was to give advantage to live at extreme altitudes. Consequently, we could argue that the primary function of *EPAS1* was not related to high altitude adaptation. High altitude adaptation in humans is an example of convergent evolution in three main populations: Himalayans, Andeans and Ethiopians (131, 133, 346). This adaptation involves both a set of shared and unshared genes between populations: five genes (*EDN1*, *EDNRA*, *EGLN1, NOS1* and *VEGFA*) but not *EPAS1* are shared across the three continental areas suggesting their essentiality in high altitude adaptation (Figure 1.1). This could suggest that *EPAS1* represents a secondary candidate to high altitude adaptation.

Despite the original function of *EPAS1*, more work is needed to understand the role of the introgressed haplotype and *EPAS1* function in high altitude adaptation. First, the differential expression of *EPAS1* and other hypoxic factors at the protein level should be investigated in cell lines with and without the introgressed haplotype under hypoxic conditions. This will need an optimised protocol for protein extraction from samples exposed to hypoxia, as hypoxia inducible factors (HIFs) get quickly degraded when exposed to normoxia via the *VHL*-mediated ubiquination pathway (347). A potential way around this would be to use proteasome inhibitors such as MG132, which is an oxygen scavenger and stabilises the hypoxic condition (347), although it can also be toxic to the cells. If the in-depth analyses of the RNA-seq data fail to identify plausible functionally important variants in the 300 kb positively selected haplotype, an alternative approach would be to sequentially delete sections of the introgressed genomic region in suitable cell lines using CRISPR-Cas9 genome editing in order to potentially narrow down to the critical genomic area containing the functional variants. Subsequently, the variants within the candidate region could be further explored at the single point mutation level using the same approach.

Many population-genetic studies have identified candidate regions/variants of positive selection in different human populations, yet their relevant biological function and mechanism in improving reproductive fitness usually remains unknown or under-studied. Large-scale functional studies of all variants positively selected in any population will provide a comprehensive picture and full understanding of the molecular mechanisms and biological interaction network of human adaption and their potential health consequences.

References

1. Darwin C. The descent of man and selection in relation to sex. London: John Murray; 1871.
2. Huxley T. Evidence as to Man's Place in Nature (Williams & Norgate, London, 1863). Google Scholar.
3. Klein RG. Darwin and the recent African origin of modern humans. Proc Natl Acad Sci U S A. 2009;106(38):16007-16009.
4. Meyer M, Kircher M, Gansauge M-T, Li H, Racimo F, Mallick S, et al. A High-Coverage Genome Sequence from an Archaic Denisovan Individual. Science. 2012;338(6104):222-226.
5. Hublin J-J, Ben-Ncer A, Bailey SE, Freidline SE, Neubauer S, Skinner MM, et al. New fossils from Jebel Irhoud, Morocco and the pan-African origin of Homo sapiens. Nature. 2017;546:289.
6. McDougall I, Brown FH, Fleagle JG. Stratigraphic placement and age of modern humans from Kibish, Ethiopia. Nature. 2005;433:733.
7. White TD, Asfaw B, DeGusta D, Gilbert H, Richards GD, Suwa G, et al. Pleistocene Homo sapiens from Middle Awash, Ethiopia. Nature. 2003;423:742.
8. Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proceedings of the National Academy of Sciences of the United States of America. 2005;102(44):15942-15947.
9. Gonder MK, Mortensen HM, Reed FA, de Sousa A, Tishkoff SA. Whole-mtDNA genome sequence analysis of ancient African lineages. Molecular biology and evolution. 2007;24(3):757-768.
10. Cruciani F, Trombetta B, Massaia A, Destro-Bisol G, Sellitto D, Scozzari R. A Revised Root for the Human Y Chromosomal Phylogenetic Tree: The Origin of Patrilineal Diversity in Africa. Am J Hum Genet. 2011;88(6):814-818.
11. Mendez FL, Krahn T, Schrack B, Krahn AM, Veeramah KR, Woerner AE, et al. An African American paternal lineage adds an extremely ancient root to the human Y chromosome phylogenetic tree. Am J Hum Genet. 2013;92(3):454-459.
12. Henn BM, Gignoux CR, Jobin M, Granka JM, Macpherson JM, Kidd JM, et al. Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. Proceedings of the National Academy of Sciences. 2011;108(13):5154-5162.
13. Scerri EML, Thomas MG, Manica A, Gunz P, Stock JT, Stringer C, et al. Did Our Species Evolve in Subdivided Populations across Africa, and Why Does It Matter? Trends in ecology & evolution. 2018;33(8):582-594.
14. López S, van Dorp L, Hellenthal G. Human Dispersal Out of Africa: A Lasting Debate. Evolutionary Bioinformatics Online. 2015;11(Suppl 2):57-68.
15. Bae CJ, Douka K, Petraglia MD. On the origin of modern humans: Asian perspectives. Science. 2017;358(6368).
16. Mercier N, Valladas H, Bar-Yosef O, Vandermeersch B, Stringer C, Joron JL. Thermoluminescence Date for the Mousterian Burial Site of Es-Skhul, Mt. Carmel. Journal of Archaeological Science. 1993;20(2):169-174.
17. Schwarcz H, Grün R, Vandermeersch B, Bar-Yosef O, Valladas H, Tchernov E. ESR dates for the hominid burial site of Qafzeh in Israel. Journal of Human Evolution. 1988;17(8):733-737.

18. Hershkovitz I, Weber GW, Quam R, Duval M, Grün R, Kinsley L, et al. The earliest modern humans outside Africa. Science. 2018;359(6374):456-459.
19. Pagani L, Schiffels S, Gurdasani D, Danecek P, Scally A, Chen Y, et al. Tracing the route of modern humans out of Africa by using 225 human genome sequences from Ethiopians and Egyptians. Am J Hum Genet. 2015;96(6):986-991.
20. Lambeck K, Purcell A, Flemming NC, Vita-Finzi C, Alsharekh AM, Bailey GN. Sea level and shoreline reconstructions for the Red Sea: isostatic and tectonic considerations and implications for hominin migration out of Africa. Quaternary Science Reviews. 2011;30(25):3542-3574.
21. Oppenheimer S. The great arc of dispersal of modern humans: Africa to Australia. Quaternary International. 2009;202(1):2-13.
22. Macaulay V, Hill C, Achilli A, Rengo C, Clarke D, Meehan W, et al. Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. Science. 2005;308(5724):1034-1036.
23. Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature. 2016;538:201.
24. Liu W, Wu X, Pei S, Wu X, Norton CJ. Huanglong Cave: A Late Pleistocene human fossil site in Hubei Province, China. Quaternary International. 2010;211(1):29-41.
25. Liu W, Jin C-Z, Zhang Y-Q, Cai Y-J, Xing S, Wu X-J, et al. Human remains from Zhirendong, South China, and modern human emergence in East Asia. Proceedings of the National Academy of Sciences. 2010;107(45):19201-19206.
26. Bae CJ, Wang W, Zhao J, Huang S, Tian F, Shen G. Modern human teeth from Late Pleistocene Luna Cave (Guangxi, China). Quaternary International. 2014;354:169-183.
27. Westaway KE, Louys J, Awe RD, Morwood MJ, Price GJ, Zhao Jx, et al. An early modern human presence in Sumatra 73,000–63,000 years ago. Nature. 2017;548:322.
28. Clarkson C, Jacobs Z, Marwick B, Fullagar R, Wallis L, Smith M, et al. Human occupation of northern Australia by 65,000 years ago. Nature. 2017;547:306.
29. Pagani L, Lawson DJ, Jagoda E, Mörseburg A, Eriksson A, Mitt M, et al. Genomic analyses inform on migration events during the peopling of Eurasia. Nature. 2016;538:238.
30. Rogers AR, Bohlender RJ, Huff CD. Early history of Neanderthals and Denisovans. Proceedings of the National Academy of Sciences. 2017;114(37):9859-9863.
31. Reich D, Green RE, Kircher M, Krause J, Patterson N, Durand EY, et al. Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature. 2010;468:1053.
32. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, et al. A Draft Sequence of the Neandertal Genome. Science. 2010;328(5979):710-722.
33. Wall JD, Yang MA, Jay F, Kim SK, Durand EY, Stevison LS, et al. Higher Levels of Neanderthal Ancestry in East Asians than in Europeans. Genetics. 2013;194(1):199-209.
34. Reich D, Patterson N, Kircher M, Delfin F, Nandineni MR, Pugach I, et al. Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania. Am J Hum Genet. 2011;89(4):516-528.
35. Skoglund P, Jakobsson M. Archaic human ancestry in East Asia. Proc Natl Acad Sci U S A. 2011;108(45):18301-18306.
36. Browning SR, Browning BL, Zhou Y, Tucci S, Akey JM. Analysis of Human Sequence Data Reveals Two Pulses of Archaic Denisovan Admixture. Cell. 2018;173(1):53-61.e9.
37. Clamp M, Fry B, Kamal M, Xie X, Cuff J, Lin MF, et al. Distinguishing protein-coding and noncoding genes in the human genome. Proceedings of the National Academy of Sciences. 2007;104(49):19428-19433.

38. Ezkurdia I, Juan D, Rodriguez JM, Frankish A, Diekhans M, Harrow J, et al. Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes. Human molecular genetics. 2014;23(22):5866-5878.
39. Chi KR. The dark side of the human genome. Nature. 2016;538:275.
40. Graves JA. Sex chromosome specialization and degeneration in mammals. Cell. 2006;124(5):901-914.
41. Eyre-Walker A, Awadalla P. Does Human mtDNA Recombine? Journal of Molecular Evolution. 2001;53(4):430-435.
42. Jobling MA. Human evolutionary genetics. New York; London: Garland Science; 2014.
43. Zhang F. Copy Number Variation in Human Health, Disease, and Evolution. 2009;10:451-481.
44. Sudmant PH. An integrated map of structural variation in 2,504 human genomes. 2015;526(7571):75-81.
45. Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, et al. Diet and the evolution of human amylase gene copy number variation. Nat Genet. 2007;39(10):1256-1260.
46. Modiano G, Morpurgo G, Terrenato L, Novelletto A, Di Rienzo A, Colombo B, et al. Protection against malaria morbidity: near-fixation of the alpha-thalassemia gene in a Nepalese population. Am J Hum Genet. 1991;48(2):390-397.
47. Pollack JR, Sorlie T, Perou CM, Rees CA, Jeffrey SS, Lonning PE, et al. Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. Proc Natl Acad Sci U S A. 2002;99(20):12963-12968.
48. Lin M, Whitmire S, Chen J, Farrel A, Shi X, Guo J-t. Effects of short indels on protein structure and function in human genomes. Scientific reports. 2017;7(1):9313.
49. Ng TK, Liang XY, Lu F, Liu DT, Yam GH, Ma L, et al. Protective effects of an HTRA1 insertion-deletion variant against age-related macular degeneration in the Chinese populations. Laboratory investigation; a journal of technical methods and pathology. 2017;97(1):43-52.
50. Hu H, Petousi N, Glusman G, Yu Y, Bohlender R, Tashi T, et al. Evolutionary history of Tibetans inferred from whole-genome sequencing. PLoS Genet. 2017;13(4):e1006675.
51. Fryxell KJ, Moon W-J. CpG Mutation Rates in the Human Genome Are Highly Dependent on Local GC Content. Molecular biology and evolution. 2005;22(3):650-8.
52. MacArthur DG, Tyler-Smith C. Loss-of-function variants in the genomes of healthy humans. Human molecular genetics. 2010;19(R2):R125-R30.
53. MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, et al. A systematic survey of loss-of-function variants in human protein-coding genes. Science. 2012;335(6070):823-828.
54. Zerbino DR, Achuthan P, Akanni W, Amode M R, Barrell D, Bhai J, et al. Ensembl 2018. Nucleic acids research. 2018;46(D1):D754-D761.
55. Slatkin M. Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. Nat Rev Genet. 2008;9(6):477-485.
56. Lonjou C, Zhang W, Collins A, Tapper WJ, Elahi E, Maniatis N, et al. Linkage disequilibrium in human populations. Proceedings of the National Academy of Sciences. 2003;100(10):6069-6074.
57. Shifman S, Kuypers J, Kokoris M, Yakir B, Darvasi A. Linkage disequilibrium patterns of the human genome across populations. Human molecular genetics. 2003;12(7):771-776.

58. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, et al. The Structure of Haplotype Blocks in the Human Genome. Science. 2002;296(5576):2225-2229.
59. Hardy GH. MENDELIAN PROPORTIONS IN A MIXED POPULATION. Science (New York, NY). 1908;28(706):49-50.
60. Weinberg W. Über den Nachweis der Vererbung beim Menschen. Jahreshefte des Vereins für vaterländische Naturkunde in Württemberg. 1908;64:368–382.
61. Masel J. Genetic drift. Current Biology. 2011;21(20):R837-R838.
62. Ridley M. Evolution: Wiley; 2003.
63. Wright S. Evolution in Mendelian Populations. Genetics. 1931;16(2):97-159.
64. Wright S. Size of population and breeding structure in relation to evolution. Science. 1938;87:430-431.
65. Lynch M, Conery JS. The Origins of Genome Complexity. Science. 2003;302(5649):1401-1404.
66. Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68-87.
67. Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, et al. Worldwide human relationships inferred from genome-wide patterns of variation. Science. 2008;319(5866):1100-1104.
68. Fu Q, Meyer M, Gao X, Stenzel U, Burbano HA, Kelso J, et al. DNA analysis of an early modern human from Tianyuan Cave, China. Proc Natl Acad Sci U S A. 2013;110(6):2223-2227.
69. Pinhasi R, Fernandes D, Sirak K, Novak M, Connell S, Alpaslan-Roodenberg S, et al. Optimal Ancient DNA Yields from the Inner Ear Part of the Human Petrous Bone. PLOS ONE. 2015;10(6):e0129102.
70. Yang MA, Fu Q. Insights into Modern Human Prehistory Using Ancient Genomes. Trends in Genetics. 2018;34(3):184-196.
71. Reich D. Who We Are and How We Got Here: Pantheon Books; 2018.
72. Slatkin M, Racimo F. Ancient DNA and human history. Proceedings of the National Academy of Sciences. 2016;113(23):6380-6387.
73. Hubisz MJ, Falush D, Stephens M, Pritchard JK. Inferring weak population structure with the assistance of sample group information. Mol Ecol Resour. 2009;9(5):1322-1332.
74. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Research. 2009;19(9):1655-1664.
75. Jolliffe IT, Cadima J. Principal component analysis: a review and recent developments. Philosophical transactions Series A, Mathematical, physical, and engineering sciences. 2016;374(2065):20150202.
76. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, et al. Genes mirror geography within Europe. Nature. 2008;456(7218):98-101.
77. Pickrell JK, Pritchard JK. Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. Plos Genetics. 2012;8(11).
78. Lawson DJ, Hellenthal G, Myers S, Falush D. Inference of Population Structure using Dense Haplotype Data. Plos Genetics. 2012;8(1).
79. Brisbin A, Bryc K, Byrnes J, Zakharia F, Omberg L, Degenhardt J, et al. PCAdmix: Principal Components-Based Assignment of Ancestry along Each Chromosome in Individuals with Admixed Ancestry from Two or More Populations. Hum Biol. 2012;84(4):343-364.

80. Maples BK, Gravel S, Kenny EE, Bustamante CD. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. Am J Hum Genet. 2013;93(2):278-288.
81. Price AL, Tandon A, Patterson N, Barnes KC, Rafaels N, Ruczinski I, et al. Sensitive Detection of Chromosomal Segments of Distinct Ancestry in Admixed Populations. PLOS Genetics. 2009;5(6):e1000519.
82. Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, et al. Ancient admixture in human history. Genetics. 2012;192(3):1065-1093.
83. Li H, Durbin R. Inference of human population history from individual whole-genome sequences. Nature. 2011;475(7357):493-496.
84. Schiffels S, Durbin R. Inferring human population size and separation history from multiple genome sequences. Nat Genet. 2014;46(8):919-925.
85. Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, et al. Positive Natural Selection in the Human Lineage. Science. 2006;312(5780):1614-1620.
86. Kimura M. Evolutionary Rate at the Molecular Level. Nature. 1968;217:624.
87. Hellmann I, Ebersberger I, Ptak SE, Paabo S, Przeworski M. A neutral explanation for the correlation of diversity with recombination rates in humans. Am J Hum Genet. 2003;72(6):1527-1535.
88. Smith JM, Haigh J. The hitch-hiking effect of a favourable gene. Genetical Research. 1974;23(1):23-35.
89. Booker TR, Jackson BC, Keightley PD. Detecting positive selection in the genome. BMC Biology. 2017;15(1):98.
90. Hermisson J, Pennings PS. Soft Sweeps: Molecular Population Genetics of Adaptation From Standing Genetic Variation. Genetics. 2005;169(4):2335-2352.
91. Pritchard JK, Pickrell JK, Coop G. The Genetics of Human Adaptation: Hard Sweeps, Soft Sweeps, and Polygenic Adaptation. Current biology : CB. 2010;20(4):R208-R215.
92. Jain K, Stephan W. Modes of Rapid Polygenic Adaptation. Molecular biology and evolution. 2017;34(12):3169-175.
93. Berg JJ, Coop G. A Population Genetic Signal of Polygenic Adaptation. PLOS Genetics. 2014;10(8):e1004412.
94. Hedrick PW. Balancing selection. Current Biology. 2007;17(7):R230-R231.
95. Evans SN, Shvets Y, Slatkin M. Non-equilibrium theory of the allele frequency spectrum. Theoretical Population Biology. 2007;71(1):109-119.
96. Wright S. The Distribution of Gene Frequencies Under Irreversible Mutation. Proceedings of the National Academy of Sciences of the United States of America. 1938;24(7):253-259.
97. Pavlidis P, Alachiotis N. A survey of methods and tools to detect recent and strong positive selection. Journal of Biological Research-Thessaloniki. 2017;24(1):7.
98. Colonna V, Ayub Q, Chen Y, Pagani L, Luisi P, Pybus M, et al. Human genomic regions with exceptionally high levels of population differentiation identified from 911 whole-genome sequences. Genome Biology. 2014;15(6):R88.
99. Yi X, Liang Y, Huerta-Sanchez E, Jin X, Cuo ZX, Pool JE, et al. Sequencing of 50 human exomes reveals adaptation to high altitude. Science. 2010;329(5987):75-78.
100. Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, et al. Genome-wide detection and characterization of positive selection in human populations. Nature. 2007;449.
101. Hamblin MT, Di Rienzo A. Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. Am J Hum Genet. 2000;66(5):1669-1679.

102. Tishkoff SA, Varkonyi R, Cahinhinan N, Abbes S, Argyropoulos G, Destro-Bisol G, et al. Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. Science. 2001;293(5529):455-462.
103. Beall CM. Andean, Tibetan, and Ethiopian patterns of adaptation to high-altitude hypoxia. Integrative and Comparative Biology. 2006;46(1):18-24.
104. Huerta-Sanchez E, Jin X, Asan, Bianba Z, Peter BM, Vinckenbosch N, et al. Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. Nature. 2014;512(7513):194-197.
105. Gayden T, Mirabal S, Cadenas AM, Lacau H, Simms TM, Morlote D, et al. Genetic insights into the origins of Tibeto-Burman populations in the Himalayas. Journal of Human Genetics. 2009;54(4):216-223.
106. Majumder PP. Genomic inferences on peopling of south Asia. Current Opinion in Genetics & Development. 2008;18(3):280-284.
107. Gayden T, Perez A, Persad PJ, Bukhari A, Chennakrishnaiah S, Simms T, et al. The Himalayas: Barrier and conduit for gene flow. American Journal of Physical Anthropology. 2013;151(2):169-182.
108. Qi X, Cui C, Peng Y, Zhang X, Yang Z, Zhong H, et al. Genetic evidence of paleolithic colonization and neolithic expansion of modern humans on the tibetan plateau. Molecular biology and evolution. 2013;30(8):1761-1778.
109. Meyer MC, Aldenderfer MS, Wang Z, Hoffmann DL, Dahl JA, Degering D, et al. Permanent human occupation of the central Tibetan Plateau in the early Holocene. Science. 2017;355(6320):64-67.
110. Aldenderfer M. Peopling the Tibetan plateau: insights from archaeology. High Alt Med Biol. 2011;12(2):141-147.
111. Kraaijenbrink T, van der Gaag KJ, Zuniga SB, Xue Y, Carvalho-Silva DR, Tyler-Smith C, et al. A Linguistically Informed Autosomal STR Survey of Human Populations Residing in the Greater Himalayan Region. Plos One. 2014;9(3):e91534.
112. van Driem G. Languages of the Himalayas: an ethnolinguistic handbook of the greater Himalayan region containing an introduction to the symbiotic theory of language. Leiden: Brill; 2001. 2 vols. (1375 S.) p.
113. Jeong C, Alkorta-Aranburu G, Basnyat B, Neupane M, Witonsky DB, Pritchard JK, et al. Admixture facilitates genetic adaptations to high altitude in Tibet. Nat Commun. 2014;5:3281.
114. Cai X, Qin Z, Wen B, Xu S, Wang Y, Lu Y, et al. Human migration through bottlenecks from Southeast Asia into East Asia during Last Glacial Maximum revealed by Y chromosomes. PloS ONE. 2011;6(8):e24282.
115. Kang L, Lu Y, Wang C, Hu K, Chen F, Liu K, et al. Y-chromosome O3 haplogroup diversity in Sino-Tibetan populations reveals two migration routes into the Eastern Himalayas. Ann Hum Genet. 2012;76(1):92-99.
116. Cole AM, Cox S, Jeong C, Petousi N, Aryal DR, Droma Y, et al. Genetic structure in the Sherpa and neighboring Nepalese populations. BMC Genomics. 2017;18(1):102.
117. Bhandari S, Zhang X, Cui C, Bianba, Liao S, Peng Y, et al. Genetic evidence of a recent Tibetan ancestry to Sherpas in the Himalayan region. Scientific reports. 2015;5:16249.
118. Zhang C, Lu Y, Feng Q, Wang X, Lou H, Liu J, et al. Differentiated demographic histories and local adaptations between Sherpas and Tibetans. Genome Biology. 2017;18(1):115.
119. Lu D, Lou H, Yuan K, Wang X, Wang Y, Zhang C, et al. Ancestral Origins and Genetic History of Tibetan Highlanders. Am J Hum Genet. 2016;99(3):580-594.

120. Jeong C, Peter BM, Basnyat B, Neupane M, Beall CM, Childs G, et al. A longitudinal cline characterizes the genetic structure of human populations in the Tibetan plateau. PLoS One. 2017;12(4):e0175885.
121. Jeong C, Ozga AT, Witonsky DB, Malmström H, Edlund H, Hofman CA, et al. Long-term genetic stability and a high-altitude East Asian origin for the peoples of the high valleys of the Himalayan arc. Proceedings of the National Academy of Sciences. 2016;113(27):7485-7490.
122. Wang HW, Li YC, Sun F, Zhao M, Mitra B, Chaudhuri TK, et al. Revisiting the role of the Himalayas in peopling Nepal: insights from mitochondrial genomes. J Hum Genet. 2012;57.
123. Dhimal M, Ahrens B, Kuch U. Malaria control in Nepal 1963-2012: challenges on the path towards elimination. Malaria Journal. 2014;13.
124. Sah OP, Subedi S, Morita K, Inone S, Kurane I, Pandey BD. Serological study of dengue virus infection in Terai region, Nepal. Nepal Medical College journal : NMCJ. 2009;11(2):104-106.
125. Parajuli RP, Umezaki M, Watanabe C. Behavioral and nutritional factors and geohelminth infection among two ethnic groups in the Terai region, Nepal. American journal of human biology : the official journal of the Human Biology Council. 2009;21(1):98-104.
126. Lorenzo FR, Huff C, Myllymaki M, Olenchock B, Swierczek S, Tashi T, et al. A genetic mechanism for Tibetan high-altitude adaptation. Nat Genet. 2014;46(9):951-956.
127. Hackinger S, Kraaijenbrink T, Xue Y, Mezzavilla M, Asan, van Driem G, et al. Wide distribution and altitude correlation of an archaic high-altitude-adaptive EPAS1 haplotype in the Himalayas. Hum Genet. 2016;135(4):393-402.
128. Penaloza D, Arias-Stella J. The heart and pulmonary circulation at high altitudes: healthy highlanders and chronic mountain sickness. Circulation. 2007;115(9):1132-1146.
129. Paralikar SJ. High altitude pulmonary edema-clinical features, pathophysiology, prevention and treatment. Indian Journal of Occupational and Environmental Medicine. 2012;16(2):59-62.
130. Beall CM. Two routes to functional adaptation: Tibetan and Andean high-altitude natives. Proceedings of the National Academy of Sciences of the United States of America. 2007;104:8655-8660.
131. Bigham AW, Brutsaert T, Julian C, Moore LG, Parra EJ, Shriver MD, et al. Natural selection at high altitude: Andean and Tibetan patterns of adaptation to an extreme environment. American Journal of Human Biology. 2012;24(2):219-220.
132. Bigham AW, Lee FS. Human high-altitude adaptation: forward genetics meets the HIF pathway. Genes & development. 2014;28(20):2189-2204.
133. Bigham AW. Genetics of human origin and evolution: high-altitude adaptations. Current Opinion in Genetics & Development. 2016;41:8-13.
134. Moore LG, Niermeyer S, Zamudio S. Human adaptation to high altitude: regional and life-cycle perspectives. Am J Phys Anthropol. 1998;Suppl 27:25-64.
135. Moore LG. Human Genetic Adaptation to High Altitude. High Altitude Medicine & Biology. 2001;2(2):257-279.
136. Eichstaedt CA, Pagani L, Antao T, Inchley CE, Cardona A, Mörseburg A, et al. Evidence of Early-Stage Selection on EPAS1 and GPR126 Genes in Andean High Altitude Populations. Scientific reports. 2017;7(1):13042.
137. Huerta-Sánchez E, DeGiorgio M, Pagani L, Tarekegn A, Ekong R, Antao T, et al. Genetic Signatures Reveal High-Altitude Adaptation in a Set of Ethiopian Populations. Molecular biology and evolution. 2013;30(8):1877-1888.

138. Yang D, Peng Y, Ouzhuluobu, Bianbazhuoma, Cui C, Bianba, et al. HMOX2 Functions as a Modifier Gene for High-Altitude Adaptation in Tibetans. Hum Mutat. 2016;37(2):216-223.
139. Kumar H, Choi D-K. Hypoxia Inducible Factor Pathway and Physiological Adaptation: A Cell Survival Pathway? Mediators of Inflammation. 2015;2015:584758.
140. Ratcliffe PJ. Oxygen sensing and hypoxia signalling pathways in animals: the implications of physiology for cancer. The Journal of physiology. 2013;591(8):2027-2042.
141. Dengler VL, Galbraith M, Espinosa JM. Transcriptional Regulation by Hypoxia Inducible Factors. Critical reviews in biochemistry and molecular biology. 2014;49(1):1-15.
142. Schödel J, Oikonomopoulos S, Ragoussis J, Pugh CW, Ratcliffe PJ, Mole DR. High-resolution genome-wide mapping of HIF-binding sites by ChIP-seq. Blood. 2011;117(23):e207-e217.
143. Takeda K, Ho VC, Takeda H, Duan LJ, Nagy A, Fong GH. Placental but not heart defects are associated with elevated hypoxia-inducible factor alpha levels in mice lacking prolyl hydroxylase domain protein 2. Molecular and cellular biology. 2006;26(22):8336-8346.
144. Minamishima YA, Moslehi J, Padera RF, Bronson RT, Liao R, Kaelin WG, Jr. A feedback loop involving the Phd3 prolyl hydroxylase tunes the mammalian hypoxic response in vivo. Molecular and cellular biology. 2009;29(21):5729-5741.
145. Beall CM, Cavalleri GL, Deng L, Elston RC, Gao Y, Knight J, et al. Natural selection on EPAS1 (HIF2α) associated with low hemoglobin concentration in Tibetan highlanders. Proceedings of the National Academy of Sciences. 2010;107(25):11459-11464.
146. Bodmer W. Genetic characterization of human populations: from ABO to a genetic map of the British people. Genetics. 2015;199(2):267-279.
147. LaFramboise T. Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. Nucleic acids research. 2009;37(13):4181-4193.
148. Pugach I, Stoneking M. Genome-wide insights into the genetic history of human populations. Investigative genetics. 2015;6:6.
149. Lachance J, Tishkoff SA. SNP ascertainment bias in population genetic analyses: why it is important, and how to correct it. BioEssays : news and reviews in molecular, cellular and developmental biology. 2013;35(9):780-786.
150. Lopez Herraez D, Bauchet M, Tang K, Theunert C, Pugach I, Li J, et al. Genetic variation and recent positive selection in worldwide human populations: evidence from nearly 1 million SNPs. PLoS One. 2009;4(11):e7888.
151. Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, Piouffre L, et al. A human genome diversity cell line panel. Science. 2002;296(5566):261-262.
152. The International HapMap Project. Nature. 2003;426(6968):789-796.
153. Fu Q, Li H, Moorjani P, Jay F, Slepchenko SM, Bondarev AA, et al. Genome sequence of a 45,000-year-old modern human from western Siberia. Nature. 2014;514(7523):445-449.
154. Arciero E, Kraaijenbrink T, Asan, Haber M, Mezzavilla M, Ayub Q, et al. Demographic history and genetic adaptation in the Himalayan region inferred from genome-wide SNP genotypes of 49 populations. Mol Biol Evol 35(8): 1916-1933.
155. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. American Journal of Human Genetics. 2007;81(3):559-575.

156. Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT. Data quality control in genetic case-control association studies. Nature protocols. 2010;5(9):1564-1573.
157. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. Plos Genetics. 2006;2(12):2074-2093.
158. Teo YY, Fry AE, Clark TG, Tai ES, Seielstad M. On the usage of HWE for identifying genotyping errors. Ann Hum Genet. 2007;71(Pt 5):701-703.
159. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics. 2006;38(8):904-909.
160. Metspalu M, Romero IG, Yunusbayev B, Chaubey G, Mallick CB, Hudjashov G, et al. Shared and Unique Components of Human Population Structure and Genome-Wide Signals of Positive Selection in South Asia. American Journal of Human Genetics. 2011;89(6):731-744.
161. Chaubey G, Metspalu M, Choi Y, Maegi R, Romero IG, Soares P, et al. Population Genetic Structure in Indian Austroasiatic Speakers: The Role of Landscape Barriers and Sex-Specific Admixture. Molecular biology and evolution. 2011;28(2):1013-1024.
162. Korneliussen TS, Albrechtsen A, Nielsen R. ANGSD: Analysis of Next Generation Sequencing Data. BMC Bioinformatics. 2014;15:356.
163. Mezzavilla M, Ghirotto S. Neon: An R Package to Estimate Human Effective Population Size and Divergence Time from Patterns of Linkage Disequilibrium between SNPS. Journal of Computer Science & Systems Biology. 2015;8(1):037-044.
164. Hayes BJ, Visscher PM, McPartlan HC, Goddard ME. Novel multilocus measure of linkage disequilibrium to estimate past effective population size. Genome Res. 2003;13(4):635-643.
165. de Roos AP, Hayes BJ, Spelman RJ, Goddard ME. Linkage disequilibrium and persistence of phase in Holstein-Friesian, Jersey and Angus cattle. Genetics. 2008;179(3):1503-1512.
166. McEvoy BP, Powell JE, Goddard ME, Visscher PM. Human population dispersal "Out of Africa" estimated from linkage disequilibrium and allele frequencies of SNPs. Genome Res. 2011;21(6):821-829.
167. Tassi F, Ghirotto S, Mezzavilla M, Vilaca ST, De Santi L, Barbujani G. Early modern human dispersal from Africa: genomic evidence for multiple waves of migration. Investigative genetics. 2015;6:13.
168. Fenner JN. Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. American Journal of Physical Anthropology. 2005;128(2):415-423.
169. Howrigan DP, Simonson MA, Keller MC. Detecting autozygosity through runs of homozygosity: A comparison of three autozygosity detection algorithms. BMC Genomics. 2011;12(1):460.
170. Dray S, Dufour A-B. The ade4 Package: Implementing the Duality Diagram for Ecologists. 2007. 2007;22(4):20.
171. Loh P-R, Lipson M, Patterson N, Moorjani P, Pickrell JK, Reich D, et al. Inferring Admixture Histories of Human Populations Using Linkage Disequilibrium. Genetics. 2013;193(4):1233-1254.
172. Reich D, Thangaraj K, Patterson N, Price AL, Singh L. Reconstructing Indian population history. Nature. 2009;461(7263):489-U50.
173. Raghavan M, Skoglund P, Graf KE, Metspalu M, Albrechtsen A, Moltke I, et al. Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. Nature. 2014;505(7481):87-91.

174. Allentoft ME, Sikora M, Sjogren K-G, Rasmussen S, Rasmussen M, Stenderup J, et al. Population genomics of Bronze Age Eurasia. Nature. 2015;522(7555):167-172.
175. Haak W, Lazaridis I, Patterson N, Rohland N, Mallick S, Llamas B, et al. Massive migration from the steppe was a source for Indo-European languages in Europe. Nature. 2015;522(7555):207-211.
176. Olalde I, Allentoft ME, Sanchez-Quinto F, Santpere G, Chiang CWK, DeGiorgio M, et al. Derived immune and ancestral pigmentation alleles in a 7,000-year-old Mesolithic European. Nature. 2014;507(7491):225-228.
177. Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68-74.
178. Reynolds J, Weir BS, Cockerham CC. Estimation of the coancestry coefficient: basis for a short-term genetic distance. Genetics. 1983;105(3):767-779.
179. Fumagalli M, Moltke I, Grarup N, Racimo F, Bjerregaard P, Jorgensen ME, et al. Greenlandic Inuit show genetic signatures of diet and climate adaptation. Science. 2015;349(6254):1343-1347.
180. Ayub Q, Mezzavilla M, Pagani L, Haber M, Mohyuddin A, Khaliq S, et al. The Kalash genetic isolate: ancient divergence, drift, and selection. Am J Hum Genet. 2015;96(5):775-783.
181. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics (Oxford, England). 2005;21(2):263-5.
182. Fisher RA. Statistical methods for research workers: Oliver and Boyd, Edinburgh; 1950. xv + 354 pp. p.
183. Coop G, Witonsky D, Di Rienzo A, Pritchard JK. Using environmental correlations to identify loci underlying local adaptation. Genetics. 2010;185(4):1411-1423.
184. Gunther T, Coop G. Robust identification of local adaptation from allele frequencies. Genetics. 2013;195(1):205-220.
185. Lischer HE, Excoffier L. PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. Bioinformatics (Oxford, England). 2012;28(2):298-299.
186. Kass RE, Raftery AE. Bayes Factors. Journal of the American Statistical Association. 1995;90(430):773-795.
187. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protocols. 2015;10(6):845-858.
188. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic acids research. 2015;43(Database issue):D447-D452.
189. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol. 2016;17(1):122.
190. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014;46(3):310-315.
191. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013;45(6):580-5.
192. Newman D, Pilson D. Increased probability of extinction due to decreased genetic effective population size: Experimental populations of Clarkia pulchella. Evolution. 1997;51(2):354-362.
193. Driem Gv. Languages of the Himalayas an ethnolinguistic handbook of the greater Himalayan region containing an introduction to the symbiotic theory of language. Leiden: Brill; 2001. 2 vols. (1375 S.) p.

194. Moorjani P, Thangaraj K, Patterson N, Lipson M, Loh PR, Govindaraj P, et al. Genetic Evidence for Recent Population Mixture in India. Am J Hum Genet. 2013;93(3):422-438.
195. Kang HM, Sul JH, Service SK, Zaitlen NA, Kong S-y, Freimer NB, et al. Variance component model to account for sample structure in genome-wide association studies. Nature Genetics. 2010;42(4):348-U110.
196. Basang Z, Wang B, Li L, Yang L, Liu L, Cui C, et al. HIF2A Variants Were Associated with Different Levels of High-Altitude Hypoxia among Native Tibetans. Plos One. 2015;10(9).
197. Petousi N, Croft QPP, Cavalleri GL, Cheng H-Y, Formenti F, Ishida K, et al. Tibetans living at sea level have a hyporesponsive hypoxia-inducible factor system and blunted physiological responses to hypoxia. Journal of Applied Physiology. 2014;116(7):893-904.
198. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The Genotype-Tissue Expression (GTEx) project. Nature Genetics. 2013;45(6):580-585.
199. Simonson TS, McClain DA, Jorde LB, Prchal JT. Genetic determinants of Tibetan high-altitude adaptation. Human genetics. 2012;131.
200. Zaka R, Dion AS, Kusnierz A, Bohensky J, Srinivas V, Freeman T, et al. Oxygen tension regulates the expression of ANK (progressive ankylosis) in an HIF-1-dependent manner in growth plate chondrocytes. Journal of bone and mineral research : the official journal of the American Society for Bone and Mineral Research. 2009;24(11):1869-1878.
201. Yang J, Jin ZB, Chen J, Huang XF, Li XM, Liang YB, et al. Genetic signatures of high-altitude adaptation in Tibetans. Proc Natl Acad Sci U S A. 2017;114(16):4189-4194.
202. Li K, Gesang L, Dan Z, Gusang L. Genome-Wide Transcriptional Analysis Reveals the Protection against Hypoxia-Induced Oxidative Injury in the Intestine of Tibetans via the Inhibition of GRB2/EGFR/PTPN11 Pathways. Oxidative medicine and cellular longevity. 2016;2016:6967396.
203. Yuan Z, Chen J, Chen D, Xu G, Xia M, Xu Y, et al. Megakaryocytic leukemia 1 (MKL1) regulates hypoxia induced pulmonary hypertension in rats. PLoS One. 2014;9(3):e83895.
204. Sankararaman S, Mallick S, Patterson N, Reich D. The Combined Landscape of Denisovan and Neanderthal Ancestry in Present-Day Humans. Current biology : CB. 2016;26(9):1241-1247.
205. Iancu CV, Zamoon J, Woo SB, Aleshin A, Choe JY. Crystal structure of a glucose/H+ symporter and its mechanism of action. Proc Natl Acad Sci U S A. 2013;110(44):17862-17867.
206. India, Census C, Mitra A, India, Superintendent of Census Operations WB. [West Bengal District census handbook]. [Delhi]: [Manager of Publications]; 1952.
207. Peng Y, Yang Z, Zhang H, Cui C, Qi X, Luo X, et al. Genetic variations in Tibetan populations and high-altitude adaptation at the Himalayas. Molecular biology and evolution. 2011;28(2):1075-1081.
208. Xu S, Li S, Yang Y, Tan J, Lou H, Jin W, et al. A genome-wide search for signals of high-altitude adaptation in Tibetans. Molecular biology and evolution. 2011;28(2):1003-1011.
209. Peng Y, Cui C, He Y, Ouzhuluobu, Zhang H, Yang D, et al. Down-Regulation of EPAS1 Transcription and Genetic Adaptation of Tibetans to High-Altitude Hypoxia. Molecular biology and evolution. 2017;34(4):818-830.
210. Foll M, Gaggiotti OE, Daub JT, Vatsiou A, Excoffier L. Widespread signals of convergent adaptation to high altitude in Asia and america. Am J Hum Genet. 2014;95(4):394-407.

211. Xiang K, Ouzhuluobu, Peng Y, Yang Z, Zhang X, Cui C, et al. Identification of a Tibetan-specific mutation in the hypoxic gene EGLN1 and its contribution to high-altitude adaptation. Molecular biology and evolution. 2013;30(8):1889-1898.
212. Skubutyte R, Markova D, Freeman TA, Anderson DG, Dion AS, Williams CJ, et al. HIF Regulation of ANK Expression in Nucleus Pulposus Cells: Possible Implications in Controlling Dystrophic Mineralization in the Intervertebral Disc. Arthritis and rheumatism. 2010;62(9):2707-2715.
213. Dick CF, Dos-Santos AL, Meyer-Fernandes JR. Inorganic phosphate as an important regulator of phosphatases. Enzyme research. 2011;2011:103980.
214. Tajima R, Kawaguchi N, Horino Y, Takahashi Y, Toriyama K, Inou K, et al. Hypoxic enhancement of type IV collagen secretion accelerates adipose conversion of 3T3-L1 fibroblasts. Biochimica et biophysica acta. 2001;1540(3):179-187.
215. Sudhakar A, Nyberg P, Keshamouni VG, Mannam AP, Li J, Sugimoto H, et al. Human alpha1 type IV collagen NC1 domain exhibits distinct antiangiogenic activity mediated by alpha1beta1 integrin. The Journal of clinical investigation. 2005;115(10):2801-2810.
216. Cen B, Selvaraj A, Burgess RC, Hitzler JK, Ma Z, Morris SW, et al. Megakaryoblastic leukemia 1, a potent transcriptional coactivator for serum response factor (SRF), is required for serum induction of SRF target genes. Molecular and cellular biology. 2003;23(18):6597-608.
217. Ghosal A, Sabui S, Said HM. Identification and characterization of the minimal 5'-regulatory region of the human riboflavin transporter-3 (SLC52A3) in intestinal epithelial cells. American journal of physiology Cell physiology. 2015;308(2):C189-196.
218. Wang YP, Wei JY, Yang JJ, Gao WN, Wu JQ, Guo CJ. Riboflavin supplementation improves energy metabolism in mice exposed to acute hypoxia. Physiological research. 2014;63(3):341-350.
219. Blanck HM, Bowman BA, Serdula MK, Khan LK, Kohn W, Woodruff BA. Angular stomatitis and riboflavin status among adolescent Bhutanese refugees living in southeastern Nepal. The American journal of clinical nutrition. 2002;76(2):430-435.
220. Ai H, Huang L, Ren J. Genetic diversity, linkage disequilibrium and selection signatures in Chinese and Western pigs revealed by genome-wide SNP markers. PLoS One. 2013;8.
221. Ai H, Yang B, Li J, Xie X, Chen H, Ren J. Population history and genomic signatures for high-altitude adaptation in Tibetan pigs. BMC Genomics. 2014;15(1):834.
222. Skov L, Hui R, Hobolth A, Scally A, Schierup MH, Durbin R. Detecting archaic introgression without archaic reference genomes. bioRxiv. 2018.
223. Terrenato L, Shrestha S, Dixit KA, Luzzatto L, Modiano G, Morpurgo G, et al. Decreased malaria morbidity in the Tharu people compared to sympatric populations in Nepal. Annals of tropical medicine and parasitology. 1988;82(1):1-11.
224. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics (Oxford, England). 2009;25(14):1754-1760.
225. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9):1297-1303.
226. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43(5):491-498.
227. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Current protocols in bioinformatics. 2013;43:11.0.1-33.

228. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics (Oxford, England). 2011;27(21):2987-2993.
229. Zheng-Bradley X, Streeter I, Fairley S, Richardson D, Clarke L, Flicek P, et al. Alignment of 1000 Genomes Project reads to reference assembly GRCh38. GigaScience. 2017;6(7):1-8.
230. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics (Oxford, England). 2011;27(15):2156-2158.
231. Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, et al. The complete genome sequence of a Neandertal from the Altai Mountains. Nature. 2014;505(7481):43-49.
232. Poznik GD, Xue Y, Mendez FL, Willems TF, Massaia A, Wilson Sayres MA, et al. Punctuated bursts in human male demography inferred from 1,244 worldwide Y-chromosome sequences. Nature Genetics. 2016;48:593.
233. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics (Oxford, England). 2009;25(16):2078-2079.
234. Weissensteiner H, Pacher D, Kloss-Brandstatter A, Forer L, Specht G, Bandelt HJ, et al. HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. Nucleic acids research. 2016;44(W1):W58-63.
235. Handsaker RE, Van Doren V, Berman JR, Genovese G, Kashin S, Boettger LM, et al. Large multiallelic copy number variations in humans. Nat Genet. 2015;47(3):296-303.
236. Carson AR, Smith EN, Matsui H, Brækkan SK, Jepsen K, Hansen JB, et al. Effective filtering strategies to improve data quality from population-based whole exome sequencing studies. BMC Bioinformatics. 2014;15:125.
237. Garrison E. Vcflib, a simple C++ library for parsing and manipulating VCF files.https://github.com/vcflib/vcflib.
238. Handsaker RE, Korn JM, Nemesh J, McCarroll SA. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. Nature Genetics. 2011;43:269.
239. The Genomes Project C, Durbin RM, Altshuler D, Durbin RM, Abecasis GR, Bentley DR, et al. A map of human genome variation from population-scale sequencing. Nature. 2010;467:1061.
240. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics (Oxford, England). 2010;26(6):841-842.
241. Sudmant PH, Mallick S, Nelson BJ, Hormozdiari F, Krumm N, Huddleston J, et al. Global diversity, population stratification, and selection of human copy number variation. Science. 2015.
242. Loh PR, Palamara PF, Price AL. Fast and accurate long-range phasing in a UK Biobank cohort. Nat Genet. 2016;48(7):811-816.
243. Loh PR, Danecek P, Palamara PF, Fuchsberger C, Y AR, H KF, et al. Reference-based phasing using the Haplotype Reference Consortium panel. Nat Genet. 2016;48(11):1443-1448.
244. Szpak M, Mezzavilla M, Ayub Q, Chen Y, Xue Y, Tyler-Smith C. FineMAV: prioritizing candidate genetic variants driving local adaptations in human populations. Genome Biology. 2018;19(1):5.
245. Browning Brian L, Browning Sharon R. Detecting Identity by Descent and Estimating Genotype Error Rates in Sequence Data. The American Journal of Human Genetics. 2013;93(5):840-851.

246. Browning SR, Browning BL. Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. The American Journal of Human Genetics. 2007;81(5):1084-1097.
247. Terhorst J, Kamm JA, Song YS. Robust and scalable inference of population history from hundreds of unphased whole-genomes. Nat Genet. 2017;49(2):303-309.
248. Poznik GD. Identifying Y-chromosome haplogroups in arbitrarily large samples of sequenced or genotyped men. bioRxiv. 2016.
249. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics (Oxford, England). 2014;30(9):1312-1313.
250. Bergström A, Nagle N, Chen Y, McCarthy S, Pollard Martin O, Ayub Q, et al. Deep Roots for Aboriginal Australian Y Chromosomes. Current Biology. 2016;26(6):809-813.
251. Prüfer K, de Filippo C, Grote S, Mafessoni F, Korlević P, Hajdinjak M, et al. A high-coverage Neandertal genome from Vindija Cave in Croatia. Science. 2017.
252. Neph S, Kuehn MS, Reynolds AP, Haugen E, Thurman RE, Johnson AK, et al. BEDOPS: high-performance genomic feature operations. Bioinformatics (Oxford, England). 2012;28(14):1919-1920.
253. Weir BS, Cockerham CC. Estimating F-Statistics for the Analysis of Population Structure. Evolution. 1984;38(6):1358-1370.
254. Kulakovskiy IV, Vorontsov IE, Yevshin IS, Sharipov RN, Fedorova AD, Rumynskiy EI, et al. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. Nucleic acids research. 2018;46(Database issue):D252-259.
255. Bailey TL, Johnson J, Grant CE, Noble WS. The MEME Suite. Nucleic acids research. 2015;43(W1):W39-W49.
256. Ambrosini G, Groux R, Bucher P. PWMScan: a fast tool for scanning entire genomes with a position-specific weight matrix. Bioinformatics (Oxford, England). 2018;34(14):2483-2484.
257. Sloan CA, Chan ET, Davidson JM, Malladi VS, Strattan JS, Hitz BC, et al. ENCODE data at the ENCODE portal. Nucleic acids research. 2016;44(Database issue):D726-732.
258. Lou H, Lu Y, Lu D, Fu R, Wang X, Feng Q, et al. A 3.4-kb Copy-Number Deletion near EPAS1 Is Significantly Enriched in High-Altitude Tibetans but Absent from the Denisovan Sequence. American Journal of Human Genetics. 2015;97(1):54-66.
259. Karlsson EK, Harris JB, Tabrizi S, Rahman A, Shlyakhter I, Patterson N, et al. Natural Selection in a Bangladeshi Population from the Cholera-Endemic Ganges River Delta. Science translational medicine. 2013;5(192):192ra86.
260. Mondal M, Casals F, Xu T, Dall'Olio GM, Pybus M, Netea MG, et al. Genomic analysis of Andamanese provides insights into ancient human migration into Asia and adaptation. Nature Genetics. 2016;48:1066.
261. Wang J, Raskin L, Samuels DC, Shyr Y, Guo Y. Genome measures used for quality control are dependent on gene function and ancestry. Bioinformatics (Oxford, England). 2015;31(3):318-323.
262. Samuels DC, Wang J, Ye F, He J, Levinson RT, Sheng Q, et al. Heterozygosity Ratio, a Robust Global Genomic Measure of Autozygosity and Its Association with Height and Disease Risk. Genetics. 2016;204(3):893-904.
263. Ceballos FC, Joshi PK, Clark DW, Ramsay M, Wilson JF. Runs of homozygosity: windows into population history and trait architecture. Nature Reviews Genetics. 2018;19:220.

264. Gusev A, Palamara PF, Aponte G, Zhuang Z, Darvasi A, Gregersen P, et al. The Architecture of Long-Range Haplotypes Shared within and across Populations. Molecular biology and evolution. 2012;29(2):473-486.
265. Yao X, Tang S, Bian B, Wu X, Chen G, Wang C-C. Improved phylogenetic resolution for Y-chromosome Haplogroup O2a1c-002611. Scientific reports. 2017;7(1):1146.
266. Shi H, Dong Y, Wen B, Xiao C, Underhill P A, Shen P, et al. Y-Chromosome Evidence of Southern Origin of the East Asian–Specific Haplogroup O3-M122. Am J Hum Genet. 2005;77(3):408-419.
267. Wang CC, Li H. Inferring human history in East Asia from Y chromosomes. Investigative genetics. 2013;4:11.
268. Sharma S, Rai E, Sharma P, Jena M, Singh S, Darvishi K, et al. The Indian origin of paternal haplogroup R1a1* substantiates the autochthonous origin of Brahmins and the caste system. J Hum Genet. 2009;54(1):47-55.
269. Derenko M. Origin and Post-Glacial Dispersal of Mitochondrial DNA Haplogroups C and D in Northern Asia. 2010;5(12).
270. Metspalu M, Kivisild T, Metspalu E, Parik J, Hudjashov G, Kaldma K, et al. Most of the extant mtDNA boundaries in South and Southwest Asia were likely shaped during the initial settlement of Eurasia by anatomically modern humans. BMC Genetics. 2004;5(1):26.
271. Li Q, Lin K, Sun H, Liu S, Huang K, Huang X, et al. Mitochondrial haplogroup M9a1a1c1b is associated with hypoxic adaptation in the Tibetans. Journal Of Human Genetics. 2016;61:1021.
272. Wolf AB, Akey JM. Outstanding questions in the study of archaic hominin admixture. PLOS Genetics. 2018;14(5):e1007349.
273. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genetics in medicine : official journal of the American College of Medical Genetics. 2015;17(5):405-424.
274. Ng D, Johnston JJ, Teer JK, Singh LN, Peller LC, Wynter JS, et al. Interpreting secondary cardiac disease variants in an exome cohort. Circ Cardiovasc Genet. 2013;6(4):337-346.
275. Duzkale H, Shen J, McLaughlin H, Alfares A, Kelly MA, Pugh TJ, et al. A systematic approach to assessing the clinical significance of genetic variants. Clin Genet. 2013;84(5):453-463.
276. Vinkemeier U, Obermann W, Weber K, Furst DO. The globular head domain of titin extends into the center of the sarcomeric M band. cDNA cloning, epitope mapping and immunoelectron microscopy of two titin-associated proteins. Journal of cell science. 1993;106 (Pt 1):319-330.
277. Shin C, Feng Y, Manley JL. Dephosphorylated SRp38 acts as a splicing repressor in response to heat shock. Nature. 2004;427(6974):553-558.
278. Ma M, Li Z, Wang DW, Wei X. Next-generation sequencing identifies novel mutations in the FBN1 gene for two Chinese families with Marfan syndrome. Mol Med Rep. 2016;14(1):151-158.
279. Rosendahl J, Bödeker H, Mössner J, Teich N. Hereditary chronic pancreatitis. Orphanet Journal of Rare Diseases. 2007;2:1.
280. Mathieson I, McVean G. Demography and the Age of Rare Variants. PLOS Genetics. 2014;10(8):e1004528.

281. Ruyssen-Witrand A, Degboé Y, Cantagrel A, Nigon D, Lukas C, Scaramuzzino S, et al. Association between RANK, RANKL and OPG polymorphisms with ACPA and erosions in rheumatoid arthritis: results from a meta-analysis involving three French cohorts. RMD Open. 2016;2(2).
282. Jeru I, Cochet E, Duquesnoy P, Hentgen V, Copin B, Mitjavila-Garcia MT, et al. Brief Report: Involvement of TNFRSF11A molecular defects in autoinflammatory disorders. Arthritis & rheumatology (Hoboken, NJ). 2014;66(9):2621-2627.
283. Boei JJWA, Vermeulen S, Klein B, Hiemstra PS, Verhoosel RM, Jennen DGJ, et al. Xenobiotic metabolism in differentiated human bronchial epithelial cells. Archives of Toxicology. 2017;91(5):2093-105.
284. Dannemann M, Prüfer K, Kelso J. Functional implications of Neandertal introgression in modern humans. Genome Biology. 2017;18(1):61.
285. Deschamps M, Laval G, Fagny M, Itan Y, Abel L, Casanova J-L, et al. Genomic Signatures of Selective Pressures and Introgression from Archaic Hominins at Human Innate Immunity Genes. The American Journal of Human Genetics. 2016;98(1):5-21.
286. Racimo F, Sankararaman S, Nielsen R, Huerta-Sánchez E. Evidence for archaic adaptive introgression in humans. Nature reviews Genetics. 2015;16(6):359-371.
287. Wali VB, Haskins JW, Gilmore-Hebert M, Platt JT, Liu Z, Stern DF. Convergent and Divergent Cellular Responses by ErbB4 Isoforms in Mammary Epithelial Cells. Molecular cancer research : MCR. 2014;12(8):1140-1155.
288. Paatero I, Jokilammi A, Heikkinen PT, Iljin K, Kallioniemi OP, Jones FE, et al. Interaction with ErbB4 promotes hypoxia-inducible factor-1alpha signaling. The Journal of biological chemistry. 2012;287(13):9659-9671.
289. Taniguchi M, Kawabata M. KIR3DL1/S1 genotypes and KIR2DS4 allelic variants in the AB KIR genotypes are associated with Plasmodium-positive individuals in malaria infection. Immunogenetics. 2009;61(11-12):717-730.
290. Hussain T, Mulherkar R. Lymphoblastoid Cell lines: a Continuous in Vitro Source of Cells to Study Carcinogen Sensitivity and DNA Repair. International Journal of Molecular and Cellular Medicine. 2012;1(2):75-87.
291. Gazdar AF, Girard L, Lockwood WW, Lam WL, Minna JD. Lung Cancer Cell Lines as Tools for Biomedical Discovery and Research. JNCI Journal of the National Cancer Institute. 2010;102(17):1310-1321.
292. Delaneau O, Marchini J, Zagury J-F. A linear complexity phasing method for thousands of genomes. Nature Methods. 2012;9(2):179-181.
293. Templeton AR, Crandall KA, Sing CF. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation. Genetics. 1992;132(2):619-633.
294. Paradis E. pegas: an R package for population genetics with an integrated-modular approach. Bioinformatics (Oxford, England). 2010;26(3):419-420.
295. The EPC, Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57.
296. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics (Oxford, England). 2013;29(1):15-21.
297. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature protocols. 2012;7(3):562-578.
298. Anders S, Huber W. Differential expression analysis for sequence count data. Genome biol. 2010;11(10):R106.
299. Loots GG, Ovcharenko I. Dcode.org anthology of comparative genomic tools. Nucleic acids research. 2005;33(suppl_2):W56-W64.

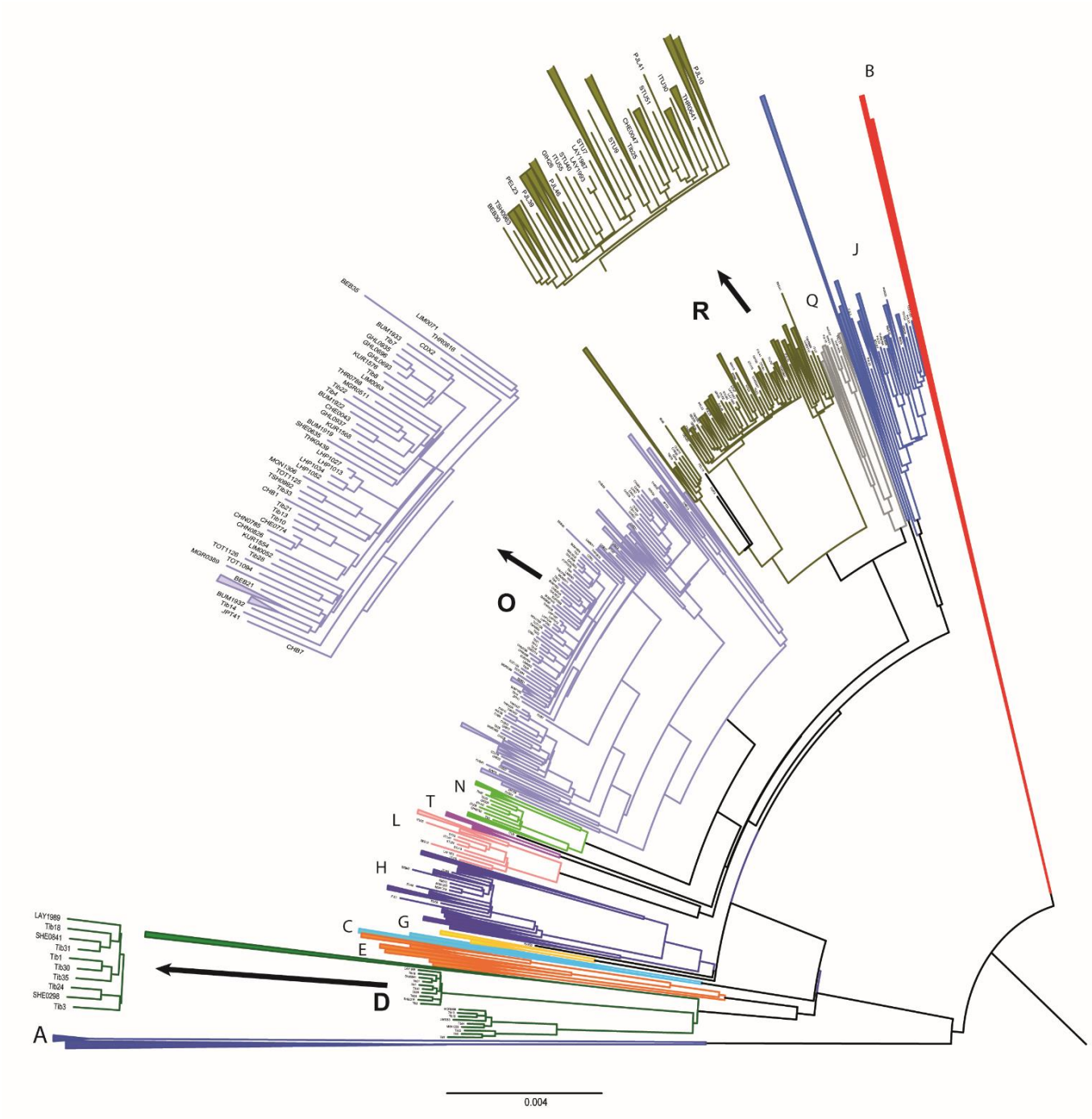
300. O'Connor C. Karyotyping for Chromosomal Abnormalities. Nature Education. 2008;1(1):27.
301. An International System for Human Cytogenetics Nomenclature: Karger; 2013.
302. Lin Q, Cong X, Yun Z. Differential hypoxic regulation of hypoxia-inducible factors 1α and 2α. Molecular cancer research : MCR. 2011;9(6):757-765.
303. Zhou W, Dosey TL, Biechele T, Moon RT, Horwitz MS, Ruohola-Baker H. Assessment of Hypoxia Inducible Factor Levels in Cancer Cell Lines upon Hypoxic Induction Using a Novel Reporter Construct. PLOS ONE. 2011;6(11):e27460.
304. Shimoda LA, Semenza GL. HIF and the Lung: Role of Hypoxia-inducible Factors in Pulmonary Development and Disease. American Journal of Respiratory and Critical Care Medicine. 2011;183(2):152-156.
305. Carithers LJ, Moore HM. The Genotype-Tissue Expression (GTEx) Project. Biopreservation and Biobanking. 2015;13(5):307-308.
306. Uchida T, Rossignol F, Matthay MA, Mounier R, Couette S, Clottes E, et al. Prolonged hypoxia differentially regulates hypoxia-inducible factor (HIF)-1alpha and HIF-2alpha expression in lung epithelial cells: implication of natural antisense HIF-1alpha. The Journal of biological chemistry. 2004;279(15):14871-1488.
307. Schmittgen TD, Livak KJ. Analyzing real-time PCR data by the comparative CT method. Nat Protocols. 2008;3(6):1101-1108.
308. Nolan T, Hands RE, Bustin SA. Quantification of mRNA using real-time RT-PCR. Nature protocols. 2006;1:1559.
309. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics. 2009;10:57.
310. Eisenberg E, Levanon EY. Human housekeeping genes are compact. Trends in Genetics. 2003;19(7):362-365.
311. Zhu J, He F, Hu S, Yu J. On the nature of human housekeeping genes. Trends in genetics : TIG. 2008;24(10):481-484.
312. Ye J, Coulouris G, Zaretskaya I, Cutcutache I, Rozen S, Madden TL. Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. BMC Bioinformatics. 2012;13:134.
313. Thornton B, Basu C. Real-time PCR (qPCR) primer design using free online software. Biochemistry and Molecular Biology Education. 2011;39(2):145-54.
314. Bustin S, Huggett J. qPCR primer design revisited. Biomolecular Detection and Quantification. 2017;14:19-28.
315. Bustin SA, Benes V, Garson JA, Hellemans J, Huggett J, Kubista M, et al. The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. Clinical chemistry. 2009;55(4):611-622.
316. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: The reference human genome annotation for The ENCODE Project. Genome Research. 2012;22(9):1760-1774.
317. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics (Oxford, England). 2016;32(19):3047-3048.
318. Delaneau O, Ongen H, Brown AA, Fort A, Panousis NI, Dermitzakis ET. A complete tool set for molecular QTL discovery and analysis. Nature Communications. 2017;8:15452.
319. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics (Oxford, England). 2014;30(7):923-930.

320. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology. 2014;15(12):550.
321. Chi J-T, Wang Z, Nuyten DSA, Rodriguez EH, Schaner ME, Salim A, et al. Gene Expression Programs in Response to Hypoxia: Cell Type Specificity and Prognostic Significance in Human Cancers. PLOS Medicine. 2006;3(3):e47.
322. Shaw K. Environmental cues like hypoxia can trigger gene expression and cancer Nature Education. 2008;1(1):198.
323. Sha Y, Phan JH, Wang MD. Effect of low-expression gene filtering on detection of differentially expressed genes in RNA-seq data. Conference proceedings : Annual International Conference of the IEEE Engineering in Medicine and Biology Society IEEE Engineering in Medicine and Biology Society Annual Conference. 2015;2015:6461-6464.
324. Tian H, Hammer RE, Matsumoto AM, Russell DW, McKnight SL. The hypoxia-responsive transcription factor EPAS1 is essential for catecholamine homeostasis and protection against heart failure during embryonic development. Genes & development. 1998;12(21):3320-3324.
325. Patel SA, Simon MC. Biology of Hypoxia-Inducible Factor-2α in Development and Disease. Cell death and differentiation. 2008;15(4):628-634.
326. Páldi E, Szalai G, Janda T, Horváth E, Rácz I, Lásztity D. Determination of Frost Tolerance in Winter Wheat and Barley at the Seedling Stage. Biologia Plantarum. 2001;44(1):145-147.
327. Chen FH, Dong GH, Zhang DJ, Liu XY, Jia X, An CB, et al. Agriculture facilitated permanent human occupation of the Tibetan Plateau after 3600 B.P. Science. 2015;347(6219):248-250.
328. Gayden T, Cadenas AM, Regueiro M, Singh NB, Zhivotovsky LA, Underhill PA, et al. The Himalayas as a directional barrier to gene flow. Am J Hum Genet. 2007;80(5):884-894.
329. Zhang XL, Ha BB, Wang SJ, Chen ZJ, Ge JY, Long H, et al. The earliest human occupation of the high-altitude Tibetan Plateau 40 thousand to 30 thousand years ago. Science. 2018;362(6418):1049-1051.
330. Kaifu Y. Archaic Hominin Populations in Asia before the Arrival of Modern Humans: Their Phylogeny and Implications for the “Southern Denisovans”. Current Anthropology. 2017;58(S17):S418-S33.
331. Ko KH. Hominin interbreeding and the evolution of human variation. Journal of Biological Research. 2016;23:17.
332. Moreno Roig E, Yaromina A, Houben R, Groot AJ, Dubois L, Vooijs M. Prognostic Role of Hypoxia-Inducible Factor-2α Tumor Cell Expression in Cancer Patients: A Meta-Analysis. Frontiers in Oncology. 2018;8:224.
333. Anand AC, Sashindran VK, Mohan L. Gastrointestinal problems at high altitude. Tropical gastroenterology : official journal of the Digestive Diseases Foundation. 2006;27(4):147-153.
334. Ma LG, Chen QH, Wang YY, Wang J, Ren ZP, Cao ZF, et al. Spatial pattern and variations in the prevalence of congenital heart disease in children aged 4-18years in the Qinghai-Tibetan Plateau. The Science of the total environment. 2018;627:158-165.
335. Choudhuri JA, Ogden LG, Ruttenber AJ, Thomas DSK, Todd JK, Simoes EAF. Effect of altitude on hospitalizations for respiratory syncytial virus infection. Pediatrics. 2006;117(2):349-356.
336. Ericsson CD, Steffen R, Basnyat B, Cumbo TA, Edelman R. Infections at High Altitude. Clinical Infectious Diseases. 2001;33(11):1887-1891.

337. Eisen S, Pealing L, Aldridge RW, Siedner MJ, Necochea A, Leybell I, et al. Effects of Ascent to High Altitude on Human Antimycobacterial Immunity. PLOS ONE. 2013;8(9):e74220.
338. Lopez-Pascual A, Bes-Rastrollo M, Sayón-Orea C, Perez-Cornago A, Díaz-Gutiérrez J, Pons JJ, et al. Living at a Geographically Higher Elevation Is Associated with Lower Risk of Metabolic Syndrome: Prospective Analysis of the SUN Cohort. Frontiers in Physiology. 2017;7(658).
339. Burtscher M. Effects of Living at Higher Altitudes on Mortality: A Narrative Review. Aging and Disease. 2014;5(4):274-280.
340. Petousi N, Robbins PA. Human adaptation to the hypoxia of high altitude: the Tibetan paradigm from the pregenomic to the postgenomic era. Journal of applied physiology (Bethesda, Md : 1985). 2014;116(7):875-884.
341. Doedens AL, Phan AT, Stradner MH, Fujimoto JK, Nguyen JV, Yang E, et al. Hypoxia-inducible factors enhance the effector responses of CD8(+) T cells to persistent antigen. Nature immunology. 2013;14(11):1173-1182.
342. Voisin S, Cieszczyk P, Pushkarev VP, Dyatlov DA, Vashlyayev BF, Shumaylov VA, et al. EPAS1 gene variants are associated with sprint/power athletic performance in two cohorts of European athletes. BMC genomics. 2014;15(1):382.
343. Yamamura K, Uruno T, Shiraishi A, Tanaka Y, Ushijima M, Nakahara T, et al. The transcription factor EPAS1 links DOCK8 deficiency to atopic skin inflammation via IL-31 induction. Nature communications. 2017;8:13946.
344. Schaffer K, Taylor CT. The impact of hypoxia on bacterial infection. The FEBS journal. 2015;282(12):2260-2266.
345. Giaccia AJ, Simon MC, Johnson R. The biology of hypoxia: the role of oxygen sensing in development, normal function, and disease. Genes & development. 2004;18(18):2183-2194.
346. Scheinfeldt LB, Soi S, Thompson S, Ranciaro A, Woldemeskel D, Beggs W, et al. Genetic adaptation to high altitude in the Ethiopian highlands. Genome Biology. 2012;13(1):R1-R1.
347. Salceda S, Caro J. Hypoxia-inducible factor 1alpha (HIF-1alpha) protein is rapidly degraded by the ubiquitin-proteasome system under normoxic conditions. Its stabilization by hypoxia depends on redox-induced changes. The Journal of biological chemistry. 1997;272(36):22642-22647.

APPENDIX A

Y chromosome tree of 1000GP and Himalayan individuals (GRCh38)


A
B
C
D
E
G
H
J
L
N
O
Q
R
T
0.004

APPENDIX B

List of Denisovan introgressed SNPs

Denisovan SNPs within Himalaya- specific introgressed tracts

CHR	- chromosome
POS	-genomic position
SNP	- SNP rs number
ALLELE	- archaic allele
HGNC	-gene name
CONSEQUENCE	- functional prediction (VEP)

Denisova introgressed SNPs					
CHR	**POS**	**SNP**	**ALLELE**	**HGNC**	**CONSEQUENCE**
1	6354421	rs568534004	G	-	open_chromatin_region
1	6367361	rs538578120	C	-	open_chromatin_region
1	6389990	rs545807707	G	-	enhancer
1	6390320	rs562430899	A	-	enhancer
1	6438420	rs577801293	A	-	CTCF_binding_site
1	6536773	rs544195746	T	-	CTCF_binding_site
1	14049383	rs74557977	T	-	promoter_flanking_region
1	14050239	rs78095306	G	-	promoter_flanking_region
1	14050615	rs77986549	A	-	promoter_flanking_region
1	14060211	rs114023548	C	-	promoter_flanking_region
1	14079247	rs532546078	C	-	promoter_flanking_region
1	14118977	rs116702257	C	-	open_chromatin_region
1	33631004	rs142594804	C	-	open_chromatin_region
1	33642134	rs3766817	C	-	open_chromatin_region
1	33752272	rs77880723	G	-	open_chromatin_region
1	33766943	rs146318909	T	-	open_chromatin_region
1	33801928	rs117765738	T	-	promoter_flanking_region
1	33804437	rs117836468	C	-	promoter_flanking_region
1	113245618	rs368840677	T	*RHOC*	splice_region_variant,intron_variant
1	113250043	rs375619655	A	-	promoter
1	167683557	rs376459065	C	-	promoter_flanking_region
1	167774578	rs150219328	C	-	enhancer
1	167786711	rs147629796	A	-	TF_binding_site
1	229477835	rs571590141	A	-	promoter
1	232645738	rs7552486	C	-	open_chromatin_region
1	232649763	rs766435961	G	*SIPA1L2*	synonymous_variant
2	21242771	rs200868559	C	*APOB*	synonymous_variant
2	21264999	rs531023775	A	-	TF_binding_site
2	21266363	rs570818403	T	-	promoter_flanking_region
2	21266980	rs547640128	T	-	promoter_flanking_region
2	23265626	rs75127632	T	-	open_chromatin_region
2	33363144	rs140066553	C	-	promoter_flanking_region
2	33370502	rs79979829	G	-	open_chromatin_region
2	33371346	rs79037232	A	-	open_chromatin_region
2	45464590	-	G,A	-	promoter_flanking_region
2	46406589	rs1968635	G	-	enhancer
2	46406619	rs1530668	G	-	enhancer
2	46409146	rs3754559	T	-	open_chromatin_region
2	46412422	rs2278773	G	-	CTCF_binding_site
2	46456816	rs79136032	A	-	promoter_flanking_region
2	46495437	rs368889159	A	-	promoter_flanking_region
2	46521270	rs373013254	T	-	CTCF_binding_site

CHR	POS	SNP	ALLELE	HGNC	CONSEQUENCE
2	46533730	rs375877310	T	-	promoter_flanking_region
2	46558530	rs375418933	C	-	promoter_flanking_region
2	46561400	rs370299814	A	-	promoter_flanking_region
2	46561659	rs368706892	T	-	promoter_flanking_region
2	46577808	rs555698019	A	-	promoter_flanking_region
2	47570709	rs554756783	C	-	CTCF_binding_site
2	47630045	rs786202645	T	-	promoter
2	47766727	rs538742577	T	-	CTCF_binding_site
2	47796858	rs370357176	A	-	promoter_flanking_region
2	47798125	rs541185621	T	-	promoter_flanking_region
2	47799856	rs570702589	G	*AC138655.1*	missense_variant
2	97007908	rs774996456	T	-	open_chromatin_region
2	101554875	rs78527344	A	-	open_chromatin_region
2	101558505	-	T,A	-	open_chromatin_region
2	109165478	rs199597739	C	-	open_chromatin_region
2	109196758	rs143717707	T	-	enhancer
2	109253205	rs201901817	T	-	promoter_flanking_region
2	167011292	rs3886534	T	-	TF_binding_site
2	211442785	-	G	-	CTCF_binding_site
2	231973651	rs760260232	T	*HTR2B*	synonymous_variant
2	231975795	rs907215603	C	-	open_chromatin_region
2	231988965	rs888680018	G	-	promoter_flanking_region
2	232011670	rs1042559328	G	-	open_chromatin_region
2	232018264	rs764590208	G	-	open_chromatin_region
2	232025455	rs1042480674	T	-	promoter_flanking_region
2	232035351	rs751952582	G	*PSMD1*	synonymous_variant
2	232037004	rs1030146827	G	-	open_chromatin_region
2	232037080	rs949733786	A	-	open_chromatin_region
2	232049910	rs955445159	T	-	enhancer
2	232063420	rs980173749	A	-	promoter
2	232077473	rs890075866	A	-	open_chromatin_region
2	232092252	rs924574717	T	-	CTCF_binding_site
2	232100085	rs370669179	T	-	open_chromatin_region
2	233791484	rs144485640	G	-	promoter_flanking_region
2	234091138	rs70940832	T	*INPP5D*	synonymous_variant
3	15351858	rs139347281	T	-	promoter_flanking_region
3	45970944	rs13099120	G	-	CTCF_binding_site
3	186710784	rs542973619	T	-	enhancer
4	3419093	rs16844311	T	*RGS12*	splice_region_variant,intron
4	186578505	rs2306704	G	-	enhancer
5	67111452	rs117153929	A	-	open_chromatin_region
5	77353203	rs117785921	T	-	open_chromatin_region
5	142077223	rs117020353	A	-	promoter_flanking_region

CHR	POS	SNP	ALLELE	HGNC	CONSEQUENCE
6	7975489	-	C	-	promoter_flanking_region
6	10521305	rs72821511	C	-	promoter_flanking_region
6	10522868	rs72821513	G	-	promoter_flanking_region
6	25015153	rs367724830	G	-	enhancer
6	30012624	rs114315683	A	-	open_chromatin_region
6	35454554	rs75480498	T	-	promoter_flanking_region
6	35461867	rs72894781	A	-	promoter_flanking_region
6	35471721	rs188389031	C	-	CTCF_binding_site
6	56620831	rs145031609	C	-	enhancer
6	99958077	rs748665547	A	*USP45*	missense_variant
6	99964134	rs986920319	T	-	promoter
6	100023574	rs330850	C	-	open_chromatin_region
6	119631639	-	G,T	-	enhancer
6	119634071	rs77683431	T	-	promoter_flanking_region
6	119634137	rs78585986	G	-	promoter_flanking_region
6	119635876	rs3798635	C	-	promoter_flanking_region
6	119636240	rs3823001	A	-	promoter_flanking_region
6	119636490	rs3798637	A	-	promoter_flanking_region
6	119636588	rs3798638	A	-	promoter_flanking_region
7	17514707	rs187739737	T	-	CTCF_binding_site
7	17514772	rs191428236	A	-	CTCF_binding_site
7	80488999	rs79725824	T	-	open_chromatin_region
7	80506633	rs138238608	A	-	TF_binding_site
7	80513534	rs186257173	G	-	promoter_flanking_region
7	149510149	rs740112	T	-	CTCF_binding_site
7	149519760	rs202144866	C	-	open_chromatin_region
7	149559072	rs200599838	T	*ZNF862*	synonymous_variant
8	10555301	rs77073793	T	*C8orf74*	missense_variant
8	21614351	rs559929741	T	-	enhancer
8	21614507	rs184078269	T	-	enhancer
8	37403652	rs375443835	A	-	promoter_flanking_region
8	38446214	rs569518627	A	-	open_chromatin_region
8	38483175	rs566392509	G	-	CTCF_binding_site
8	53303768	rs1026771210	A	-	TF_binding_site
9	36867495	rs17391491	A	-	CTCF_binding_site
9	36867507	rs77776435	A	-	CTCF_binding_site
9	100871756	rs1035579538	G	-	promoter_flanking_region
9	100880885	-	A	-	promoter
9	113732569	rs16915602	T	-	open_chromatin_region
9	138133766	rs940329833	T	-	CTCF_binding_site
10	80893337	rs80183004	T	-	enhancer
10	112506454	rs531624200	C	-	CTCF_binding_site
10	119223386	rs560418517	C	-	open_chromatin_region

CHR	POS	SNP	ALLELE	HGNC	CONSEQUENCE
10	119236269	rs533194625	T	-	open_chromatin_region
10	126236440	rs17695019	C	-	open_chromatin_region
11	8615738	-	T	-	promoter
11	8634657	-	A	-	open_chromatin_region
11	8634687	rs181255316	C	-	open_chromatin_region
11	29287690	rs183865464	A	-	TF_binding_site
11	36399086	rs191868665	T	-	promoter
11	36402967	rs117520408	G	-	promoter_flanking_region
11	36406842	rs117153510	A	-	open_chromatin_region
11	83424833	rs80154930	T	-	TF_binding_site
11	83436260	rs180851175	G	-	CTCF_binding_site
11	110045286	rs1040021791	G	-	promoter_flanking_region
11	113954215	rs118050126	C	-	promoter_flanking_region
11	125812061	-	G,A	-	TF_binding_site
11	131618541	rs570541277	C	-	CTCF_binding_site
11	131618935	rs530598977	A	-	open_chromatin_region
11	131619057	rs546924141	G	-	open_chromatin_region
12	6184143	rs2239155	A	-	promoter_flanking_region
12	6187788	rs117115831	A	-	enhancer
13	21063533	rs202040300	G	-	CTCF_binding_site
13	114880593	rs117295582	T	-	enhancer
13	114912023	rs12585992	T	-	CTCF_binding_site
13	114944785	rs118122842	T	-	CTCF_binding_site
14	53618860	rs561056322	A	-	promoter
14	79977041	rs77934216	G	-	open_chromatin_region
14	91164759	rs78720229	T	-	enhancer
14	96728122	rs34334635	A	-	open_chromatin_region
15	28340920	rs180681661	T	-	promoter_flanking_region
15	71510985	rs543912832	T	-	open_chromatin_region
15	85877075	-	A,T	-	enhancer
16	3026680	rs4149788	A	*PKMYT1*	synonymous_variant
16	3027039	rs4149785	A	*PKMYT1*	splice_region_variant,intron_variant
16	87712807	rs936651169	T	-	CTCF_binding_site
16	87715900	-	G	-	open_chromatin_region
16	87717790	rs753264216	G	*JPH3*	synonymous_variant
16	87732196	-	T	-	TF_binding_site
16	87736136	rs1027754692	A	-	promoter_flanking_region
17	7034452	-	A,C	-	promoter_flanking_region
18	22585593	rs79111688	G	-	open_chromatin_region
18	47638207	rs138261729	G	-	CTCF_binding_site
18	47678309	rs371235559	T	-	open_chromatin_region
18	47688484	rs374840270	T	-	CTCF_binding_site
18	60021761	rs35211496	T	*TNFRSF11A*	missense_variant

CHR	POS	SNP	ALLELE	HGNC	CONSEQUENCE
19	2043107	rs146470954	C	-	promoter_flanking_region
19	41622495	rs76335535	A	*CYP2F1*	missense_variant
19	41626271	rs139705553	T	*CYP2F1*	synonymous_variant
20	2723275	rs574706705	G	-	CTCF_binding_site
20	18393782	rs143400891	C	-	promoter_flanking_region
20	18395092	rs139700575	T	*DZANK1*	missense_variant
21	17923179	rs78195113	G	-	open_chromatin_region
21	17923777	rs79615460	A	-	open_chromatin_region

APPENDIX C

List of Fine*MAV* candidates

100 top scoring SNPs detected by Fine*MAV*

CHR - chromosome

POS -genomic position

SNP - SNP rs number

HGNC -gene name

ANC_ALLELE -ancestral allele

DER_ALLELE -derived allele

Fine*MAV_POP* *-FineMAV score in relevant population*

CONSEQUENCE - functional prediction (VEP)

High versus low altitude Himalayans
High altitude

CHR	POS	SNP	HGNC	ANC_ALLELE	DER_ALLELE	Fine*MAV*_HIM_H	CONSEQUENCE
1	1417587	rs181312132	*ATAD3B*	.	T	2.598	missense_variant
1	9009451	rs2274329	*CA6*	G	C	2.4202	missense_variant
1	158549492	rs863362	*OR10X1*	C	T	2.4901	stop_gained
1	230745328	-	*RP11-543E8.2*	C	T	2.492	downstream_gene_variant
1	236195728	rs78916782	*NID1*	C	T	3.234	missense_variant
2	3358295	rs12612251	*TSSC1*	T	C	2.4011	intron_variant
2	34525050	rs181481277	*AC009499.1*	G	A	2.4622	downstream_gene_variant
2	46552202	rs149594770	*EPAS1*	T	A	3.3245	intron_variant
2	46553044	rs113305133	*EPAS1*	A	G	2.4851	intron_variant
2	46553781	rs369553078	*EPAS1*	T	A	3.219	intron_variant
2	46558530	rs375418933	*EPAS1*	G	C	2.5049	intron_variant
2	46568084	rs150049928	*EPAS1*	C	T	2.4673	intron_variant
2	46569770	rs73926265	*EPAS1*	G	A	3.2876	intron_variant
2	46576918	rs76242811	*EPAS1*	T	C	2.9492	intron_variant
2	46577251	rs188801636	*EPAS1*	T	C	6.2474	intron_variant
2	46577808	rs555698019	*EPAS1*	T	A	7.3718	intron_variant
2	46583581	rs189807021	*EPAS1*	G	A	3.1901	intron_variant
2	46588019	rs150877473	*EPAS1*	C	G	4.6218	splice_region_variant,intron_variant
2	46589032	rs74898705	*EPAS1*	C	T	2.9138	intron_variant

CHR	POS	SNP	HGNC	ANC_ALLELE	DER_ALLELE	Fine*MAV*_HIM_H	CONSEQUENCE
2	46598025	rs141426873	*EPAS1*	C	G	3.2968	intron_variant
2	4660030	rs116611511	*EPAS1*	A	G	4.6302	intron_variant
2	46600358	rs369097672	*EPAS1*	A	G	6.6004	intron_variant
2	46656647	rs56048837	*TMEM247*	G	C	2.736	upstream_gene_variant
2	46675505	rs77111769	*TMEM247*	T	C	4.9108	intron_variant,non_coding_transcript_variant
2	46681989	rs117115595	*TMEM247*	C	T	3.9203	intron_variant,non_coding_transcript_variant
2	46689216	rs115350785	*TMEM247*	A	G	2.5145	intron_variant,non_coding_transcript_variant
2	46690452	rs1868082	*TMEM247*	C	A	2.4279	intron_variant,non_coding_transcript_variant
2	46693330	rs182127341	*TMEM247*	C	A	3.0291	intron_variant,non_coding_transcript_variant
2	46693993	rs116871724	*TMEM247*	T	A	3.2155	intron_variant,non_coding_transcript_variant
2	46699639	rs58143719	*TMEM247*	T	C	2.7582	intron_variant,non_coding_transcript_variant
2	46702159	rs113190671	*TMEM247*	T	G	3.3463	intron_variant,non_coding_transcript_variant
2	46706765	rs192690066	*TMEM247*	C	T	2.6154	intron_variant,non_coding_transcript_variant
2	46707674	rs116983452	*TMEM247*	C	T	4.6567	intron_variant,non_coding_transcript_variant
2	46710530	rs79542054	*TMEM247*	C	T	5.9521	intron_variant,non_coding_transcript_variant
2	46718090	rs12986653	*ATP6V1E2*	G	A	6.1156	intron_variant,non_coding_transcript_variant
2	46720010	rs75483237	*ATP6V1E2*	T	G	3.0471	intron_variant,non_coding_transcript_variant
2	46720473	rs117677853	*ATP6V1E2*	C	T	2.6394	intron_variant,non_coding_transcript_variant
2	46720802	rs117446572	*ATP6V1E2*	G	A	2.8891	intron_variant,non_coding_transcript_variant
2	46733738	rs117128262	*ATP6V1E2*	T	C	7.11	intron_variant,non_coding_transcript_variant
2	46743164	rs117024627	*ATP6V1E2*	T	C	2.8849	intron_variant

CHR	POS	SNP	HGNC	ANC_ALLELE	DER_ALLELE	Fine*MAV*_HIM_H	CONSEQUENCE
2	46745024	rs13010097	*ATP6V1E2*	G	A	2.8206	intron_variant
2	46752292	rs78082841	*ATP6V1E2*	G	A	2.6724	intron_variant
2	46752876	rs150798075	*ATP6V1E2*	G	T	4.4883	intron_variant
2	46754702	rs138754673	*ATP6V1E2*	C	T	4.1626	intron_variant
2	46768078	rs75498296	*RHOQ*	C	T	6.1009	upstream_gene_variant
2	46772997	rs74869223	*RHOQ*	A	G	3.7979	intron_variant
2	46792633	rs139501481	*RHOQ*	T	A	3.2163	intron_variant
2	46797843	rs78123734	*RHOQ*	C	G	3.0458	intron_variant
2	46809046	rs17035318	*RHOQ*	C	T	2.7835	3_prime_UTR_variant
2	46840552	rs118024480	*CRIPT*	T	C	3.0555	upstream_gene_variant
2	46846131	rs117720191	*CRIPT*	C	T	2.7119	intron_variant
2	46863851	rs75553031	-	A	C	4.2158	intergenic_variant
2	59732957	rs17050199	*AC007131.2*	C	T	2.4261	intron_variant,non_coding_transcript_variant
2	146866679	rs73001881	-	G	T	2.7397	intergenic_variant
2	146963473	rs2552554	-	A	C	3.3969	intergenic_variant
2	161767360	rs2138438	-	A	C	2.6471	regulatory_region_variant
3	9870862	rs2290304	*TTLL3*	G	A	2.9988	missense_variant
3	130031522	-	-	A	C	2.7672	intergenic_variant
3	130031531	rs1045982388	-	C	T	2.7363	intergenic_variant
3	130361856	rs16830494	*COL6A6*	G	A	2.8614	missense_variant
4	164547367	rs9990679	*36951*	G	C	3.0541	intron_variant

CHR	POS	SNP	HGNC	ANC_ALLELE	DER_ALLELE	Fine*MAV*_HIM_H	CONSEQUENCE
6	10070806	rs370734563	*OFCC1*	-	A	4.3426	intron_variant,non_coding_transcript_variant
6	10587038	rs539351	*GCNT2*	G	C	2.5057	missense_variant
6	31140909	rs79527728	*POU5F1*	T	G	3.6341	upstream_gene_variant
6	31502767	rs3131628	*SNORD117*	T	C	2.8806	downstream_gene_variant
6	31506624	-	*SNORD117*	C	T	2.6672	upstream_gene_variant
6	32609212	rs1142326	*HLA-DQA1*	.	T	3.3247	missense_variant
6	32628022	rs1140347	*HLA-DQB1*	.	G	2.6002	synonymous_variant
6	33053609	rs11551421	*HLA-DPB1*	G	A	5.6718	missense_variant
6	33068772	rs79470430	*HLA-DPA2*	G	A	2.9873	upstream_gene_variant
6	33071744	rs61074882	*COL11A2P1*	C	T	3.7319	non_coding_transcript_exon_variant
6	33071839	rs61705355	*COL11A2P1*	C	A	4.8989	intron_variant,non_coding_transcript_variant
6	45709477	rs57630563	-	G	T	2.4582	regulatory_region_variant
6	134385155	rs78562617	-	C	G	2.8132	regulatory_region_variant
6	143168431	rs79504391	*HIVEP2*	T	C	2.595	intron_variant
7	9257189	rs1285418	-	G	A	2.5189	intergenic_variant
9	89165089	-	-	T	G	3.6425	regulatory_region_variant
9	129383900	rs10760442	*LMX1B*	G	A	2.9083	intron_variant
10	115086017	rs79267391	-	G	A	3.2478	regulatory_region_variant
10	115096759	rs12573692	-	G	A	2.6265	intergenic_variant
10	124936990	rs78590159	-	G	A	2.5057	intergenic_variant
11	96757271	rs17129673	-	A	G	2.7382	intergenic_variant

CHR	POS	SNP	HGNC	ANC_ALLELE	DER_ALLELE	Fine*MAV*_HIM_H	CONSEQUENCE
12	57626018	rs11557166	*SHMT2*	C	T	2.5305	synonymous_variant
12	110350819	rs773936923	*TCHP*	A	G	2.4872	missense_variant
14	37154111	rs201299512	*SLC25A21*	A	C	2.8607	stop_gained
15	37809660	rs113706235	-	C	T	2.7775	intergenic_variant
15	90023558	rs17807723	*RHCG*	G	A	3.2258	missense_variant
17	5960915	rs12944495	*WSCD1*	A	G	2.8401	upstream_gene_variant
18	58221665	-	*AC010928.1*	C	A	2.5594	upstream_gene_variant
18	72573324	rs17055773	*ZNF407*	A	G	2.5544	intron_variant
18	72578312	rs7227571	*ZNF407*	C	T	2.5691	intron_variant
19	56030150	-	*SSC5D*	-	C	2.5428	stop_gained
22	24468386	rs17854874	*CABIN1*	G	A	2.5506	missense_variant
22	44608457	rs139203	*PARVG*	A	C	2.4746	downstream_gene_variant
Y	4733000	-	-	.	A	2.4336	intergenic_variant
Y	6932163	rs373532788	*TBL1Y*	.	A	2.9657	missense_variant

High versus low altitude Himalayans **Low altitude**							
CHR	**POS**	**SNP**	**HGNC**	**ANC_ALLELE**	**DER_ALLELE**	**Fine*MAV*_TIB_L**	**CONSEQUENCE**
1	1423286	rs571285142	*ATAD3B*	.	C	2.2791	stop_gained
1	8418541	rs200525338	*RERE*	C	T	2.5369	missense_variant
1	15770022	COSM897576	*CTRC*	.	C	2.3754	3_prime_UTR_variant
1	23418153	rs10799790	*LUZP1*	T	C	2.2911	missense_variant
1	40773150	rs12077871	*COL9A2*	.	G	2.1929	downstream_gene_variant
1	1.52E+08	rs150014958	*TCHHL1*	.	G	2.1533	stop_gained
1	1.88E+08	rs16827680	-	A	G	3.163	intergenic_variant
1	2.04E+08	rs11584773	-	T	C	2.4775	intergenic_variant
1	2.04E+08	rs11586883	-	T	G	2.5548	intergenic_variant
2	18697640	rs16985174	-	C	G	2.3117	regulatory_region_variant
2	33701632	rs17594260	*RASGRP3*	C	T	3.3745	intron_variant
2	46747867	rs2346416	*ATP6V1E2*	A	G	2.3372	upstream_gene_variant
2	47039991	rs935376	*LINC01118*	G	A	2.9429	upstream_gene_variant
2	90259979	rs201899294	*IGKV1D-8*	.	C	2.2158	stop_gained
2	97860471	rs150846613	*ANKRD36*	.	T	2.2178	missense_variant
2	1.62E+08	rs55762147	-	T	C	2.2391	intergenic_variant
2	1.79E+08	rs4894028	*TTN*	C	T	3.1114	missense_variant
2	1.79E+08	rs62621236	*TTN*	A	G	2.7474	missense_variant
2	1.79E+08	rs747122	*TTN*	C	T	2.435	missense_variant

CHR	POS	SNP	HGNC	ANC_ALLELE	DER_ALLELE	Fine*MAV*_TIB_L	CONSEQUENCE
2	1.79E+08	rs11897366	*TTN*	A	C	2.1655	synonymous_variant
2	1.79E+08	rs10164753	*TTN*	C	T	2.903	missense_variant
2	1.79E+08	rs10497517	*TTN*	A	G	2.5016	synonymous_variant
2	1.79E+08	rs16866425	*TTN*	T	C	2.1862	synonymous_variant
2	1.8E+08	rs4893852	*TTN*	A	G	2.6875	intron_variant
2	1.8E+08	rs13398235	*TTN*	G	A	2.6125	intron_variant
2	1.8E+08	rs11888217	*TTN*	C	T	2.2375	intron_variant
2	1.8E+08	rs4893853	*TTN*	T	C	2.17	missense_variant
2	1.8E+08	rs4894048	*TTN*	C	T	2.3025	missense_variant
2	2.13E+08	rs13390226	*ERBB4*	G	A	3.399	intron_variant
3	37562369	rs142719179	*ITGA9*	C	T	2.2588	intron_variant
3	1.14E+08	rs1025399	*GRAMD1C*	A	G	2.2304	intron_variant
3	1.61E+08	rs223125	*B3GALNT1*	C	T	2.1773	missense_variant
4	1563853	rs111807331	-	T	C	2.1587	intergenic_variant
4	1564086	rs113334316	-	C	T	2.1587	intergenic_variant
4	9386295	rs35948112	*RP11-1396O13.13*	.	A	3.65	stop_gained
4	48514580	rs550020018	*FRYL*	C	T	2.2902	missense_variant
4	1.57E+08	rs2705453	-	G	A	2.2468	intergenic_variant
5	1.09E+08	rs13181131	*PJA2*	T	C	2.4641	intron_variant
5	1.77E+08	rs148853192	*DOK3*	G	A	3.6507	upstream_gene_variant
6	10070806	rs370734563	*OFCC1*	-	G	4.1136	intron_variant,non_coding_transcript_variant

CHR	POS	SNP	HGNC	ANC_ALLELE	DER_ALLELE	Fine*MAV*_TIB_L	CONSEQUENCE
6	25273653	rs77918027	*RP3-522P13.2*	T	C	2.1578	splice_donor_variant,non_coding_transcript_variant
6	32184835	rs151131761	*NOTCH4*	C	T	2.2516	missense_variant
6	32573713	rs9270974	-	.	C	2.2307	intergenic_variant
6	32578040	rs9271174	-	.	C	2.2664	regulatory_region_variant
6	32628022	rs1140347	*HLA-DQB1*	.	A	2.8671	synonymous_variant
6	1.62E+08	rs9355339	*AGPAT4*	T	C	2.3496	intron_variant
8	39912400	rs2929103	-	G	T	2.4331	regulatory_region_variant
8	90507811	-	-	C	T	4.2429	intergenic_variant
8	1.11E+08	rs78254607	*SYBU*	C	A	2.3738	intron_variant
8	1.22E+08	rs76075043	-	C	T	2.4696	intergenic_variant
9	17922765	rs13294429	-	T	C	2.2668	intergenic_variant
9	25530564	rs2183901	-	T	C	2.5563	intergenic_variant
9	25538211	rs557419369	-	T	A	2.9447	intergenic_variant
9	27062721	rs3429	*IFT74*	C	T	2.8875	missense_variant
9	34306410	rs17350674	*KIF24*	C	A	2.7489	missense_variant
9	97321410	rs541509845	*FBP2*	.	C	2.2158	missense_variant
10	25777562	rs1547685	*GPR158*	A	G	2.2212	intron_variant
10	34582735	rs11009717	*PARD3*	T	A	3.4654	intron_variant
10	34599236	rs11009725	*PARD3*	A	C	2.8055	downstream_gene_variant
10	34604895	rs11009730	*PARD3*	T	G	2.948	intron_variant
10	34626113	rs10508802	*PARD3*	A	G	2.5882	intron_variant

CHR	POS	SNP	HGNC	ANC_ALLELE	DER_ALLELE	Fine*MAV*_TIB_L	CONSEQUENCE
10	77874502	rs182098476	*C10orf11*	C	T	2.3887	intron_variant
10	78784075	rs7922133	*KCNMA1*	C	T	2.3913	intron_variant
11	20101744	rs1867115	*NAV2*	A	C	3.0006	synonymous_variant
11	38827333	rs78267989	-	T	A	2.5183	intergenic_variant
11	61643375	rs371188401	*FADS3*	.	C	2.1828	missense_variant
11	1.26E+08	rs115533243	*CDON*	T	C	2.6109	missense_variant
12	20720679	rs12369747	*PDE3A*	C	T	2.6788	intron_variant
12	49304447	rs58829932	*CCDC65*	T	C	2.3363	intron_variant
12	49333799	rs2228417	*ARF3*	G	A	2.8613	synonymous_variant
12	54491224	rs12819039	*Y_RNA*	A	G	2.5423	downstream_gene_variant
12	1.1E+08	rs34725387	*MYO1H*	C	T	2.243	missense_variant
12	1.16E+08	rs11614480	-	T	C	3.3112	intergenic_variant
12	1.16E+08	rs11615848	-	G	T	2.4407	intergenic_variant
12	1.21E+08	rs206968	*RP1-166H1.2*	G	C	2.6672	downstream_gene_variant
13	73675865	rs914799	*FABP5P1*	A	T	2.5125	upstream_gene_variant
13	73681090	rs2483174	-	G	T	2.3963	intergenic_variant
14	37154111	rs201299512	*SLC25A21*	A	C	2.889	stop_gained
14	98920402	rs11622215	-	A	G	2.215	intergenic_variant
15	53782668	rs12916434	-	T	A	2.1604	intergenic_variant
15	53811633	rs12902770	*WDR72*	T	C	2.7249	intron_variant
15	68628163	rs2306022	*ITGA11*	C	T	2.2775	missense_variant

CHR	POS	SNP	HGNC	ANC_ALLELE	DER_ALLELE	Fine*MAV*_TIB_L	CONSEQUENCE
16	6675913	rs11648207	*RBFOX1*	C	T	2.1976	intron_variant
16	31099000	rs201075024	*PRSS53*	C	T	2.7295	missense_variant
16	75532994	rs385847	*CHST6*	C	T	2.3838	upstream_gene_variant
16	75532995	rs116974915	*CHST6*	C	A	2.2713	upstream_gene_variant
17	1340129	rs145983279	*CRK*	C	T	2.3596	missense_variant
17	56387461	rs372749390	*BZRAP1*	C	T	2.1514	missense_variant
17	76010121	rs76402227	*TNRC6C*	C	T	2.3887	intron_variant
18	23395887	rs117034580	*RN7SL97P*	C	T	2.2664	downstream_gene_variant
18	54697922	rs2848967	*WDR7*	A	G	2.55	3_prime_UTR_variant
18	66178930	rs17804596	-	C	A	2.7036	intergenic_variant
19	10685705	rs375010828	*AP1M2*	-	C	2.247	missense_variant,splice_region_variant
19	43233421	rs202207386	*PSG3*	N	C	2.2158	stop_gained
19	52380625	rs149543099	*ZNF577*	.	G	2.247	stop_gained
19	56030150	-	*SSC5D*	-	C	2.568	stop_gained
X	82751254	rs12860283	-	C	T	2.3331	intergenic_variant

APPENDIX D

Experimental protocols used in this study

- Hypoxic protocol
- RT-qPCR protocol
- SYBR Green qPCR Protocol

Hypoxic protocol

Before starting:

- Prepare conditioned medium:
    - Culture cells in appropriate size flask (T75, T150) under normal culture media (21%O_2, 37^{0}C, 5% CO_2).
    - Collect the medium from cultured cells (handle the flask carefully to avoid taking too many cells with the medium) and transfer it to a Falcon tube.
    - Centrifuge the medium from cultured cells at 1500 rpm for 4 minutes to remove any cell debris, etc.
    - Collect supernatant from the medium of cultured cells (leave ~1 of medium in the Falcon tube with the pellet).
    - Mix the same volume (1:1) of the medium from cultured cells with fresh medium (RMPI 1640 [Thermo Fisher cat num: 11875093] + 15/20% FBS [Thermo Fisher cat num: 10082139] + PenStrep [cat num: 10378016).
    - Filter the conditioned medium using a 0.2µm filter [Whatman-SIGMA cat num: 10462200] in a new Falcon tube.
    - The conditioned medium can be stored in the fridge. Use it within few days (3-4 days).

- Automatic count of cell lines for the experiment using trypan blue or AOPI (Cellometer 2000). Check cell viability is above 90%.

Experimental set up:

- Two conditions: hypoxia (1% O_2, 37^{0}C, 5% CO_2) and normoxia (21%O_2, 37^{0}C, 5% CO_2).
- Time points: 4h, 16h, 24h, 48h, 72h (time points can vary).
- Replicates: 2 per sample in each condition.
- Samples labelling: from 1 to total number of samples, replicates indicated as "A" and "B". Maximise randomisation.

Experimental procedure:

Day 1

- Label six-well plates according to experimental set up, indicating the time point, sample and replicate (important to be able to process hypoxic samples quickly and reduce exposure to normoxia).
- Prepare six-well plates with conditioned medium: fill each well of the six-well plate with 3ml of conditioned medium. Total amount of medium: six wells per plate x number of six well plates in the experiment (Note: consider all samples, different time points and replicates).
- Place plates with conditioned medium for hypoxic conditions in the hypoxic incubator (1% O_2, 37^{0}C, 5% CO_2) and leave overnight to degasify.
- Number of cells for each cell line: ~1.5x10^6 cells/ml.

- Take the right amount of total cells for each sample (Note: consider all samples, different time points and replicates) and place them in a new T25 flask with some fresh medium (this is a critical step if cells are not growing well) and leave them overnight to adjust.

Day 2

- Centrifuge gently (500 rpm, 4 minutes) cells in T25 flasks from Day 1.
- **Note:** Total volume for each well: 3 ml of conditioned medium (prepared in Day 1) + 0.7ml of cells for 72h time point, 1 ml of cells for 24h and 48h time points and 1.5 ml for 4h time point (the small volume of cells is important to reduce re-gasification of the hypoxic medium).
- Re-suspend pellet of each cell lines in the right amount of conditioned medium for all samples.
- Take the six-well plates in the normoxic incubator. Seed plates with 1.5 ml of cells for 4h time point and with 1ml of cells for 24h and 48h plates.
- Take the six-well plates in the hypoxic incubator one at the time. Seed plates with 1.5 ml of cells for 4h time point, with 1ml of cells for 24h and 48h plates and with 0.7ml for 72h time point.
- Once all plates are seeded, wait 1h to allow the hypoxic medium to degasify completely again and start counting the time points.

Day3, Day4, Day5

- When each time point is reached:
    - Label 15ml Falcon tubes with sample name, time point and replicate.
    - Start processing hypoxic samples.
    - Take plates from incubator and transfer quickly samples to the Falcon tubes.
    - Centrifuge cells at 500 rcf (~2000 rpm) for 2/3 minutes.
    - Remove supernatant and re-suspend cells in QIAGEN Buffer RLT [QIAGEN cat num: 79216]: pipette well up and down and vortex each tube to be sure cells are completely lysed. Work on ice.
    - Freeze samples at -80^{0}C.
    - RNA extraction: look RNeasy Mini kit protocol [QIAGEN cat num: 74104].

RT-qPCR protocol

RNA extraction

Performed with the Qiagen RNeasy mini RNA extraction kit with DNase treatment.
RNeasy mini kit: 74104
DNase set: 79254

Use ~1-2x10^6 cells per 350 µL buffer RLT.

Reverse Transcriptase Protocol

Aim: To convert total RNA into cDNA for use in quantitative PCR

Nanodrop of total RNA

Measure the total RNA on the nanodrop machine to get the concentration and also to determine the purity of the RNA isolated.

The 260/280 ratio estimates the amount of protein contamination (measured at 280nm) compared to the nucleotide (measured at 260nm) concentration. A good RNA extraction should have a value of above 1.8.

The 260/230 ratio gives the amount of other impurities (measured at 230nm), such as ethanol or salt. A clean extraction should have a value of above 1.8.

Low values for the 260/280 or 260/230 ratio may cause a slight loss in efficiency

Procedure for reverse transcribing 500ng RNA into cDNA for use with quantitative PCR

To convert the total RNA into cDNA, the protocol uses random primers (a mix of different hexamers); these will bind to different parts of the RNA and prime the conversion of the single stranded RNA into complementary DNA (cDNA) which can then be used for quantitative PCR.

Reagents

RNA	
Random Primer	Promega C1181
dNTP	Life Tech 10 mM
RNASE OUT	Invitrogen 10777-019
Superscript II	Invitrogen 18064071

Step 1 - Denaturation of RNA and Primer

Thaw RNA samples on ice to avoid degradation. After thawing of all reagents, mix the tubes by vortexing briefly and then spin down in a micro centrifuge.

Per reaction

RNA	500ng made up to a total volume of 11.875uL with nuclease free water

Random Primer	0.5uL
dNTP	1uL

1. Add the appropriate amount of nuclease free water, random primer and dNTP (primer and dNTP can be added together as a mix) to each tube.

2. Add 500ng of RNA to each tube and spin down to make sure everything is at the bottom.

3. Incubate at 65^{0}C in a heat block for 5min, then snap cool on ice. This step is to denature any secondary structure in the RNA and also to denature the random primer; the snap cooling ensures that the structure does not reform.

4. Spin tubes down and place in a rack at room temperature.

Step 2 - Reverse Transcription of RNA

Per reaction – make as a master mix for appropriate number of samples.

5x 1st strand buffer	4uL
0.1M DTT	2uL
RNase out	0.5uL
Superscript II	0.125uL

1. Add 6.625uL of the master mix to the side of each tube and then spin down. The final reaction volume is 20uL.

2. Incubate in heat block/PCR machine.

   25^{0}C x 10min – primer annealing step
   42^{0}C x 50min – extension step
   70^{0}C x 15min – inactivation of enzyme.

Dilute the cDNA to a final volume of 500uL with nuclease free water prior to use for qPCR.

Use 5 µL of diluted cDNA per 20 µL qPCR reaction.

SYBR Green qPCR Protocol

Based on using ABI StepOne Plus and KAPA SYBR Fast 2x (KK4604)
Plates: MicroAmp Fast optical 96-well plates (4346906) and some MicroAmp optical adhesive film (4311971).

20 µL reaction

10 µL 2x KAPA SYBR Fast ABI
0.8 µL 10 µM primer mix
4.2 µL nuclease-free water
5 µL dilutes cDNA

ABI StepOne Plus cycling

95 °C – 3minutes
95 °C – 20 seconds
61 °C – 20 seconds } x40
72 °C – 20 seconds
+melt curve to check for single products (you may want to sequence these amplicons to check for specificity).