University College London

Detecting signals of selection in the genomes of Native Americans and admixed Latin Americans

UCL Genetics Institute (UGI)

A Thesis submitted for the degree of doctor of philosophy

Author:
Javier Mendoza-Revilla

Supervisors:
Professor Andres Ruiz-Linares
Professor Francois Balloux
Dr Garrett Hellenthal

August, 2018

Declaration

I, Javier Alberto Mendoza Revilla confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm this has been indicated in the thesis.

Publications arising from this thesis:

- Adhikari & **Mendoza-Revilla J** et al., (2018). Genome-wide Association Scan in Latin Americans highlight the convergent of lighter skin pigmentation in Eurasia. In press. Nature Communications.
- **Mendoza-Revilla J** et al., (2018). Genome-wide selection scan in Native American populations reveal instances of local adaptation. In advanced preparation.
- **Mendoza-Revilla J** et al., (2018). Inferring selection post-admixture in admixed Latin Americans. In advanced preparation.

Publications not directly related to this thesis:

- L. van Dorp, S. Lowes, J. Weigel, N. Ansari-Pour, S. Lopez, **Mendoza-Revilla J**, et al. The Genetic Legacy of State Centralization in the Kuba Kingdom of the Democratic Republic of the Congo. Submitted PNAS.
- Montalva N, Adhikari K, Liebert A, **Mendoza-Revilla J**, et al. Adaptation to milking pastoralism in Chilean goat herders and nutritional benefit of lactase persistence. In press. Annals of Human Genetics.
- Chacon-Duque, J.C, Adhikari K, Fuentes-Guajardo M, **Mendoza-Revilla J**, et al., 2018. Latin Americans show wide-spread *Converso* ancestry and the imprint of local Native ancestry on physical appearance. Submitted Nature Communications.
- Adhikari K, Chacon-Duque JC, **Mendoza-Revilla J**, Fuentes-Guajardo M, Ruiz-Linares A. 2017. Genetic Diversity in the Americas. Annual Review of Genomics and Human Genetics. Volume 18.
- Arauna LR, **Mendoza-Revilla J**, et al., (2017) Recent historical migrations have shaped the gene pool of Arabs and Berbers in North Africa. Molecular Biology and Evolution. 34 (2): 318-329

- Adhikari K, Fuentes-Guajardo M, Quinto-Sanchez M, **Mendoza-Revilla J**, et al., 2016. A genome-wide association scan implicates *DCHS2*, *RUNX2*, *GLI3*, *PAX1* and *EDAR* in human facial variation. Nature Communications. 7:11616

- Adhikari K, Fontanil T, Cal S, **Mendoza-Revilla J**, et al,. 2016. A genome-wide association scan in admixed Latin Americans identifies loci influencing facial and scalp hair fea-tures. Nature Communications. 7:10815

- Adhikari K, **Mendoza-Revilla J**, Chacon-Duque JC, Fuentes-Guajardo M, Ruiz-Linares A. 2016. Admixture in Latin America. Current Opinion in Genetics & Development, Volume 41, Pages 106-114.

Statement of work

Chapter 1

Some of the text present in Chapter 1, Section 1.3 is based on and adapted from two reviews I jointly published with Dr Kaustubh Adhikari, Dr Juan Chacon-Duque, Ms. Macarena Fuentes-Guajardo and Prof. Andres Ruiz-Linares. This work is published in Adhikari et al. (2016) and Adhikari et al. (2017) and uses text adapted from it, some of which was originally written jointly by myself, Dr Kaustubh Adhikari, Dr Juan Camilo Chacon-Duque, Ms. Macarena Fuentes-Guajardo and Prof. Andres Ruiz-Linares.

Chapter 2

Methods refer to in this chapter were described exclusively by myself.

Chapter 3

Work in Chapter 3 was undertaken exclusively by myself.

Chapter 4

Methods referred to in this chapter (Section 4.4.2) were written by Dr Garrett Hellenthal and developed jointly by Dr Garrett Hellenthal and myself. Some consultation was undertaken with Dr Alicia R. Martin, Dr Cesar Fortes-Lima and Dr Brian Maples for the RFMix analysis described in Section 4.3.5.

Chapter 5

Work in Chapter 5 was undertaken jointly with Dr Kaustubh Adhikari. The work conducted by Dr Kaustubh Adhikari included parts of Section 5.3.2 – 5.3.8. All other work was performed by myself. This work is published in Adhikari K & Mendoza-Revilla J et al. (2018) and uses text adapted from it, some of which was originally written jointly by myself, Dr Kaustubh Adhikari and Prof. Andres Ruiz-Linares and includes input from all other authors on this publication.

Chapter 6

Work in Chapter 6 was undertaken predominantly by myself. An early dataset merge of the worldwide populations was done by Dr Kaustubh Adhikari prior to publication. This work is published in Adhikari K & Mendoza-Revilla J et al. (2018) and uses text adapted from it, some of which was originally written jointly by myself, Dr Kaustubh Adhikari and Prof. Andres Ruiz-Linares and includes input from all other authors on this publication. Some consultation was undertaken with Dr Aida Andres and Dr Louise

Ormond for the Approximate Bayesian Computation (ABC) analysis described in Section 6.3.5.

Dataset contribution

Genomic data from admixed Latin Americans was collected by the Consortium for the Analysis and Diversity and Evolution (CANDELA). I am grateful to those involved in the sample collection, documentation and processing of the genomic and phenotypic data. I am also very grateful to all the volunteers who agreed to take part in the CANDELA study. The rest of the samples were obtained from publicly available datasets as acknowledged in the text.

Acknowledgements

First of all I would like to thank my primary supervisor Prof. Andres Ruiz-Linares who allowed me to be part of the CANDELA Consortium. I have greatly benefited from his knowledge and guidance throughout my PhD. I would also like to thank my secondary and third supervisors, Prof. Francois Balloux and Dr Garrett Hellenthal, who allowed me to become another member of their respective labs. I have benefited incredibly from them both as well as from all their lab members. I would also like to wholeheartedly thank Dr Kaustubh Adhikari and Dr Matteo Fumagalli for their constant support throughout my PhD. I have learned a lot from these two outstanding researchers. I am also very grateful to all the current and past members of the Ruiz-Linares lab, especially to Dr. Juan Camilo Chacon-Duque and to Macarena Fuentes-Guajardo as well as Victor Acuna-Alonso, Caio Cesar Silva de Cerqueira and Celia Cintas, all who greatly enriched my PhD experience. I would also like to thank everyone at the UGI, especially Dr Lucy van Dorp, Dr Liam Shaw, Dr Saioa Lopez, Dr Stephen J. Price, Carla Bardua, Dr Florent Lassalle and many others! I am also very grateful to all my friends outside UCL, who have made the experience of doing a PhD in London such an incredible journey. I would also like to thank my family for all their love and for believing that I could become a Dr! Final thanks to my partner Clem, for her constant and unconditional support, especially on these last few months... I could not be more grateful. *Merci!*

Abstract

The peopling of the Americas represents the last major expansion of human populations worldwide. As the first humans moved into the continent they were exposed to new environments requiring them to adapt. The subsequent colonization of the continent by Europeans, along with the African slave trade, involved a major admixture process that was accompanied by new selective pressures, most notably exposure to new pathogens. Applying current and novel methods to genome-wide SNP data of Native and admixed Latin Americans, this PhD thesis provides an analysis of the adaptive history in the Americas. I show that prior to the European contact, candidate regions of selection in Native Americans include genes associated with metabolic traits, highlighting a possible adaptation to dietary changes. Using novel and existing methods to detect selection post-admixture, I show that genes related to immune response were probably under selection in admixed Latin Americans. As an example on the evolution of an adaptive trait, I also conduct a Genome Wide Association Study on a sample of over 6,000 Latin Americans for skin, eye and hair pigmentation. I report eighteen independent genome-wide significant signals of association, including five novel variants. One of the novel variants associated to skin pigmentation is common in East Asians and Native Americans, but is almost absent everywhere else in the world. I show that this variant was selected in East Asians after their split from Europeans, and likely carried by the first Americans to the Americas.

Impact Statement

Over the last decade, there has been a dramatic increase in the number human genomic datasets being generated that have enabled us to study the historic and adaptive history of human populations. Importantly, the identification of variants subject to selection, which are likely to be of functional relevance, can lead to insights into how genes affect human phenotypic variation. Such variants include those that predispose or protect individuals to disease, and might therefore inform the development of therapeutic and prevention strategies and/or aid biologically-informed drug discovery. In this PhD thesis, I apply current and novel methods to detect signals of selection in Native and admixed Latin Americans, and report regions of the genome with evidence of selection that may represent important candidate genes with potentially functional relevance.

I also conduct a genome-wide association scan of pigmentation phenotypes in Latin Americans. This work has several implications. First, human genetic studies have been mainly conducted in populations of European descent, raising the question about the transferability of these findings to other human populations. By analyzing Latin Americans, a vastly underrepresented population in human genetic studies, I have addressed this question directly, by reporting the effect of pigmentation associated variants in this population. Second, my work shows that the diversity in phenotypic and genetic variability in Latin Americans affords ample opportunity to discover novel genetic associations, and highlights how genetic variants that contribute to pigmentation phenotypes in other worldwide populations are yet to be explored. Third, since the history of Latin Americans involved extensive admixture of Native American, Europeans and Africans, my work shows how genetic association studies in Latin Americans can also inform us about the impact of associated variants in these parental populations. Finally, these findings have a direct practical implication, as the better understanding of human skin pigmentation genetics is highly relevant for medical studies due to the shared role between pigmentation-associated loci and many type of skin cancers, and to forensic applications through the development and application of pigmentation phenotype prediction based on DNA variants.

Contents

1 Introduction **27**

1.1 Overview . . . 27
1.2 Natural selection . . . 29
1.2.1 Types of selection . . . 29
1.2.2 Detecting positive natural selection . . . 30
1.2.3 Challenges to detecting selection . . . 33
1.3 Genetic history of the Americas . . . 34
1.3.1 Genetic history of Native Americans . . . 35
1.3.2 Genetic history of admixed Latin Americans . . . 38
1.4 Recent human adaptation . . . 43
1.4.1 Previous studies on detecting selection in Native Americans . . . 43
1.4.2 Previous studies on detecting selection in admixed Latin Americans 45
1.5 Human pigmentation variation . . . 46
1.5.1 The biology of human pigmentation . . . 47
1.5.2 The evolution of human pigmentation . . . 48
1.5.3 The genetic determinants of human pigmentation . . . 52
1.6 Finding genetic associations . . . 57
1.6.1 Genome Wide Association Study (GWAS) — rationale and scientific basis . . . 57
1.7 Genomics is failing on diversity . . . 61
1.8 Consortium for the Analysis of the Diversity and Evolution of Latin America - CANDELA . . . 62
1.9 Summary . . . 64

2 Methods **67**

2.1 Overview . . . 67
2.2 SNP-based approaches to detect selection . . . 67
2.2.1 Allele frequency differentiation based approaches . . . 67
2.2.2 Haplotype-based approaches . . . 69
2.2.3 Time scales for the signatures of selections . . . 72
2.3 Combining selection signals from many loci . . . 72
2.4 Using environmental data to identify loci underlying local adaptation . . . 74
2.5 SNP-based approaches to detect natural selection in admixed populations . 75
2.5.1 Local ancestry deviations . . . 75

2.6 Inferring the starting time and intensity of selection via Approximate Bayesian Computation (ABC) . . . 76
2.7 Genome Wide Association Studies (GWAS) . . . 77
2.7.1 Single marker associations . . . 77
2.7.2 Correcting for population structure . . . 77
2.8 Summary . . . 79

3 Detecting signatures of selection in Native Americans 80
3.1 Overview . . . 80
3.2 Background . . . 80
3.3 Materials and methods . . . 84
3.3.1 Description of Data . . . 84
3.3.2 Quality control . . . 84
3.3.3 Selecting individuals without post-Columbian admixture . . . 85
3.3.4 Selection scans in Native American individuals without post-Columbian admixture . . . 85
3.3.5 Identification of Latin American individuals with specific Native American ancestry components . . . 86
3.3.6 Gene set enrichment analysis using biological pathways . . . 87
3.3.7 Gene Ontology (GO) enrichment analysis . . . 88
3.3.8 Phenotypic association analysis . . . 88
3.3.9 Worldwide allele frequencies . . . 89
3.4 Results . . . 89
3.4.1 Overview . . . 89
3.4.2 Selection signals in Native Americans . . . 90
3.4.3 Association testing with metabolic and anthropometric phenotypes in top SNPs in Native Americans without post-Columbian admixture 97
3.4.4 Gene set enrichment analysis using biological pathways and Gene Ontology (GO) categories in Native Americans . . . 99
3.4.5 Selection signals in three distinct Native American populations . . . 99
3.4.6 Gene set enrichment analysis using biological pathways and GO categories in three distinct Native American populations . . . 105
3.5 Discussion and limitations . . . 105
3.5.1 Dietary adaptations in Native Americans . . . 105
3.5.2 Immune adaptations in Native Americans . . . 106
3.5.3 Adaptations at previously reported genes under selection . . . 108
3.6 Summary . . . 111

4 Detecting Post-Columbian signals of selection in Latin Americans 112
4.1 Overview . . . 112
4.2 Background . . . 112
4.3 Materials and methods . . . 115
4.3.1 Description of the genomic data . . . 115
4.3.2 Quality control . . . 115

4.3.3 ADMIXTURE and PCA analysis . . . 115
4.3.4 A new statistical model to detect selection post-admixture . . . 116
4.3.5 Local ancestry deviation analysis . . . 118
4.4 Results . . . 119
4.4.1 Description of the new beta-binomial model to detect selection post-admixture . . . 119
4.4.2 Admixture proportions and genetic drift estimates in five admixed Latin American populations . . . 121
4.4.3 Candidate regions of selection post-admixture identified by the new beta-binomial model in admixed Latin Americans . . . 121
4.4.4 Candidate regions of selection post-admixture based on local ancestry deviations . . . 126
4.5 Discussion . . . 126
4.6 Summary . . . 135

5 Genetic determinants of pigmentation in Latin Americans 136
5.1 Overview . . . 136
5.2 Background . . . 136
5.2.1 Previous studies . . . 137
5.3 Materials and methods . . . 138
5.3.1 Study subjects . . . 138
5.3.2 DNA genotyping and quality control . . . 138
5.3.3 Description of pigmentation phenotypes . . . 139
5.3.4 Phasing and imputation . . . 140
5.3.5 Narrow sense heritability . . . 140
5.3.6 ADMIXTURE analysis . . . 141
5.3.7 Association analysis . . . 141
5.3.8 Meta-analysis . . . 142
5.3.9 SNP x SNP interaction of genome-wide associated SNPs . . . 142
5.4 Results . . . 142
5.4.1 Distributing of pigmentation phenotypes in Latin Americans . . . 142
5.5 Correlation between pigmentation phenotypes and covariables. . . . 144
5.5.1 Heritability of pigmentation phenotypes in Latin Americans . . . 148
5.5.2 Genomic regions showing signals of association . . . 148
5.5.3 Meta-analysis . . . 151
5.5.4 Allelic heterogeneity at OCA2/HERC2 and GRM5/TYR . . . 154
5.5.5 Interactions between SNPs independently associated to pigmentation 159
5.6 Discussion and limitations . . . 159
5.7 Summary . . . 163

6 Exploring the convergent evolution of lighter skin pigmentation in Eurasia 164
6.1 Overview . . . 164
6.2 Background . . . 164

6.2.1 Previous studies . . . 165
6.3 Materials and methods . . . 166
6.3.1 Description of data . . . 166
6.3.2 Selection signals at skin pigmentation-associated genomic regions . . 168
6.3.3 Enrichment analysis of selection signals at pigmentation-associated genomic regions . . . 169
6.3.4 Using solar radiation data to identify pigmentation loci under selection 169
6.3.5 Approximate Bayesian Computation (ABC) analysis . . . 170
6.4 Results . . . 171
6.4.1 Selection has shaped the genetic diversity at pigmentation-associated regions . . . 171
6.4.2 Loci underlying local adaptation through solar radiation exposure . 174
6.4.3 *MFSD12* is a novel candidate gene for the convergent evolution of lighter skin pigmentation in East Asians . . . 176
6.5 Discussion . . . 176
6.6 Summary . . . 179

7 Conclusions **180**
7.1 Future directions and significance in studies of human adaptation . . . 181
7.2 Future directions and significance in GWAS . . . 184

A Detecting signatures of selection in Native Americans **247**

B Detecting signals of selection post-admixture in Latin Americans **259**

C Genetic determinants of pigmentation in Latin Americans **261**

D Exploring the convergent evolution of lighter skin pigmentation in Eurasia **273**

List of Figures

1.1 Examples of recent human local adaptation. 30
1.2 Schematic representation of genomic signatures of positive selection. 32
1.3 Polygenic scores across Eurasian populations for different GWAS data sets. 34
1.4 Major human migrations across the world inferred through genomic data. . 38
1.5 Estimated size of the Native American population at the time of Columbus's first landing on the Americans. 40
1.6 Estimated number of African slaves transported to the Americas. 41
1.7 Proportion of African, European and Native American ancestry estimated with mtDNA, Y-chromosome, X-chromosome and autosomal data in thirteen Latin American populations. 42
1.8 Proportion of African, European and Native American ancestry from samples from countries and dependencies across the American continent. 42
1.9 Schematic of signal of selection post-admixture. 46
1.10 Skin pigmentation in world-wide human populations 47
1.11 Melanin synthesis and histology of different skin types 48
1.12 Map of predicted skin pigmentation 50
1.13 Spurious association due to unaccounted population structure 59
1.14 GWAS SNP-trait discovery time line 60
1.15 Omnigenic model for complex phenotypic traits 61
1.16 Proportions of volunteers of different ancestries in Genome-Wide Association Studies (GWAS) in 2009 and 2016. 63
1.17 Birthplace locations of CANDELA volunteers 64

2.1 Schematic representation of the Population Branch Statistic. 69

3.1 The geographic distribution of the putatively selected FADS haplotype in Native American populations. 82
3.2 The geographic origin of the selected EDAR haplotype. 82
3.3 Native American reference population samples and ancestry estimates for the CANDELA sample. 84
3.4 Genome-wide scan for selection in Native Americans. 90
3.5 Selection for four selection tests in candidate region surrounding top selected SNP rs7631391 in Native Americans. 93
3.6 Selection for four selection tests in candidate region surrounding top selected SNP rs5996039 in Native Americans. 94

3.7 Selection for four selection tests in candidate region surrounding top selected SNP rs139553 in Native Americans. 95
3.8 Selection for four selection tests in candidate region surrounding top selected SNP rs8021638 in Native Americans. 96
3.9 Regional Manhattan plot focused on *ADAMTS9* within a 5Kb radius around the gene for nine anthropometric phenotypes. 97
3.10 Regional Manhattan plot focused on *ADAMTS9* within a 5Kb radius around the gene for six metabolic phenotypes. 98
3.11 Genome-wide scan for selection in Meso-Americans, Andeans and Mapuche Native American populations. 102
3.12 Worldwide allele frequencies of the top PBS SNPs detected in Native Americans within *ADAMTS9* and *NPAS2*. 107

4.1 Schematic representation of a selection post-admixture event. 113
4.2 Schematic of the new model used to identify variants under selection post-admixture in admixed populations. 120
4.3 Genome-wide scan of selection post-admixture in admixed Latin American populations. 124
4.4 Local ancestry deviation in BRA admixed population. 127
4.5 Local ancestry deviation in CHL admixed population. 127
4.6 Local ancestry deviation in COL admixed population. 128
4.7 Local ancestry deviation in MEX admixed population. 128
4.8 Local ancestry deviation in PER admixed population. 129
4.9 Signals of selection post-admixture at 10q22 in the Peruvian population. . . 132
4.10 Signals of selection post-admixture at the MHC region in the Mexican population. 134

5.1 Distribution of skin, hair and categorical eye pigmentation phenotypes in the CANDELA sample. 143
5.2 Quantitative eye pigmentation phenotypes examined in the CANDELA sample. 145
5.3 Summary of GWAS findings. 150
5.4 Meta-analysis for 6 index SNPs representing novel associations to pigmentation traits. 152
5.5 Worldwide allele frequencies of novel variants associated to skin pigmentation. 155
5.6 Regional association (LocusZoom) plots for SNPs in the five genomic regions showing novel genome-wide significant associations to pigmentation traits. . 156
5.7 Heatmaps of statistical interactions between the 18 index SNPs showing genome-wide significant associations to pigmentation phenotypes. 160

6.1 Distribution of allele frequencies for SNPs rs1800414 and rs74653330 at *OCA2* in East Asia. 166
6.2 Global allele frequency distribution of SNP rs885479 at *MC1R*. 167

6.3 Estimation of the start of selection and selection coefficient at *MFSD12* gene region. . . . 176
6.4 Evidence for selection in the *MFSD12* gene region. . . . 178

A.1 Correlation between maximum and mean PBS score at each gene region. . . 252
A.2 Distribution of anthropometric phenotypes in the CANDELA sample. . . . 253
A.3 Distribution of anthropometric phenotypes in Mexican volunteers from CANDELA sample. . . . 254
A.4 Principal Component Analysis (PCA) of admixed Mexican individuals from the CANDELA sample. . . . 255
A.5 Principal Component Analysis (PCA) of admixed Mexican individuals from the CANDELA sample. . . . 256
A.6 Correlation between selection statistics. . . . 257
A.7 Worldwide allele frequencies of the top PBS SNPs detected in Native Americans. . . . 258

B.1 Principal Component Analysis (PCA) of admixed Latin American individuals and continental reference panels. . . . 260

C.1 Continental ancestry in the CANDELA sample . . . 267
C.2 Distribution of Melanin Index variability in the CANDELA sample. . . . 268
C.3 Continental ancestry in the CANDELA sample . . . 269
C.4 GWAS quantile-quantile (QQ) plots of pigmentation phenotypes . . . 270
C.5 Genomic annotation in the 10q26 intergenic region around SNP rs11198112 271
C.6 Phenotypic effects (regression beta coefficients) and derived allele frequencies for the associated SNPs to pigmentation phenotypes in the CANDELA sample. . . . 272

D.1 Selection scans around candidate gene *SLC45A2* at SNP rs16891982 in Eurasian populations. . . . 277
D.2 Selection scans around candidate gene *IRF4* at SNP rs12203592 in Eurasian populations. . . . 278
D.3 Selection scans around candidate gene *EMX2* at SNP rs11198112 in Eurasian populations. . . . 279
D.4 Selection scans around candidate gene *TYR* at SNP rs1042602 in Eurasian populations. . . . 280
D.5 Selection scans around candidate gene *TYR* at SNP rs1126809 in Eurasian populations. . . . 281
D.6 Selection scans around candidate gene *GRM5* at SNP rs7118677 in Eurasian populations. . . . 282
D.7 Selection scans around candidate gene *OCA2* at SNP rs1800404 in Eurasian populations. . . . 283
D.8 Selection scans around candidate gene *OCA2* at SNP rs1800407 in Eurasian populations. . . . 284

D.9 Selection scans around candidate gene *OCA2* at SNP rs4778219 in Eurasian populations. 285
D.10 Selection scans around candidate gene *HERC2* at SNP rs12913832 in Eurasian populations. 286
D.11 Selection scans around candidate gene *SLC24A5* at SNP rs1426654 in Eurasian populations. 287
D.12 Selection scans around candidate gene *MC1R* at SNP rs885479 in Eurasian populations. 288
D.13 Selection scans around candidate gene *MFSD12* at SNP rs2240751 in Eurasian populations. 289
D.14 RMSE plots. 290
D.15 Joint estimation of the starting time of selection (T) and selection coefficient (s) at the *MFSD12* gene region. 291

Acronyms

1KG 1000 Genomes Project.

ABC Approximate Bayesian Computation.

aDNA ancient DNA.

AMR Ad Mixed ameRican (from the 1KG).

BMI Body Mass Index.

BCE Before Common Era.

BF Bayes Factor.

C Chroma (from the HCL color space).

CE Common Era.

CANDELA Consortium for the Analysis of the Diversity and Evolution of Latin America.

CI Confidence Interval.

cM centiMorgan distance.

DNA DeoxyriboNucleic Acid.

EHH Extended Haplotype Homozygosity.

EM Expectation Maximisation algorithm.

GO Gene Ontology.

GWAS Genome Wide Association Study.

H Hue (from the HCL color space)

HC Hip Circumference.

HCL Hue, Chroma, Lightness color space.

HGDP Human Genome Diversity Panel.

HLA Human Leukocyte Antigen.

IBD Identity by Descent.

IBS Identity by State.

iHS integrated Haplotype Score.

Kb Kilo-base pair.

Kya Thousand years ago.

L Lightness (from the HCL color space).

LD Linkage Disequilibrium.

LGM Last Glacial Maximum.

MAF Minor Allele Frequency.

MCMC Markov Chain Monte Carlo.

MHC Major Histocompatibility Complex.

ML Machine Learning.

MI Melanin Index.

mtDNA mitochondrial DNA.

nSL number of Segregating sites by Length.

PCA Principal Component Analysis.

PBS Population Branch Statistic.

QC Quality Control.

SGDP Simons Genome Diversity Project.

SNP Single Nucleotide Polymorphism.

WC Waist Circumference.

WES Whole Exome Sequencing.

WHR Waist to Hip Ratio.

WGS Whole Genome Sequencing.

XP-EHH Cross Population Extended Haplotype Homozygosity.

ya years ago.

Glossary

Adaptation
Heritable changes in genotype or phenotype that result in increased fitness.

Admixture
Gene-flow between previously isolated populations.

Allele
A variant form of a gene located at a specific location on a specific chromosome.

Bonferroni correction
When multiple hypotheses are tested, the Bonferroni correction to the overall desired significance level (α) is obtained by dividing it by the number of independent tests (k), so that each hypothesis is rejected if P-value $< \alpha/k$.

Fitness
A measure of the capacity of an individual to survive and reproduce.

Genetic drift
Changes in allele frequency in a population due to random sampling from generation to generation.

Genotype
Combination of alleles at a particular locus.

Genome Wide Association Study (GWAS)
A study of a genome-wide set of genetic variants in unrelated individuals to determine if any variant is associated with a phenoype.

Haplotype
A set of alleles that are inherited together from a single source.

Heritability
The proportion of phenotypic variation that can be attributed to any genetic variation (broad-sense heritability) or to additive genetic variation (narrow-sense heritability (h^2)).

IMPUTE
A software to infer unobserved genotypes using known haplotype information.

Linkage Disequilibrium
The non-random association of alleles at different loci.

Locus
The location of a gene (or of a significant variant) on a chromosome. Plural loci.

Meta-analysis
The combination of the results of multiple scientific studies that address the same, or similar, hypotheses.

Mutation
Permanent change in the nucleotide sequence of the genome of an organism.

Phase
The original allelic combinations that an individual received from its parents. When know or inferred this is referred to as phased data.

Population stratification
Refers to a situation in which the population of interest includes subgroups of individuals that are on average more related to each other than to other members of the wider population.

Polygenic adaptation
Refers to a situation in which adaptation occurs by simultaneous selection on variants at many loci (perhaps tens or hundreds or more) of (usually small) genetic effect.

Positive selection
Selection acting upon new advantageous mutations in a population.

Recombination
Genetic exchange of DNA segments between maternally and paternally inherited copies of a chromosome.

Selection
Process by which certain phenotypes become more prevalent in a population than other phenotypes resulting in a change in allele frequency over generations.

Selective sweep
Process by which a new advantageous mutation eliminates or reduces variation in linked neutral sites as it increases in frequency in the population.

SHAPEIT
A software to infer haplotype phase from genotype data.

SNP
Single Nucleotide Polymorphism. Single change in a nucleotide occurring at a locus in the genome.

Chapter 1

Introduction

1.1 Overview

Modern humans emerged in Africa approximate 200,000 years ago (ya). Although the exact routes of colonization remain controversial, by 15,000 ya modern humans had spread over all continents of the earth (with the exception of Antarctica). The different environments (raging from tropical to arctic, low-lands to high-lands and even toxic environments) that these early migrants encountered imposed new selective pressures that lead to novel adaptations. As these humans spread out of Africa, they also encountered extinct hominins with whom they interbred. It is now well established that non-African populations share on average between 1-2% Neanderthal ancestry (Green et al., 2010; Wall et al., 2013; Prüfer et al., 2014) and East Asians and Melanesians around 0.2% and 3% Denisovan ancestry (Reich et al., 2010, 2011; Meyer et al., 2012; Prüfer et al., 2014), respectively. Highly divergent haplotypes have also been found in African genomes, suggesting archaic introgression from an unknown population (Lachance et al., 2012). This archaic introgression event may have provided a faster rate of adaptation and many populations are now thought to have benefited from it (Huerta-Sánchez et al., 2014; Vernot and Akey, 2014; Sankararaman et al., 2014; Racimo et al., 2015; Deschamps et al., 2016; Vernot et al., 2016; Dannemann et al., 2017; Dannemann and Kelso, 2017; Racimo et al., 2017; Browning et al., 2018). At the end of the Neolithic revolution ($\sim 10,000$ ya) humans had transitioned from a hunter-gatherer to a more sedentary life style that included agriculture and pastoralism. This change in subsistence strategy, which involved the availability of a new dietary resource (mainly milk and its derivates), prompted a strong selection pressure for the ability to digest it. Consequently, lactase persistance into adulthood currently represents one of the strongest signals of selection in the human genome (Bersaglieri et al., 2004; Tishkoff et al., 2007; Gerbault et al., 2009; Gallego Romero et al., 2012; Schlebusch et al., 2013; Sverrisdóttir et al., 2014; Allentoft et al., 2015). However, this transition was also accompanied by a massive growth in population size that involved being in constant proximity to other humans and animals that lead to an increase and spread of infectious diseases (Bocquet-Appel, 2011). Infectious diseases are arguably one of the strongest selective forces exerted over human populations for over thousands of years and consequently, immune-related genes have been shown to be major targets of selection (Fumagalli et al., 2011; Karlsson et al., 2014). Although more controversial, there is now also increasing evidence that several physiological and life-history, and body size traits, such as age at

first birth in females, age of menopause, weight, and height, are associated with differential reproductive success in modern post-industrial societies (Nettle, 2002; Stearns et al., 2010; Sanjak et al., 2018). Overall, it is now clear that past and ongoing selective pressures have and are still shaping the genetic and phenotypic diversity of human populations.

Many of these documented adaptive events have been made feasible due to the advent of sequencing and large-scale genotyping technologies. Genome-wide scans for signatures of selection have now been conducted in modern and ancient human populations providing a fuller picture of local human adaptation (Voight et al., 2006; Grossman et al., 2010; Sabeti et al., 2007; Mathieson et al., 2015; Fan et al., 2016; Field et al., 2016a). Although the majority of genomic scans of selection have mainly been conducted as a means to explore human history (and molecular evolution in general), identifying candidate variants have also provided important and biologically meaningful information. The rationale for this observation is simple: variants in the genome that are under positive selection must be of functional importance, otherwise selection could not be acting on them (Nielsen et al., 2007). Consequently, current efforts of adaptation studies are now relying on the integration of phenotypic, functional and environmental data providing a better description of physiological impact and evolutionary history of candidate variants (Crawford et al., 2017a; Martin et al., 2017b; Key et al., 2018; Ilardo et al., 2018).

Throughout the course of this thesis I will present analysis on the adaptive history of Native Americans and admixed Latin Americans. I have performed this analysis primarily through the application of current and newly developed genomic tools for the analysis of high density SNP data. I have also integrated the results from my selection scan with available phenotypic data available for this same sample of Latin Americans, in order to better understand the phenotypic impact of the candidate variants found. Additionally, I have also conducted an association scan in the same sample of Latin Americans on pigmentation phenotypes, as an example of the evolution and discovery of novel variants affecting an adaptive trait. Specifically, my analyses have provided novel inferences regarding:

1. Signatures of adaptation in Native Americans and admixed Latin Americans.
2. Genetic determinants of skin and eye pigmentation in admixed Latin Americans.
3. The convergent evolution of lighter skin pigmentation in Eurasian populations.

I start this chapter with a brief description of types of selection, current statistical approaches for detecting instances of positive selection, and the main limitations and challenges of their use. I then present the evolutionary and demographic history of the Americas focusing on the peopling of the continent and the admixture process that shaped the current genomic make-up of modern Latin Americans. I then describe previous studies on detecting selection in Native Americans and admixed Latin Americans. I also briefly describe the biological basis of human pigmentation variation, followed by its evolutionary history and the main genetic variants affecting its variability. Finally, I describe the main methodological aspects of a Genome Wide Association Study (GWAS) and a detailed

description of the sample used for the upcoming analysis.

1.2 Natural selection

Natural selection as defined by Darwin and later elaborated by Fisher is the differential reproduction of individuals with distinct genotypes in subsequent generations (Hartl et al., 1997; Jobling et al., 2013). In other words, individuals with different genotypes will have a differential ability to survive and reproduce in different environments. In order to contribute genetically to the next generation, an individual must survive until reproductive age and then reproduce. Thus, selection can occur at any stage from the formation of an individual's genotype during fertilization to that individual generating their own progeny. The ability of an individual's genotype to survive and reproduce is defined as its fitness, which partly depends on the environment. When modeling natural selection, fitness can simply be represented as the probability to survive and reproduce (Relethford, 2012). The quantity of interest here, is the proportional change for each genotype from one generation to the next, which is known as the absolute fitness (W). However, when considering the evolution under natural selection of different genotypes, rather than considering the specific absolute fitness value it is more important to consider the absolute fitness of a genotype relative to the other genotypes (Relethford, 2012). This value is known as the relative fitness (w), which expresses (by convention) the absolute fitness relative to the fittest genotype. Typically, relative fitness w has a subscript to refer to the different genotypes. In the case of a bi-allelic variant with alleles A and a, the symbol w_{AA} is used to refer to the relative fitness of genotype of AA. The relative fitness is simply computed by dividing each absolute fitness value to the highest absolute fitness value, which sets the highest relative fitness to 1. Relative fitness is usually expressed in terms of the selection coefficient (s), which measures the fitness advantage or disadvantage of a particular genotype. By defining the relative fitness in terms of the selection coefficient, it is easier to model different types or modes of natural selection by assigning specific selection coefficient to particular genotypes, some of which are discussed below.

1.2.1 Types of selection

Natural selection can act in different ways. Random mutations are more likely to be deleterious rather than beneficial, and thus the majority of novel variants are removed from the gene pool. These type of mutations that reduce the fitness of an individual are subject to negative selection (also called purifying selection). The ongoing process of negative selection is referred to as background selection, producing long stretches of conserved genomic regions, as many linked variants with the non-beneficial mutations are also removed (Charlesworth et al., 1995; Hudson and Kaplan, 1995). Alternatively, mutations can also result in beneficial variants, in which the allele is favoured by positive selection and so increases in frequency in consecutive generations (Mitchell-Olds et al., 2007). The dynamics of the selection process in diploid organism such as humans, will not only depend on the advantage or disadvantage of individual alleles, but on their interaction, which in turn afffects the efficacy of natural selection to act on specific genotypes. One such type of selection

is balancing selection, in which both alleles (in a diploid organism) are maintained in the population (Richman, 2000). This process may happen during overdominance selection (also known as heterozygote advantage) creating balancing polymorphism. Alternatively, underdominance selection will operate when novel alleles reduce the fitness of heterozygotic individuals. Additionally, frequency dependent selection, in which the frequency of the genotype will determine its advantage, can also result in balanced polymorphism. If the alleles being maintained in the population result in opposing phenotypic effects, then this phenomenom is called diversifying (or disruptive) selection. By contrast, if the intermediate phenotype values are favoured the phenomenom is called stabilizing selection.

Despite the diversity of different types of selection processes, the majority of research has been focused on the development of methods to detect instances of positive selection (Fan et al., 2016). One reason is practical, as detecting positive selection can be easier to detect due to its more conspicuous signature left on the genome (Vitti et al., 2013) (see Section 1.2.2). Another reason might be that positive selection is assumed to be the main driver of local adaptation, particularly in human history (Figure 1.1) (Vitti et al., 2013; Fan et al., 2016).

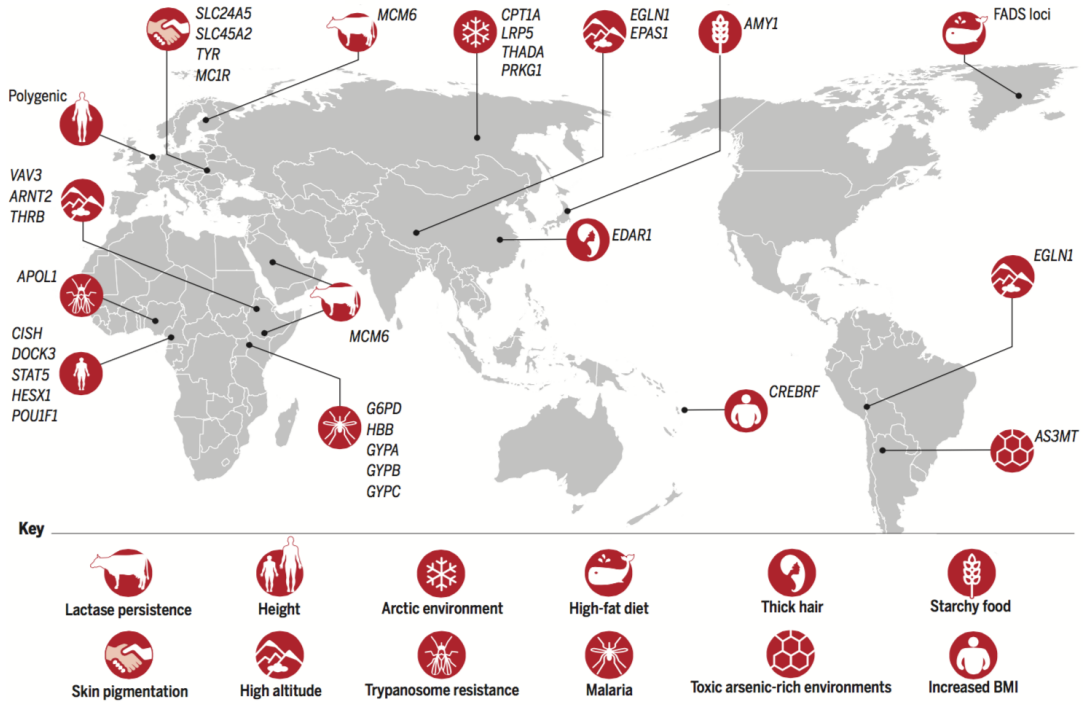

Figure 1.1: Examples of recent human local adaptation. Each example includes the candidate gene under selection, the phenotype and/or the selective pressure. From Fan et al. (2016).

1.2.2 Detecting positive natural selection

When a beneficial mutation is subjected to positive selection it rises up rapidly to high frequency or fixation (i.e. 100% prevalence) within a population. Nearby linked sites tend also to rise up in frequency, a process called "genetic hitchhiking". Depending on its

selection coefficient, the effect of selection can be faster than the ability of recombination and mutation to break down the haplotypes carrying the beneficial mutations, drastically reducing the genetic variability — a hallmark of a selective sweep (Figure 1.2). In the most simple scenario a selective sweep occurs when positive selection acts on a novel mutation, i.e. a "hard sweep" (Figure 1.2A). The beneficial mutation increases in frequency and so does the haplotypic background it is associated with, thereby reducing the genetic variation at this region in the population. A so-called "soft sweep" (Hermisson and Pennings, 2005; Pennings and Hermisson, 2006a,b) describes two slightly different scenarios (Figure 1.2B). In one scenario, independent mutations at a single locus can be subjected to positive selection and increase in frequency simultaneously. If the alleles are similarly advantageous none of them would fix during the selective event. In the second scenario, due to a change in selective pressure, a variant that was previously segregating in the population becomes advantageous and rises in frequency. Since the beneficial variant is usually present in distinct haplotypic backgrounds, this type of selective event does not result in a complete loss of genetic variation at the region. Thus, a "soft sweep" tends to be more difficult to detect than a "hard sweep". A third type of positive selection, termed polygenic adaptation, refers to a process in which a population adapts through small changes in allele frequencies at several loci. It is important to note that this does not mean that a polygenic phenotype can only be targeted by natural selection via a polygenic adaptative event, as any single locus (e.g. one with a strong genetic effect) can also be targeted by a selective sweep, and consequently affect the evolution of a polygenic phenotype. (Figure 1.2C). Importantly, this process is in line with classical models of natural and artificial selection in quantitative genetics, where it is assumed that most traits are polygenic i.e. controlled by many loci (Falconer, 1960). The common scenario of a polygenic adaptative event as envisaged by Pritchard et al. (2010) involves a quantitative phenotype that is affected by many alleles, each with a small genetic effect. Under a novel environmental pressure, selection might favour a new phenotypic optimum and consequently the population will adapt by allele frequency shifts at all loci controlling the phenotype. Importantly, this allele frequency shift will not necessary push alleles upwards in frequency. The shift in allele frequency is directional, favouring (or disfavouring) alleles towards the new phenotypic optimum. The result will be that not many alleles will reach fixation and therefore, such events are hard to detect using classical methods for selective sweeps.

A fourth process that can result in positive selection is gene-flow resulting from inter- or intra-species admixture events (Figure 1.2D). In this scenario, following an admixture event the beneficial mutation coming from one of the source populations rises up in frequency in an haplotypic background not present in the receiving population. In the specific case of human evolutionary genetics when the admixture event involved two populations with an old divergence time (such as between modern humans and archaic hominins), this process has been termed adaptive introgression (Racimo et al., 2017). In the case where this admixture event happened between human populations that have not exchanged migrants for several generations, such as the case between Africans, Europeans and Native Americans in the formation of modern admixed Latin American populations, the process has usually been referred to as selection post-admixture (Tang et al., 2007).

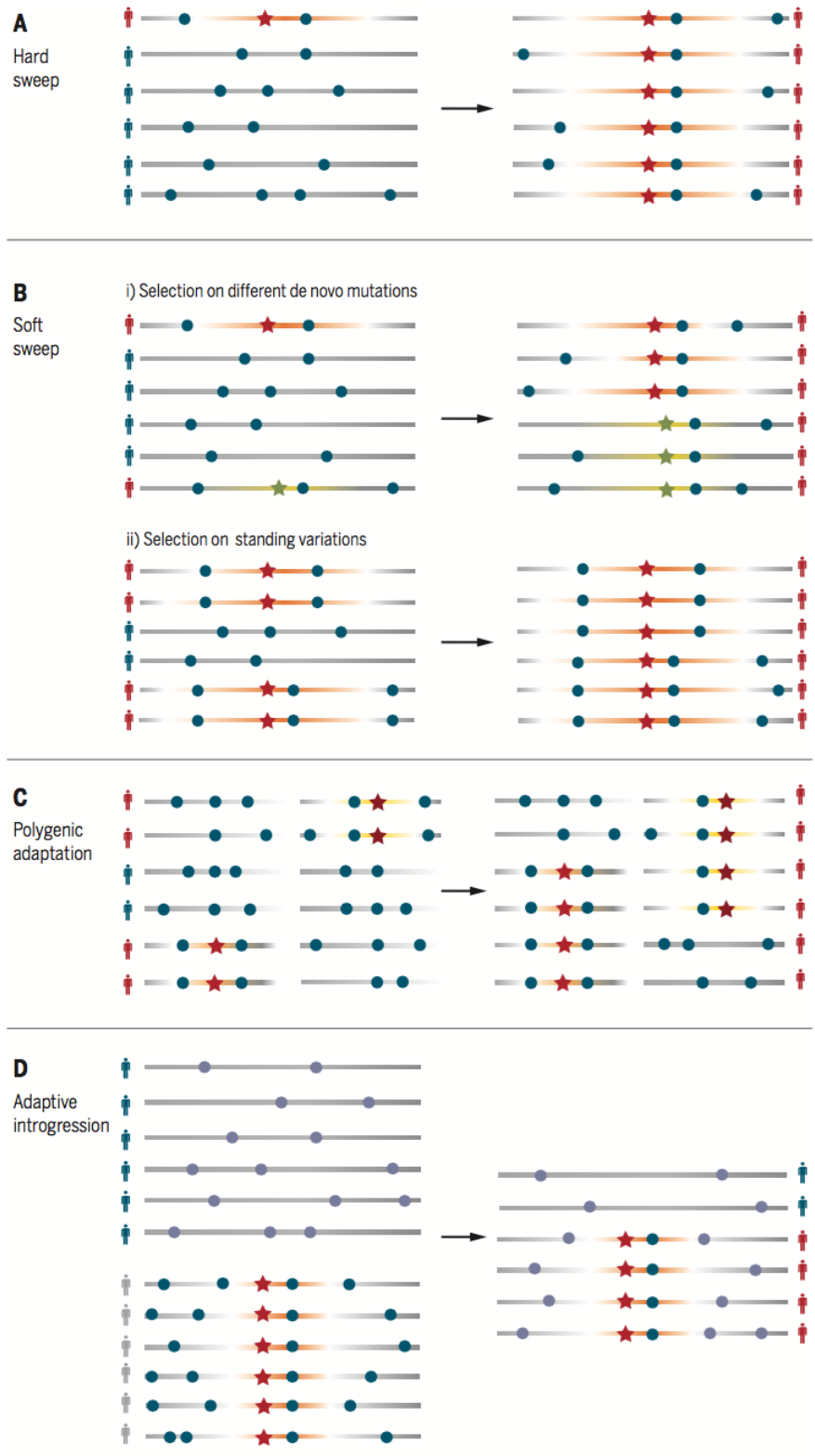

Figure 1.2: Schematic representation of genomic signatures of positive selection. A) Hard sweep scenario in which a novel mutation is selected and rises up in frequency. B) A soft sweep scenario in which selection is acting on i) two different de-novo mutations or ii) standing mutations segregating in the population. C) A polygenic adaptation scenario in which many loci associated to a particular phenotype are under selection. D) An adaptive introgression scenario in which a novel mutation present in a different population is selected after the admixture event. Each horizontal bar represents a haplotype .The orange segment is graded to indicate the strength of LD between beneficial (stars) and neutral (dots) loci. From Fan et al. (2016).

1.2.3 Challenges to detecting selection

Although the explosion of genomic data generation and of novel statistical tools to search for evidence of selection has allowed us to discover many instances of past adaptive events (Vitti et al., 2013; Fan et al., 2016), one of the consequences of this revolution has been a shift from hypothesis testing to a hypothesis generation framework (Vitti et al. 2013). This has resulted in the majority of selection statistics being evaluated under an outlier approach, where the significance is assessed by comparing the selection statistic score to its genome-wide distribution. This approach has the benefit of accounting for demographic processes that could mimic selective signals, as true selection events are expected to act only on specific regions of the genome (Oleksyk et al., 2010; Vitti et al., 2013). However, choosing an adequate threshold, such as defining a percentile, remains an arbitrary choice among studies and there is not an established consensus. A second approach involves constructing a neutral model of the demographic evolution of the population being studied, and assessing the significance of the observed selection score to the scores obtained from simulations under neutral evolution. This requires a good knowledge of the demographic history of the target population. Additionally, there will be a possibility that a demographic neutral model that was not considered could be a better fit to the data, leading to an incorrect conclusion. Another limitation that is likely to be pervasive among genomic scans of selection, is that of ascertainment bias as the majority of genomic data is still based on SNP data. This can result in a biased representation of the true genetic diversity of the tested population, as the majority of SNPs will fail to capture rare variants, affecting inferences on populations that are largely underrepresented in genomic studies, leading to spurious or low-powered inferences. Perhaps the most challenging aspect of establishing instances of adaptation involves not only detecting a true adaptive locus, but also determining the phenotype under selection and ideally, the selection pressure driving the adaptation (Fan et al., 2016). In this regard, with the advent of "polygenic scores", which aim to predict an individual's phenotype using GWAS data, recent studies have provided important clues regarding the adaptation of complex phenotypes, with perhaps the clearest example being that of selection for height in Europe (Turchin et al., 2012; Berg and Coop, 2014; Robinson et al., 2015; Zoledziewska et al., 2015; Berg et al., 2017; Racimo et al., 2018b; Guo et al., 2018). Specifically, these studies have found that polygenic scores for height increase from a south-to-north gradient in Europe. Notably, two preprints recently showed that these previously reported signals of selection on height, either did not replicate or are significantly lower when using effect sizes obtained using the UK-Biobank sample (Bycroft et al., 2018), which constitutes a much larger and homogenous cohort than those previously used for these type of analyses (Berg et al., 2018; Sohail et al., 2018) (Figure 1.3). These two studies thus highlight a need of caution when conducting analyses of polygenic adaptation using polygenic scores across populations (Novembre and Barton, 2018). Finally, perhaps the ultimate assessment to understand and complement a study of selection, will more likely involve some type of functional analysis e.g. in vitro cell lines or model organisms (Vitti et al. 2013; Fan et al., 2016). However, this approach will also have to be carefully assessed, as many variants are likely to possess a pleiotropic effect and therefore the screening of many phenotypes will be needed before concluding the effect on

a specific phenotype. Assessing the underlying selective force will also likely involve the integration of environmental data. In this case, it will also be necessary to assume both that the environmental variable used represents a good proxy for the past environmental variable and that the selective pressure has not in fact been driven by another highly correlated environmental variable.

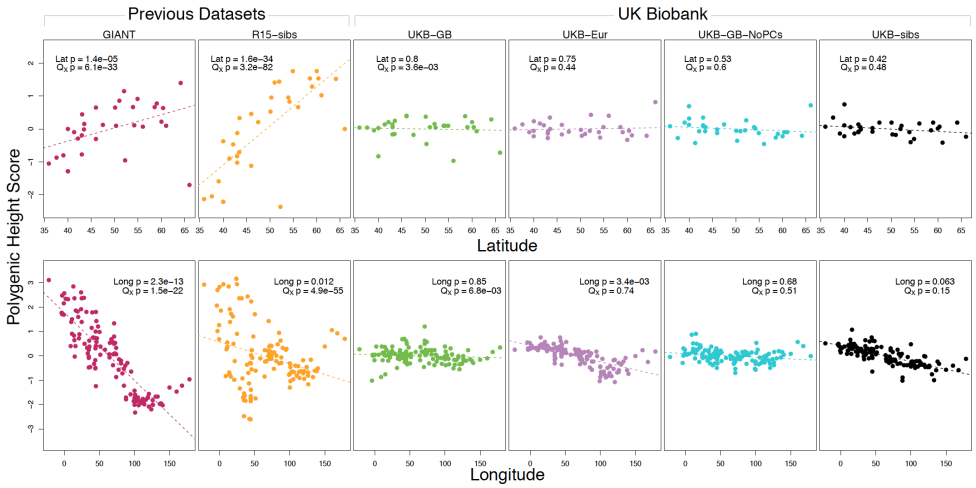

Figure 1.3: Polygenic scores across Eurasian populations for different GWAS data sets. The top row shows polygenic scores of height obtained in European populations from a combined dataset using the 1000 Genomes Project (1000 Genomes Project Consortium et al., 2015) and Human Origins populations (Patterson et al., 2012) plotted against latitude. The bottom row shows polygenic scores of height obtained in Eurasians populations from the same two datasets plotted against longitude. While there is a strong significantly correlation between polygenic scores of height using datasets previously used to report adaptation of height in humans across geographical clines (first two panels), the correlation is absent or greatly attenuated in the much more homogeneously designed UK-Biobank study dataset (third to fifth panel). From Berg et al. (2018).

1.3 Genetic history of the Americas

In order to understand the selective pressures that could have shaped the genetic diversity of the present-day inhabitants of the Americas, it is important to understand the evolution and recent history of the Americas. In this section I briefly describe recent genomic studies regarding the peopling of the Americas. The major migratory events to the Americas described in the text are represented in Figure 1.4. I then describe genomic studies of admixed Latin American populations. Finally, I describe recent results on the population structure of the Consortium for the Analysis of the Diversity and Evolution of Latin America (CANDELA) sample from Chacon-Duque et al. (2018), which is the main sample used in this thesis.

1.3.1 Genetic history of Native Americans

The American continent represents the last major landmass to have been settled by people migrating from the northeastern tip of Asia through Beringia — a land bridge connecting Siberia to Alaska (Reich et al., 2012; Raghavan et al., 2015). Archeological data shows that northern eastern Siberia was occupied by at least 28,000 ya (Nikolskiy and Pitulko, 2013), a period that coincided with the Last Glacial Maximum (LGM) marked by drastic climatic conditions and glacial barriers across northern latitudes (Hoffecker et al., 2016). This observation led to the proposal that these early colonisers might have been isolated for extended periods of time with limited dispersal across the region — a hypothesis known as the Beringian Standstill or Beringian Incubation model (Tamm et al., 2007). Only after the end of the LGM did the ice retreat from parts of the Pacific coast (circa. 16,000 ya), raising the possibility for a coastal migration and later an ice-free corridor through the center of the continent that permitted a second route for colonization to the Americas (Heintzman et al., 2016). The first unambiguous evidence of presence in the Americas dates from around 14,000 to 15,000 ya including occupation in southern Chile around 14,000 ya (Guidon and Delibrias, 1986; Dillehay and Collins, 1988; Parenti et al., 1990). Together, these studies indicate that the colonization of most of the Americas was a rapid process that could have happened in less than 2,000 years (Tamm et al., 2007; Brandini et al., 2018). Current genomic efforts are focused on estimating whether and for how long early Native Americans were isolated in Beringia, the number of migrations and timings of the entry to the Americas, the timing of the divergence between Native Americans groups, and additional genetic affinities with other populations apart from Northern East Asians (Wang et al., 2007b; Reich et al., 2012; Raghavan et al., 2015; Skoglund et al., 2015; Skoglund and Reich, 2016; Moreno-Mayar et al., 2018).

Genomic studies based on mitochondrial DNA (mtDNA) were the first to show evidence of the profound founder effect in Native Americans (Wallace et al., 1985). These studies showed that only a limited number of individuals contributed to the genomic make-up of most modern Native American groups. The estimated time to the most recent common ancestors (TMRCA) from the five observed founding maternal lineages have been dated to around 15,000 to 18,000 ya, indicating a probable bottleneck around that time (Torroni et al., 1992; Horai et al., 1993; Torroni et al., 1993). Subsequent mtDNA studies have also confirmed that the ancestors of Native Americans most likely paused when they reached Beringia, a period that lasted many thousands of years with little shared gene-flow of mtDNA from non-Native Americans, during which the mtDNA lineages differentiated from their East Asian sister-clades (Tamm et al., 2007). In contrast to mtDNA studies, Y-chromosome studies have been more challenging due to the drastic reduction of indigenous Americans Y-chromosome lineages as a consequence of the Spanish conquest,which was characterized by a male-bias in gene-flow (Lell et al., 1997; Bianchi et al., 1998; Karafet et al., 1999; Ruiz-Linares et al., 1999; Bortolini et al., 2003). Nonetheless, Y-chromosome studies have found 2 major lineages that account for 6% and 75% of Native American Y chromosomes, respectively (Underhill et al., 1996; Karafet et al., 1997; Lell et al., 1997; Bianchi et al., 1998; Karafet et al., 1999; Bortolini et al., 2003; Zegura

et al., 2004; Battaglia et al., 2013). Given the highly distinct frequency of these founding haplogroups and the difference in geographic distribution of their sub-clades, it has been suggested that their entry occurred at different times. The most prevalent haplogroup is thought to have arisen 24,5000 ya (Zegura et al., 2004), whereas the less prevalent seems to to have originated 7,100 to 16,700 ya (Karafet et al., 2002). More recent Y-chromosome studies have also confirmed that the differentiation of these major Y haplogroups most likely occurred in Beringia (Battaglia et al., 2013).

In comparison with studies based on uni-parental markers, approaches using autosomal DNA are now starting to provide a finer resolution of the history of Native American populations. However, it still remains unknown whether one or more early migrations gave rise to the founding population of Native Americans (Wang et al., 2007b; Reich et al., 2012; Raghavan et al., 2015; Skoglund et al., 2015; von Cramon-Taubadel et al., 2017; Moreno-Mayar et al., 2018). The first comprehensive study based on $\sim$ 700 microsatellite genotype data included over 400 individuals representing 24 Native American groups (Wang et al., 2007). This study found evidence that supported a single main colonization event from Siberia. Additionally, they suggested an scenario in which coastal routes facilitated migration of early Native American founders in comparison with inland routes, and a partial agreement between the genetic similarity and the linguistic classification of Native American groups. A second major study based on genome-wide SNP data from over 50 Native American populations suggested that Native Americans descended from at least 3 migratory events coming from Northern East Asia (Reich et al., 2012). Specifically, Reich et al. (2012) showed that Native Americans from Central and South America descended from a single ancestral population. Eskimo-Aleut speaking populations from the Arctic however, seemed to have inherit almost 50% of their ancestry from a second migratory event from East Asia, and the Athabaskan-speaking Chipewyan populations were estimated to harbour approximately 10% of ancestry from a third migratory event. Additionally, they suggested that the peopling of the Americas most likely followed an initial southward expansion along the coast with a subsequent split and little gene flow after the divergence, especially in South America, as suggested previously by Wang et al. (2007). However, a third major study from Native American populations based on WGS data claimed that Native American populations descended most likely from two migratory events (Raghavan et al., 2015). Similar to Reich et al. (2012) the authors found evidence that Inuit populations originated from a separate migration, but concludad that Northern Native Americans including Athabaskan-speaking populations and Central and South American Native Americans descended from a single migration that occurred no later than 23,000 ya (Raghavan et al., 2015). The authors further dated the divergence time between these northern and southern Native Americans groups to about 13,000 ya, most likely within the American continent. Another recent study with contradictory results to these two previous studies further tested whether all groups of Central and South American descended from a single migration event by analyzing novel genomic data from previously uncharacterized Amazonian populations from Brazil. Skolund et al. (2015) detected a strong signal linking these Amazonian populations to present-day Melanesians,

New Guineans and Andaman Islanders, and thus providing evidence for two founding lineages for Central and South America. By carrying different types of modelling analysis, the authors further showed that the patterns of variation could be explained by approximately 2% admixture from Australian related populations or alternatively, from a larger admixture component between 2 to 85% of ancestry from a population that existed in a substructured North East Asia population.

Novel findings relevant for the history of Native American populations have also arisen from studies of ancient individuals from the Americas and Siberia. One of the most important findings come from the remains of a child associated with the Clovis culture in western Montana known as Anzick-1 and directly dated to 12,600 before present (BP) (Rasmussen et al., 2014). The authors showed that this individual was more closely related to Central and South Americans than to some North American populations. This result suggested that the present day structuring of Northern and Southern (i.e. Central and South American) lineages dates back to at least 12,600 years (Rasmussen et al., 2015). Another major ancient DNA study also revealed important observations regarding the major admixture process that lead to the formation of the ancestral populations of Native Americans (Raghavan et al., 2014). The analysis of a 24,000 year old individual from Central Siberia showed affinities to both European (West Eurasian) and Native American populations, but without close affinities to East Asian groups (Raghavan et al., 2014). The authors estimated that up to 38% of Native American ancestry may have originated through gene flow from this ancient population which this Central Siberian individual was part of. Additionally, in line with what has been observed by Reich et al. (2012) and Rasmussen et al. (2015), genomic evidence from a tuft of hair dated to 4,000 ya from Greenland showed that the population which this individual belonged to (termed Paleo-Eskimos) had migrated from Siberia to the Northern America Arctic region independently of the Inuit migration, but that it was largely replaced by the Inuits approximately 700 ya (Rasmussen et al., 2010). Thus, excluding the Inuit population, it seems that the genetic evidence seems points to a highly structured Beringian population that contributed to the formation of present day Native Americans (Skoglund et al., 2016). These comprise populations related to East Asians, populations related a Central Siberian population, and populations with ancestry related to present day Australo-Melanesians and Andamanese Islanders (Skoglund and Reich, 2016). The existence of this highly structured Beringian population was further confirmed by a recent study of one ancient genome from the Upward Sun River in Alaska dated to about 11,500 ya (Moreno-Mayar et al., 2018). The first mtDNA genetic data from two individuals from the same site (only one individual was analyzed by Moreno-Mayar et al. (2018)) showed that the two different mitochondrial lineages were not typical of the modern people inhabiting this region (Tackney et al., 2015). The authors hypothesised that this population might represent the descendants of an ancient Benringian population, but were not able to test this hypothesis without autosomal data. Moreno-Mayar et al. (2018) showed that this individual is basal to all ancient and modern Native Americans and therefore represents an ancient Beringian population

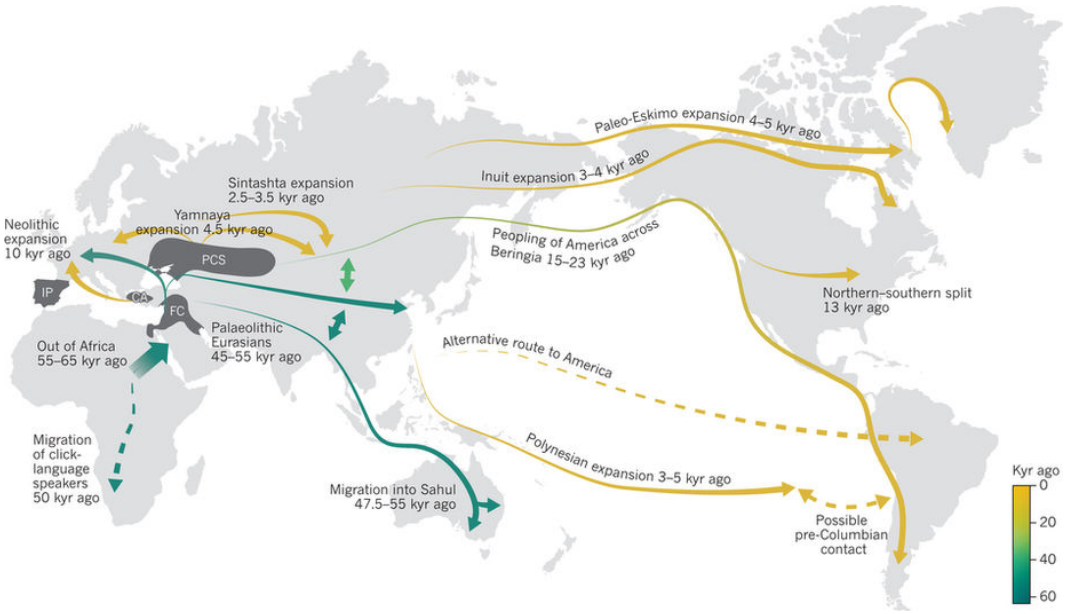

Figure 1.4: Major human migrations across the world inferred through genomic data. Migration routes to the Americas as described in the text that are highly accepted are represented in solid lines and more controversial in dashed lines. Abbreviations: CA, Central Anatolia; FC, Fertile Crescent; IP, Iberian Peninsula; PCS, Pontic-Caspian steppe. From Nielsen et al. (2017).

1.3.2 Genetic history of admixed Latin Americans

The genetic history of current admixed Latin American populations stems mainly from the encounter of Native American, Europeans (mainly from the Iberian Peninsula) and Western Africans. Estimates of the total size of the Native American populations at the time of the arrival of first European "conquistadores" vary between 10 to 100 million, but most place it around 40 million (Sanchez-Albornoz, 1974; Thornton, 1987; Denevan, 1992). Although definite estimates are not currently available for the sizes of different Native American populations, it is likely that they varied greatly across the continent reflecting the highly heterogenous mode of subsistance and social organization that ranged from densely populated urban areas (such as the Inca or Aztec empires) to scarcely populated regions occupied mainly by hunter-gatherers (such as the Amazons and parts of Patagonia and of North America) (Bellwood et al., 2007) (Figure 1.5). The European colonization that was set in motion after Columbus's arrival in America was mainly of Spanish and Portuguese origin and was only later followed by other European nations such as the British and French, and to a lesser extent the Dutch. It has been estimated that during the colonial period up until the 19th century, approximately half a million Spanish and half a million Portuguese migrated to the Americas. This migration was predomenantly characterized by males, which comprised up to 80% of the total migration. The total number of British, French and Dutch immigrants that arrived in their respective American colonies is though to have been smaller and estimated to be around 1 million. In contrast to the Iberian migration, they also included a considerable number of families. The trans-Atlantic slave trade that caused the introduction of Africans into the Americas is thought to have been mainly prompted by the strong decline of the indigenous Native

American populations (Curtin, 1969; Thomas, 1997). It has been estimated that up to 90% of the indigenous population perished after the arrival of the first Europeans, with highest decline in areas with small Native American population densities such as the Antilles and parts of North America. The number of Africans introduced to the Americas is thought to have been around 10 million, with great heterogeneity among the different colonies (Curtin, 1969) (Figure 1.6). Highest numbers of Africans are thought to have arrived in Brazil (42%), followed by the British (25%), the Spanish (15%), and the French colonies (14%). After the independence of the majority of American countries, the abolition of the slave trade stopped the massive African flow, but there was still a continued migration of over 100 million individuals from several European countries, mainly to the US. Several million of Europeans also settled in former Ibero-American colonies, particularly in South American countries such as Chile, Argentina, Uruguay, and Brazil (Baily and Miguez, 2003; Kent, 2016). The majority of these were of Spanish and Portuguese origin, followed by Italians and Germans. In addition to Europeans, vast number of migrants from Asian countries also migrated to the American continent, with the vast majority of them settling in the US, although non-negligible numbers of Japanese and Chinese migrants also settled in Brazil and Peru, respectively (Baily and Miguez, 2003).

The encounter of Native Americans with Europeans and Africans throughout the continent lead to extensive admixture. Early mtDNA and Y-chromosome studies in Latin America showed a contrasting pattern where most of the paternal ancestry could be traced back to Europeans, whereas the maternal ancestry was mainly of Native American or African origin (Bedoya et al., 2006; Carvajal-Carmona et al., 2000, 2003; Wang et al., 2008) (Figure 1.7). This result points to a strong sex-bias gene flow that involved mainly men of European origin and Native American or African women, consistent with the documented historical records. Following the results using uni-parental markers, genome-wide analysis also identified an enrichment of Native American or African ancestry in the X-chromosome in comparison with the autosomes, consistent with a gene flow sex-bias (Wang et al., 2008) (Figure 1.7). These more recent studies based on high-density genome-wide data also showed the extensive variation in Native American, European and African ancestry between as well as within countries throughout Latin America (Figure 1.8). Thus, far from being a genetically homogenous population, Latin Americans have extensive population structure, with individual admixture estimates frequently ranging from 0 to 100% of the three main continental ancestries. Estimates of the time since admixture also illustrate this complex and varying admixture process in Latin America. For example, analysis of populations from the Caribbean populations estimated an admixture event approximately 16 generations ago (ga) (i.e. about 500 years considering 30 years per generation) for the island populations, and about 13 ga for the continental populations (i.e. about 400 years considering 30 years per generation), consistent with historical records (Moreno-Estrada et al., 2013). Other genetic analysis in Latin Americans populations, also demonstrated extensive gene-flow for longer periods of time (Bedoya et al., 2006), or multiple gene-flow events (Kehdy et al., 2015; Homburger et al., 2015; Chacon-Duque et al., 2018).

In addition to estimating the varying ancestral proportions of Native American, Eu-

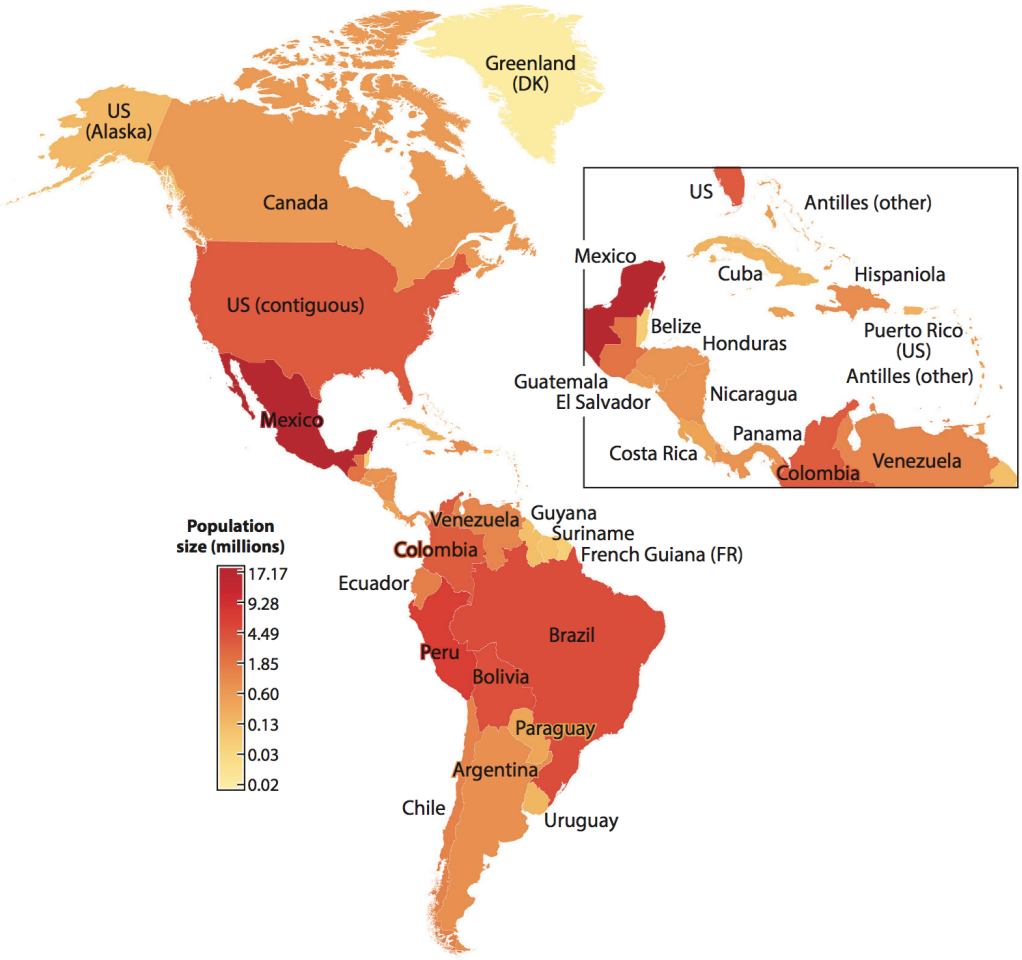

Figure 1.5: Estimated size of the Native American population at the time of Columbus's first landing on the Americas. From Adhikari et al. (2017).

ropean or African ancestry, high-density SNP genome-wide data and better approaches that leverages the correlation between SNPs (i.e. use of haplotypes instead to independent SNPs) have permitted the inference of fine-scale population structure (i.e. sub-continental ancestry), producing novel results. For instance, Moreno-Estrada et al. (2014) showed that the Native American component in individuals from the south east of Mexico has highest genetic affinity to the Maya, whereas in individuals from central Mexico the highest genetic affinity is to the Nahua. Similarly, Homburger et al. (2015) explored the sub-continental population structure in five Latin American countries (Colombia, Ecuador, Peru, Chile and Argentina) and found a strong gradient of Native American ancestry that was associated with the geographical location of local indigenous populations. Furthermore, their analysis of the European ancestry showed that while most of the European ancestry is from the Iberian Peninsula, many individuals showed highest affinity to Italy, especially those in Argentina. Additionally, small levels of East Asian ancestry in the Peruvian samples were detected that had not been reported previously. In Brazil, a similar analysis of various populations throughout the country showed that, for example, the African component has highest genetic affinity to Western African populations (Kehdy et al., 2015).

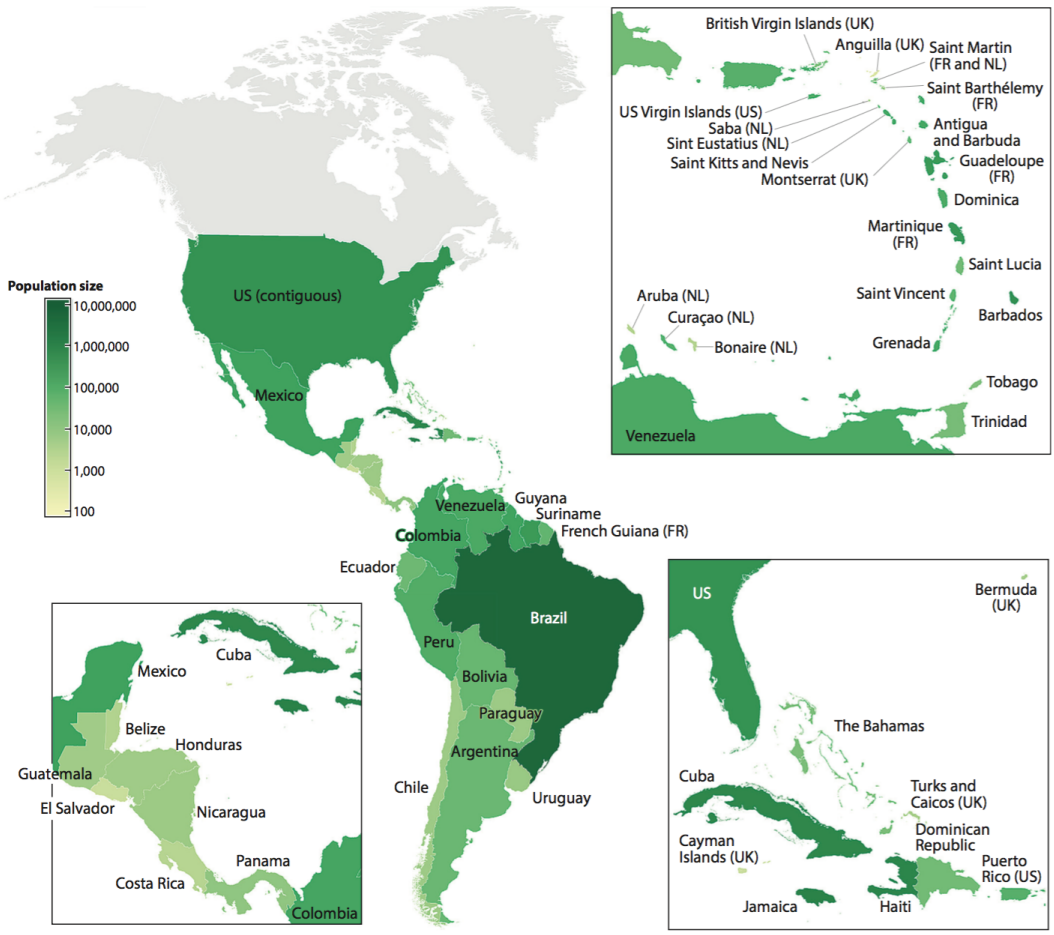

Figure 1.6: Estimated number of African slaves transported to the American continent. From Adhikari et al. (2017).

Finally, a recent analysis of the CANDELA cohort (Section 1.8), coupled with a novel haplotype-based method, permitted the detection of further sub-continental ancestry patterns (Chacon-Duque et al., 2018). Similar to other studies, pre-Columbian Native American genetic structure was associated to the local indigenous populations. However, the authors also detected a Sephardic ancestry component across Latin America, that they hypothesised steamed mainly from the clandestine colonial migration of Christian converts of non-European origin, known as *Conversos*. Estimates of Sephardic ancestry present in Latin American individuals were higher than those present in the sample of modern Spanish individuals, which suggested the migration of individuals with higher levels of that ancestry to the New World (Chacon-Duque et al., 2018). Interestingly, this period coincides with the known expulsion of Jews from Spanish territory. Although this migration must have been clandestine, as migration to the colonies was forbidden to *Conversos*, historical records have documented that a low number of individuals did arrive to the Americas. This result not only supports these records, but also suggests that this was substantially more prevalent that previously thought (Chacon-Duque et al., 2018). Altogether, these observations emphasize the high level of fine-scale population structure present across Latin America that is just starting to be explored and understand with

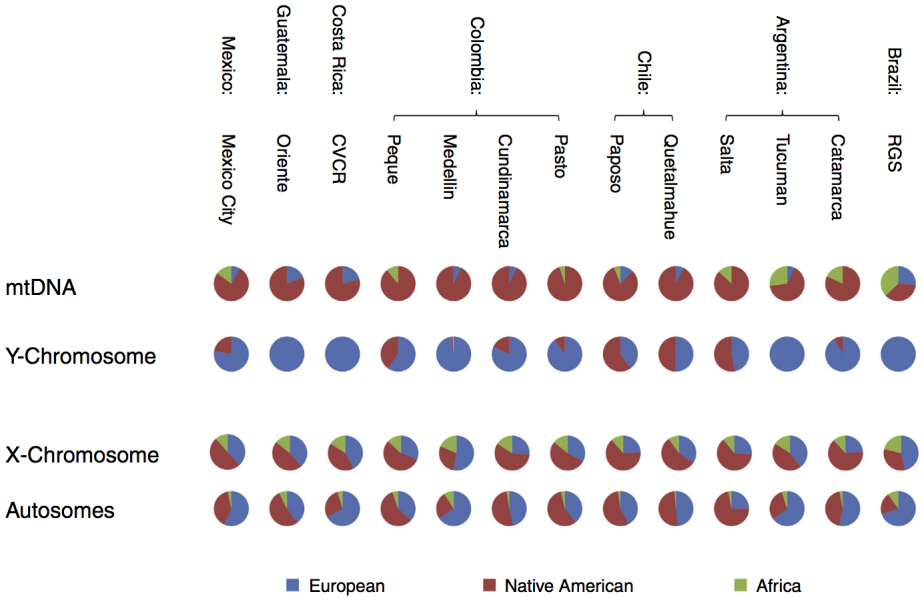

Figure 1.7: Proportion of African, European and Native American ancestry estimated with mtDNA, Y-chromosome, X-chromosome and autosomal data in thirteen Latin American populations. Native American (and African ancestry) is higher in the mt-DNA. Conversely, European ancestry is higher in the Y-chromosome. This pattern is consistent with a sex-bias admixture process involving mainly European men and Native American or African women. Consistent with this pattern the amount of Native American and African ancestry is higher in the X-chromosome and lower in the autosomes. From Adhikari et al. (2017). Adapted from Wang et al. 2008

more comprehensive sampling efforts and more sophisticated genomic approaches.

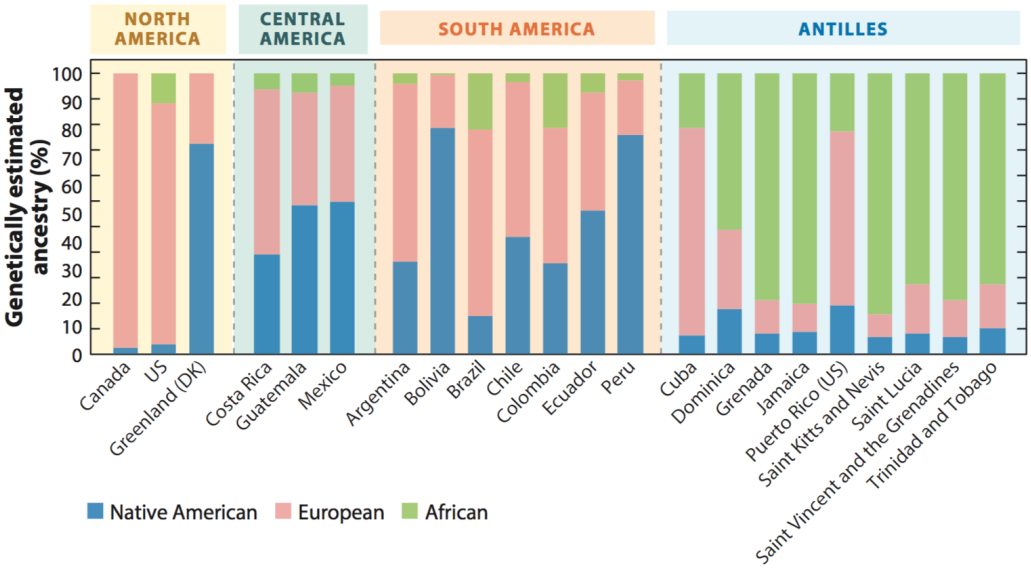

Figure 1.8: Proportion of African, European and Native American ancestry from samples from countries and dependencies across the American continent. There is a great variation of continental ancestries between different population throughout the Americas. From Adhikari et al. (2017).

1.4 Recent human adaptation

In the next section I describe past efforts of detecting adaptive events conducted exclusively in Native Americans and on admixed Latin Americans. The review of this literature will serve as a basis for the analysis on Native Americans conducted in Chapter 3 and on admixed Latin Americans conducted in Chapter 4.

1.4.1 Previous studies on detecting selection in Native Americans

Compared to other human populations, there are currently fewer studies that have explored signals of natural selection in Native Americans (Fan et al., 2016). An exception, are the populations of the Andean altiplano, which have been the focus of extensive research to understand the biological basis for high altitude adaptation (Beall et al., 1997; Bigham et al., 2009, 2010; Zhou et al., 2013a; Bigham et al., 2014; Eichstaedt et al., 2014; Foll et al., 2014; Eichstaedt et al., 2015a; Valverde et al., 2015; Fehren-Schmitz and Georges, 2016; Bigham, 2016; Crawford et al., 2017a). One of the genetic pathways that have been extensively implicated is the Hypoxia Inducible Factor (HIF) a regulator involved in the activation of genes responsible for cellular hipoxia, although other non-HIF candidate genes have also been reported in several Andean populations (Bigham et al., 2009, 2010; Zhou et al., 2013a; Eichstaedt et al., 2014, 2015a; Crawford et al., 2017a). Other candidate gene loci in Andean populations have also been linked to growth and birth outcomes (Bigham et al. 2014) and oxygen saturation (Bigham et al., 2014). A study using whole genome sequencing (WGS) data has also revealed that genes related to erythropoiesis and cancer appear to be associated with chronic mountain sickness (Zhou et al., 2013a). A recent study, using WGS data from of an Aymara Andean population has also found among the strongest genes showing evidence for selection genes not related to the HIF pathways (Crawford et al., 2017a). Although all these recent advances have produced interesting candidate genes, there is still a need to link the candidate genomic loci to specific phenotypes, and to further characterize the functional effect of these variants.

Another interesting case results from the higher prevalence of metabolic and obesity related traits (including Type 2 Diabetes) observed in Native Americans and admixed Latin Americans (Mulligan et al., 2004; Cossrow and Falkner, 2004; Aguilar Salinas et al., 2007). Early studies suggested that this suite of metabolic phenotypes arose as an adaptation for more efficient food storage utilization in post-Beringia environments, commonly known as the "thrifty genotype hypothesis" (Neel, 1962). Interestingly, a study conducted in over 4,000 Native American individuals from distinct populations examined the levels of signatures of selection in a common variant in the *ABCA1* (ATP-binding cassette transporter A1) gene that has been associated to low high-density lipoprotein cholesterol (HDL-C) (Villarreal-Molina et al., 2008). They found that this region showed strong signals of adaptations based on haplotype-based and allele-frequency differentiation methods (Acuña-Alonzo et al., 2010).

In another study conducted in more than 53 indigenous human populations including Native American individuals, Pickrell et al. (2009) found that among the most differentiated regions between Mexican Mayas and Siberians (used as a proxy for the ancestor of modern Native Americans) was a gene highly expressed in lymphoblasts (*DPP3*) and a cluster of interleukin receptors. Another recent study in Native American populations from San Antonio de los Cobres in Argentina provided the first evidence for an adaptive region related to arsenic-rich environments (Schlebusch et al., 2015). The authors found a potentially protective regulatory variant at a gene involved in arsenic methylation; this gene was likely to be under selection based on haplotype-based and allele-frequency differentiation statistics (Schlebusch et al., 2015). In two Native American populations living in the Amazonian tropical forest Amorim et al. (2015) found signals of positive selection in genes related to lipid metabolism, and the immune system as well as body development. Another recent study aimed to find signals of adaptation shared by Native American populations explored the genomes of more than 44 Native American populations living across the American continent and found strong evidence of adaptation in the fatty acid desaturases (FADS) gene cluster (Amorim et al., 2017). Interestingly, the same region was previously found to be under strong selection in Greenlandic Inuit populations (Fumagalli et al., 2015) and therefore the authors suggested that this adaptation probably arose due to a single adaptive event most likely in Beringia, before the expansions of the first American into the continent. It has recently been proposed that the Native Americans populations living in Beringia during this time probably experienced strong selection in the FADS genes cluster and in the ectodysplasin A receptor (*EDAR*) gene, because of the advantage these variants conferred in transmitting nutrients from mother to infant through breast milk under conditions of extremely low solar radiation exposure levels (Hlusko et al., 2018).

Finally, studies based on ancient DNA on Native American populations, although scarce, have also confirmed and revealed interesting adaptive candidate loci. A recent candidate gene study on known loci related to adaptation to high altitude in Andean populations on more than 100 ancient samples found that a gradual allele frequency shift across their samples time frame between 8,500 to 560 BP could be explained by invoking natural selection compared to random genetic drift (Fehren-Schmitz and Georges, 2016). Similarly, a study based on whole-genome exome data conducted in ancient and modern individuals from the Northwest cost of North America, dating from before and after the European conquest, identified strong signals of positive selection on the human leucocyte antigen (HLA) gene *HLA-DQA1* (Lindo et al., 2016). The authors concluded that the signals observed were consistent with the European-borne epidemics of the 1800s in the Northwest Coast region. Interestingly, this result is in line with a recent study that found evidence from *Salmonella enterica* in ancient samples from southern Mexico (Vågene et al., 2018). The authors proposed this finding could represent a strong candidate for the cause of population decline after the Spanish conquest that occurred in Meso America.

1.4.2 Previous studies on detecting selection in admixed Latin Americans

The early scans of selection post-admixture in admixed populations were usually based on a small number of genetic markers (Workman et al., 1963; Reed, 1969). With the advent of better local ancestry inference algorithms and high-density SNP data it is now possible to more accurately infer ancestry along the genome. In this context local ancestry simply refers to the ancestral origin for a particular genomic region in an individual's chromosome. In the local ancestry paradigm (Falush et al., 2003; Patterson et al., 2004; Tang et al., 2006; Sankararaman et al., 2008; Price et al., 2009; Baran et al., 2012; Maples et al., 2013; Guan, 2014) one can imagine each individual's genome as partitioned into chromosomal segments, each with a specific ancestral origin. The goal here is to define the segment boundaries and assign a particular ancestry to each genomic region. This is different to the global ancestry paradigm (Pritchard et al., 2000; Tang et al., 2005; Alexander et al., 2009; Lawson et al., 2012). The first genomic scan of selection post-admixture in an admixed Latin American population was conducted by Tang et al. (2007) in a small sample of Puerto Ricans. Using a novel local ancestry inference software, Tang et al. (2007) were able to infer regions of the genome that significantly deviated from the genome-wide ancestry average — a hallmark of selection post-admixture. The rationale is illustrated in Figure 1.9 and is closely related to an association mapping technique called admixture mapping (Winkler et al., 2010). Under evolutionary neutrality it is expected that the mean local ancestry at a particular genomic region (averaged across all individuals) should follow the genome-wide ancestry average. However, the local ancestry proportion at a genomic region can deviate from the expectation for a number of reasons: i) sampling error in the ancestral reference or admixed population, ii) genetic drift and iii) selection. A significantly strong deviation is usually suggested as being caused by some form of selection. Tang et al. (2007) reported three genomic regions showing strong evidence of recent selection post-admixture including the human leukocyte antigen (HLA) region, which harbours various genes with a known function in immune response, including resistance and susceptibility to a broad range of infectious diseases (Hill, 1998, 2001). Given that African ancestry was enriched at this region, the authors suggested that certain African alleles could have conferred a selective advantage to certain infectious diseases most likely brought by Europeans. Other recent studies have since then independently replicated the adaptive signal at the HLA region in admixed Latin American populations such as Mexico (Zhou et al., 2016; Deng et al., 2016), Colombia (Rishishwar et al., 2015; Deng et al., 2016), Argentina (Deng et al., 2016), and Costa Rica (Deng et al., 2016). Interestingly, other genome-wide scans in Mexico (Basu et al., 2008) and Brazil (Ettinger et al., 2009) however did not replicate this signal, although this could have been due to the small sample size and/or low number of loci used in these studies.

It has also been cautioned, however, that many of these signals of selection might have been artifacts due to wrongly assigned ancestries caused by unmodeled long range admixture LD and innaccurate populations used as reference populations (Price et al., 2007) Additionally, it is also important to consider the number of independent tests when

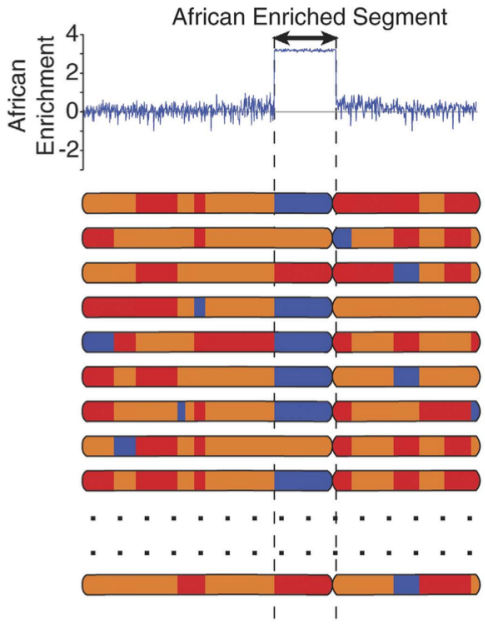

Figure 1.9: Schematic of signal of selection post-admixture. Chromosomes from an admixed populations are a mixture of ancestry segments depicted here with blue, orange and red colors. Under neutrality the proportion of an ancestry at a locus should be similar to the genome-wide population average. However, under selection the proportion of that ancestry will deviate from the genome-wide average producing a deviation or an enrichment signal. From Rishishwar et al. (2015).

performing a genome-wide scan of selection. In this case, the number of independent test will not be the number of genetic markers employed, but the number of ancestral genomic regions across the genome. These, will be determined by the evolutionary history of the admixed population, specifically by the time since the admixture event, as younger admixed populations will have longer admixture-LD along the genomes than older admixed populations.

1.5 Human pigmentation variation

Human pigmentation variation in skin, eye, and hair represents one of the most striking aspect of human variation. The distribution of pigmentation traits shows a remarkable variation between geographic regions (Figure 1.10) that stands in sharp contrast to other phenotypic traits and variation at the genomic level (Relethford, 2002). For example, Relethford (2002) estimated that ∼88% of the total variation in skin pigmentation can be explained by differences between major geographical regions. The estimated variation at the genomic level (depending on the genomic marker used) is usually between 10-15% and that of other traits such as craniometric differences is around 13% (Lewontin, 1972; Tishkoff and Kidd, 2004), two examples that highlight the atypical pattern of pigmentation phenotypes considering the recent origin of modern humans and its expansion throughout the world. The molecular, genetic, and evolutionary basis of this variation has been

subject to extensive research by a variety of researchers in distinct fields across the life sciences (Jablonski, 2008, 2012) and it is still of great interest today (Martin et al., 2017b; Wollstein et al., 2017; Rawofi et al., 2017; Crawford et al., 2017b; Mathieson et al., 2018; Brace et al., 2018; Hysi et al., 2018). In this section I will briefly describe the biological basis of human pigmentation variation, followed by its evolutionary history and the main genetic variants affecting its variability.

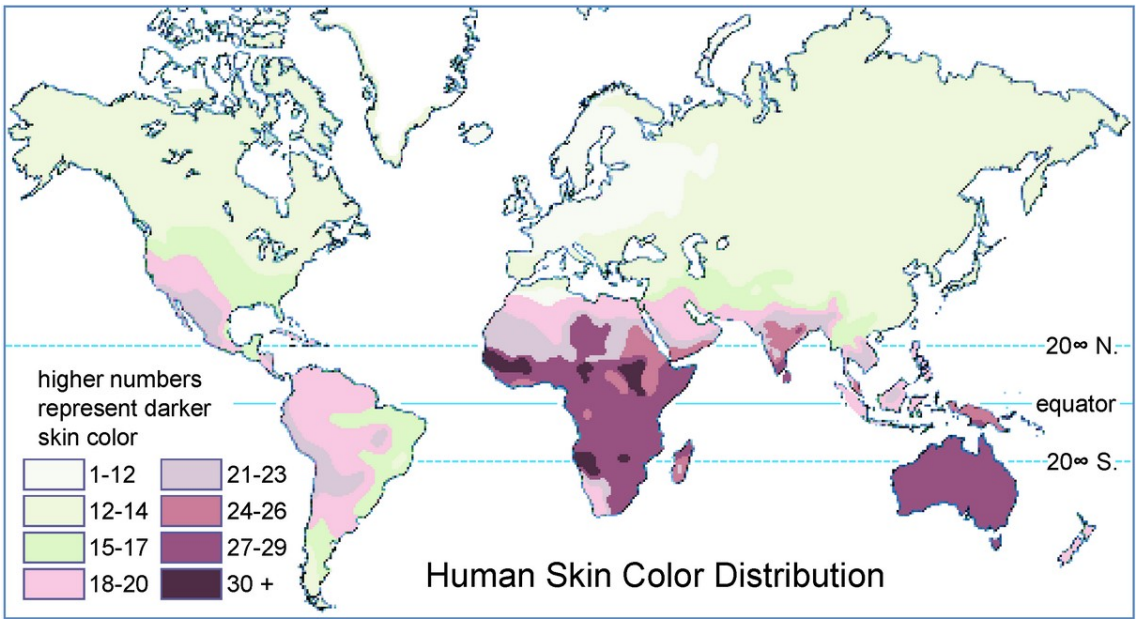

Figure 1.10: Skin pigmentation in world-wide human populations. The variation of human skin pigmentation shows great differences between major geographical regions. This map is based on the work from the Italian geographer Biasutti. From Parra (2007).

1.5.1 The biology of human pigmentation

The color of human skin, eye and hair is largely determined by the melanin pigment (Barsh, 2003; Parra, 2007; Sturm, 2009). Other chromophores such as hemoglobin also affect skin color, although with only a minor role (Parra, 2007; Sturm, 2009). Melanin is produced within melanocytes, which are located in the basal layer of the epidermis, iris, and hair bulb, and are also present in the inner ear, vaginal epithelium, meninges, bones and heart. Within melanocytes, lysosome-like organelles called melanosomes synthesise melanin. There are two major types of melanin: eumelanin and pheomelanin. The key steps in melanin synthesis involve using the primary substrate tyrosine to form a key intermediary molecule dopaquinone. In the abscence of the amino acid cysteine, dopaquinone gives rise to the dark-brown eumelanin pigment, whereas in the presence of cysteine, dopamine gives rise to the red-yellow pheomelanin (Bolognia and Orlow, 2003; Meredith and Sarna, 2006) (Figure 1.11A). The regulation and distribution of melanin synthesis also differs between skin, eye and hair. In the skin, melanosomes located within the basal layer of epidermis are transferred to the surrounding keratinocytes, which then migrate to the upper layer of the epidermis (Rees, 2003). Similarly, in the hair, melanosomes located in the hair bulb are transferred to keratinocytes that will migrate to the hair shaft (Slomin-

ski et al., 2005). In contrast to skin and hair, however, melanosomes present in the iris do not migrate (Sturm and Frudakis, 2004). Although melanocytes can vary in different parts of the body (Whiteman et al., 1999), the number or density of melanocytes does not seem to affect differences between lighter and darkly pigmented individuals. Rather, the main drivers of variation are the type (i.e. the ratio of eumelanin to pheomelanin), the amount of melanin, as well as the shape and distribution of melanosomes (Parra, 2007; Sturm, 2009). Darkly pigmented skin has more melanin, is enriched in eumelanin and the melanosomes are larger and distributed as single units (Szabó et al., 1969; Alaluf et al., 2002). Lightly pigmented skin, on the other hand has a higher amounts of pheomelanin, and the melanosomes tend to be less pigmented, smaller and packaged into groups (Szabó et al., 1969; Alaluf et al., 2002) (Figure 1.11). Darker hair has also the highest eumelanin-to-pheomelanin ratio compared to lighter hair (Rees, 2003). Darker irises have larger amounts of melanin and high numbers of melanosomes, whereas lighter eyes have low melanin content and a lower number of melanosomes (Sturm and Frudakis, 2004).

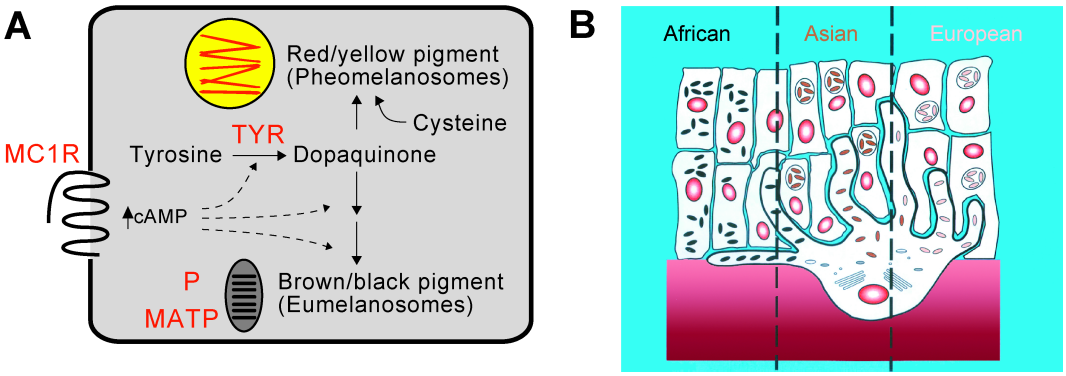

Figure 1.11: Melanin synthesis and histology of different skin types. A. Schematic of melanin synthesis. The key step in melanin synthesis involves the primary substrate tyrosine to form a key intermediary molecule dopaquinone. In the absence of the aminoacid cysteine, dopaquinone gives rise to the dark-brown eumelanin pigment, whereas in the presence of cysteine, dopamine gives rise to the red-yellow pheomelanin pigment. B. Variation in melanosome structure and distribution among populations with varying degrees of skin pigmentation. A single skin melanosome is partitioned into three sections showing the stages of melanosome formation until the migration into the surroundings keratynocytes. In heavily pigmented individuals (African), the melanosomes remain as singular pigmented units distributed across the keratynocytes, whereas in lightly pigmented people (Asians and Europeans), the melanosomes cluster in membrane bound organelles. From Barsh (2003).

1.5.2 The evolution of human pigmentation

One of the most distinguishable features of humans compared to the other great apes is an almost complete absence of fur (Held, 2010). Because skin is not preserved in the fossil record, research regarding loss of body hair has relied on comparative anatomical, paleoecological, climatological and physiological evidence. This process has been proposed as an adaptive process for enhanced thermoregulation during high physical activity, for example by allowing the gain in sweat glands to increase heat dissipation (Bramble and

Lieberman, 2004; Lieberman and Bramble, 2007). However, this event was also accompanied by a disadvantage due to reduced protection from solar radiation. Comparative studies have revealed that many human accelerated regions (HARs) include genes related to keratinization and epidermal differentiation, in line with this hypothesis (Waterson et al., 2005; Toulza et al., 2007; Gautam et al., 2014). As our closest relatives such as chimpanzees have lighter skin pigmentation under their fur, the ancestral skin color of the human lineage has been assumed to be of a similar lighter shade (Jablonski, 2008, 2012)). Strikingly, the timing of darkly pigmented skin from genomic studies of different human populations at the human melanocortin 1 receptor (*MC1R*) gene was placed around the time of body hair loss (Rogers et al., 2004), consistent with strong selective pressure at this period of human evolution. Additionally, the absence of functional variants on this gene seem to have been maintained by strong purifying selection (Harding et al., 2000b), further supporting the reproductive advantage for darkly pigmented skin in strong solar radiation environments (Jablonski and Chaplin, 2017).

The strongest hypothesis for the evolution and maintenance of dark skin pigmentation in zones with high solar radiation exposure is due to the lower photodegradation of cutaneous and systemic folate. Folate deficiencies have been associated with greater neonatal malformations such as neural tube defect and also with lower male fertility (Fleming and Copp, 1998; Lucock, 2000; Tamura and Picciano, 2006) and thus have a direct association to reproductive success, that other competing hypotheses lack (Jablonski and Chaplin, 2017). For example, the hypothesis that the evolution of darkly pigmented skin was as an adaptation against sunburns and skin cancer (Greaves, 2014) does not provide an association with a reduced reproductive success, as skin cancer rarely causes death during peak reproductive age (Jablonski and Chaplin, 2014).

As modern humans moved out of Africa, they moved from a zone with constant high amounts of solar radiation to zones with variable amounts of solar radiation. This change in environment relaxed the selective pressure for darkly pigmented skin, but was accompanied by other selective pressures. Currently, the most popular hypothesis for prevailing lighter human skin color variation in northern latitudes is the vitamin D hypothesis (Loomis, 1967; Murray, 1934; Jablonski and Chaplin, 2000). Vitamin D production is determined by the amount of solar radiation received on the skin and its deficiency can result in rickets and an increase of miscarriages (Robins, 2009). Consistent with this hypothesis, the distribution of skin color variation is highly correlated to solar radiation ($r = 0.93$) (Figure 1.12). Patterns of solar radiation are relevant because only some wavelengths of solar radiation (those between 270 and 300 nm) can activate vitamin D production. At higher latitudes, average solar radiation along those wavelengths is low and highly variable, and darker skin thus has a fitness disadvantage (Murray, 1934; Loomis, 1967; Jablonski and Chaplin, 2000). Vitamin D production still occurs in darkly pigmented people living at these higher latitudes, but low doses of solar radiation do not facilitate its production to physiologically adequate levels (Jablonski and Chaplin, 2017). There is ample evidence that people with darkly pigmented skin and living north or south of the 43° do not receive

sufficient amounts of solar radiation to satisfy the required levels of vitamin D (Jablonski and Chaplin, 2018). The end result would have been a strong selective pressure for reduced skin pigmentation at higher latitudes in order to allow for a long term habitation in these regions.

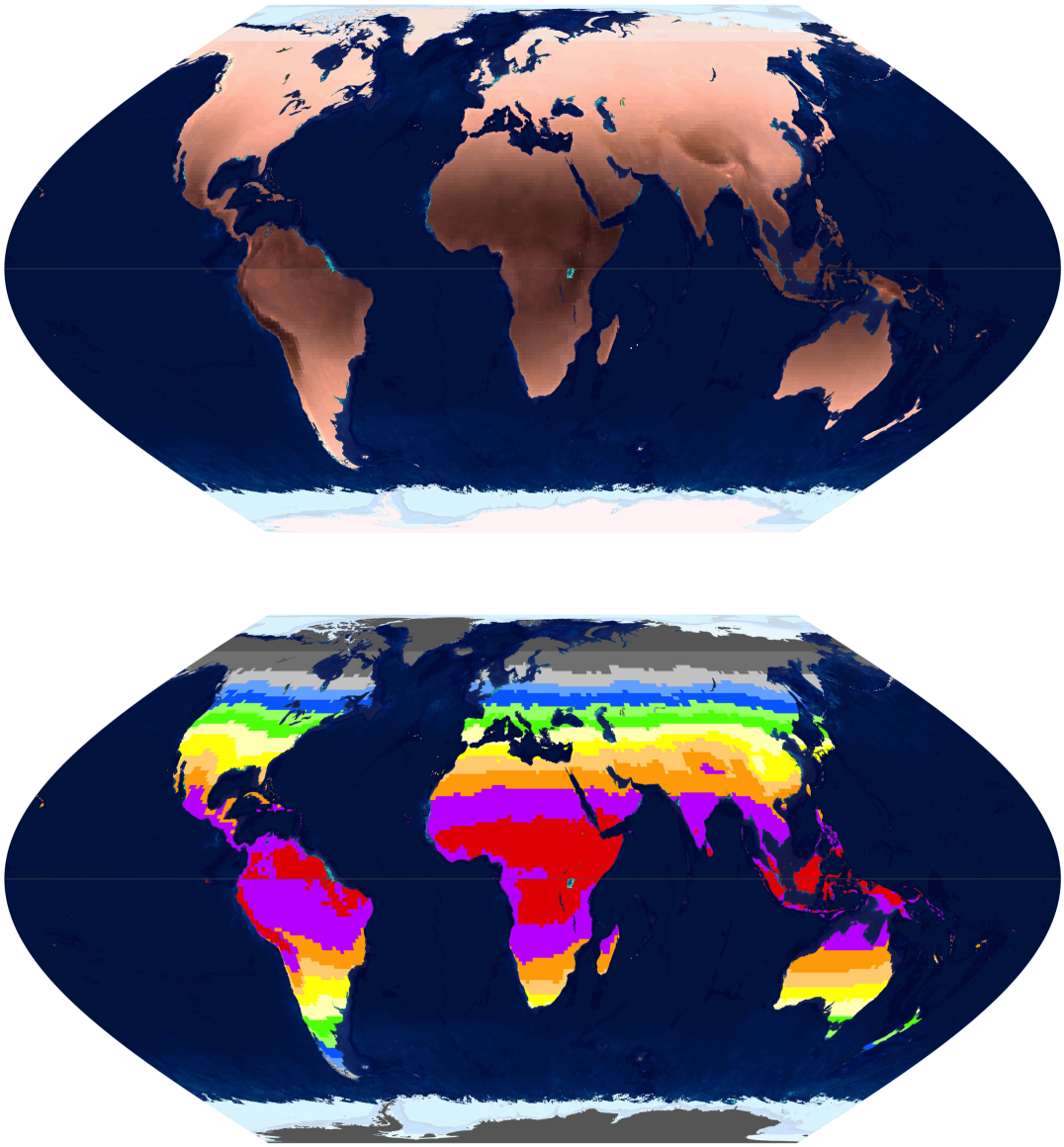

Figure 1.12: Map of predicted skin pigmentation. Top panel. Predicted human skin pigmentation. Bottom panel. Solar radiation (Autumnal UV). Skin pigmentation can be almost fully modelled as an effect of solar radiation ($r = 0.93; p < 0.0001$). Reproduced from http://sites.psu.edu/ninajablonski/educational-resources-2/. Based on Jablonski and Chaplin (2000).

Therefore, according to these two hypotheses skin pigmentation variation would have been determined by two opposing forces: one selecting for darkly pigmented skin in zones with high and constant amounts of solar radiation, and one for reduced skin pigmentation at higher latitudes. Skin pigmentation at intermediate latitudes on the other hand seems to have evolved a greater capacity for tanning (Nan et al., 2009; Quillen, 2015). The

indigenous inhabitants of the Americas with lighter skin pigmentation than other populations living at these latitudes seem to be explained by the fact that these populations have inhabited these regions for shorter periods of time (at least < 15,000 to 20,000 ya) and to better cultural adaptation such as the making of better shelters (Jablonski, 2012). It is also important to note that these hypotheses do not necessarily posit invariant skin pigmentation for populations living at zones with extreme amount of solar radiation such as Africans populations living in the tropics. African individuals, although showing on average a greater amount of skin pigmentation, have a much greater variation in skin pigmentation compared to other non-African populations (Martin et al., 2017b; Crawford et al., 2017b). For example, Martin et al. (2017) reported that Ghanians have a variance that is ten times higher than Irish. This result can still be explained by the vitamin-D folate hypothesis, and is consistent with a "melanin threshold" model (Chaplin, 2004; Norton et al., 2006) where selection for dark skin pigmentation would act until a photoprotective minimum is reached (in order to protect extensive folate degradation) above which populations can vary (Lasisi and Shriver, 2018).

The role of sexual selection in skin pigmentation, which was initially proposed as early as 1871 by Charles Darwin (Darwin, 1871) and later revived by others (Frost, 2006, 2007), seems not to have sufficient explanatory power to explain the global pattern of human skin pigmentation variation and its strong correlation with latitude (Jablonski and Chaplin, 2000, 2017). However, it seems that sexual selection could have had a role in increasing the degree of sexual dimorphism (Jablonski, 2012; Jablonski and Chaplin, 2017). Lighter skin pigmentation in females in many populations has been observed and it may possibly be related to the greater vitamin D requirements during pregnancy (Jablonski, 2008, 2012).

In contrast to skin pigmentation, where solar radiation might have imposed a strong selective pressure, variation in hair and eye color is more difficult to explain. There is not an observable correlation between hair or eye color and solar radiation or latitude (Liu et al., 2013b). In addition, variation in hair color is consistent across populations with the exception of Europe. Similarly, variation in eye color is mainly present in Europe (and its neighbour regions such as North Africa and parts of the Middle East) and in Central and South East Asia (Westgate et al., 2013; Norton et al., 2016). Although dark hair is considered to be more protective from solar radiation compared to lighter hair color, this protection seems to be minor given the low coverage of hair in human bodies (Held, 2010). The back of the iris is also darkly pigmented for all types of iris color and therefore provides a similar protection for all type of eye color (Liu et al., 2013b). Blue eye color, however, has been associated with greater intraocular light scattering and higher levels of melatonin supression, which might have been adaptive in higher latitudes (Higuchi et al., 2007). Additionally, experimental data on color-based mate choice preference in fish has been shown (Amundsen and Forsgren, 2001) and similarly in humans, blue-eyed woman have been shown to be preferred by blue-eye men (Laeng et al., 2007). This result has been interpreted as a preference because of the phenotypically-based assurance of paternity (Laeng et al., 2007), suggesting mate-choice preference could have played a role in the

evolution of eye color variation (Jablonski and Chaplin, 2017). However, given the highly pleiotropic effect of many genetic variants contributing to skin, hair and eye pigmentation, it is also possible that strong selection for skin pigmentation phenotypes has affected variation and hair and eye color simply as a by-product (Wilde et al., 2014; Jablonski and Chaplin, 2017).

More recently, studies of ancient genomes, mainly from populations from Western Eurasia have also started to contribute to the understanding of the evolutionary history of pigmentation phenotypes (Wilde et al., 2014; Mathieson et al., 2015, 2018). Ancient DNA samples from the Eneolithic (ca. 6,500 to 5,000 ya) and Bronze Age (ca. 5,000 to 4,000 ya) showed a strong support for strong positive selection for lighter pigmentation phenotypes operating over the last 5,000 years (Wilde et al., 2014). Similarly, ancient genomes from Anatolian Neolithic farmers in Western Eurasia (6,500 to 300 ya) showed that selection for lighter skin pigmentation was operating since at least 6,500 to 4,000 ya (Mathieson et al., 2015). These results, which point to selection for lighter skin pigmentation only after the Neolithic revolution, might have resulted from a switch from more rich and varied diet of hunter-gatherer populations (including vitamin D rich foods) to the poorer diet of early farmers (Richards et al., 2005; Mathieson et al., 2015). This would also be in line with the observation that hunter-gatherers had mainly darkly pigmented skin (and distinctive lighter eye color) (Olalde et al., 2014; Mathieson et al., 2015, 2018). Additionally, the light (particularly blue) eye color genetic variant has also been observed to have a higher frequency in Scandinavian and Latvian hunter-gatherers compared to hunter-gatherers from Ukraine and the border of present-day Romania and Serbia, which could indicate the possibility of long-term balancing selection (Mathieson et al., 2018).

1.5.3 The genetic determinants of human pigmentation

Human skin, hair and eye pigmentation are complex polygenic traits. Unlike other complex polygenic traits, such as height, where hundreds of genetic variants with small genetic effects are involved (Lango Allen et al., 2010), pigmentation phenotypes are affected by genetic variants that can explain large proportions of the variation as well as large differences between major geographical regions. Other genes have more subtle effects and many are still yet to be discovered, especially those affecting pigmentation variation outside Western Europe (Martin et al., 2017b; Crawford et al., 2017b), but the vast majority of genetic variants discovered so far have been characterized as mainly affecting the melanin synthesis pathway.

The melanocortin 1 receptor (*MC1R*) is one of the most exhaustibly studied pigmentation genes (Valverde et al., 1995; Schiöth et al., 1996; Bastiaens et al., 2001; Ha et al., 2003; Naysmith et al., 2004; Rees, 2000, 2004; Mundy et al., 2003; Ringholm et al., 2004; Sánchez-Más et al., 2005; García-Borrón et al., 2005; D'Orazio et al., 2006; Haitina et al., 2007; Savage et al., 2008; Yamaguchi et al., 2012; Jarrett et al., 2015; Hernando et al., 2016; Liu et al., 2016). Its role in normal pigmentation variation in humans was discovered

through an association with fair skin and red hair Valverde et al. (1995). *MC1R* has also been associated with freckles, but unlike other pigmentation genes it has not been found to affect eye color (Bastiaens et al., 2001). *MC1R* has also been recently associated with different types of skin cancer such as melanoma (Ransohoff et al., 2017) and basal cell carcinoma (Chahal et al., 2016), as well as skin aging (Law et al., 2017) and to vitiligo (Jin et al., 2016). *MC1R* is a member of the family of protein-coupled receptors and has an important role in switching between the synthesis of eumelanin and pheomelanin. Binding of the melanocyte-stimulating hormone (α-MSH) results in eumelanin production whereas binding of its antagonist protein (*ASIP*) results in pheomelanin synthesis. *MC1R* also shows an interesting pattern of variation across worldwide populations. Many non-synonymous variants have been found to be nearly absent in several African populations, especially in sub Saharan populations, but many polymorphisms have been found outside Africa (Rana et al., 1999; Harding et al., 2000a; Makova and Norton, 2005; Nakayama et al., 2006). Interestingly, a low number of polymorphisms have also been found to be present in other darkly skin pigmented populations, such as Papuans and South Asians (Rana et al., 1999; Harding et al., 2000a; Nakayama et al., 2006). This pattern of diversity on *MC1R* has been interpreted by strong purifying selection acting at this locus in zones with high incidence of solar radiation in order to remove variants that lead to less darkly pigmented pheomelanin (Rana et al., 1999; Harding et al., 2000a; Parra, 2007). In European populations many different genetic variants have been associated with lighter skin and hair pigmentation Duffy et al. (2004); Sulem et al. (2007); Eriksson et al. (2010); Lin et al. (2015); Liu et al. (2015), as well as tanning response (Nan et al., 2009). Additionally, many variants at this gene have also been shown to lead to a partial or complete loss of function, which could explain the largely recessive inheritance of red hair color (Ringholm et al., 2004). In Asia, different non-synonymous variants, which show higher derived allele frequencies in East Asians and are almost absent in Europeans, seem to have contributed to the independent evolution for lighter skin pigmentation in Western and Eastern Eurasians (Nakayama et al., 2006; Yamaguchi et al., 2012). A related protein to *MC1R*, is the agouti signalling protein *ASIP* that antagonizes the interaction between *MC1R* and $\alpha - MSH$ resulting in pheomelanin production. SNPs at *ASIP* has been associated with skin and hair color in European and European admixed populations such as Latin Americans Eriksson et al. (2010); Liu et al. (2015); Sulem et al. (2007); Hernandez-Pacheco et al. (2017). *ASIP* has also been associated to pigmentation-related phenotypes such as freckles (Sulem et al., 2008), tanning (Zhang et al., 2013a), melanoma (Ransohoff et al., 2017), cutaneous squamous cell carcinoma (Chahal et al., 2016) and vitiligo (Jin et al., 2016).

The so called "golden" gene (*SLC24A5*) was first shown to be associated with the golden color of the zebrafish as well as with skin pigmentation in admixed African Americans and African Caribbeans (Lamason et al., 2005). *SLC24A5* encodes a $Na^+/Ca^{2+}/K^+$ exchanger 5 (NCKX5) protein, an intracellular membrane protein that regulates the calcium concentration in the melanosome (Lamason et al., 2005). A non-synonymous polymorphism (SNP rs1426654) was found to have the greatest allele frequency differentiation

in populations from Europe compared to East Asians and Africans and thus suggested to be have undergone strong selection in Europeans (Lamason et al., 2005). In addition, the derived allele at this same SNP was found to be significantly associated with lower skin pigmentation and to account for between 28 to 38% of differences in skin pigmentation between European and African ancestry (Lamason et al., 2005). Interestingly, in South Asian populations rs1426654 was also strongly suggested as being the casual variant for reduced skin pigmentation (Stokowski et al., 2007). In addition, Norton et al (2007) studied the global distribution of this variant and found that higher frequency of the derived allele was mainly found in Europe and its neighbour populations including North Africa, the Middle East and Pakistan, and almost absent everywhere else. SNP rs1426654 has also been associated with lighter skin pigmentation in African (Martin et al., 2017b; Crawford et al., 2017b) and African admixed populations (Beleza et al., 2013a; Lloyd-Jones et al., 2017). In admixed Latin American populations rs1426654 has also been recently associated to skin (Hernandez-Pacheco et al., 2017) and hair pigmentation (Adhikari et al., 2016a).

The solute carrier family 45 member 2 (*SLC45A2*) is another transporter protein that mediates melanin synthesis. *SLC45A2* is a melanocyte differentiation antigen highly expressed in melanoma cell lines (Harada et al., 2001). Mutations at this gene are known to cause oculocutaneous albinism type 4 (OCA4) (Newton et al., 2001; Rundshagen et al., 2004; Inagaki et al., 2004) and have also been recently associated to melanoma (Ransohoff et al., 2017)) and squamous cell carcinoma (Asgari et al., 2016). Non-synonymous polymorphisms in this gene show a remarkable pattern of frequency across the world. Alleles associated with lighter pigmentation are mainly restricted to European and its neighbour populations and nearly absent everywhere else. Interestingly, the derived allele at SNP rs16891982 is associated with darker skin and eye color in European (Graf et al., 2005, 2007; Han et al., 2008) and South Asian populations (Stokowski et al., 2007). Similarly, the derived allele has been associated to darker hair and eye color (Eriksson et al., 2010). In Latin American admixed populations rs16891982 has also been recently associated with skin and hair pigmentation (Adhikari et al., 2016a; Hernandez-Pacheco et al., 2017).

The HECT and RLD domain containing the E3 ubiquitin protein ligase 2 (*HERC2*) and oculocutaneous albinism 2 (*OCA2*) genes are two neighbouring genes that are of greatest importance determining skin and eye pigmentation. The *OCA2* gene (also known as the *P* gene) encodes the P-protein that assists tyrosinase trafficking and processing, melanosomal pH and glutathionine metabolism (Park et al., 2015). It has also been recently been shown to assist in anion transport, increasing chloride conduction from the melanosome (Bellono et al., 2014). The key determinant SNP (rs12913832) located 21kb upstream of *OCA2* and within intron 86 of *HERC2* shows the strongest association for lighter eye color with the derived allele in European populations (Sulem et al., 2007; Sturm, 2009; Liu et al., 2010; Eriksson et al., 2010; Candille et al., 2012; Zhang et al., 2013a; Wollstein et al., 2017). SNP rs12913832 has also been associated with lighter skin and hair pigmentation in European populations (Han et al., 2008; Zhang et al., 2013a; Liu et al., 2015). *HERC2*

rs12913832 SNP was shown experimentally to function as an enhancer regulating *OCA2* transcription by modulating chromatin folding (Visser et al., 2012). Specifically, molecular approaches showed that *HERC2* rs12913832 communicates with the *OCA2* promoter via a long-range chromatin loop that is modulated by several transcription factors, including the Melanogenesis Associated Transcription Factor (MITF) (Visser et al., 2012). SNPs at *OCA2* independently of *HERC2* have also shown strong association with eye color. Particularly, when adjusting for *HERC2* rs12913832, *OCA2* rs1800407 SNP still showed a significant association with eye color whereas the effect of other *OCA2* SNPs was reduced (Liu et al., 2009). Studies of interaction between *HERC2* and *OCA2* have also been reported to affect eye pigmentation (Branicki et al., 2009; Liu et al., 2010; Pośpiech et al., 2011), consistent with the regulatory action of *HERC2* over *OCA2*. In Asian populations two SNPs rs74653330 and rs1800414 have been associated to lighter skin pigmentation Edwards et al. (2010); Abe et al. (2013) and are found at very low frequency outside of Asia, consistent with the convergent evolution of lighter skin pigmentation in East Asia (Murray et al., 2015). Interestingly, the distribution of these SNPs within Asia has shown to be quite different within eastern Eurasia. While the derived allele of rs1800414 has high frequencies in the broad East-Asian region, the derived allele of rs74653330 is primarily restricted to northern East Asia (Murray et al., 2015). This result suggest that these variants may have been selected independently in different regions of East Asia (Murray et al., 2015). Recently, SNP rs1800414 has also been associated with lighter eye pigmentation and iris heterochromia in East Asian populations (Rawofi et al., 2017). Outside of Eurasia, SNP rs1800404 has also been associated with skin pigmentation in African populations (Crawford et al., 2017b) and with hair pigmentation in admixed Latin Americans (Adhikari et al., 2016a).

The *TYR* gene encodes the Tyrosinase enzyme, a key enzyme for controlling melanin synthesis (Kwon et al., 1987). As described above (Figure 1.11), *TYR* catalyses the first step required for melanin synthesis: the oxidation of tyrosine and dopa to form the intermediate molecule dopaquinone, which is used as substrate to synthesise the two main pigments eumelanin and pheomelanin. *TYR* was firstly associated with oculocutaneous albinism type 1 (OCA1) in European populations (Rooryck et al., 2008). Common SNPs have also been associated with skin, eye and hair pigmentation variation in different European populations Sulem et al. (2007); Liu et al. (2010); Eriksson et al. (2010). The strongest associated SNPs (rs1042602 and rs1126809) have high derived allele frequency mainly in European populations and are largely absent in other continental populations, similar to the geographical pattern observed for other well established pigmentation genes as described above. In African American and African Caribbean populations highest association was found at SNP rs1042602 with skin pigmentation (Shriver et al., 2003). In admixed European-African admixed populations, SNP rs10831496 was associated with skin pigmentation (Beleza et al., 2013a; Lloyd-Jones et al., 2017) and in admixed populations from Latin America another SNP rs598952 showed highest association with hair color (Adhikari et al., 2016a). Other SNPs at *TYR* have also been associated with pigmentation-related phenotypes such as presence of freckles (Sulem et al., 2007), sensitivity to sunburns

(Zhang et al., 2013a), as well as to melanoma (Barrett et al., 2011; Ransohoff et al., 2017), cutaneous cell carcinoma (Chahal et al., 2016) and vitiligo Jin et al. (2016).

The *TYRP1* gene also encodes a key enzyme in the melanin synthesis pathway (Del Marmol and Beermann, 1996). Mutations at *TYRP1* lead to a complete loss of function that results in oculocutaneous albinism type 3 (OCA3) or rufous albinism (Rooryck et al., 2006; Chiang et al., 2009). One of the functions of *TYRP1* is to stabilise Tyrosinase and form heterodimeric complexes within the melanosome (Kobayashi et al., 1998; Kobayashi and Hearing, 2007). SNP rs1408799 shows association with eye pigmentation (Sulem et al., 2008; Zhang et al., 2013a). A suggestive association has also been found with hair pigmentation in European populations (Sulem et al., 2008). A notable example of convergent lighter hair color evolution at this gene was reported in Solomon Islanders populations: a non-synonymous SNP rs387907171 at *TYRP1* was found to be associated with blonde hair and the associated causal variant was not present in any other population outside Oceania (Kenny et al., 2012). Recently, another SNP in *TYRP1* has been associated to skin pigmentation in African populations (Martin et al., 2017b).

The interferon regulatory factor 4 is a protein encoded by the *IRF4* gene. A single SNP (rs12203592) has been associated with eye and hair pigmentation in European populations (Han et al., 2008; Eriksson et al., 2010; Zhang et al., 2013a; Liu et al., 2015). The associated variant at SNP rs12203592 is only found in Europe (and European admixed populations) and shows a north-south gradient across the continent (Moskvina et al., 2010; Walsh et al., 2011). SNP rs12203592 has also been associated with hair color in admixed Latin American populations (Adhikari et al., 2016a). *IRF4* SNPs have also been associated with pigmentation-related phenotypes such as freckling (Eriksson et al., 2010), tanning (Zhang et al., 2013), skin aging Law et al. (2017), sensitivity to sunburns (Zhang et al., 2013) and different types of skin cancer (Zhang et al., 2013, Chahal et al., 2016; Asgari et al., 2016). Recently, SNP rs12203592 has also been associated with the first time to hair greying (Adhikari et al., 2016a).

A newly recognized skin pigmentation gene is the major facilitator superfamily domain containing 12 (*MFSD12*) gene, which encodes a lysosomal transmembrane solute transporter (Crawford et al., 2017b). Functional experiments on model organisms showed that *MFSD12* significantly affected pigmentation. In knocked-out zebrafish and mice, red-yellow pigments were lost and the mice's light brown fur turned grey. Two derived alleles, only found in African populations, were associated with darker skin pigmentation (Crawford et al., 2017b). The observation that the derived variants in African populations confer darker skin color support the hypothesis that darker pigmentation is probably a derived trait that originated in the human lineage after humans lost most of their protective body hair, as discussed above (Section 1.5.2).

Other genes with more subtle effects compared to the ones listed above have also been shown to be associated with different pigmentation phenotypes. These genes include:

KITLG , *DCT*, *TPCN2*, *LYST*, and *BCN2*. *KITLG* encodes the c-KIT ligand that affects melanocyte proliferation and activates keratinocytes (Cario-André et al., 2006). Mutations at *KITLG* have been associated with familial progressive hyper- and hypo-pigmentation (Wang et al., 2009; Amyere et al., 2011). Common polymorphisms have also been associated with eye and hair pigmentation in European populations (Sulem et al., 2007; Zhang et al., 2013)and skin pigmentation in African Americans (Miller et al., 2007). *DCT* is a key enzyme involved in the formation of eumelanin that is highly and exclusively expressed in the skin. Particular *DCT* haplotypes have been associated with eye pigmentation in Europeans (Frudakis et al., 2003), but the effect of particular SNPs is still uncertain (Ainger et al., 2017). Variants at the two pore segment channel 2 *TPCN2* gene have been associated with hair pigmentation in European populations (Sulem et al. 2008; Eriksson et al., 2010) and with cutaneous malignant melanoma (Law et al., 2015). *LYST* is a vesicular transport protein that affects lysosome-related organelles, including melanosomes (Jackson, 1997). In melanocytes, *LYST* may be regulated by MITF transcription factor (Hoek et al., 2008). *LYST* has been associated only with human eye pigmentation in European populations (Liu et al., 2010). *BCN2* encodes a conserved zinc finger protein expressed in many tissues with a potential regulatory function in keratinocytes (Romano et al., 2004). *BCN2* has been associated with skin pigmentation in Europeans and East Asians (Visser et al., 2014; Jacobs et al., 2015).

1.6 Finding genetic associations

1.6.1 Genome Wide Association Study (GWAS) — rationale and scientific basis

A Genome Wide Association Study (GWAS) is an experimental design that seeks to discover variants associated to traits in a sample from a population. GWAS typically focus on SNPs, although there have been many reported associations between traits and other genetic variants, such as copy-number variants (CNVs) (Zhang et al., 2009), and the underlying methodological aspects remain essentially the same. Current GWAS rely and exploit the linkage disequilibrium (LD) patterns that exist along the genomes of human populations that arise due to a variety of factors including: i) mutation, ii) genetic drift, iii) recombination, and iv) natural selection. The statistical power to detect an association between a genetic variant and a trait will depend on several assumptions, such as: the underlying genetic architecture of the phenotype (i.e. the joint distribution of the allele frequency and effect sizes), the sample size, and the LD between the genotyped genetic variant and the true causal variant (Chapman et al., 2003; Spencer et al., 2009; Visscher et al., 2017). Given that the LD structure of the human genome is complicated and known to vary between different human populations, this effect cannot be easily captured and statistical power is recommended to be calculated mainly via simulation (Spencer et al., 2009). Nonetheless, there is an analytical formula to assess the power to detect the association of a genetic variant and a trait using the non-centrality parameter (NCP) (i.e. the value of the statistical test under the alternative hypothesis) assuming Hardy-Weinberg

equilibrium (HWE) at the genetic locus (Visscher et al., 2017):

$$NCP = N2p(1-p)\beta^2 r^2 \tag{1.1}$$

where N is the sample size, p is the allele frequency of the tested variant, β^2 is the effect size of tested variant in standard deviation units, and r^2 the LD between the genotyped variant and the casual variant. Although in reality estimating the power in a specific GWAS study might be more complicated (Spencer et al. 2009) this equation does provide useful intuition. For example, when the effect size β is large, the LD between the genotyped variant and the true casual variant r^2 may only need to be weak, whereas if the effect size β is small, the LD between the genotyped and true casual variant may need to be strong. A further complication arises due to the fact that LD measured in this way can only be large if the allele frequencies at both loci are similar (Wray, 2005). For example, a variant that is rare (<0.01 in frequency in the population) will be in low LD with a common variant (>0.05 in frequency in the population) even if they are both present on the same haplotype block (Wray, 2005). This is specially relevant in studies based on SNP arrays, given that most of the genotyping platform capture common SNPs. This problem can be reduced by relying on imputation, which can be accurate even for variants with frequencies < 0.0001 (McCarthy et al., 2016), or by using whole genome sequence (WGS) data, in which r^2 will be equal to 1. The major remaining limitation will be the frequency of the tested variant; where in the case for very rare variants, even with strong effect sizes, required samples sizes can turn out be prohibitively large, even in the order of millions of individuals (Visscher et al. 2017).

In a GWAS, the desired reason for a significant result is a causal association between the genetic variant and the phenotype of interest. However, the result of this association can be confounded due to unaccounted population stratification, such as population structure, admixture and/or cryptic relatedness (Astle et al., 2009). Neglecting or not accounting for population stratification can therefore lead to false positives or spurious associations (Balding, 2006). The problem of spurious association due to unaccounted population structure is illustrated in Figure 1.13. In an hypothetical example of a case-control study, the study population consists of two distinct subpopulations that differ genetically. In this example, the blue allele of a bi-allelic SNP (represented here as circles) has a higher frequency in subpopulation 1. Conducting an association on this structured sampled population will likely produce a spurious (i.e. false positive) association with the phenotype under study. The association at this SNP will occur simply because the cases are more frequent in subpopulation 1, and not because the genetic variant is truly associated with the disease. Because GWAS are conducted with over thousands of markers, failing to account for the underlying population structure can lead to hundreds of genetic variants being falsely associated to the phenotype. A similar problem will arise in an admixed population, because the cases and controls are composed of individuals with different degrees of ancestries, and the cases (or controls) are overrepresented with individuals with higher proportion of a particular ancestry. In this case, genetic variants that have highly divergent allele frequencies between the ancestral populations (i.e. the

populations that mixed to gave rise to the admixed population) will be falsly associated to the phenotype. This will be extremely important in populations where the disease prevalence differs between individuals of different ancestries, such as in Latin America, where Native American ancestry is associated with higher prevalence of different diseases partly due to its association to lower socio-economic status (Ruiz-Linares et al. 2014). Accounting for population structure is therefore of utmost importance in GWAS. I discuss three of the most common methods applied to account for this problem in the Methods section (Section 2.7.2)

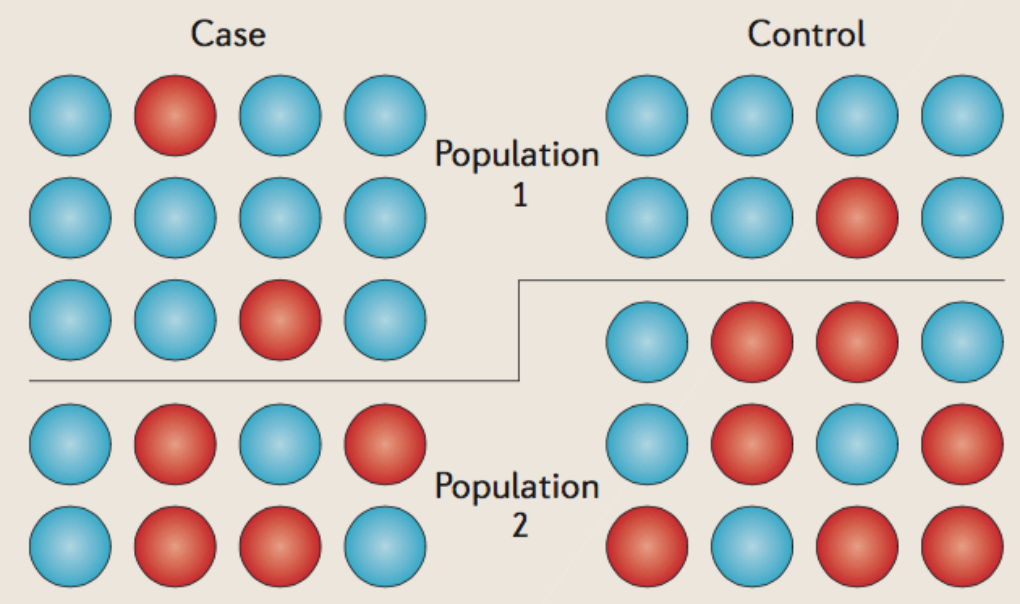

Figure 1.13: Spurious association due to unaccounted population structure. In an hypothetical example of a case-control study, the study population consist of two distinct subpopulations that differ genetically. Here, the blue allele of a bi-allelic SNP (represented here as circles) has a higher frequency in subpopulation 1. Conducting an association on this structured sampled population will likely produce a spurious (i.e. false positive) association with the phenotype under study. From Balding (2006).

Despite some of these limitations, GWAS have been proven to be remarkably successful. As of February 2018 the National Human Genome Research Institute (NHGRI) and European Bioinformatics Institute (EBI) GWAS catalogue (Welter et al., 2013) currently reports $\sim 10,000$ robust associations (defined as having P-value $< 5 \times 10^{-8}$) with one or more complex phenotypes. Figure 1.14 shows a cumulative GWAS SNP-trait discovery timeline from the year 2008 up to 2016 showing many robustly associated genetic variants for a variety of traits. Additionally, for particular phenotypes GWAS have been shown to be highly replicable within and between populations (Torgerson et al., 2011; Marigorta and Navarro, 2013). One of the early observations arising from GWAS studies was the minor contribution of the most significant genetic variants on the phenotypic variability. This observation was refered to as the mystery of the "missing heritability" (Manolio et al., 2009), but has largely been resolved by showing that many genetic variants with genetic effect sizes below the genome-wide significant association threshold account for most of the "missing heritability", as exemplified in the study of height (Yang et al., 2012; Shi et al.,

2016). Nonetheless, it has also been shown that many rare genetic variants with larger genetic effect sizes can also explain a large proportion of phenotypic variability (Marouli et al., 2017), especially in diseases with strong impact of fitness (Simons et al., 2014).

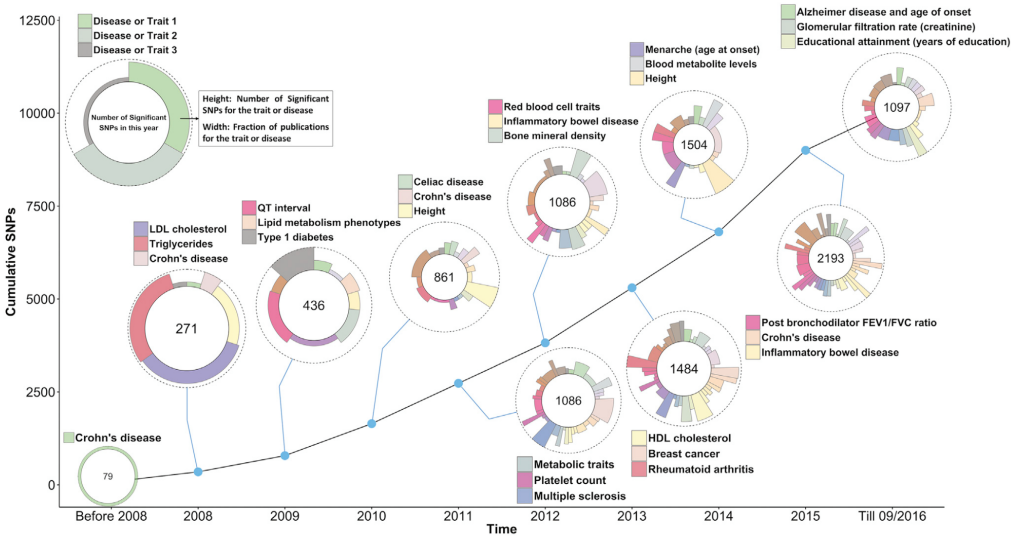

Figure 1.14: GWAS SNP-trait discovery time line. From Visscher et al. (2017).

Another important observation was that many complex phenotypic traits are mainly driven by genetic associations falling outside gene regions (Pickrell, 2014; Li et al., 2016), which is in sharp contrast to Mendelian diseases (Botstein and Risch, 2003). Many of these genetic associations have been shown to fall within active chromatin genomic regions, such as promoters and enhancers (Maurano et al., 2012; Farh et al., 2015; Roadmap Epigenomics Consortium et al., 2015). These two observation, lead Boyle et al. (2017) to readdress the view on complex phenotypic traits, changing from a polygenic to an omnigenic model (Figure 1.15). The authors stated that for a given complex phenotype, only a limited number of genes (termed "core genes") will have a direct and consequently non-trivial genetic effect on the phenotype. However, given that cell regulatory networks (including transcriptional networks, post-translational modifications, protein-protein interactions, and intercellular signaling, etc.) are highly interconnected, the small-world property suggest that most genes will only be a few nodes away from the core genes and therefore will also have a non-trivial genetic effect on the phenotype. Further, since the core genes are hugely outnumbered by the peripheral genes, a much larger fraction of the total genetic variability will therefore be explained by genetic variants at these non-core genes. It has repeatedly been found that the same genetic variants can be associated with different phenotypic traits. This had been already noticed in the study of many Mendelian disorders where specific mutations caused a specific syndrome or disease associated with various phenotypes (Hamosh et al., 2005). Furthermore, analytical methods that estimate the correlation between different GWAS have also provided additional evidence for this widespread pleiotropy (Solovieff et al., 2013; Bulik-Sullivan et al., 2015a). Finally, although GWAS have provided huge insights in understanding the links between genetic variations and specific phenotypes including disease, there is still a gap of the translation

between GWAS results into medicine. For example, there is currently only 8 examples between GWAS discoveries and drug targets, with the best examples probably arising only from three particular examples: type 2 diabetes, auto-immune diseases, and schizophrenia (Visscher et al., 2017). Large biobank efforts such as the UK BioBank (Sudlow et al., 2015) and others will undoubtedly provide many novel insights that will arise not only due to the increasing size of the cohorts, but also due to better analytical methods, better phenotype sampling and a move towards long-term studies including repeated measures over time. It is therefore of the utmost importance to expand this technology and efforts to all different populations throughout the world (Need and Goldstein, 2009; Bustamante et al., 2011; Popejoy and Fullerton, 2016).

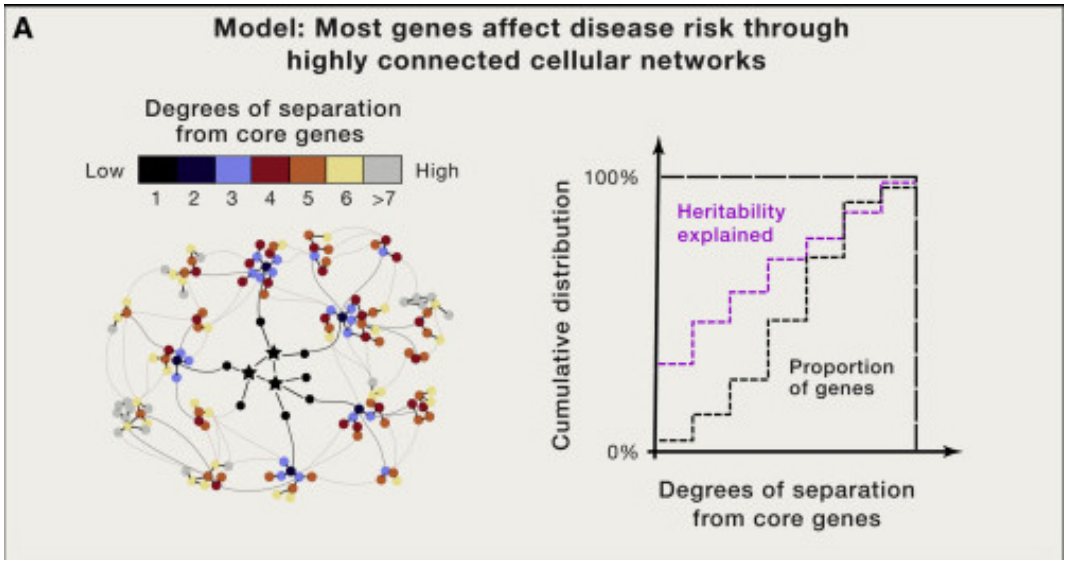

Figure 1.15: The omnigenic model for complex phenotypic traits. Left panel. A gene-network showing genes associated with complex phenotypic trait. Under the omnigenic model. only a small number of core genes (e.g. those possessing genetic variants showing genome-wide significant; black stars) will have a direct effect on the phenotypic trait. However, due to the small world property of networks, the majority of genes will be only a few steps aways (nodes) from the core genes and thus, will have a non-zero effect on the phenotypic trait. Right panel. Cumulative distribution of the heritability explained for a complex phenotypic traits as a function of the degree of separation from core genes. Since core genes only represent a small proportion of all genes, the vast majority of the heritability can be explained by the effect of many genes indirectly related to the cores genes. From Boyle et al. (2017).

1.7 Genomics is failing on diversity

As described in the previous section, over the last years a remarkable range of discoveries have been prompted by the use of GWAS methodology. Nonetheless, the vast majority of GWAS conducted to date have been exclusively performed in European descendant population and have therefore failed to explore the vast majority of genomic diversity worldwide and its potential for novel discovery (Need and Goldstein, 2009; Bustamante et al., 2011; Popejoy and Fullerton, 2016). Need & Goldstein (2009) reported that $>$ 96% of par-

ticipants in GWAS were of European ancestry (Figure 1.16). These findings prompted the response of some members of the scientific community to advocate for more inclusive studies. Bustamante et al. (2011) warned:

"Geneticists worldwide must investigate a much broader ensemble of populations, including racial and ethnic minorities. If we do not, a biased picture will emerge of which variants are important, and genomic medicine will largely benefit a privileged few."

A recent reanalysis of the number of participants from different ethnicities showed that while the percentages of participants of non-European ancestry had increased from 4% to 19%, approximately 14% of this was due to the inclusion of participants of Asian ancestry, while the inclusion of participants from other ancestries had barely changed or even decreased (Figure 1.16) (Popejoy and Fullerton, 2016). For example, the number of Latin American individuals in GWAS up to 2016 currently constitutes 0.5% and that of indigenous Natives only 0.05%, compared to 0.06% and 0.06% in 2009. If this bias continues, research efforts will still fail to capture most of the genetic diversity of human populations and thus fail to discover associations that could benefit the populations which currently have the greatest health disparities (Bustamante et al., 2011). Although some of the persistent bias may be due to logistical and systemic biases, e.g. due to the costs related to the recruiting of volunteers in order to have enough statistical power, or the logistics of conducting studies in countries with low resources (Bustamante et al., 2011; Popejoy and Fullerton, 2016), conducting GWAS in these settings can still provide to be cost-effective and/or produce novel biological insights. For example, the aforementioned discovery of a novel variant in Inuit populations associated to height and cholesterol levels (Fumagalli et al., 2015), showed that this variant had also an effect on height in European populations, a conclusion that would have otherwise been missed given the extremely low frequency of this variant in Europeans. In a more recent example, Crawford et al. (2017) discovered a novel variant associated with skin pigmentation in a sample of $\sim$ 2,000 individuals, even though skin pigmentation has been extensively researched in European populations with bigger sample sizes. Finally, a number of big consortia such as the Population Architecture using Genomics and Epidemiology (PAGE) Consortium have started to use novel strategies for multi-ethnic analysis, including studying admixed populations, and have shown strong evidence of the genetic effect-size heterogeneity across ancestries for previously published associations (Wojcik et al., 2017). These results highlight the need for novel and larger genomic efforts in diverse human populations.

1.8 Consortium for the Analysis of the Diversity and Evolution of Latin America - CANDELA

The Consortium for the Analysis of the Diversity and Evolution of Latin America (CANDELA) is a research effort focused on the evolutionary history and phenotypic variation of Latin America (Ruiz-Linares et al., 2014). The sample consists of over 6,000 Latin American individuals from five different locations: Porto Alegre in Brazil, Arica in Chile,

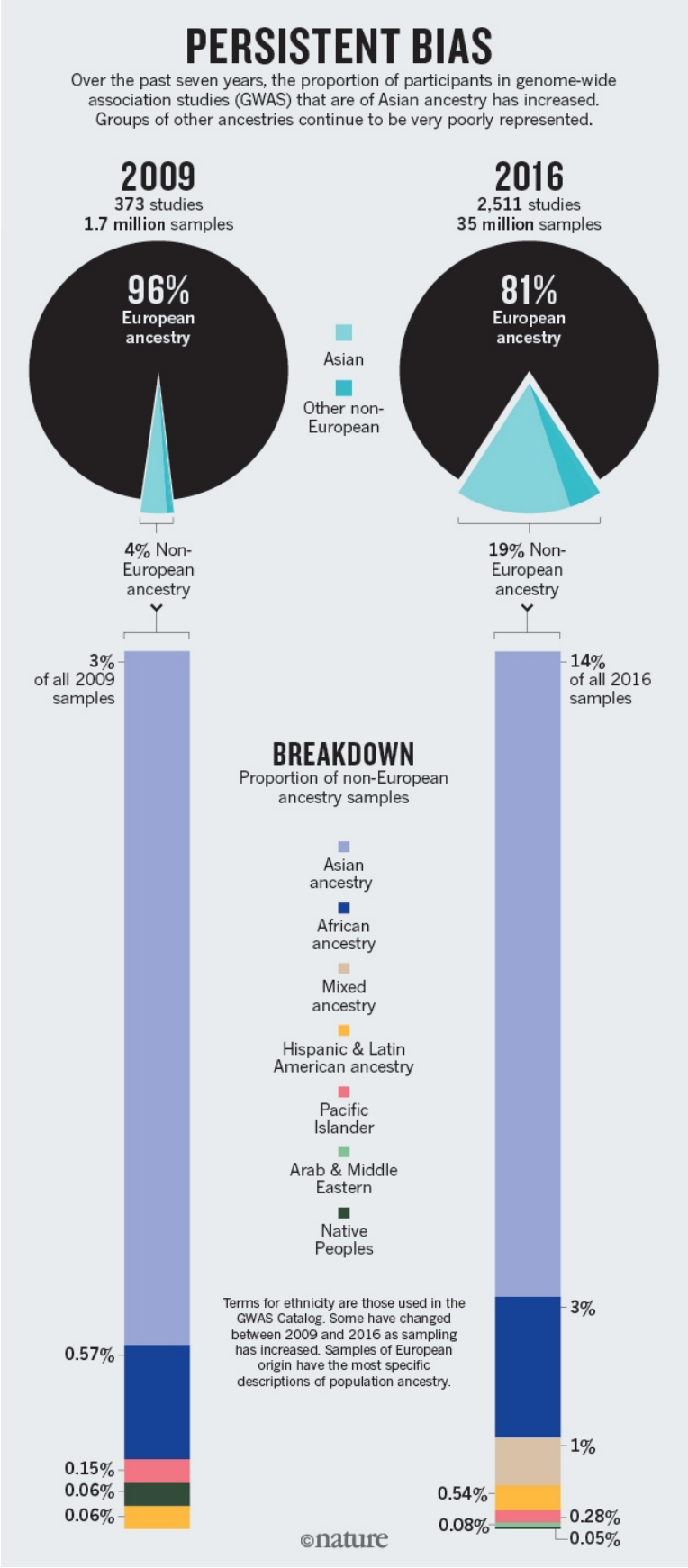

Figure 1.16: Proportions of volunteers of different ancestries in Genome-Wide Association Studies (GWAS) in 2009 and 2016.. During the past years, the proportion of volunteers used in GWAS from participants who are not of European ancestry has increased fivefold. However, $\sim 78\%$ of this growth is due to an increase in the number of volunteers from Asian ancestry. From Popejoy and Fullerton (2016).

Medellin in Colombia, Mexico City in Mexico and Lima in Peru (Figure 1.17). Adult individuals of both sexes were recruited from these sites with the condition that their four grandparents had been born in the same country. Information on a range of variables was obtained for each volunteer, including phenotypic data, socio-economic information and self-perception of ancestry. Genomic data was obtained using the Illumina Omni express bead chip consisting of 730,525 SNPs. An extensive description of this sample can be found in Ruiz-Linares et al. (2014) and Adhikari et al. (2016a).

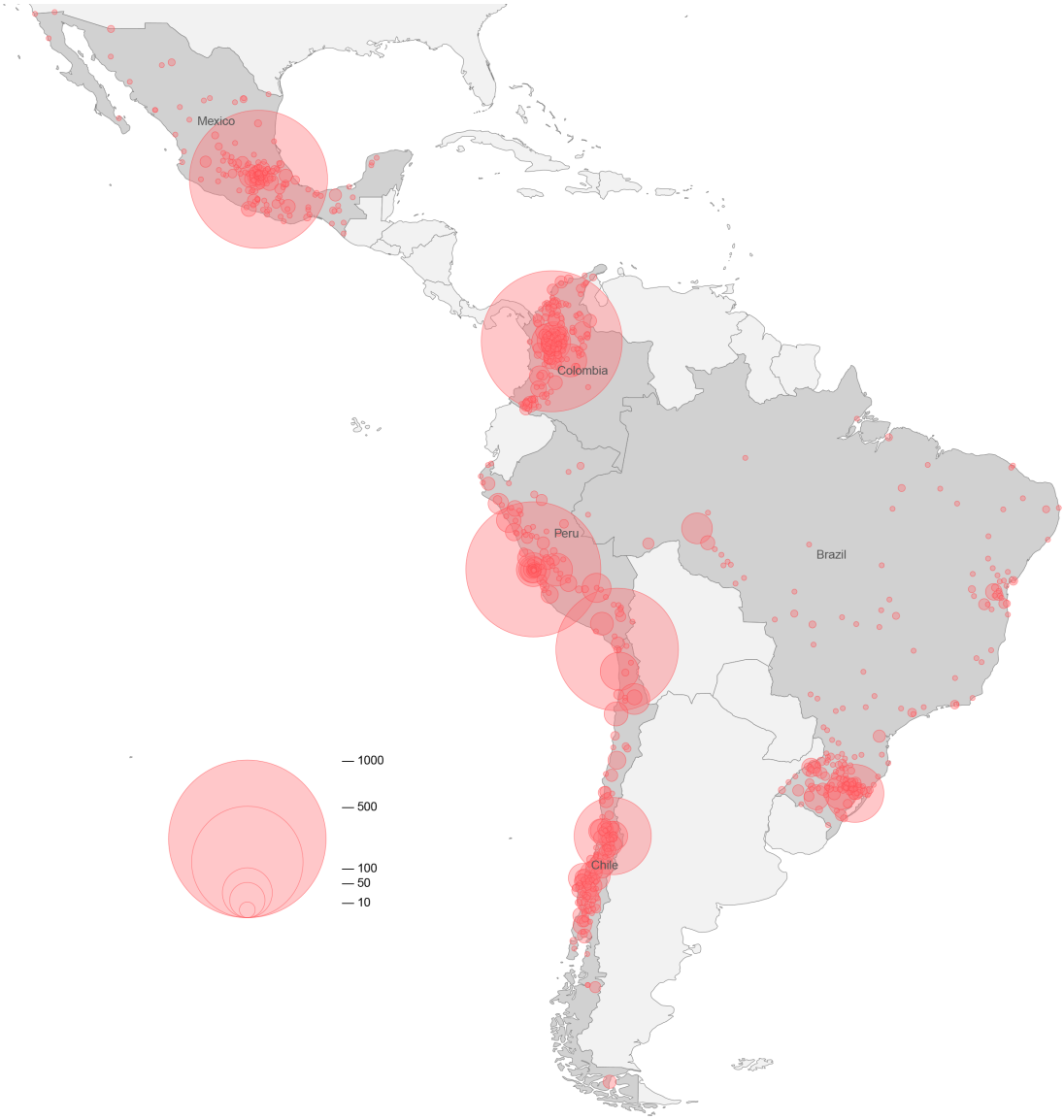

Figure 1.17: Birthplace locations of CANDELA volunteers. Adapted from Ruiz-Linares et al. (2014).

1.9 Summary

In this introduction I have described what is currently known about the evolutionary demographic and adaptive history of modern populations in the Americas, including Native Americans and admixed Latin Americans. I have also presented the evolutionary history of pigmentation variation in humans, and the known genetic determinant that influence

this trait. I have also described the rationale and scientific basis of GWAS and highlighted the importance of including underrepresented human populations in genetic association studies. In short, this thesis aims to shed further light on the adaptive history of Native Americans and admixed Latin American populations and discover novel variants associated to pigmentation phenotypes trough a GWAS in a large sample of admixed Latin Americans.

In chapter 2 I outline some of the commonly used methods currently applied to genome-wide autosomal data in order to detect natural selection. I describe a variety of methods including those that aim to discover signals of selection at a particular locus or a biological pathway. I also describe selection methods that rely on external data, such as environmental variables in order to get insight into the selective pressures driving the adaptive process. Additionally, I also present the methodological aspects of a GWAS and various commonly applied methods to account for population structure, a feature that is present in the CANDELA sample. Some of these methods are referred to in subsequent chapters within this thesis.

In chapter 3 I conduct a genome-wide scan of selection in Native Americans. I provide important candidate genes that were likely beneficial in the ancestral population of Native Americans in Beringia, prior to their entry into the Americas. I show that some of the selected variants are shared with several Arctic populations and found at high frequency, consistent with a shared adaptive event. In this chapter I also report candidate genes in local Native American populations by leveraging the predominant Native American ancestry present in admixed Latin Americans from the CANDELA cohort. I report candidate regions of selection impacting on immune-related genes that probably resulted from an adaptation to local pathogens in the Americas or perhaps to diseases brought after European contact.

In chapter 4 I conduct a genome-wide scan of selection in five admixed Latin American population from the CANDELA sample. I do this via a novel statistical model that detects signals of selection post-admixture. I report a strong signal of selection post-admixture in the Peruvian sample at a genomic region associated with glucose metabolism. In addition, I also report a significant increase of African ancestry at the MHC in the Chilean and Mexican populations. The genes at MHC involved in infectious disease resistance might have been selected due to diseases brought from the Old World after European contact.

In chapter 5 I report novel variants associated to skin and eye pigmentation in admixed Latin Americans. The results highlight the complex genetic architecture of pigmentation in Latin Americans, as evidenced by independent variants at different gene regions as well as multiple independent variants within gene regions. The novel associations using quantitative eye color variables show the greater statistical power obtained by using sensible color models. I also report a novel associated variant in the *MFSD12* gene, which represents a potential East Asian and Native American specific skin pigmentation locus.

In chapter 6 I provide evidence that the novel variant in *MFSD12* played a role in shaping lighter skin pigmentation in East Asians but not in Europeans. I further show that the distribution of the derived allele frequency of this variant seemed to have been affected by the solar radiation intensity in East Asia, supporting the role of natural selec-

tion in shaping skin pigmentation variation and the convergent evolution of lighter skin pigmentation that occurred in Eurasia.

In the final chapter I discuss and conclude the work presented in this thesis. I end by describing the possible directions and significance of research in studies of human population adaptation and genetic association studies.

In each of the above chapters tables and figures are included as part of the chapter. Where this is not the case, but that additional tables and figures make important contributions to the chapter narrative, they are included in the thesis appendices and referred to with a letter A to D.

Chapter 2

Methods

2.1 Overview

There is an increasing interest in genomic regions that have been targeted by natural selection. The interest stems from the desire to learn more about evolutionary processes and from the realization that inferences regarding selection can provide important functional information. The effect of natural selection can leave distinctive patterns of DNA variation along the genome and many statistics have been developed to detect these signals. In this chapter, I review methods commonly applied to detect signals of selection. Methods to detect selection can be broadly classified between approaches that detect selection at the macroevolutionary level (i.e. at the species level) and at the microevolutionary level (i.e. at the population level). I describe methods aimed at detecting instances of selection at the population level that are based on sequencing and genotype data. I end this chapter by describing the experimental design of Genome Wide Association Studies (GWASs). In recent years GWASs have led to a remarkable range of discoveries in human genetics that have facilitated not only the understanding of the biology of diseases, but also to a better understanding of the genetic architecture of complex traits. I present the underlying methodology in a GWAS by describing the statistical framework with an emphasis on how to account and correct for population structure.

2.2 SNP-based approaches to detect selection

2.2.1 Allele frequency differentiation based approaches

Allele frequency differentiation methods rely on the principle that allele frequency differences between populations can arise due to differences in environmental pressures. That is, if selection is acting on one population, but not the other, then selection on a beneficial allele will tend to increase the levels of genetic differentiation between populations at that particular locus.

One of the most common used statistics for population differentiation is Wright's fixation index (F_{ST}) (Wright, 1949). F_{ST} can be defined in different ways (Bhatia et al., 2013), with one typical implementation simply comparing the variance of allele frequencies

within and between populations and usually defined as:

$$F_{ST} = \frac{\delta_S^2}{\delta_T^2} \tag{2.1}$$

$$F_{ST} = \frac{\delta_S^2}{\overline{p} - (1 - \overline{p})} \tag{2.2}$$

Where $\overline{p}$ is the mean frequency of an allele in the total population, δ_S^2 the variance in allele frequency of the allele between populations (weighted by population size), and δ_T^2 the variance in the total population. Large values of F_{ST} will indicate strong differentiation between populations, e.g. positive selection in one population. Values close to zero can indicate that the population are genetically homogenous, which could also indicate balancing or positive selection at both populations. One of the advantages of population differentiation methods is that they can detect different types of selection, including hard or soft-sweeps. However, one disadvantage is that F_{ST} does not offers directionality so it is not possible to identify the population which selection has been acting on.

There are different derivatives from F_{ST} that address this issue, and in turn improve its power by analysing more than two populations. The Locus Specific Brach Length (LSBL) (Shriver et al., 2004) uses pairwise F_{ST} measurements between three populations, and is capable of discerning population specific allele frequency changes. A very similar method that incorporates data from three populations is the Population Branch Statistic (PBS) (Yi et al., 2010). It involves converting the F_{ST} estimates into branch lengths (T) via a log-transformation (Cavalli-Sforza, 1969):

$$T = -log(1 - F_{ST}) \tag{2.3}$$

Using the above log-transformation has the benefit of obtaining additive distances that place branches of different magnitudes on the same scale. PBS is then defined as:

$$PBS = \frac{T^{TargetControl} + T^{TargetOutgroup} - T^{ControlOutgroup}}{2} \tag{2.4}$$

where $T^{TargetControl}$ represents the branch length between a target population and a control population, $T^{TargetOutgroup}$ the branch length between a target population and an outgroup population, and $T^{ControlOutgroup}$ the branch length between the control and outgroup populations. A PBS value for a target population thus represents an estimate from the amount of allele frequency change at a locus since the divergence from the other two populations (Yi et al., 2010) (Figure 2.1). Additionally, because one of the reference populations used in the analysis can be a population with a very short divergence time, PBS has greater power to detect incomplete sweeps compared to other classic neutrality test methods (e.g. Tajima's D) (Yi et al., 2010). Recently, a modified version of the PBS statistics termed PBS_{n1} (Crawford et al., 2017a) has been proposed where a rescaling of the PBS scores are computed in order to account for saturation (i.e. when F_{ST} are high or low between all populations) avoiding artificially high PBS values.

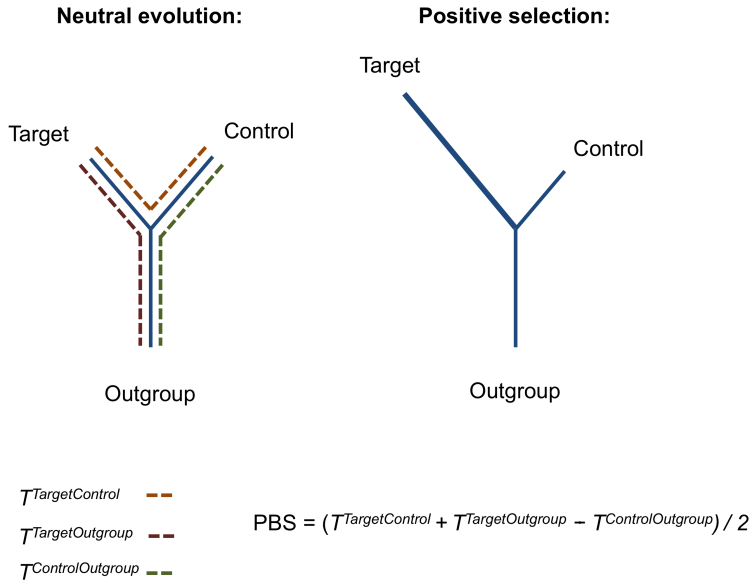

Figure 2.1: Schematic representation of the Population Branch Statistic. An unrooted tree showing the relationship between a Target, Control and Outgroup population under neutral evolution. Dotted lines along the branches represent the estimated branch lengths (T) between each pair of populations (left panel). An unrooteed tree of the same three populations showing an instance of positive selection in the Target population (right panel). The Population Branch Statistic (PBS) equation estimates the length of the branch leading to the Target population since the divergence from the Control and Outgroup population (bottom panel). Adapted from a figure provided by Matteo Fumagalli.

2.2.2 Haplotype-based approaches

Haplotype-based methods rely on Linkage Disequilibrium (LD) structure along genomes. Specifically, the genetic diversity at or near a locus under selection is reduced, and nearby linked SNPs tend to be homozygous (and identical by descent). The reduction in genetic variation at a locus under selection has been termed genetic hitchhiking (Smith and Haigh, 1974) and different types of statistics that have been developed to exploit this genetic signature. Most are based on a model of a hard-sweep where a de novo mutation arises in a particular haplotypic background and quickly rises up in frequency due to selection. A similar model, termed soft-sweep, occurs when the selected variant was previously segregating in the populations in different haplotypic backgrounds before the selection event. Depending on the selection intensity, this process can occur at a faster rate than the recombination rate leaving a distinct pattern along the genome. One of the earliest methods developed is the Extended Haplotype Homozygosity (EHH) statistic (Sabeti et al., 2002a). This statistic measures the probability that two random chromosomal segments with the same "core" haplotype (i.e. a haplotype with a locus of interest) are homozygous for an extended interval from this region. Formally, for a sample on n chromosomes, if C denotes all possible distinct haplotypes at a core locus and $C(x_i)$ all possible distinct haplotypes extending (upstream and downstream) i number of markers from the locus of interest,

following Szpiech and Hernandez (2014), EHH can be defined as:

$$EHH(x_i) = \sum_{h \in C} \frac{\binom{n_h}{2}}{\binom{n}{2}} \tag{2.5}$$

where n_h represents the number of observed haplotypes of type $h \in C(x_i)$. When a core haplotype containing an adaptive locus has recently reached or is reaching towards fixation, $EHH(x_i)$ is expected to decay much slower than a core haplotype evolving under neutrality. Thus, to formally assess for selection at a locus a relative EHH (rEHH) is usually calculated by comparing the EHH at a locus of interest to the EHH of other haplotypes randomly selected across the genome.

2.2.2.1 Integrated Haplotype Score (iHS)

Building from this same premise, many derivatives of the EHH statistic have also been developed. The integrated Haplotype Score (iHS) statistic (Voight et al., 2006) computes the EHH for locus of interest and estimates the area under EHH for the haplotypes carrying the ancestral (0) and derived alleles (1), separately. Because the areas are integrated with respect to genetic distances, these estimates are referred to as integrated Haplotype Homozygosity (iHH) scores. Formally, and following Szpiech and Hernandez (2014), iHH can be defined as:

$$iHH = \sum_{i=1}^{D} \frac{1}{2}(EHH(x_{i-1}) + EHH(x_i))g(x_{i-1}, x_i) + \sum_{i=1}^{U} \frac{1}{2}(EHH(x_{i-1}) + EHH(x_i))g(x_{i-1}, x_i) \tag{2.6}$$

where D is the set of loci downstream of the locus of interest, such that $x_i \in D$ denotes the i$^{\text{th}}$ closest downstream locus from the locus of interest (x_o); U and $x_i \in U$ are defined similarly; and $g(x_{i-1}, x_i)$ is the genetic distance between two adjacent loci. The unstandardized iHS is computed as a log-ratio of the ancestral and derived iHH scores:

$$iHS_{(unstandardized)} = \ln\left(\frac{IHH_1}{IHH_0}\right) \tag{2.7}$$

Under this formulation, unstandarized iHS close to zero indicate that the rate of decay of EHH is the same on the ancestral and derived haplotypes and there is therefore, low evidence for selection. Large negative values indicate longer than expected haplotypes carrying the derived allele and positive values indicate the converse, and are indicative of selection. Because unstandardized iHS can be correlated with allele frequency (e.g. low frequency variants are expected to be young and reside in longer haplotypes), the final standardized iHS score (i.e. mean zero and unit standard deviation [SD]) is computed using frequency bins across the genome. Formally:

$$iHS = \frac{\ln\left(\frac{IHH_1}{IHH_0}\right) - E_p[\ln\left(\frac{IHH_1}{IHH_0}\right)]}{SD_p[\ln\left(\frac{IHH_1}{IHH_0}\right)]} \tag{2.8}$$

where $E_p[\ln\left(\frac{IHH_1}{IHH_0}\right)]$ and $SD_p[\ln\left(\frac{IHH_1}{IHH_0}\right)]$ is the mean or expected and SD of the $iHS_{(unstandardized)}$

at frequency bin p, respectively. Although large negative iHS scores (indicating that the haplotype carrying the derived allele has swept up in frequency) would suggest selection, Voight et al. (2006) showed that large positive iHS scores can also arise at nearby SNPs if the haplotype carrying the ancestral allele has been hitchhiked with the selected variant. Therefore, it is common to search for high absolute iHS values (i.e. both negative and positive iHS values) as signals of selection.

2.2.2.2 Number of Segregating Sites by Length (nSL)

The number of Segregating Sites by Length (nSL) statistic (Ferrer-Admetlla et al., 2014) is another approach to detect long stretches of haplotypes in homozygosity that is highly related to iHS. The main difference is the use of physical distance instead of genetic map information. While iHS estimates iHH by using the genetic distance between adjacent SNPs, nSL measures the estimates based on the number of segregating sites. Because a genetic map is not required, this statistic is more robust towards recombination and mutation rates (Ferrer-Admetlla et al., 2014). Similar to iHS, the log-ratio of iHH scores at derived and ancestral haplotypes are computed and standardized in frequency bins across the genomes to obtain a final standardize nSL score. In contrast to iHS, nSL has been shown to have greater power to detect sweeps from standing variation (i.e. soft-sweeps) and incomplete sweeps (Ferrer-Admetlla et al., 2014).

2.2.2.3 Cross Population Extended Haplotype Homozygosity (XP-EHH)

Another haplotype-based method, called the Cross-Population Extended Haplotype Homozygosity (XP-EHH) statistic (Sabeti et al., 2007), has been developed to compare haplotype homozygosity patterns between a target and a reference population. Similarly to iHS, iHH (based on genetic distance) is calculated for each population separately, and the log ratio of these values is estimated to obtain an unstandardized XP-EHH score. The unstandardized XP-EHH scores are typically standardized with respective to the genome-wide distribution of XP-EHH scores. In contrast to iHS and nSL, however, the XP-EHH scores are directional, and positive values (or negative) are indicative of longer haplotypes in the target (or reference) populations depending on the formulation of the log-ratio. Due to the use of a target and reference population, XP-EHH scores have been shown to possess a higher statistical power to detect selection at SNPs where the frequency has approached or reached fixation at a site compared to other haplotype-based methods (Sabeti et al., 2007).

One major limitation of haplotype-based methods is the effect of recombination rates and especially recombination hot-spots on haplotype homozygosity patterns. Although, some statistics, like nSL, may be more robust to variation in recombination rates, it is still not clear how a genetic map estimated from one population can be applied to other human populations (Hinch et al., 2011). Additionally, population genetics theory (Pritchard et al., 2010) as well as recent empirical data based on genomic scans of selection across different populations (Schrider and Kern, 2016, 2017a) both suggest that soft-sweeps are the

dominant mode of adaptation in the human genome. Many of the methods describe above, still lack enough statistical power to detect soft-sweeps, although novel methods with the specific aim of detecting this pattern have also recently been developed (Garud et al., 2015). Novel methods, based on powerful machine learning approaches with high power to detect and differentiate between different type of selective sweeps have also started to been developed with great promise (Pavlidis et al., 2010; Lin et al., 2011a; Ronen et al., 2013; Pybus et al., 2015; Schrider and Kern, 2016; Sheehan and Song, 2016a; Flagel et al., 2018; Sugden et al., 2018).

2.2.3 Time scales for the signatures of selections

Finally, while more powerful methods might be able to better detect and differentiate between different type of selective events, it is important to consider that the different type of genomic signatures that these methods aim to detect will persist over varying different time scales, which will ultimately define their statistical power (Smith and O'Brien, 2005; Sabeti et al., 2006; Oleksyk et al., 2010). Allele-frequency differentiation approaches (Section 2.2.1), rely on the degree of reproductive isolation between populations, which in the case of human populations, will limit their power to detect events that occurred >75,000 years ago (Smith and O'Brien, 2005; Sabeti et al., 2006). In the case of haplotype-based approaches (Section 2.2.2), the major limitations is due to the rapid break down of haplotypes over generations. For example, after approximately 30,000 years a chromosome is estimated to have undergone more than one recombination ever per 100kb, leaving haplotypic fragments too short to detect for current haplotype-based selection statistics.

2.3 Combining selection signals from many loci

Phenotypic traits that are controlled by two or more genes are called polygenic phenotypes or polygenic traits (Stranger et al., 2011). When a polygenic trait is undergoing adaptation, a process called polygenic adaptation, the signature of adaptation tends to be more diffused and thus, no specific region along the genome shows a strong signature of selection (Pritchard et al., 2010). The selection process however, affects the distribution of allele frequency at these loci, thus, producing an adaptive signature across the genome (Latta, 1998, 2004; Pritchard et al., 2010; Le Corre and Kremer, 2012; Berg and Coop, 2014; Stephan, 2016). Although the majority of selection statistics are not well suited to capture this type of selection, using the information on specific genes or variants from different databases such as biological pathways (Daub et al., 2013, 2015; Hsieh et al., 2016; Polimanti et al., 2016; Daub et al., 2017; Owers et al., 2017; Bergey et al., 2018) or Genome Wide Association Studies (GWAS) (Turchin et al., 2012; Corona et al., 2013; Fraser, 2013; Berg and Coop, 2014; Adhikari et al., 2016a; Berg et al., 2017) can produce statistics to detect this type of selection.

Using information from GWAS, recent studies have tried to detect instances of polygenic adaptation by looking for coordinated shifts in the allele frequency distributions of

different polygenic phenotypes (Turchin et al., 2012; Fraser, 2013; Corona et al., 2013). Berg and Coop (2014) presented an improvement over these works, by being able to detect instances of polygenic adaptation using several populations with an arbitrary underlying population structure. They proposed an excess of variance test that compares the Polygenic Risk Scores (PRS) (i.e. the sum of population allele frequencies weighted by effect size) across populations, to the distribution of PRS based on a null model constructed using randomly chosen loci which should therefore capture the underlying population history between populations. The distance of the observed PRS is then compared to the distribution of the null model to assess the significance of a polygenic adaptive event for a particular trait (Berg and Coop, 2014). More recently, Racimo et al. (2018a) developed a method to detect polygenic adaptation in an admixture graph, which is a representation of the historical splits and admixture events between different populations through time. Notably, their new method can not only detect which populations have evidence of polygenic adaptation, but also where in the history of these populations adaptation occurred.

Other commonly used methods that exploit signals of selection across loci are gene set enrichment approaches. This type of approach involves testing whether the distribution of a selection statistic computed across genes from a particular biological category (e.g. Gene Ontology [GO] (Ashburner et al., 2000) or biological pathways (Kanehisa and Goto, 2000)) statistically differs from the genome-wide expectation. This method differs from those aimed at detecting signals of polygenic adaptation using GWAS data, in that the individual selection scores computed at each locus are used only to assign selection scores to each gene, and no information regarding whether the particular locus is associated to a phenotype (or its genetic effect on that particular phenotype) is taken into account. Nonetheless, this information can also be easily incorporated in this kind of approaches, for example by simply defining a set of genes on the basis of whether they have variants that have been previously associated to a particular phenotype, as has been done previously in (Adhikari et al., 2016a) and in this thesis (Section 6.3.3). One of the first methods, termed set enrichment analysis (GSEA) (Subramanian et al., 2005), simply employs a weighted Kolmogorov-Smirnov test to assess enrichment of a biological pathway. Building up from this approach several other variation have also been developed (Subramanian et al., 2007; Wang et al., 2007a; Holden et al., 2008; Nam et al., 2010; Zhang et al., 2010a; Kofler and Schlötterer, 2012; Rosenberger et al., 2015; Schmid et al., 2016; Yoon et al., 2018). A method developed by Daub et al. (2013) computes the sum of a selection statistic associated to each gene in a given biological pathway (termed the SUMSTAT score (Tintle et al., 2009)) and compares it to the SUMSTAT score using a set of random genes. After correcting for the number of variants used to compute each selection statistic per gene, a significantly different SUMSTAT score for a given biological pathway can be used as evidence for polygenic adaptation (Daub et al. 2013; Daub et al. 2016). Other very similar implementations involve conducting non-parametric tests, such a Mann-Whitney U test, to compare if the distribution of selection statistics for a given biological pathway is significantly different to the distribution for all other genes across the genome (Adhikari et al., 2016a; Hsieh et al., 2016; Owers et al., 2017).

Because the majority of phenotypic traits are highly polygenic (Falconer, 1960) it is likely that the majority of selection at these traits will occur due to polygenic adaptation, rather than by hard selective sweeps (Pritchard et al. 2010). It is therefore expected that with the increase in number of GWAS and/or better annotation of current genomic regions, tests for polygenic adaptation will discover many more novel instances of human adaptation.

2.4 Using environmental data to identify loci underlying local adaptation

Comparing allele frequencies between populations that differ in environmental conditions was one of the earliest methods developed to detect selection, and approaches were usually based on allele frequency differentiation estimates such as F_{ST} measures (Cavalli-Sforza, 1966; Lewontin and Krakauer, 1973; Endler, 1983). Other approaches, although similar in principle, differed from these early efforts by looking at associations between allele frequencies and environmental variables, in order to identify the environmental pressures (or a highly correlated variable) driving the selection (Haldane, 1948; Slatkin, 1973; Mullen and Hoekstra, 2008). Modern implementations include conducting a logistic regression between the counted allele data and environmental variables (Joost et al., 2007). However, conducting a standard logistic regression assumes that the allele counts across populations are independent and thus fails to capture the correlation between populations due to their shared genetic history. Another method to detect correlations between genetic and environmental variables involves conducting a partial Mantel test. Under this approach, the partial Mantel test is first used to estimate the relationship between allele frequencies and geographical distances between populations, and then to test for the effect of an environmental variable above and beyond isolation by distance (Nadkarni et al., 2005; Balloux et al., 2009). Because this approach involves the use of 3 matrices, significance can be assessed via permutations in which rows and the corresponding columns are shuffled at random.

A more modern implementation of this principle uses a Bayesian framework to test the fit of a model with a linear relationship between allele frequencies and an environmental variable over a null model in which the allele frequencies are dependent on population structure alone (Coop et al., 2010; Günther and Coop, 2013). Specifically, given a set of populations a null model is first constructed using a set of unlinked variants to estimate how allele frequencies covary across populations. Sample allele frequencies are then drawn from a set of underlying population frequencies assumed to be distributed according to a multivariate normal distribution around a transformed global allele frequency, with a variance-covariance matrix representing population structure across populations. The Markov chain Monte Carlo (MCMC) algorithm is then used to sample from the posterior distribution of the covariance matrix given the ancestral allele frequencies (α) and pop-

ulation allele frequencies. The alternative model is then constructed to allow the allele frequency to be dependent on an environmental variable. Specifically, the allele frequency at a tested locus is allowed to have a deviation from its ancestral allele frequency that is linearly proportional to the tested environmental variable with a coefficient β. Formally:

$$P(\theta|\Omega, \alpha, \beta) \sim N(\alpha + \beta Y, \alpha(1-\alpha)\Omega) \tag{2.9}$$

Where θ is the transformed allele frequency, α is the ancestral allele frequency, Y is the environmental variable, and Ω is the variance-covariance matrix from a single draw from the posterior estimated in the null model. A Bayes Factor (BF) is constructed to estimate the support for the alternative over the null model. This method has greater power compared to other similar methods, mainly because it accounts for population structure (Coop et al., 2010; Günther and Coop, 2013; De Mita et al., 2013). It has also been applied to humans producing strong candidate loci thought to be under selection (Hancock et al., 2008, 2011; Fumagalli et al., 2011; Kita and Fraser, 2016). Other similar methods that estimate the covariance matrix by explicitly modelling isolation by distance have been shown to produce considerable gains in computational time (Guillot et al., 2014). Additionally, other methods based on a latent factor mixed model that estimates the effect of population history and environmental correlations simultaneously have also been devoloped (Frichot et al., 2013). Finally, models that were originally developed to compare traits across species accounting for phylogenetic autocorrelation, such as the Phylogenetic Regression (Grafen, 1989), have also recently been applied to human data by computing a phylogenetic tree based on genome-wide F_{ST} values (Key et al., 2018).

2.5 SNP-based approaches to detect natural selection in admixed populations

2.5.1 Local ancestry deviations

Under evolutionary neutrality it is expected that the mean local ancestry at a particular genomic region (averaged across all individuals) should follow the genome-wide ancestry average. However, the local ancestry proportion at a genomic region can deviate from the expectation due to sampling error in the ancestral reference or admixed population, genetic drift and/or selection. A significantly strong deviation is usually suggested as a being caused by some type of selection. Tang et al. (2007) formally tested this by introducing a new statistic called delta-ancestry (δ_k^m). Specifically, for a particular ancestry k at a genomic locus m, delta ancestry is defined as:

$$\delta_k^m = \overline{q}_k^m - \overline{q}_k \tag{2.10}$$

where $\overline{q}_k^m$ is the mean of ancestry k at genomic locus m averaged over all individuals, and $\overline{q}_k$ is the proportion of ancestry k averaged over all individuals and the entire genome. In order to assess significance it is straightforward to perform a permutation test (Tang

et al., 2007) or, under a normal approximation, compute associated P-values using the corresponding standard deviations at each site (Zhou et al., 2016).

2.6 Inferring the starting time and intensity of selection via Approximate Bayesian Computation (ABC)

The methods discussed above have been mainly developed to detect genomic regions under selection. However, another major goal in population genetics in recent years has been to infer the parameters underlying this adaptive process, such as the time when a selected variant arose, the time when the variant started to be selected and the selection coefficient (Beaumont et al., 2002; Wegmann et al., 2010; Peter et al., 2012). One of the methodologies developed to estimate these parameters in complex demographic models is the Approximate Bayesian Computation (ABC) framework (Beaumont et al., 2002; Fagundes et al., 2007; Blum and François, 2010; Csilléry et al., 2010; Wegmann et al., 2010). ABC is a rejection sampling algorithm used to estimate the posterior distribution of a parameter (θ) under a given model, commonly used when the likelihood cannot be computed analytically. Formally, in the ABC inference framework, a parameter value θ_i is sampled from a prior distribution to simulate a dataset y_i, for $i = 1...n$, where n denotes the number of simulations. A set of summary statistics $S(y_i)$ is then computed from the simulated data and compared to the set of summary statistics obtained from the actual data $S(y_0)$. If these set of summary statistics are sufficient (i.e capture all the information present in the simulated data), this step is exact (Peter et al., 2012). However, in reality, computed summary statistics do not capture the full information present in the data and comparing the set of simulated summary statistics $S(y_i)$ to the observed $S(y_0)$ results in an approximation step (Peter et al., 2012). The condition of an exact match is relaxed and therefore, some distance measure (δ) (such as the absolute Euclidean distance) is considered, i.e.

$$|\delta(S(y_i), S(y_0)| < \epsilon \tag{2.11}$$

and only simulations with some arbitrarily, but conservative small distance ϵ are accepted. The accepted θ_i represent a sample from an approximation of the posterior distribution. Point estimates and intervals, such as the Maximum A Posteriori (MAP) and Bayesian Credible Intervals (BCI) can be computed. It is also common to perform some post-sampling adjustment to correct for the approximation step, e.g. by using regression techniques (Blum and François, 2010).

While computing a large number of summary statistics can indeed increase the amount of information extracted from the data, an important phenomenon that can arise is the so-called "curse of dimensionality" (Wegmann et al., 2010). When there is a large summary statistics space, it becomes extremely difficult to obtain simulations that will closely match the observed data, which in turn affects the threshold used in the rejection step. Additionally, the larger the number of summary statistics, the larger the noise that is basically included in the posterior estimation (Joyce and Marjoram, 2008). To solve this

problem, Wegmann et al. (2010) proposed to transform the summary statistics via Partial Least Squares (PLS) (Boulesteix and Strimmer, 2006), to obtain a number of orthogonal linear combinations of the summary statistics that best explain the variance in the model parameter space.

2.7 Genome Wide Association Studies (GWAS)

2.7.1 Single marker associations

Univariate and multivariate regression analysis using generalized linear models (GLM) are a suitable framework to detect genetic associations between a genotyped variant and a phenotype of interest (Stram, 2016). Linear or logistic regression analysis of a continuous (i.e. a quantitative phenotype) or a discrete (i.e. case-control status) thus represent natural statistical tools to test for associations (Balding, 2006). In this section I will consider the case where the data used for the genetic association include genotype and phenotype data of a sample of n individuals. y_i denotes the phenotype value (i.e. the dependent variable) of the i^{th} individual. This value can be either continuous (in the case of a quantitative variable) or discrete with values of 0 or 1 (in the case of a case-control study). I will only consider a bi-allelic SNP marker with alleles A and a as the independent variable. This variable is usually presented as an indicator variable, where g_i can be either 0, 1 or 2, depending on the number of copies of the allele A present in individual i. Additional covariates (e.g. age and/or sex) will be represented as x_{ij} with $j = 1, ..., r$. Formally, when the phenotype y is a continuous variable, the standard regression analysis can be presented as:

$$y_i = \beta_0 + \beta_1 g_i + \beta_2 x_{i1} + ... + \epsilon \tag{2.12}$$

where ϵ is the noise variable, i.e. the part of the phenotype variable y that is not explained by the SNPs or the additional covariates (e.g. environmental effects). When the y variable is discrete, a logistic regression analysis can be applied by the transformation $logit(\pi) = \log(\pi/(1-\pi))$, where π_i is the disease risk of the i^{th} individual. Other statistical tests can also be performed besides GLM (e.g. see Balding (2006) and Stram (2016)).

2.7.2 Correcting for population structure

2.7.2.1 Genomic control

In a GWAS, the desired reason for a significant result is a causal association between the genetic variant and the phenotype of interest. However, the result of this association can be confounded due to unaccounted population stratification (Astle et al., 2009). Neglecting or not accounting for population stratification can therefore lead to false positives or spurious associations (Balding, 2006). In this section I describe three methods that account for population stratification commonly used in GWAS.

In Genomic Control (GC) (Devlin and Roeder, 1999), the Armitage trend statistic is computed at each SNP, and compared to its expectation under the null hypothesis, which follows a χ^2_1 distribution. An inflation factor λ is calculated by dividing the median of the Armitage trend statistic computed at each SNP to the median of the χ^2_1. If there is no population structure, the distribution of the Armitage trend statistic computed at each SNP should follow the null and thus λ should be ≈ 1. If there is population structure, the distribution will deviate from χ^2_1 due to an inflated variance and thus λ is expected to be > 1. Devlin and Roeder (1999) proposed to account for population structure, in the case $\lambda > 1.12$, by dividing the test statistic used for the association by this λ inflation factor to cancel the population structure effect. However, it has been shown that GC can be either too conservative or too anticonservative in different settings (Marchini et al., 2004).

2.7.2.2 Regression-based adjustment for leading Principal Components

Principal Component Analysis (PCA) probably represents the most widely used method to identify and account for population structure in GWAS (Zhang et al., 2003; Patterson et al., 2006). This method involves applying PCA to the genotype data to infer continuous axes of variation to reduce this data to a number of small dimensions that explain as much variability as possible. Here, the idea is that the genotypes of the individuals should not be correlated in the absence of population structure. Since the PCs are linear combinations of the original genotypes, plotting the location of these individuals along the major axes of variation should not show any clustering. Formally, if Z denotes a matrix with n rows corresponding to the number of individuals and l corresponding to the number of SNPs, standardized to have zero mean and a unit variance, with i^{th}, l^{th} elements, for the ith individual and lth SNP, i.e.

$$Z_{il} = \frac{g_i^l - \hat{p}_l}{\sqrt{\hat{p}_l(1-\hat{p}_l)}} \tag{2.13}$$

where $\hat{p_l}$ is the estimated population allele frequency at SNP l, an estimated kinship matrix $\hat{K}$ can be obtained by:

$$\hat{K} = \frac{1}{l} Z Z^T \tag{2.14}$$

PCA is then performed by obtaining the eigendecomposition of the kinship matrix $\hat{K}$. In practice, the l SNPs must be in the orders of thousands and should not be in strong LD between each other (i.e. should be independent). The top PCs are then used as covariates, for example in a linear regression analysis, to account for the underlying population structure:

$$y_i = \beta_0 + \beta_1 g_i + \beta_2 x_{i1} + \beta_3 PC_{i1} + \beta_3 PC_{i2} + ... + \epsilon \tag{2.15}$$

2.7.2.3 Mixed Linear Models

Recently, linear mixed models (LMMs) have been proposed as a powerful approach to detect genetic association in a sample with related individuals. Following the same notation as above, the model is constructed as:

$$Y = W\alpha + g\beta + u + \epsilon \tag{2.16}$$

where W is an $n \times c$ matrix of covariates (fixed effects) including a column of 1s for the intercept, α is a c-vector of corresponding coefficients including the intercept, u is a n vector of random effects with $u \sim N(0, \sigma_g^2 K)$ where σ_g^2 represents the additive genetic variance, and ϵ is an n random vector with $\epsilon \sim N(0, \sigma_e^2 I)$ where σ_e^2 represents non-genetic variance and I the identity. In contrast to PCA, typically 100,000 or more SNPs are needed to construct the kinship matrix $\hat{K}$ in order to capture fine scale population structure, such as cryptic or familial relatedness (Astle et al., 2009). A number of LMMs have been proposed an implemented to conduct fast and robust GWAS (Kang et al., 2010; Zhang et al., 2010b; Lippert et al., 2011; Zhou et al., 2013b).

Finally, it is important to note that, although the most common cause of confounding in GWAS is due to unaccounted population structure, if the trait under study is polygenic i.e. affected by the (usually small) genetic effect of many variants, the distribution of the estimated test statistic can also be inflated (Bulik-Sullivan et al., 2015b). Briefly, if a phenotype is determined by many genetic variants along the genome, genetic variants that are in LD with these causal variants will also show a strong association and consequently affect the distribution of the test statistic (Bulik-Sullivan et al., 2015b). One solution that has been proposed in order to differentiate between these two types of confounding is the LD-score regression method, which can also provide a more powerful and accurate correction factor than the GC (Bulik-Sullivan et al., 2015b). However, because the method is based on the amount of LD between the genetic variants, this method is currently not appropriate for admixed populations, as these can show long-range LD due to their shared ancestry (Price et al., 2008; Bulik-Sullivan et al., 2015b).

2.8 Summary

In this chapter I have described commonly used methods to detect genomic regions under selection such as allele-frequency differentiation and haplotype-based methods. I have also described methods that rely on external data, such as environmental data in order to get insight into the selective pressures driving the adaptive process, or biological pathways that can help detect instances of polygenic adaptation. I ended by describing the methodological aspects of a GWAS and the various methods used to account for population structure, a feature that is present in admixed samples, such as the CANDELA sample used in this thesis.

Chapter 3

Detecting signatures of selection in Native Americans

3.1 Overview

In this chapter I conduct a genome-wide scan of selection on a large dataset of Native Americans. By considering samples from throughout the American continent I explore selective pressures imposed on the common ancestral population of Native Americans. Here I show that some of the candidate regions with the strongest selection signatures have functions that were likely beneficial for the climatic and dietary conditions in Beringia, prior to the range expansion into the American continent. Some of the variants with the strongest selection signatures are found in high frequency in several Arctic populations, consistent with a shared adaptive event as reported in previous studies. In addition, I also conduct a genome-wide selection scan on three Native American populations to explore instances of local adaptation in the Americas. I use a large sample of admixed Latin American individuals that derive most of their ancestry from these three Native American groups. As these individuals have extensive non-Native American ancestry, I correct for this admixture by computing pseudounadmixed allele frequencies. I report selection signals at immune-related genes in these Native American populations that probably resulted from an adaptation to local pathogens in the Americas or to diseases brought after European contact. I also report selection signals at genes with an important adaptive interest that have been previously reported in other Native American populations and other human populations highlighting the utility of this approach.

3.2 Background

The American continent represents the last major landmass to have been settled by people migrating from the northeastern tip of Asia through Beringia — a land bridge connecting Siberia to Alaska. Genomic studies have elucidated the complex demographic history of Native American populations, particularly those involving the distinct migratory episodes from East Asia, the time and location of the basal split between the main northern and southern Native American branches, and the migratory routes used to colonize the American continent (Wang et al., 2007b; Reich et al., 2012; Moreno-Estrada et al., 2014; Raghavan et al., 2015; Skoglund et al., 2015; de la Fuente et al., 2018; Moreno-Mayar et al., 2018;

Scheib et al., 2018; Schroeder et al., 2018). In addition, recent studies have also supported the long-term habitation of ancestral Native Americans in Beringia prior to their expansion to the American continent, an hypothesis known as the Beringian "standstill model" (Tamm et al., 2007; Raghavan et al., 2015; Llamas et al., 2016; Moreno-Mayar et al., 2018). Importantly, this long term isolation from other human groups in Beringia before entering the Americas seems to have imposed strong selective pressures in these ancestral Native American populations. Amorim et al. (2017) recently reported that one of the strongest signals of selection shared by Native Americans was present on the *FADS* gene cluster. This gene is involved in the metabolism of omega-3 polyunsaturated fatty acids (Lattka et al., 2010) and was previously found to be under strong selection in Greenlandic Inuit populations (Fumagalli et al., 2015). By analyzing the geographic distribution of the putatively selected haplotype, Amorim et al. (2017) concluded that a single adaptive event most likely occurred in Beringia before the expansion of the first Americans throughout the American continent (Figure 3.1). Building up from these findings, Hlusko et al. (2018) has recently suggested that the genetic adaption in the *FADS* gene cluster was likely accompanied by selection on the Ectodysplasin A (EDA) receptor (*EDAR*) gene. *EDAR* is known to have several pleiotropic effects as it influences ectodermally derived structures, such as hair, teeth hair, and mammary gland ductal branching (Fujimoto et al., 2008; Mou et al., 2008; Kimura et al., 2009; Park et al., 2012; Tan et al., 2013, 2014; Peng et al., 2016). This gene has been shown to be under strong positive selection in East Asians (Sabeti et al., 2007; Grossman et al., 2010), and a recent study showed that selection at *EDAR* most likely occurred more than 30,000 years BP in eastern China, and thus prior to the entry to the Americas (Kamberov et al., 2013) (Figure 3.2). In addition, the authors suggested that *EDAR* was possibly selected due to its effect for modulation of thermoregulatory sweating. The recent study of Hlusko et al. (2018) however, provided an alternative hypothesis for the selection of *EDAR* in Native Americans. The authors hypothesised that selection at *EDAR* was likely due to its effect on ductal branching in the mammary gland, thereby amplifying the transfer of important nutrients (particularly vitamin D) from mothers to infants under an extreme low UV environment, such as Beringia. They also suggested that the joint selection at *FADS* was likely due to its role in modulating lipid profiles transmitted to milk from a vitamin D-rich diet high in omega-3 fatty acids.

While this provides evidence for a strong, shared signal among Native American populations, there is still a lack of studies on local adaptive events in different Native American populations. Perhaps the exception are Andean highlanders, who have been the focus of extensive research to understand the biological basis for high altitude adaptation (Beall et al., 1997; Bigham et al., 2009, 2010; Zhou et al., 2013a; Bigham et al., 2014; Eichstaedt et al., 2014; Foll et al., 2014; Eichstaedt et al., 2015a; Valverde et al., 2015; Fehren-Schmitz and Georges, 2016; Bigham, 2016; Crawford et al., 2017a). The American continent shows extensive variation in climatic environments as its territory extends along a North-South axis that must have imposed strong subsistence and environmental constraints on Native Americans. Other studies in Native American populations include the adaptation to arsenic-rich environments in Andeans (Schlebusch et al., 2015; Eichstaedt et al., 2015b; Apata et al., 2017), adaptations to lipid metabolism and body development in Amazonians

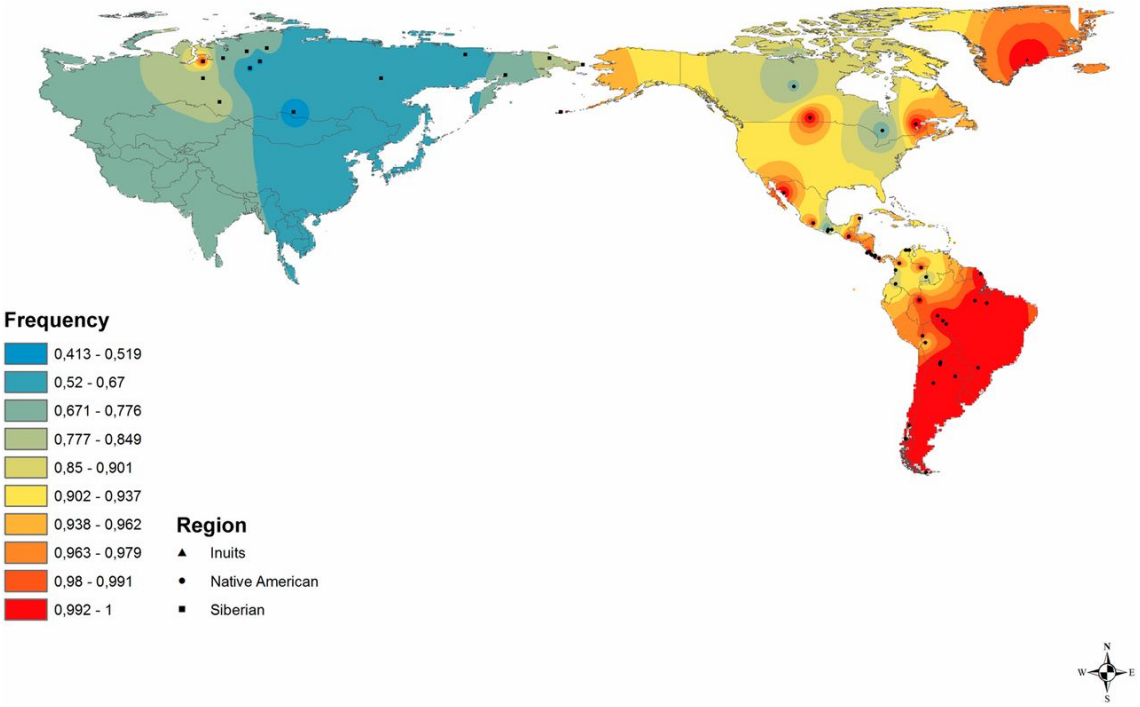

Figure 3.1: The geographic distribution of the putatively selected FADS haplotype in Native American populations. The signal of natural selection at the fatty acid desaturases (FADS) genes is not only present in Arctic populations, as was previously suggested, but throughout the American continent, suggesting a shared and strong adaptive event that occurred in Beringia. From Amorim et al. (2017).

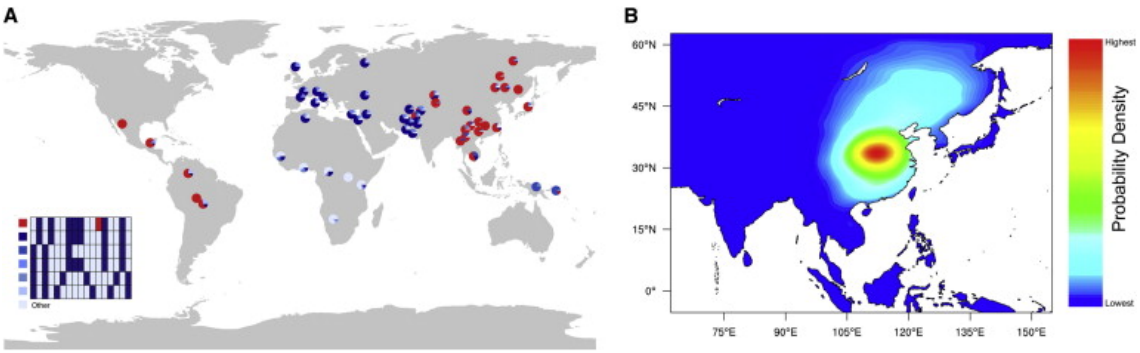

Figure 3.2: The geographic origin of the selected EDAR haplotype. A) Worldwide geographic distribution of the EDAR haplotype. The inset figure shows the six most common haplotypes, including the haplotype carrying the derived variant of the putatively selected allele (in red). B) The Posterior Probability density for the geographic origin of the putatively selected allele at the EDAR gene obtained by Approximate Bayesian Computation simulations. From Kamberov et al. (2013).

(Amorim et al., 2015), and immune related adaptations in distinct Native American populations attributed to diseases brought after the European contact (Fehren-Schmitz and Georges, 2016; Lindo et al., 2016; Vågene et al., 2018). These studies show that Native Americans have adapted to their distinct environments throughout the Americas, despite its recent and rapid settlement (Tamm et al., 2007; Reich et al., 2012). A complication in studying Native American adaptive history is the extensive admixture with Europeans and Africans (Wang et al., 2007; Reich et al., 2012; Moreno-Estrada et al., 2013; Chacon-Duque et al., 2018). This admixture can be a challenge as non-Native American ancestry can affect the statistics used to detect adaptive signals. To address this issue, two approaches have been employed. The first approach, which is the more common one, involves restricting the analysis to individuals carrying only Native American ancestry. Nonetheless, this can drastically reduce the sample size and consequently reduce the power to detect positive selection. Additionally, this approach would invalidate investigating Native American populations where all individuals show non-negligible amount of non-Native American ancestry, which is common for many Native American populations. The second, less common approach, involves accounting for non-Native American in admixed Native Americans by producing estimates of pseudounadmixed allele frequencies using the admixture ancestry proportions and allele frequencies from the populations contributing to the admixed Native American populations. This approach has been employed to successfully detect signals of positive selection in recently admixed populations, such African Americans (Bhatia et al., 2011), Ethiopian highlanders (Huerta-Sánchez et al., 2013) and more recently, in Andeans highlanders (Crawford et al., 2017a). Crawford et al. (2017a) conducted a genome-wide scan of selection using a modified version of the Population Branch Statistic (PBS; see Section 2.2.1 for a detailed description of the method), using a closely related population (lowland Native Americans), and an outgroup (Europeans). To account for the non-Native American ancestry, the authors estimated admixture proportions and admixture-corrected allele frequencies using NGadmix (Skotte et al., 2013), a software program that implements a model very similar to that of ADMIXTURE (Alexander et al., 2009). Notably, the authors found strong evidence of selection at genes related to cardiovascular development, and further demonstrated that the putatively selected haplotypes was associated with phenotypic variations related to cardiovascular health, thus highlighting the utility of this approach.

Recently, Chacon-Duque et al. (2018) examined the ancestry of more than 6,500 Latin Americans from Brazil, Chile, Colombia, Mexico and Peru (denoted the CANDELA sample, Section 1.8). Based on a novel haplotype-based analysis, and using data from several modern samples that contributed to the ancestry of present-day Latin Americans, the authors were able to infer distinct Native American ancestry components present in the CANDELA sample (Figure 3.3). The availability of these distinct Native American ancestry components in the CANDELA sample therefore offers the opportunity to explore instances of natural selection in different Native American groups using admixed Latin Americans.

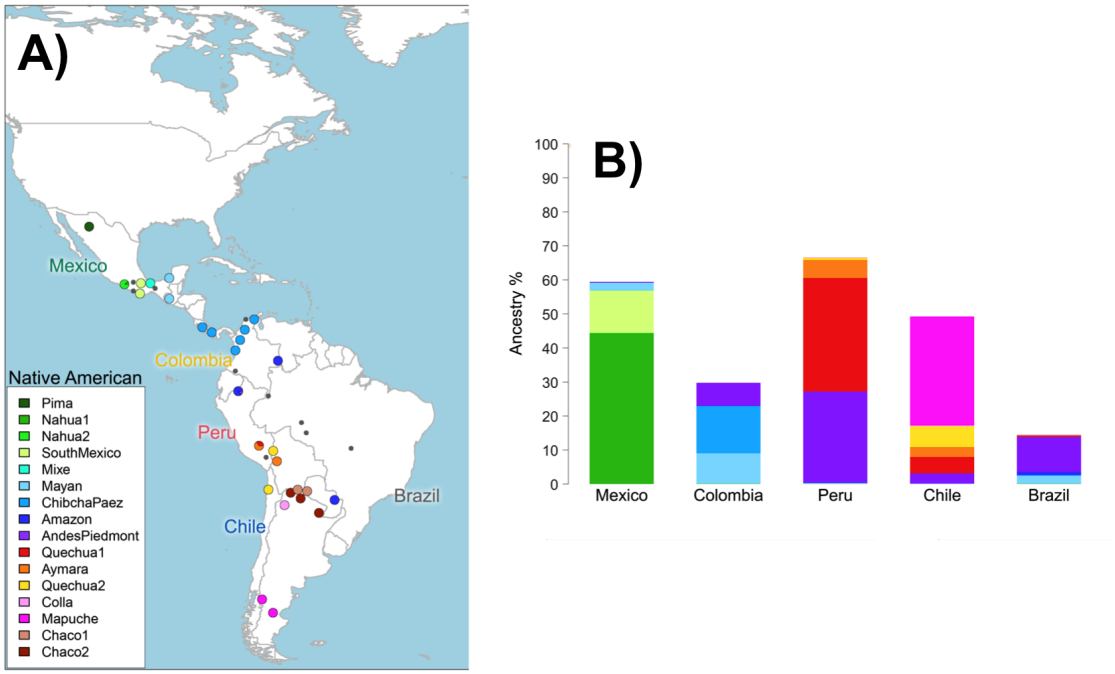

Figure 3.3: Native American reference population samples and ancestry estimates for the CANDELA sample. A) Colored pies and grey dots indicate the approximate geographic location of the 38 Native American reference populations. B) The estimated proportion of Native American components in the CANDELA sample. Adapted from Chacon-Duque et al. (2018).

3.3 Materials and methods

3.3.1 Description of Data

The individual samples analyzed here were part of several publicly available datasets. I started by using admixed Latin Americans sampled from the CANDELA Consortium (Ruiz-Linares et al., 2014 and Section 1.8). This data includes a total of 6,630 volunteers sampled in five Latin American countries: Brazil, Chile, Colombia, Mexico and Peru. I then combined this data with 231 Native American individuals from Chacon-Duque et al. (2018), 148 CLM (Colombians, from Medellin, Colombia), 107 MXL (Mexicans from Los Angeles, United States of America) and 130 PEL (Peruvians, from Lima, Peru) individuals from the 1000 Genomes Project (1KG) (1000 Genomes Project Consortium et al., 2015), 50 Native American individuals from western Argentinian from (Eichstaedt et al., 2014) and 28 Native American individuals from the Simons Genome Diversity Project (SGDP) (Mallick et al., 2016).

3.3.2 Quality control

I used PLINK v1.9 (Chang et al., 2015) to perform quality control (QC) analyses. I excluded SNPs and individuals with more than 1% missing data and retained only autosomal SNPs. After performing LD pruning the PLINK inferred IBD coefficient was calculated across all pairs of individuals within each population. Individuals with a IBD higher than

0.125 (i.e. third degree relatives) were removed. For the Native Americans population, I used the methodology described in Chacon-Duque et al. (2018), where individuals with more than 10% from the median IBD value were discarded. This is due to the lower effective population size present in this population that can affect IBD estimates based on population allele frequencies (Manichaikul et al., 2010). After applying these filters 679,855 autosomal SNPs and 6,589 individuals were retained for further analysis.

3.3.3 Selecting individuals without post-Columbian admixture

Previous studies have identified substantial amounts of European and African ancestry in modern Latin Americans (Wang et al., 2007; Reich et al., 2012; Moreno-Estrada et al., 2013; Chacon-Duque et al., 2018). This admixture is extensive across the American continent and involves not only Natives (i.e. the indigenous habitants), but also the general population, or what is now usually referred to as Latin Americans (Ruiz-Linares, 2014). To restrict the selection analysis on individuals without evidence of European or African admixture, I excluded all individuals with $> 1\%$ cluster membership of the European and African ancestry component based on an unsupervised ADMIXTURE (Alexander et al., 2009) using $K = 3$ components. This resulted in a total of 168 individuals (hereinafter referred to as Native Americans) without European or African ancestry that were used for the selection analysis described below (Section 3.3.4).

3.3.4 Selection scans in Native American individuals without post-Columbian admixture

To explore signals of selection in Native Americans I computed the Population Branch Statistic (PBS), which represents the amount of allele frequency change at a SNP of a target population compared to two other reference populations (Yi et al., 2010 and Section 2.2.1). For this analysis I used East Asians (CHB; Han Chinese in Beijing, China) and Northern Europeans (CEU; Utah Residents with Northern and Western European Ancestry) from the 1KG as reference populations. Pairwise F_{ST} were estimated using using Hudson's estimator as in equation 9 of Bhatia et al. (2013). The branch length (T) between two populations was then computed as $T = -\log_{10}(1 - F_{ST})$ (Cavalli-Sforza, 1969). The PBS combines the pairwise branch lengths between these three populations, which was computed as:

$$PBS_{NAM} = \frac{T^{NAM,CHB} + T^{NAM,CEU} - T^{CHB,CEU}}{2} \tag{3.1}$$

where NAM is the target population and CHB and CEU the two reference populations. To remove signals that could be driven by a single SNP that might be due to genotyping errors I excluded SNPs that are above the 99.99^{th} percentile of PBS scores that do not have any other significant neighbour SNP within $\pm$100Kb (i.e. 200Kb window).

I also complemented the PBS analysis by computing three additional haplotype-based statistics: the Integrated Haplotype Score (iHS) (Voight et al., 2006), Cross Population Extended Haplotype Homozygosity (XP-EHH) (Sabeti et al., 2007) and number of Segregating sites by Length (nSL) (Ferrer-Admetlla et al., 2014) to capture patterns of haplotype homozygosity based on a model of a hard selective sweep. iHS and XP-EHH have power to detect selective sweeps that have reached moderate and high frequency respectively, while nSL retains some power to detect soft selection sweeps, thus making these three statistics complementary. As these selection statistics are haplotype-based tests that employ adjacent SNPs to compute a per-SNP selection score, I used lenient QC threshold by removing SNPs with $> 5\%$ missing data. This is expected to increase the power of haplotype-based selection scans. iHS and nSL were estimated as in Voight et al. (2006) retaining all SNPs with a MAF $> 5\%$ and standardizing the scores by binning the SNPs by allele frequencies and subtracting the mean and dividing by the standard deviation to obtain a final normalized statistic with a mean of 0 and variance of 1. The frequency bins were defined by 1% frequency increments. Both of these statistics require the ancestral and the derived allele states to be specified for each SNP analyzed. To obtain this, I used the Human Ancestral Sequence FASTA file from the 1KG Project retaining both low and high confidence calls. This sequence file is based on a 6-way primate whole-genome alignment. XP-EHH statistic requires the definition of a reference population for which I use the CHB as reference population. The standardization of XP-EHH scores was conducted such that the set of all XP-EHH scores had a mean of 0 and variance of 1 as in Sabeti et al. (2007). Since XP-EHH scores are directional (Sabeti et al., 2007), I only retained positive scores, which would be indicative of positive selection in the Native American population. For the three haplotype-based selection statistics, the HapMap GRh37 genetic map (International HapMap Consortium, 2003) was used to estimate genetic distances between SNPs. Importantly, allele frequency differentiation and haplotype based approaches, exploit different genomic signatures that persist over varying times scales (Section 2.2.3), and thus I do not expect to find selection signals at a genomic locus across all different tests.

3.3.5 Identification of Latin American individuals with specific Native American ancestry components

To select admixed individuals carrying a specific Native American ancestry component, I used the inferred Native American ancestry proportions previously identified by Chacon-Duque et al. (2018) in the CANDELA sample (Figure 3.3). Combining closely related Native American ancestries, I selected admixed Latin American that derive most of their Native American ancestry to three distinct Native American populations. Specifically, I merged the "Quechua", "Colla" and "Aymara" Native American ancestries into a "Andean" component, the "Nahua", "South Mexico" and "Mixe" Native American ancestries into a "Meso-American" component and the "Mapuche" ancestry into one component. I then selected CANDELA individuals with $> 10\%$ inferred ancestry from a particular Native American ancestry component (i.e. Meso-American, Andean or Mapuche), with $< 1\%$ combined inferred ancestry from all other Native American groups, and additionally

$< 1\%$ inferred East Asian and North African ancestry. Thus, each group of admixed individuals were composed exclusively of Native American from a particular Native American group, European and African ancestry. The final number of individual per Native American population was $N = 689$ for Meso-Americans, $N = 375$ for Andeans, and $N = 427$ for Mapuche.

3.3.5.1 Estimating Native American allele frequencies in admixed Latin American samples

To estimate allele frequencies for each of the Native American groups (described in Section 3.3.5) I corrected observed allele frequencies by the amount of non-Native American ancestry. Similar approaches have been employed to infer the population or adaptive history of admixed populations by obtaining admixture-corrected allele frequencies based on admixture proportions estimated from the whole genome (Bhatia et al., 2011; Huerta-Sánchez et al., 2013; Moltke et al., 2015). Following these previous studies, I computed the admixture-corrected (or pseudounadmixed) Native American allele frequency f_{Nat} at each SNP as:

$$f_{Nat} = \frac{f_{Adm} - f_{Eur}\alpha_{Eur} - f_{Afr}\alpha_{Afr}}{1 - \alpha_{Eur} - \alpha_{Afr}} \tag{3.2}$$

where f_{Adm} is the admixed allele frequency, f_{Eur} is the European allele frequency, f_{Afr} is the African allele frequency; α_{Eur} is the European ancestry proportion and α_{Afr} is the African ancestry proportion. The allele frequencies at each SNP for f_{Eur} and f_{Afr} were estimated using the IBS and YRI populations from the 1KG. The ancestry proportion for α_{Eur} and α_{Afr} were taken from the genome-wide average estimates for the admixed individuals in Chacon-Duque et al. (2018).

3.3.5.2 Genome-wide scan of selection using admixture-corrected allele frequencies

To detect SNPs under positive selection I calculated the PBS as in equation 3.1 using the admixture-corrected allele frequencies. For each Native American group I used the two other Native American populations as reference populations in order to detect instances of local adaptation in these three Native American groups. PBS scores were estimated only at SNPs that were polymorphic in at least two populations. To remove signals driven by a single SNP that might be due to genotyping errors I removed SNPs with PBS scores above the 99.99^{th} percentile that do not have any other significant neighbour SNP within $\pm$100Kb (i.e. 200Kb window).

3.3.6 Gene set enrichment analysis using biological pathways

To test for signals of polygenic adaptation, I assessed if particular gene sets were enriched with high PBS scores. For each gene I assigned the highest PBS score in a $\pm$2kb region

surrounding each gene based on the UCSC RefSeq annotation, retaining genes with more than 2 SNPs. I also estimated the mean PBS score for each gene, and found that there was a positive correlation between the highest and the mean PBS score, indicating that the highest scoring SNP is a good representative of the selection pattern in a gene (Figure A.1). As the assigned PBS scores in a gene region can be correlated with SNP density (i.e. the number of SNPs in a gene region), I corrected for this bias by binning gene regions with a similar amount of SNPs: $<$ 10, 10 — 19, 20 — 29, 30 — 39, 40 — 49, 50 — 59, 60 — 69, 70 — 79, 80 — 89, 90 — 99 and $\geq$100, and standardizing the PBS scores within each bin. I then downloaded the curated human gene sets from the NCBI Biosystems database (https://www.ncbi.nlm.nih.gov/biosystems/; as of November 2017) and discarded all genes from the gene sets that could not be mapped to the genes based on the RefSeq annotation. I then further excluded any gene set that contained less than 10 genes, and this resulted in $\sim 1,900$ gene sets depending on the Native American population tested. For each gene set I obtained the distribution of maximum PBS scores for genes included in the gene set and the distribution of maximum PBS scores of genes in the rest of the genome. To assess significance, I compared these two distributions using a one-sided Mann-Whitney U test and reported Bonferroni adjusted P-values.

3.3.7 Gene Ontology (GO) enrichment analysis

To test for GO categories containing genes that were enriched with high PBS scores, I performed an enrichment test on the PBS-ranked gene regions, using the minimum hypergeometric (mHG) score method for ranked lists implemented in the web-based application (GOrilla) (Eden et al., 2009). I reported significant GO categories (P-values$<$0.001), with low false discovery rate q-values (q$<$0.1) and high enrichment scores (S$>$5) as in Fumagalli et al. (2015). The enrichment score is equal to $(b/n)/(B/N)$, where N is the total number of genes, B is the number of genes associated with a GO category, n is the number of genes at the top of the PBS-ranked list (as defined by the mHG method) and b is the number of genes in the intersection of n and B.

3.3.8 Phenotypic association analysis

I used the CANDELA Consortium (Ruiz-Linares et al., 2014; Section 1.8) phenotypic data to perform association analysis. All volunteers underwent anthropometric measurements including: height (cm), weight (kg), Body Mass Index (BMI) (kg/m^2), hip-circumference (HC) (cm), waist-circumference (WC) (cm), waist-to-hip ratio (WHR), HC adjusted for BMI (HCadjBMI), WC adjusted for BMI (WCadjBMI) and WHR adjusted for BMI (WHRadjBMI). Obese individuals defined as having BMI $>$ 35 and individuals older than 45 years were excluded. All anthropometric phenotypes appeared normally distributed and therefore no transformation was applied (Figure A.2). Genetic association analyses were performed using a multivariate linear regression approach with an additive genetic model incorporating age, sex and the first 6 genetic PCs as covariates as conducted in previous GWAS in the CANDELA sample (Adhikari et al., 2015, 2016a,b). Additionally, 1,172 volunteers recruited in Mexico were analyzed for 6 biochemical measurements. The measurements were performed with commercially available standardized methods

in blood samples obtained after a 12 hour fast as described in Villarreal-Molina et al. (2007). The biochemical measurements included: fasting glucose (mg/dl), total cholesterol (mg/dl), triglycerides (mg/dl), high-density lipoprotein (HDL) cholesterol (mg/dl), low-density lipoprotein (LDL) cholesterol (mg/dl) and non-HDL cholesterol (mg/dl). All biochemical measurements were approximately normally distributed (Figure A.3). Genetic association analyses were performed using a multivariate linear regression approach with an additive genetic model incorporating age, sex, BMI and the first 2 genetic PCs as covariates. Only the first two PCs were used for the genetic association of metabolic phenotypes as they captured most of the population structure caused by admixture in the Mexican sample (Figure A.5).

3.3.9 Worldwide allele frequencies

To explore the geographic distribution of the variants showing signals of selection I compiled a dataset from several human populations. This dataset included 10 populations from Africa (Schlebusch et al., 2012; 1000 Genomes Project Consortium et al., 2015), 18 populations from Europe (1000 Genomes Project Consortium et al., 2015; Chacon-Duque et al., 2018), 4 populations from North Africa and the Middle East (Chacon-Duque et al., 2018), 20 populations from East, South and South East Asia (1000 Genomes Project Consortium et al., 2015; Mallick et al., 2016; Mörseburg et al., 2016), 4 populations from Siberia (Cardona et al., 2014) and 7 populations from the Americas (Eichstaedt et al., 2014; 1000 Genomes Project Consortium et al., 2015; Chacon-Duque et al., 2018). This dataset is described in detail in Table A.1. I detail the sources of the data, number of individuals per population and geographic coordinates of the sampling locations.

3.4 Results

3.4.1 Overview

A limitation in studying Native American adaptive history is the high levels of admixture with European and African populations (Wang et al., 2007; Reich et al., 2012; Moreno-Estrada et al., 2013; Chacon-Duque et al., 2018). To address this complication, I used two different approaches in this chapter. The first approach involved restricting the analysis to individuals across the Americas without evidence of post-Columbian admixture. This allowed me to detect candidate regions of selection that occurred after the split of the ancestral Native Americans from Asians. In addition, this also allowed me to use haplotype-based approaches, as these can only be applied to non-recently admixed populations. Further, by using a larger set of individuals, I am also expected to obtain a higher statistical power to detect selective events. Second, in order to explore possible instances of local adaptation in the Americas, I used a large sample of admixed Latin Americans that derive most of their Native American ancestry to three particular Native American groups and estimated pseudounadmixed allele frequencies by accounting for non-Native American ancestry. For convenience, throughout this section I refer to these three Native American

components simply as Andean, Meso-American and Mapuche Native Americans.

3.4.2 Selection signals in Native Americans

To study selection signals in Native Americans shared across the American continent I collated genome-wide data from 168 Native Americans individuals present in several public databases. I used the Population Branch Statistic (PBS) statistic, which identifies alleles that have experienced strong changes in allele frequency in one target population (Native Americans) relative to two reference populations (East Asians [CHB] and Europeans [CEU] from the 1KG Project). The PBS analysis revealed 11 candidate regions of selection (Figure 3.4 and Table 3.1), defined as being composed of at least two adjacent SNPs with PBS scores above the 99.99th percentile of the empirical distribution. Throughout the text I refer to these SNPs as *top selected SNPs*. Within the 11 candidate regions of selection, two contained intergenic non-protein coding RNAs (*LINC00641* and *LINC00871*) without any known function and five contained no genes. To complement the PBS selection analysis, I also computed three haplotype-based selection statistics (see Section 3.3.4) at each candidate genomic region (i.e. those with SNPs with PBS scores above the 99.99th percentile of the empirical distribution). Given that these four selection statistics are moderately to strongly correlated (rho ranging from 0.12 to 0.84) (Figure A.6), as these have different power to detect different type of selection signals, combining different selection statistics can provide further evidence for selection at these candidate regions.

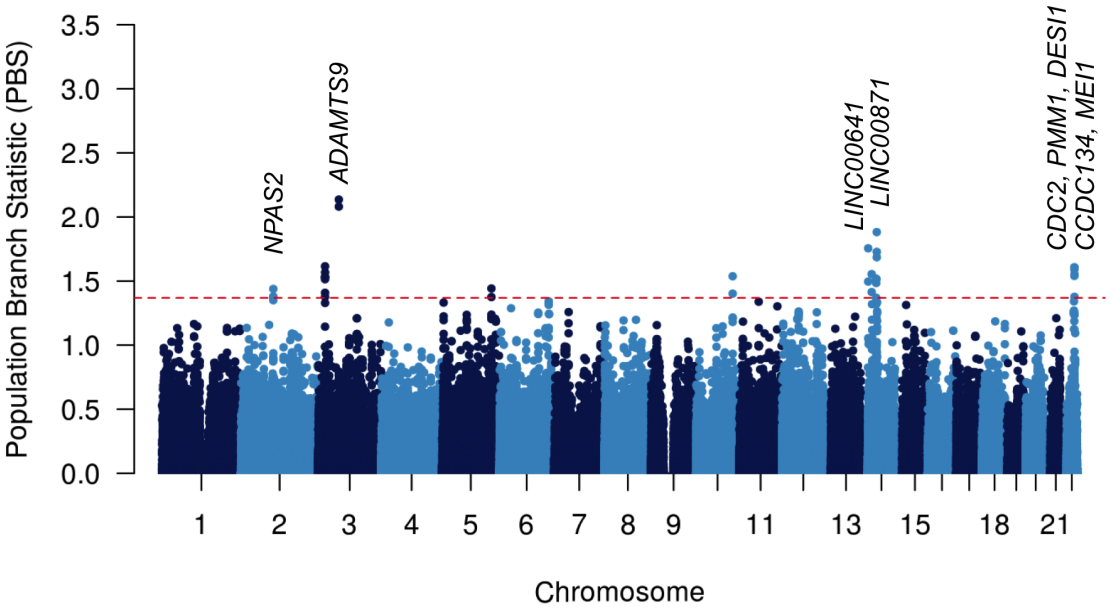

Figure 3.4: Genome-wide scan for selection in Native Americans. Population Branch Statistic (PBS) selection scores per-SNP were computed for 168 Native Americans using CHB (East Asians) and CEU (North Western Europeans) from the 1000 Genomes Project as reference populations. The red dashed line represents the 99.99th percentile. Regions with fewer than two SNPs within 200Kb above the 99.99th percentile were excluded. Names of genes associated with the highest peaks are shown.

Table 3.1: Candidate regions under selection in 168 Native Americans based on the Population Branch Statistic (PBS) selection statistic.

Genomic coordinates[a]	Length (bp)	PBS value (highest SNPs[b])	Candidate targeted genes
Chr3:64514393-64525420	11,027	2.14 (rs7631391)	*ADAMTS9*
Chr14:48581772-48633622	51,850	1.88 (rs8021638)	-
Chr14:21671316-21674214	2,898	1.76 (rs1243370)	*LINC00641*
Chr3:21260370-21354506	94,136	1.62 (rs7628403)	-
Chr22:41961831-41995335	33,504	1.61 (rs5996039, rs8139993)	*CSDC2,PMM1,DESI1*
Chr14:32635572-32688214	52,642	1.55 (rs7151991)	-
Chr22:42187199-42218856	31,657	1.55 (rs139553)	*MEI1,CCDC134*
Chr10:115180454-115239602	59,148	1.54 (rs12414053)	-
Chr14:46933384-46963280	29,896	1.52 (rs17740937, rs754960, rs2642103)	*LINC00871*
Chr2:101540415-101549843	9,428	1.44 (rs356652)	*NPAS2*
Chr5:155331080-155361116	30,036	1.44 (rs1432734)	-

[a] Candidate regions are defined as SNPs with PBS values above the 99.99^{th} percentile of the empirical distribution. SNPs within 200Kb are merged into a single entry.

[b] rs ID for the SNP showing the highest PBS value is reported in parenthesis.

The strongest signal is located at 3p14 within the ADAM Metallopeptidase With Thrombospondin Type 1 Motif 9 (*ADAMTS9*) gene (Figure 3.5). At this genomic region, none of the haplotype-based selection statistics showed strong signals of selection. Members of the ADAMTS family are implicated in the cleavage of proteoglycans, the control of organ shape during development and the inhibition of angiogenesis (Porter et al., 2005; Brocker et al., 2009). Notably, this gene has been associated with HDL cholesterol and fasting insulin levels (Liu et al., 2013a), as well as several anthropometric phenotypes including body fat distribution, body mass index (BMI), hip circumference, waist-to-hip ratio (WHR), and also with type 2 diabetes (T2D) (Zeggini et al., 2008; Heid et al., 2010; Fox et al., 2012; Morris et al., 2012; Liu et al., 2013a; Locke et al., 2015; Shungin et al., 2015; Graff et al., 2017; Justice et al., 2017; Ng et al., 2017; Zhao et al., 2017; Bonàs-Guarch et al., 2018; van der Harst and Verweij, 2018). From the GTEx database, the highest expression of *ADAMTS9* is reported to be in visceral adipose tissue (GTEx Consortium, 2013). Two other candidate regions with strong signals of selection are located at 22q13. The strongest of these encompasses three genes: the Cold Shock Domain Containing C2 (*CSDC2*), the Phosphomannomutase 1 (*PMM1*), and the Desumoylating Isopeptidase 1 (*DESI1*) genes (Figure 3.6). At this genomic region, none of the haplotype-based selection statistics showed strong signals of selection. Of potential adaptive interest is the *PMM1* gene that is a phosphomannomutase enzyme that catalyzes the conversion between D-mannose 6-phosphate and D-mannose 1-phosphate, which is a substrate for GDP-mannose synthesis (Schollen et al., 1998; Heykants et al., 2001). Diseases associated with phosphomannomutase enzymes include Congenital Disorder Of Glycosylation Type Ia (OMIM: 212065). Recently, it has been shown that *PMM1* is significantly upregulated in the liver of obese subjects compared with that of lean subjects (Lee et al., 2016). The second candidate region within chromosome 22 encompasses two genes: the Meiotic Double-Stranded Break Formation Protein 1 (*MEI1*) gene and the Coiled-Coil Domain Containing 134 (*CCDC134*) gene (Figure 3.7). At this genomic region, none of the haplotype-based selection statistics showed strong signals of selection. Of potentially adaptive interest, is the *CCDC134* that is implicated in immune function. The coded protein promotes proliferation and activation of CD8(+) T cells, suggesting a cytokine-like function and shows strong anti-tumor effects (Huang et al., 2014). The selection signal at 2q11 encompassess the Neuronal PAS Domain Protein 2 (*NPAS2*) gene. Notably, the XP-EHH selection statistics, showed also several SNPs with strong signals of selection at this genomic region (Figure 3.8). This gene encodes a transcription factor which is thought to be the major transcriptional regulators of the circadian clock mechanism in mammals (DeBruyne et al., 2007). The circadian clock, is an internal time-keeping system that regulates distinct physiological processes through the generation of $\sim$ 24 hour circadian rhythms in gene expression, which are translated into rhythms in metabolism and behavior (DeBruyne et al., 2007). Notably, epidemiological studies have linked *NPAS2* with a variety of psychological disorders, including winter depression (Partonen et al., 2007).

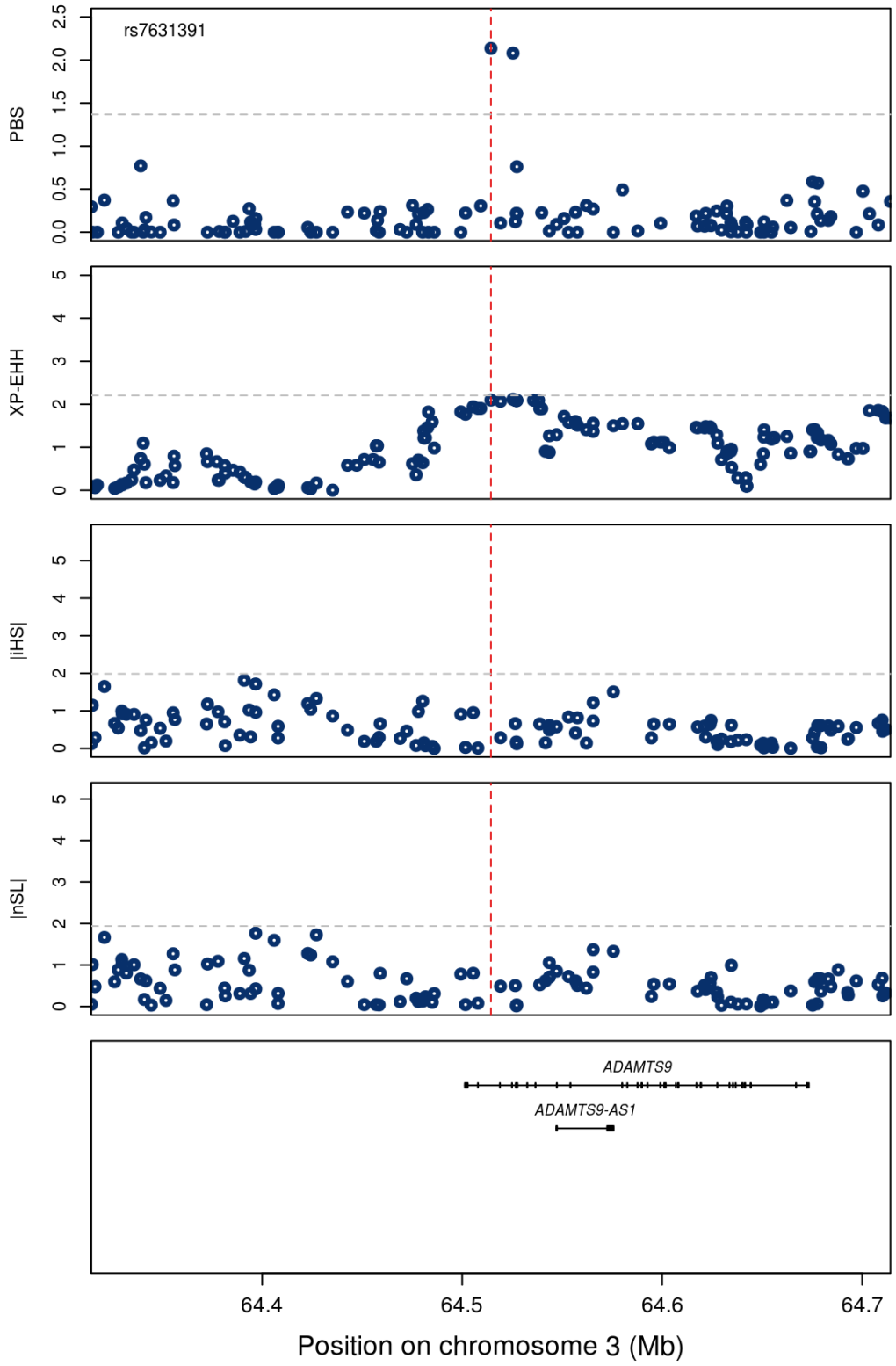

Figure 3.5: Selection for four selection tests in candidate region surrounding top selected SNP rs7631391 in Native Americans. The first four panel show the score for four individual selection statistics score: PBS, XP-EHH, iHS and nSL. The bottom panel shows the UCSC RefSeq genes within the genomic region (in hg19 coordinates). The red dotted line indicates the position of the top selected SNP. The grey dotted line in the PBS panel denotes the 99.99^{th} percentile of the empirical distribution and the grey dotted lines in the other panels denotes the 95^{th} percentile of the empirical distribution for each selection statistic.

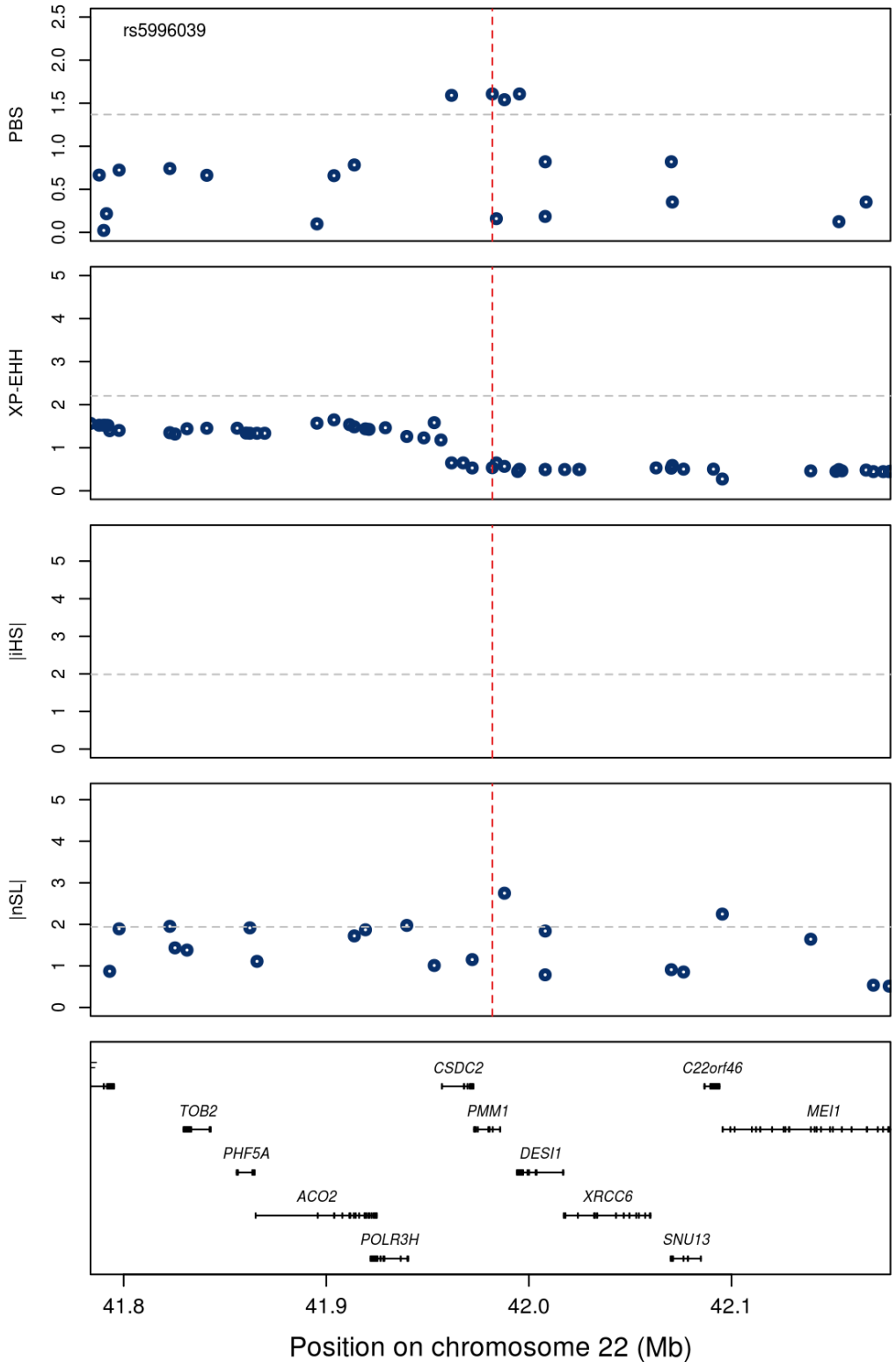

Figure 3.6: Selection for four selection tests in candidate region surrounding top selected SNP rs5996039 in Native Americans. The first four panel show the score for four individual selection statistics score: PBS, XP-EHH, iHS and nSL. The bottom panel shows the UCSC RefSeq genes within the genomic region (in hg19 coordinates). The red dotted line indicates the position of the top selected SNP. The grey dotted line in the PBS panel denotes the 99.99^{th} percentile of the empirical distribution and the grey dotted lines in the other panels denotes the 95^{th} percentile of the empirical distribution for each selection statistic.

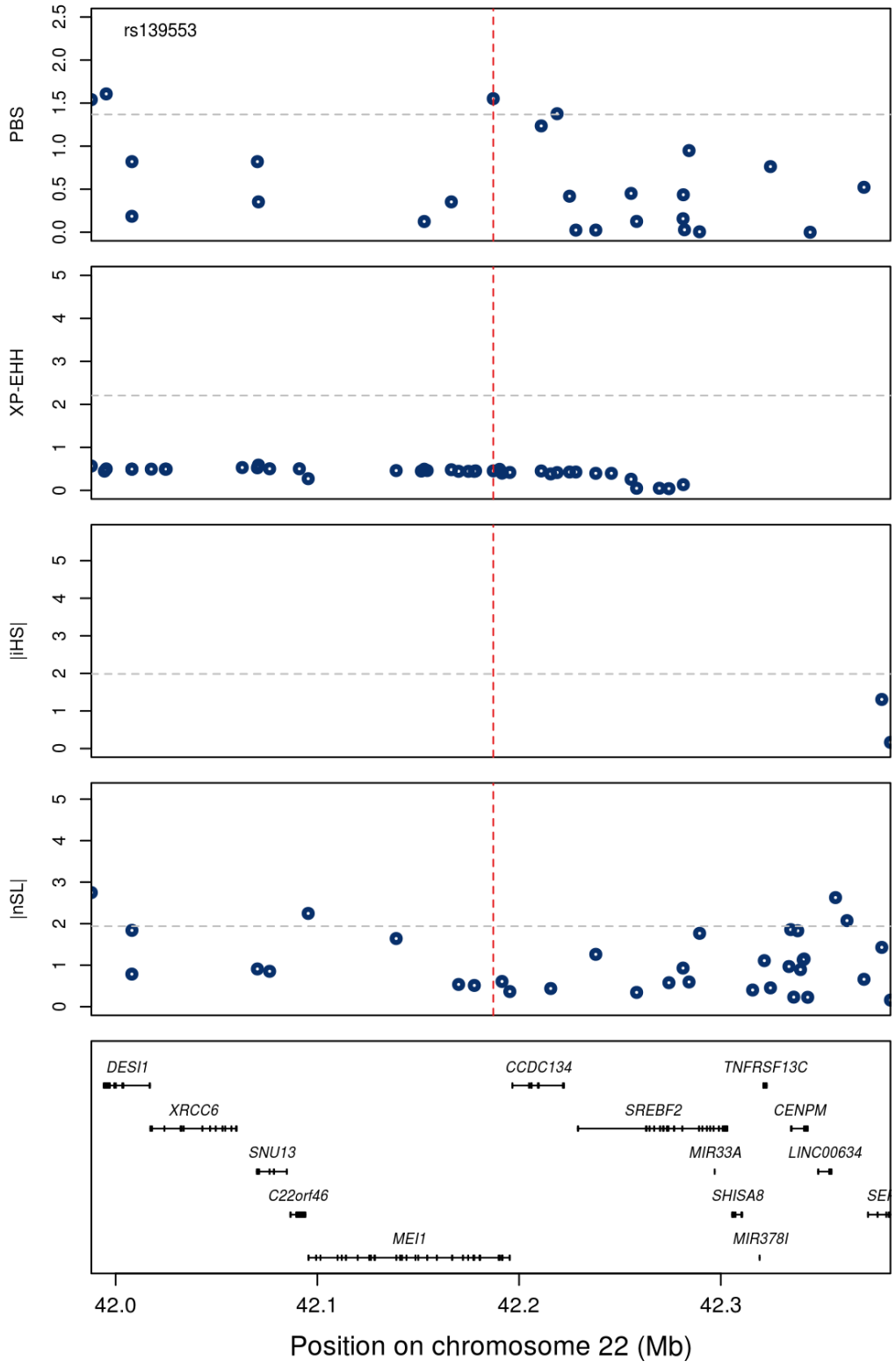

Figure 3.7: Selection for four selection tests in candidate region surrounding top selected SNP rs139553 in Native Americans. The first four panel show the score for four individual selection statistics score: PBS, XP-EHH, iHS and nSL. The bottom panel shows the UCSC RefSeq genes within the genomic region (in hg19 coordinates). The red dotted line indicates the position of the top selected SNP. The grey dotted line in the PBS panel denotes the 99.99^{th} percentile of the empirical distribution and the grey dotted lines in the other panels denotes the 95^{th} percentile of the empirical distribution for each selection statistic.

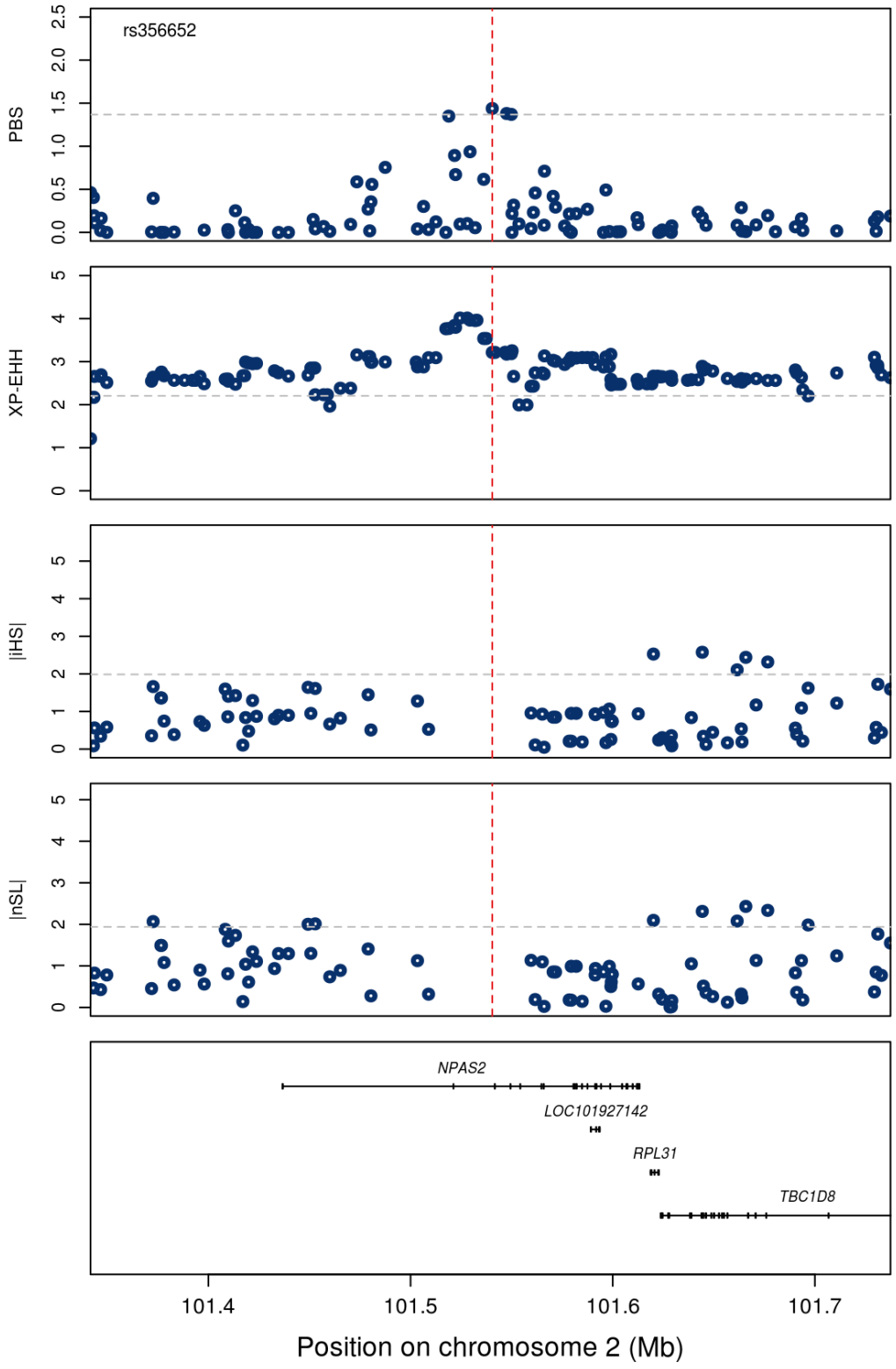

Figure 3.8: Selection for four selection tests in candidate region surrounding top selected SNP rs356652 in Native Americans. The first four panel show the score for four individual selection statistics score: PBS, XP-EHH, iHS and nSL. The bottom panel shows the UCSC RefSeq genes within the genomic region (in hg19 coordinates). The red dotted line indicates the position of the top selected SNP. The grey dotted line in the PBS panel denotes the 99.99^{th} percentile of the empirical distribution and the grey dotted lines in the other panels denotes the 95^{th} percentile of the empirical distribution for each selection statistic.

3.4.3 Association testing with metabolic and anthropometric phenotypes in top SNPs in Native Americans without post-Columbian admixture

Currently, the best-documented example of selection in Native Americans is through dietary adaptations (Acuña-Alonzo et al., 2010; Amorim et al., 2017; Hlusko et al., 2018). I therefore assessed whether SNPs within a 5Kb radius around *ADAMTS9* (a gene which has been previously associated to different anthropometric and metabolic phenotypes) were associated associated to metabolic and anthropometric phenotypes in the CANDELA sample. The association analysis did not reveal any significant association, after Bonferroni correction for testing for association considering the 72 SNPs within the tested genomic region (Figure 3.9 and 3.10). The top two candidate SNPs with the strongest PBS scores at *ADAMTS9* revealed only marginally significant associations with plasma glucose levels (rs76311391, P-value = 0.04; rs17070941,P-value = 0.03).

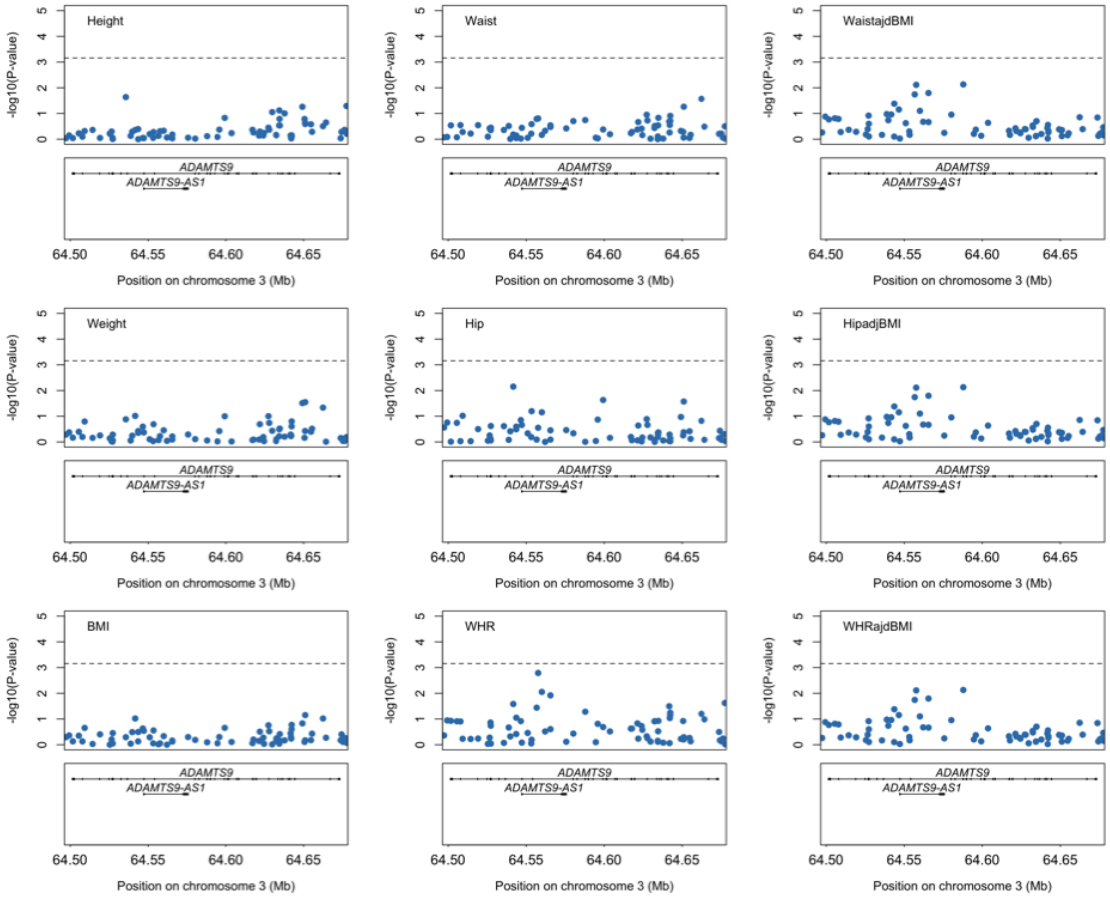

Figure 3.9: Regional Manhattan plot focused on *ADAMTS9* within a 5Kb radius around the gene for nine anthropometric phenotypes. Each plot is composed of two panels. The upper panel shows the -log10(P-values) of the association analysis. The bottom panel shows the shows the UCSC RefSeq genes within the genomic region (in hg19 coordinates). The black dotted line indicates the Bonferroni corrected threshold of 3.15, which was computed after considering 72 SNPs within the tested genomic region.

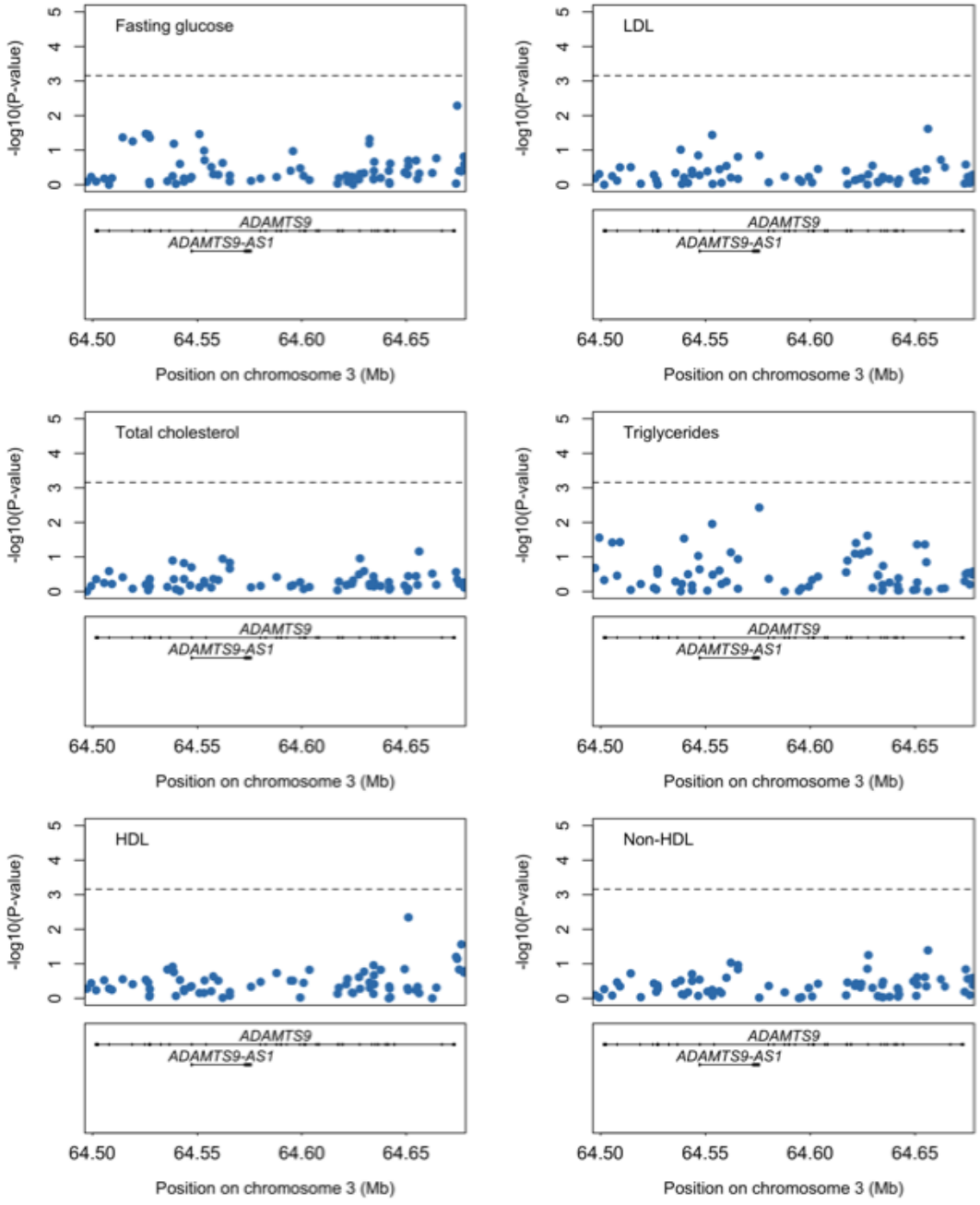

Figure 3.10: Regional Manhattan plot focused on *ADAMTS9* within a 5Kb radius around the gene for six metabolic phenotypes. Each plot is composed of two panels. The upper panel shows the -log10(P-values) of the association analysis. The bottom panel shows the shows the UCSC RefSeq genes within the genomic region (in hg19 coordinates). The black dotted line indicates the Bonferroni corrected threshold of 3.15, which was computed after considering 72 SNPs within the tested genomic region.

3.4.4 Gene set enrichment analysis using biological pathways and Gene Ontology (GO) categories in Native Americans

Enrichment of selection scores in biologically relevant pathways can be used to detect instances of polygenic adaptation. I used the gene sets of biological pathways from the NCBI Biosystems Database, and tested whether the PBS scores distribution in a set of genes from a particular biological pathway were significantly shifted toward larger PBS values than that of the remaining data. None of the biological pathways showed significance after adjusting for multiple testing (i.e. Bonferroni-adjusted P-values > 0.05). Additionally, I conducted an enrichment analysis of Gene Ontology (GO) categories using the GOrilla web-based application (Eden et al., 2009), and found three significant GO categories (Table A.2). However, the three GO categories are likely significant because they all contain several members of the protocadherin gamma gene cluster that are contiguous at 5q13.

3.4.5 Selection signals in three distinct Native American populations

3.4.5.1 Meso-American Native Americans

In the Meso-American Native American population, the PBS analysis revealed 13 candidate regions of selection (Figure 3.11 and Table 3.2), of which seven contained no genes. The strongest signal is located at 7q32 and encompasses the Interferon Regulatory Factor 5 (*IRF5*) and Deoxyribonuclease 1 Like 3 (*DNASE1L3*) genes. Of potential adaptive interest is the *IRF5* gene, which is part of a group of transcription factors with diverse roles, including virus-mediated activation of interferon and immune system activity (Yanai et al., 2007). Interestingly, this gene has been associated with autoimmune diseases, such as systemic lupus in different populations including Latin Americans (Bentham et al., 2015; Alarcón-Riquelme et al., 2016; Morris et al., 2016; Márquez et al., 2017). The candidate region located at 1p36 encompasses 8 genes: *PLEKHG5, NOL9, TAS1R1, KLHL21, ZBTB48,PHF13, THAP3* and *DNAJC11*. The SNP (rs4243829) with the highest PBS score within this region is an intronic variant of the Pleckstrin Homology And RhoGEF Domain Containing G5 (*PLEKHG5*) gene. This gene encodes a protein that activates the nuclear factor kappa B (NFKB1) signaling pathway (Matsuda et al., 2003). Mutation in this gene have been associated with autosomal recessive distal spinal muscular atrophy-4 (OMIM; 611067) (Maystadt et al., 2007) and with intermediate Charcot-Marie-Tooth disease C (OMIM; 615376), an autosomal recessive peripheral neuropathy resulting in walking difficulties due to muscle weakness and atrophy (Azzedine et al., 2013). Three candidate regions of selection are located within 12q12 encompassing a total of 5 genes: *CUX2*, *BRAP*, *ACAD10*, *ALDH2*, and *RPH3A*. Of potential adaptive interest is the Aldehyde Dehydrogenase 2 Family (Mitochondrial) (*ALDH2*) gene that encodes the second enzyme of the major oxidative pathway of alcohol metabolism. Variants within *ALDH2* have been shown to be strongly associated to alcohol consumption (Luczak et al., 2006; Eng et al., 2007; Baik et al., 2011; Takeuchi et al., 2011; Quillen et al., 2014; Wall et al., 2016; Gelernter et al., 2018) and esophageal cancer (Cui et al., 2009) in Asian populations. Notably, this gene shows also strong signals of selection in Asian populations (Okada et al., 2018).

The candidate region at 17q21 encompasses the Carbonic Anhydrase 10 (*CA10*) gene. This gene encodes a protein that belongs to the carbonic anhydrase family of zinc metalloenzymes, which catalyze the reversible hydration of carbon dioxide in various biological processes (Mori et al., 2009). Interestingly, variants within this gene have been associated to age at first menstrual bleeding in different GWAS conducted in European populations (Elks et al., 2010; Perry et al., 2014; Pickrell et al., 2016).

3.4.5.2 Andean Native Americans

In the Andean Native American group the PBS analysis revealed 9 candidate regions of selection (Figure 3.11 and Table 3.3), of which three contained no genes. The strongest selection signal is located at 10q24 spanning ∼700Kb and encompassing several genes. One of the genes within this region is the Arsenite Methyltransferase *AS3MT* gene, which encodes an enzyme that is essential for the efficient metabolization of arsenic in humans (Sumi and Himeno, 2012). Notably, this gene has been shown to under selection in Andean populations, probably as an adaptation to environments with high levels of arsenic that can be found in the Andean altiplano (Eichstaedt et al., 2015b; Schlebusch et al., 2015; Apata et al., 2017). The second strongest signal was located in the 10q11 genomic region and encompasses four genes: *STOX1*, *DDX21*, *DDX50* and *KIF1BP*. Of potential adaptive interest is the Storkhead Box 1 (*STOX1*) gene, where mutations within this gene have been associated to preeclampsia (van Dijk et al., 2005, 2010), a pathology of pregnancy characterized by high blood pressure and signs of damage to another organ system, that can be lethal for the mother and for the fetus (Sibai, 2005). The third strongest signal, also located at 10q24, encompasses the Nucleolar And Coiled-Body Phosphoprotein 1 (*NOLC1*). This gene encondes a nuclear localization signal binding protein (Meier and Blobel, 1990, 1992). The candidate region at 6p22 is located within the MHC region, which possesses important function for the immune system (Hill, 1998, 2001; Shiina et al., 2009; Mosaad, 2015). The strongest SNP was located within the *TRIM31* gene. Members of the TRIM superfamily are expressed in response to interferons and are also involved in a broad range of biological processes that are associated with innate immunity (Rajsbaum et al., 2008; Reymond et al., 2001). The candidate region at 11p15 encompasses the Signal Peptide, CUB Domain And EGF Like Domain Containing 2 (*SCUBE2*). This gene codes for a secreted cell-surface glycoprotein and is predominantly expressed in vascular endothelial cells (Tsai et al., 2009). *SCUBE2* has also been reported to act as a tumor suppresor in human breast cancer (Cheng et al., 2009; Lin et al., 2011b; Song et al., 2015). The candidate region at 15q21 encompasses two genes: *SORD* and *DUOX2*. Of adaptive interest is the Sorbitol Dehydrogenase (*SORD*) gene, which encodes an enzyme with an important function in the sorbitol pathway (part of carbohydrate metabolism) and has been suggested to play a role in the development of diabetic complications (Carr and Markham, 1995).

3.4.5.3 Mapuche Native Americans

In the Mapuche Native American group, the PBS analysis revealed 7 candidate regions of selection (Figure 3.11 and Table 3.4), of which two contained no genes. The candidate

region at 2p25 encompasses the Lipin 1 (*LPIN1*) gene. This gene encodes an enzyme involved in triglyceride synthesis (Langner et al., 1989; Reue et al., 2000; Reue, 2009) and is highly expressed in adipose tissue and skeletal muscle (Phan and Reue, 2005). Physiological studies in humans have also demonstrated a correlation between *LPIN1* expression levels in adipose tissue and insulin sensitivity (Frayn, 2002; Reue and Zhang, 2008). Additionally, variants in *LPIN1* have shown association with different metabolic phenotypes including fasting serum insulin, BMI, WC, obesity and T2D (Suviolahti et al., 2006; Loos et al., 2007; Wiedmann et al., 2008; Chang et al., 2010; Zhang et al., 2013b). The candidate region at 6p22 impacted at the MHC region that encompasses three MHC Complex Class I genes including *HLA-F*, *HLA-F-AS1*, and *HLA-G*. The region at 1p22 encompasses the Synapse Defective Rho GTPase Homolog 2 (*SYDE2*) gene. This gene and its paralog *SYDE1* are the mamalian orthologs of SYD-1, which is required for axonal guidance in *Caenorhabditis elegans*, and *Syd-1*, which regulates pre- and postsynaptic maturation in Drosophila (Hallam et al., 2002). The candidate region at 11q14 encompasses the Discs Large MAGUK Scaffold Protein 2 (*DLG2*) gene. This gene encodes proteins belonging to the membrane-associated guanylate kinase (MAGUK) superfamily, which are located in the postsynaptic density of glutamatergic excitatory brain synapses (Zhu et al., 2016). This gene has been associated with developmental disorders and intellectual disability (Reggiani et al., 2017).

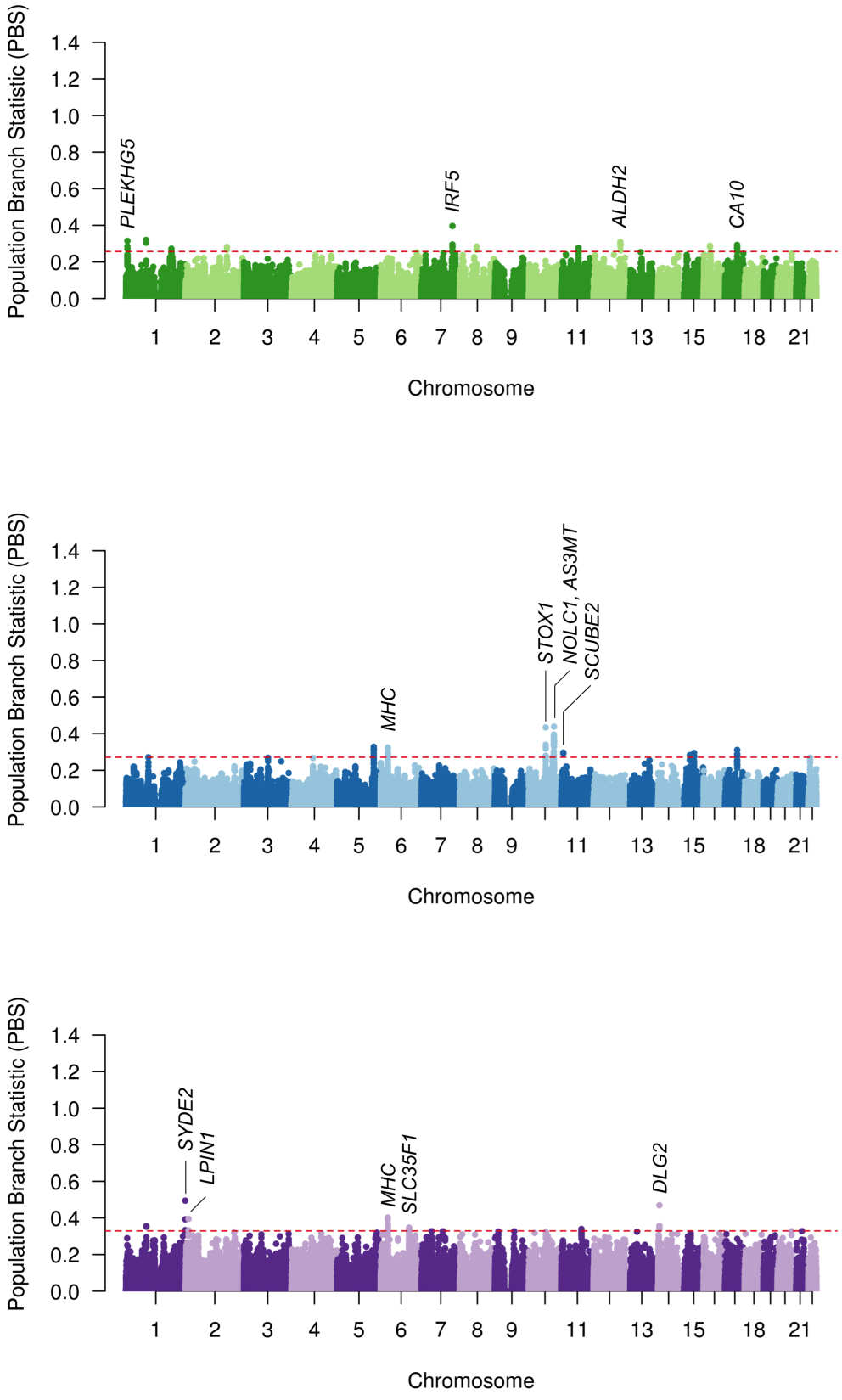

Figure 3.11: Genome-wide scan for selection in Meso-Americans, Andeans and Mapuche Native American populations. Population Branch Statistic (PBS) values per SNP. The red dashed line represents the 99.99$^{\text{th}}$ percentile. Names of genes associated with the highest peaks and discussed in the text are shown.

Table 3.2: Candidate regions under selection in Meso-American Native Americans based on the Population Branch Statistic (PBS) selection statistic.

Genomic coordinates (hg19)	**Length (bp)**	**PBS value (highest SNPs)**	***Candidate targeted genes***
Chr7:128560761-128773770	21,3009	0.40 (rs10488631)	*IRF5, DNASE1L3*
Chr1:6577765-6705944	128,179	0.32 (rs4243829)	*PLEKHG5, NOL9, TAS1R1, KLHL21, ZBTB48*
			PHF13, THAP3, DNAJC11
Chr1:84485512-84514988	29,476	0.32 (rs615352)	-
Chr12:111706877-111754597	47,720	0.31 (rs7300860)	*CUX2*
Chr12:112123284-112245170	121,886	0.30 (rs2238151)	*BRAP, ACAD10, ALDH2*
Chr12:112985328-113024793	39,465	0.29 (rs741334)	*RPH3A*
Chr16:26639714-26642059	2,345	0.29 (rs237135)	-
Chr17:49984205-50508394	35,109	0.29 (rs203075)	*CA10*
Chr17:50462226-50508394	46,168	0.29 (rs16951420)	-
Chr2:172098069-172127920	29,851	0.28 (rs4667682)	-
Chr8:70240509-70243940	3,431	0.28 (rs12542665)	-
Chr11:72390640-72394706	4,066	0.28 (rs341053)	-
Chr1:189918985-189931733	12,748	0.27 (rs815742)	-

Table 3.3: Candidate regions under selection in Andean Native Americans based on the Population Branch Statistic (PBS) selection statistic.

Genomic coordinates (hg19)	Length (bp)	PBS value (highest SNPs)	Candidate targeted genes
Chr10:104487382-105211432	724,050	0.44 (rs10883869)	*SFXN2, WBP1L, CYP17A1, C10Orf32, AS3MT,*
			CNNM2, C10Orf32-AS3MT, NT5C2, LOC729020, BC040734,
			PCGF6, INA, TAF5, CALHM2, USMG5
Chr10:70613280-70771895	15,8615	0.43 (rs10998460)	*STOX1, DDX21, DDX50, KIF1BP*
Chr10:103922896-103931947	9,051	0.34 (rs7897)	*NOLC1*
Chr6:30033884-30077967	44,083	0.32 (rs3132680)	*PPP1R11, TRIM31, RNF39, TRIM31-AS1*
Chr17:50299021-50376564	77,543	0.31 (rs1927589)	-
Chr5:151921873-152030714	10,8841	0.29 (rs2964243)	-
Chr15:62646690-62737452	90,762	0.29 (rs289151)	-
Chr11:9093486-9109287	15,801	0.30 (rs2647528)	*SCUBE2*
Chr15:45358846-45385916	27,070	0.28 (rs10851420)	*SORD, DUOX2*

Table 3.4: Candidate regions under selection in Mapuche Native Americans based on the Population Branch Statistic (PBS) selection statistic.

Genomic coordinates (hg19)	Length (bp)	PBS value (highest SNPs)	Candidate targeted genes
Chr1:247507887-247513049	5,162	0.49 (rs10924990)	-
Chr14:25646203-25698081	51,878	0.47 (rs1461556)	-
Chr2:11844839-11853964	9,125	0.40 (rs4640359)	*LPIN1*
Chr6:29616607-29828660	21,2053	0.40 (rs1611704)	*MOG, HLA-F, HLA-F-AS1, HCG4, LOC554223, HLA-G*
Chr1:85662851-85704193	41,342	0.36 (rs7524465)	*SYDE2*
Chr6:118608535-118692498	83,963	0.35 (rs7764272)	*SLC35F1*
Chr11:84627035-84708782	81,747	0.34 (rs10898329)	*DLG2*

3.4.6 Gene set enrichment analysis using biological pathways and GO categories in three distinct Native American populations

Applying the same methodology to test for polygenic adaptation as described in Section 3.4.4 to the three distinct Native American populations did not reveal any biological pathways enriched in PBS selection scores. The enrichment analysis of GO categories revealed 8 significant categories in the Andean Native American population (Table A.2). However, all of these were likely significant because they all contain several members of the UDP Glucuronosyltransferase gene family that are contiguous at 2q37.

3.5 Discussion and limitations

In this chapter I have performed a genome-wide selection scan based on the PBS selection statistic among Native Americans. By considering samples from throughout the American continent I was able to identify candidate regions of selection in the common ancestral population of Native Americans that likely resulted from selective pressures imposed on this population prior to the range expansion into the American continent. In addition to detecting signals of adaptation shared across the Americas, I also conducted a PBS selection scan based on pseudounadmixed allele frequencies for Meso-American, Andean and Mapuche Native American populations that I estimated using admixed Latin Americans individuals who derive most of their Native American ancestry from these populations. This allowed me to explore instances of local adaptation in the Americas. Some of the candidate regions of selection did not contain any genes and for some of them it is less clear what the selective pressure may have been. I therefore discuss the functional significance of the genes within the candidate regions based mainly on prior biological studies and association with normal and disease phenotypes. I organize the discussion of candidate genes below into two broad functional categories: metabolism and defense against pathogens. I then end by discussing candidate selected genes that have been reported to be under selection in other studies including Native Americans and the main limitations of these findings.

3.5.1 Dietary adaptations in Native Americans

Currently, the best-documented examples of selection in Native Americans is through dietary adaptations (Acuña-Alonzo et al., 2010; Amorim et al., 2017; Mychaleckyj et al., 2017; Hlusko et al., 2018). In line with these previous findings, I detected the strongest signal of selection across Native Americans at the *ADAMTS9* gene. This gene has been previously associated to different anthropometric and metabolic phenotypes in various human populations. Notably, a recent selection scan conducted in admixed North-Eastern Brazilians showed that, after accounting for non-Native American ancestry, *ADAMTS9* possessed the strongest selection signal across the genome (Mychaleckyj et al., 2017). Given that no North-Eastern Brazilian Native Americans were included in the selection analysis performed here, this finding further supports that *ADAMTS9* might have represented an important adaptive gene in the common ancestral population of Native Americans. Recently, Amorim et al. (2017) suggested that the strong signal of selection shared

across Native American, and that had been previously reported to be under selection Inuit Greenlanders (Fumagalli et al., 2015), was consistent with a strong adaptation in Beringia. To test whether the selection signal at *ADAMTS9* was also in line with this finding, I explored the frequency of the top selected SNP (rs763139) in a world-wide set of human populations that included Arctic populations (Table A.1). Notably, SNP rs763139 is not only at its highest frequency in Native Americans, as expected, but also in several Arctic populations, with the highest frequency among these populations observed in the Eskimo (Figure 3.12). Eskimos are the result of an admixture process between the Ancestors of the first Americans colonizers and North Asians (Reich et al., 2012). Multidisciplinary studies have also suggested that Eskimos are descendants of ancient Beringians (González-José et al., 2008; Bortolini et al., 2014). The high frequency of the selected variant observed here shared with Eskimos, is reminiscent of the signal found in the *FADS2* and also in line with the recent suggested adaptive role of *EDAR* (Hlusko et al., 2018), and therefore also consistent with a scenario of selection in Beringia. Importantly however, the SNPs within *ADAMTS9* do not seem to affect the different anthropometric or metabolic phenotypes in admixed Latin Americans. It will be necessary to conduct genetic association analysis in Native American populations to further elucidate the phenotypic effect of *ADAMTS9* in this population. In addition, although many of the top SNPs in the selection analysis for all Native Americans showed moderate to high frequency in Siberian populations (Figure A.7), the top SNP located within the *NPAS2* gene showed one of the highest allele frequencies in the four Siberian populations analyzed here (Figure 3.12). The *NPAS2* gene encodes a transcription factor thought to be the major transcriptional regulators of the circadian clock mechanism in mammals (DeBruyne et al., 2007), an internal time-keeping system that regulates distinct physiological processes including feeding behaviour, lipid and carbohydrate metabolism, sleep, and blood pressure control (Dunlap et al., 2004). Studies on knockout mice in circadian clock genes have also shown associated changes in physiology, which include altered insulin/glucose responsiveness, obesity and arrhythmicity (van der Horst et al., 1999; Bunger et al., 2000; Rudic et al., 2004; Staels, 2006). Further, in humans, epidemiological studies have also linked *NPAS2* with a variety of psychological disorders, including winter depression (Partonen et al., 2007). Notably, in a recent selection scan of circadian clock genes in several worldwide human populations, one variant within *NPAS2* was shown to possess extreme allele frequency differences between populations living at different latitudes, and hence going through different modes of seasonal fluctuations including photoperiods (Dall'Ara et al., 2016). It is interesting to speculate as to whether longterm habitation at high latitudes in Beringia also prompted selection for better adaptation to extreme photoperiods in the ancestors of Native Americans.

3.5.2 Immune adaptations in Native Americans

In addition to genes potentially related to dietary adaptations, it is intriguing that immune-related genes comprise several of the top candidate regions of selection in Native American populations. In the Meso-American Native Americans the strongest signal of selection is

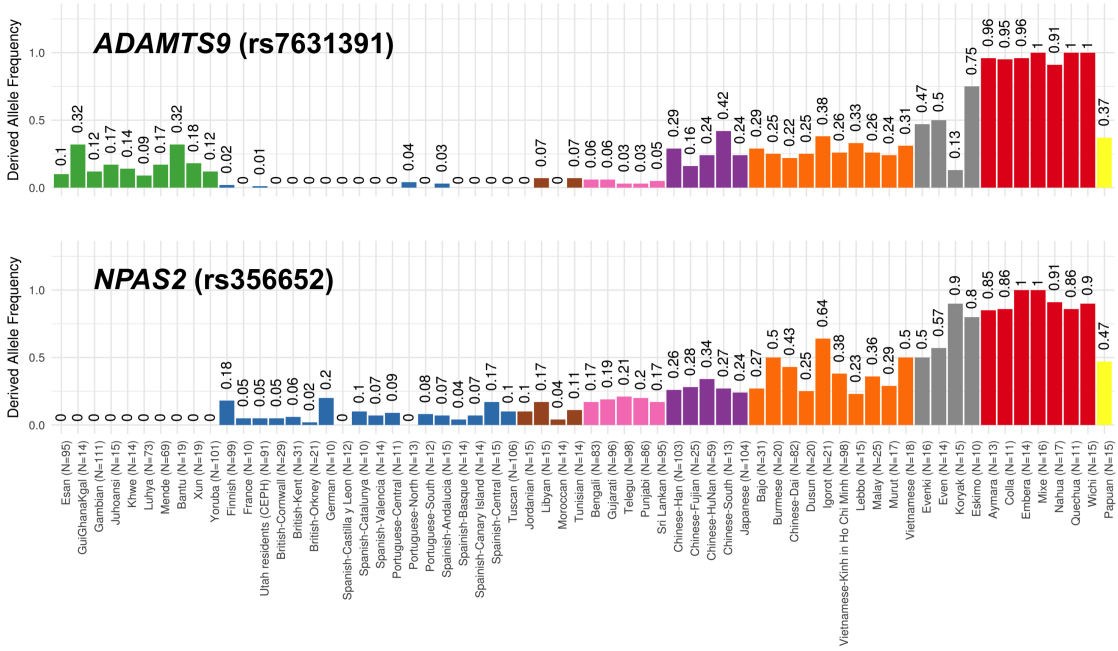

Figure 3.12: Worldwide allele frequencies of the top PBS SNPs detected in Native Americans within *ADAMTS9* and *NPAS2*. The allele frequencies are estimated from 2,391 unrelated individuals collated from several public databases. The colors of the bars reflect the geographic origin of the populations for which the allele frequencies were estimated: Africa (green), Europe (blue), Middle East and North Africa (brown), South Asia (pink), East Asia (purple), South East Asia (orange), Siberian(grey), America (red) and Oceania (yellow). The number of individuals in each population (N) is given next to the population name and the derived allele frequency is shown at the top of each bar.

located in the *IRF5* gene, which is part of a group of transcription factors with diverse roles, including virus-mediated activation of interferon and immune system activity (Yanai et al., 2007). In the Andean and Mapuche Native American populations, the strongest signals of selection are located in the MHC genomic region, which plays an important role in the immune system (Hill, 1998, 2001; Shiina et al., 2009; Mosaad, 2015). Additionally, in the selection analysis comprising all Native American samples, one of the top candidate regions encompasses the *CCDC134* gene that is implicated in immune function likely through a cytokine-like function (Huang et al., 2014). Notably, there is evidence showing that Native American populations are immunologically different from other populations (Bhatia et al., 1995; Lindenau et al., 2014a,b; Augusto et al., 2015; Lindenau et al., 2016). It has been shown that there are a reduced number of human leukocyte antigen (HLA) alleles (the gene complex encoding the majority of the MCH) in Native American populations compared to non-Native American populations, which has been suggested as one of the explanations for their differentiated susceptibility to introduced diseases from the Old World (Bhatia et al., 1995; Lindenau et al., 2014b; Augusto et al., 2015; Lindenau et al., 2016). Importantly, because these selection analyses were conducted on living Native Americans who represent surviving individuals of populations affected by European colonization, it is not possible to determine whether the immune-related genes found here were selected before or after European contact. Interestingly however, in a recent study of ancient individuals from the Northwest Coast of North America dating from before European contact, Lindo et al. (2016) showed that the strongest signals of selection derived from the MHC region. It is thus likely that at least some of adaptive signals found here support a hypothesis of Native American populations adapting to local pathogens in the Americas and not only to diseases brought after the European contact.

3.5.3 Adaptations at previously reported genes under selection

Another gene of adaptive interest found in the Meso-American Native American population may be *ALDH2*. This gene encodes the second enzyme of the major oxidative pathway of alcohol metabolism (Ohta et al., 2004). Specific variants within this gene are involved in alcohol metabolism and are strongly associated to alcohol consumption (Luczak et al., 2006; Eng et al., 2007; Baik et al., 2011; Takeuchi et al., 2011; Quillen et al., 2014; Wall et al., 2016; Gelernter et al., 2018) and esophageal cancer (Cui et al., 2009) in Asian populations. Interestingly, in a sample of Native Americans, Long et al. (1998) found evidence for genetic linkage of alcohol dependence in the vicinity of the alcohol dehydrogenases (ADH) gene cluster. Similarly, Mulligan et al. (2003) showed associations of ADH alleles with an increased risk of alcohol dependence and binge drinking in a Native American sample. The genes underlying human alcohol metabolism provide a fascinating example of how genetic variants can contribute to a complex phenotype through distinct physiological and behavioural processes. During alcohol metabolism, alcohol is first oxidized by alcohol dehydrogenase to acetaldehyde, which is then oxidized to acetate by acetaldehyde dehydrogenase (Eng et al., 2007; Wall et al., 2016). These enzymes occur in several distinct isozyme forms encoded by multigene families. Specific variants at these

loci, including *ALDH2*, can produce physiological reactions including increased levels of blood flow, dizziness, increased heart rate, sweating and nausea, which in combination are usually referred to as "flushing response" (Harada et al., 1981; Eng et al., 2007; Wall et al., 2016). Individuals suffering from these physiological response are usually protected from alcoholism due to the undesirable feeling associated to heavy alcohol consumption (Matsushita and Higuchi, 2017). Although it would be difficult to assert whether alcohol consumption was the causal phenotype that derived selective pressure, the fact that heavy alcohol consumption is also a major risk factor for developing esophageal cancer, selection at *ALDH2* in Meso-American Native American for protection to alcoholism is a hypothesis worth considering.

In the Andean Native American group the highest selection signal encompasses the *AS3MT* gene, which encodes an enzyme that is essential for the efficient metabolization of arsenic in humans (Sumi and Himeno, 2012). Although high levels of arsenic in water resources can be found all over the world, there is evidence that high levels of arsenic have been present in different water resources in the Andean region for several thousands of years (Concha et al., 2006; López et al., 2012). Long term exposure to arsenic can result in a range of ailments that include skin lesions, cancer, cardiovascular and pulmonary diseases (Abernathy et al., 2003; Rahman et al., 2009; Banerjee et al., 2013; Sun et al., 2014), and can also hinder foetal development as arsenic can cross the placental barrier (Fry et al., 2007; Raqib et al., 2009). Notably, previous studies in different Andean Native American populations have detected selection signatures within *AS3MT* (Engström et al., 2013; Eichstaedt et al., 2015b; Apata et al., 2017). In light of these results, it seems that high arsenic levels have probably imposed a strong selective pressure throughout the Andean altiplano and has not been restricted to only some specific Andean Native American groups. It would be interesting to analyze whether the selected haplotype is shared between several Andean populations or has resulted through independent selection. Another gene of adaptive interest in the Andean Native American population is the *STOX1* gene. Mutations in this gene have been previously associated to preeclampsia in European women (van Dijk et al., 2005, 2010). Preeclampsia is a pathology of pregnancy characterized by high blood pressure and signs of damage to another organ system, that can be lethal for the mother and for the fetus (Sibai, 2005; Agius et al., 2017). Interestingly, this condition has been reported to be higher in Hispanic women (Wolf et al., 2004) and exacerbated in high altitude regions such as the Andes (Moore et al., 1982, 2004; Julian, 2011). Because environmental factors such as high altitude might affect the susceptibility to develop preeclampsia it is possible that selection has acted on genes related to this condition. Notably, there is evidence showing that endothelin-1 (ET-1), a peptide hormone with potent vasoconstrictor properties, is differentially regulated by pregnancy and chronic hypoxia in Andeans compared to European residents of high altitude (Moore et al., 2004). The authors suggested that this findings support the hypothesis that long-term inhabitants of high altitude may be protected from the effect of chronic hypoxia on vascular responses to pregnancy via ET-1. It is therefore possible that selection has acted on genes related to pregnancy disorders affected by high altitude habitation such

as *STOX1* in Andean populations. In order to elucidate the adaptive importance of this gene, for example to test whether variants in this gene confer a protection to preeclampsia in Native American Andean women, it will be necessary to assess whether this gene is associated with preeclampsia in this population.

One key limitation of this study when considering samples from throughout the American continent is that I could not differentiate between selective events that occurred only in the ancestral population of Native Americans, or in the ancestral populations of Native Americans and Siberians, as the PBS selection analysis conducted here included East Asians and Europeans as reference groups. To overcome this limitation, I could have incorporated northern Siberians and East Asians in the PBS selection analysis. I would expect the strong signals of selection observed at *ADAMTS9* and *NPAS2* to disappear, as the derived allele of the SNPs with high PBS scores at these locus showed also high frequencies in Siberian populations (Figure 3.12), but to find novel signals of selection that occurred probably after Native Americans entered the continent. Further, this same rationale could also be applied to the XP-EHH analysis, using northern Siberians instead of East Asians as a reference group. An important limitation of the PBS analysis conducted in the different Native American populations using pseudo-unadmixed allele frequencies is that it assumes that both the Spanish and Western African population used here, were good proxies for the ancestral populations that admixed with modern Latin Americans, and that the allele frequencies in these populations today, have not largely changed since the admixture event. If these two assumptions are strongly violated then it is possible that the pseudo-unadmixed allele frequencies for the different Native American populations have not been correctly estimated, which could affect the PBS selection analysis by producing spurious signals of positive selection. Finally, for both the selection analysis conducted in samples from throughout the Americas and in the different Native American populations, I have limited my discussion based on the functional significance of the genes within the candidate regions based mainly on prior association studies. However, many of the SNPs with the highest selection signals were found in intergenic regions Table 3.1, 3.2, 3.3 and 3.4). It possible that many of the selection signals were driven by distal regulatory effects and not by changes affecting the structure of the gene. One way to address this limitation could have been to query the Genotype-Tissue Expression (GTEx) database (GTEx Consortium, 2013) for associations between genotypes and gene expression across a large panel of human tissues for evidence of cis- or trans-expression quantitative trait loci (eQTLs). In addition, it would have also been possible to explore the functional consequences of genetic variation at the SNPs with the strongest selection signals using the Combined Annotation Dependent Depletion (CADD) scores (Kircher et al., 2014), which provides an estimate of the functional severity of each SNP. Finally, to further investigate the possible phenotypic effect of these SNPs, I could have taken advantage from other resources besides the CANDELA cohort, such as the Biobank cohort (Bycroft et al., 2018), which although does not include individuals of Latin American, provides a rich variety of phenotypic and health-related information.

3.6 Summary

In this chapter I have conducted a genome-wide scan of selection in Native Americans and identified important candidate genes that I hypothesize were likely beneficial in the ancestral population of Native American in Beringia prior to the entry into the American continent, particularly adaptation to diet. In addition, I also explored instances of local adaptation in the Americas using admixed Latin American individuals that derive most of their ancestry from distinct Native American populations. I reported selection signals at immune-related genes in these Native American populations that probably resulted from an adaptation to local pathogens in the Americas or perhaps to diseases brought after European contact.

Chapter 4

Detecting Post-Columbian signals of selection in Latin Americans

4.1 Overview

In this chapter I conduct a genome-wide scan of selection post-admixture in five admixed Latin American populations from Brazil, Chile, Colombia, Mexico and Peru. I introduce a novel statistical model that identifies loci that have been selected after an admixture event by modelling the admixture proportions and genetic drift in an admixed population. Using this new method, I report a strong signal of selection post-admixture in the Peruvian samples at a genomic region associated to metabolic-related phenotypes. I also test for selection post-admixture by searching genomic regions with unusually high or low levels of Native American, European or African ancestry. This analysis shows a highly significant increase in African ancestry in the Chilean and Mexican populations, expanding over a large genomic region that encompasses the Major Histocompatibility Complex (MHC), which harbours various genes with a known function in immune response, including resistance and susceptibility to a broad range of infectious diseases. I discuss the potential advantages and limitations of detecting selection post-admixture by employing these two different approaches.

4.2 Background

Admixed populations offer a unique opportunity to detect recent selection. In addition to the canonical view where selection acts on a novel mutation, or a variant already present in the population (known as selection on de-novo mutations or on standing variation, respectively), another potential source of genetic variation can arise from an admixture event (Figure 4.1). In the human lineage, genomic studies have demonstrated the pervasiveness of admixture events in the history of the vast majority of human populations (Green et al., 2010; Patterson et al., 2012; Hellenthal et al., 2014; Lazaridis et al., 2014; Prüfer et al., 2014). By leveraging the identification of the genomic segments of the donor ancestral populations, several recent studies have demonstrated the importance of this process in having contributed to genetic adaptations among human populations. Notable examples include the transfer of a protective allele in the Duffy blood group gene providing resistance to *Plasmodium vivax* malaria in Malagasy population from Bantu-speaking

Africans (Hodgson et al., 2014; Pierron et al., 2014, 2018) and the transmission of the lactase persistence allele in the Fula pastoralists from Western Eurasians (Busby et al., 2017).

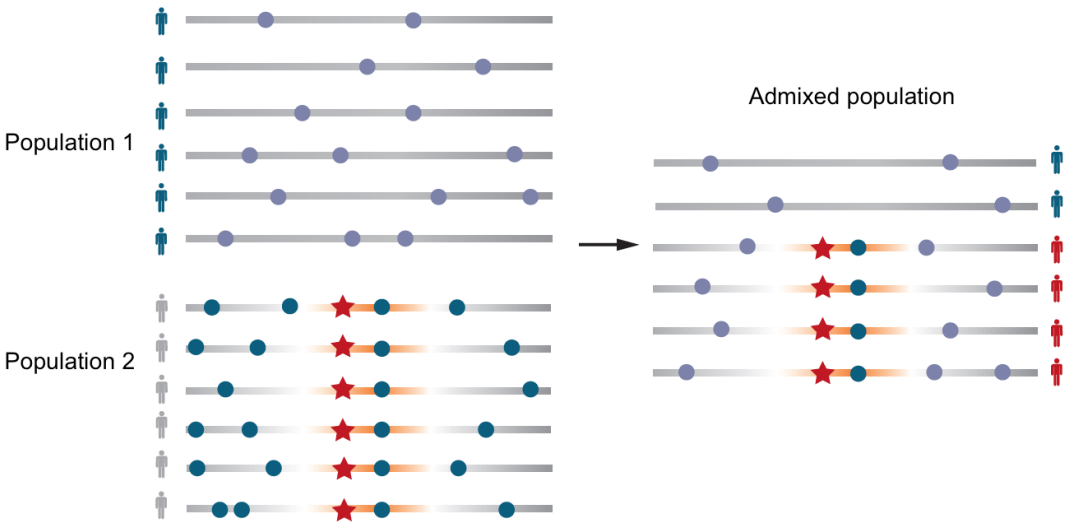

Figure 4.1: Schematic representation of a selection post-admixture event. Each horizontal bar represents an haplotype in two different populations. The selected variant is represented as a star and neutral variants as dots. Following an admixture event, the beneficial variant can rise in frequency in the admixed population due to selection. Adapted from Fan et al. (2016).

An ideal setting in which to test whether admixture contributed to genetic adaptation is Latin America. The genetic make-up of the majority of present day Latin Americans stems mainly from three distinct ancestral populations: indigenous Native Americans, Europeans (mainly from the Iberian Peninsula), and West Africans (Wang et al., 2008; Moreno-Estrada et al., 2013, 2014; Homburger et al., 2015; Chacon-Duque et al., 2018). The admixed genomes of Latin Americans are thus the result of an intermixing process between human populations that had been evolving independently for several tens-of-thousands of years and that were suddenly brought together in a new environment. In this new environment, the ancestral genomes were quickly subjected to novel environmental challenges that were largely unfamiliar from where they first evolved. Therefore, it is expected that some variants from the different ancestral populations may be more beneficial and hence increase in frequency to the detriment of the others because of selection. Motivated by this scenario several studies have explored the genomes of admixed Latin Americans in search for signatures of selection post-admixture (Tang et al., 2007; Basu et al., 2008; Ettinger et al., 2009; Guan, 2014; Rishishwar et al., 2015; Deng et al., 2016; Zhou et al., 2016). These studies have relied on an approach similar to that of admixture mapping, where the ancestry of a genomic region is assigned to a particular ancestral population, averaged across all individuals, and compared to the genome-wide population average. A significantly deviated ancestry from the genome-wide population average is then assumed to have evolved under some form of selection (Tang et al., 2007). The first study exploring selection post-admixture in Latin Americans that used genome-wide SNP data examined a small cohort of Puerto Rican individuals and reported three genomic

regions thought to be under strong recent selection post-admixture. The strongest signal was found at chromosome 6q at MHC. MHC harbors various genes with known function in immune response, including resistance to different pathogens (Hill, 1998, 2001; Frodsham and Hill, 2004). The increase in African ancestry was suggested to have conferred a selective advantage from infectious diseases likely brought from the Old World to the Americas after European colonization. In a more recent study, Rishishwar et al. (2015) examined a small sample of Colombian individuals and reported several genomic regions thought to be under strong recent selection post-admixture, using an approach similar to that of ancestry deviations. The strongest genomic regions with evidence of selection post-admixture included MHC, as well as the genomic region harbouring *SLC45A2*, which has been associated to skin pigmentation, and the ectodysplasin A receptor *EDAR*, which is involved in a range of phenotypic traits, including sweat gland density, incisor shoveling, and mammary gland ductal branching. Rishishwar et al. (2015) suggested that MHC could confer immune resistance to different endemic hosts in the Colombian population such as malaria, that the increase in European ancestry at *SCL4A5* could be related to preference for partners of lighter skin pigmentation (i.e. assortative mating), and that the increase in Native American ancestry in *EDAR* was involved in adaptation to the tropical environment given the association with sweat gland density. Finally, in a more recent study, Zhou et al. (2016) analysed genome-wide SNP data from over 3,000 Mexican individuals and reported a single genomic region under strong recent selection, as evidenced by an elevated amount of African ancestry that also impacted on the MHC region. Similar to previous reports, the authors interpreted the increase in African ancestry as occurring through an enrichment of African alleles that conferred protection from diseases likely brought from the Old World. Nonetheless, this approach has been criticized due to long-range LD (i.e. admixture-induced LD) (Price et al., 2008), inaccurate ancestral populations (Baran et al., 2012; Pasaniuc et al., 2013; Deng et al., 2016; Zhou et al., 2016) and systematic biases in local-ancestry inference (Baran et al., 2012; Pasaniuc et al., 2013; Zhou et al., 2016). Furthermore, sampling error and genetic drift have also been suggested to cause large differences in ancestry proportions along the genome (Long, 1991; Bhatia et al., 2014; Mathieson et al., 2015). Finally, detecting selection post-admixture can also be problematic in cases where admixture occurred only a few generations ago, as there may not have been enough time for selection to generate a detectable signal in the distribution of ancestry across the genome (Pierron et al., 2018).

In this chapter I present a novel statistical model for identifying SNPs under selection post-admixture using genome-wide data from five different Latin American populations collected as part from the CANDELA Consortium (Ruiz-Linares et al., 2014) and the 1000 Genomes Project (hence forth 1KG) (1000 Genomes Project Consortium et al., 2015). In contrast to previous studies, this approach is based on allele frequencies and as such does not require correct confident assignment of local ancestry along the genome to identify regions to detect selection post-admixture. In order to contrast the findings here with previous studies that have suggested the action of selection post-admixture at some loci in Latin Americans populations, I complemented the allele-frequency based analysis by

searching for local ancestry deviations along the genome. I report signals of selection post-admixture at genes related to metabolic function and the immune system. I end this chapter by discussing the advantages and limitations of these two different approaches.

4.3 Materials and methods

4.3.1 Description of the genomic data

The Latin American individuals analyzed here are part of the CANDELA (Ruiz-Linares et al., 2014) and the 1KG (1000 Genomes Project Consortium et al., 2015) data collections. The CANDELA sample included data sampled from 6,630 volunteers from five countries (Brazil, Chile, Colombia, Mexico and Peru) (Section 1.8). All participants were genotyped on the Illumina HumanOmniExpress chip at 730,525 SNPs. The genotype data from CLM (Colombians, from Medellin, Colombia), MXL (Mexicans from Los Angeles, USA) and PEL (Peruvians, from Lima, Peru) individuals from the 1KG Project were downloaded from the 1KG FTP site available at ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/.

4.3.2 Quality control

I used PLINK v1.9 (Chang et al., 2015) to perform quality control (QC) analyses. For the 1KG data I excluded insertions or deletions (indels) and further retained only bi-allelic SNPs. I then merged these two datasets and excluded SNPs and individuals with more than 1% missing data or that showed evidence of genetic relatedness. Due to the admixed nature of Latin Americans, there is an inflation in Hardy-Weinberg P-values and therefore I did not exclude SNPs based on Hardy-Weinberg deviation. This resulted in a total of 476,840 autosomal SNPs and 6,352 individuals.

4.3.3 ADMIXTURE and PCA analysis

Ancestry values were estimated from a set of LD pruned SNPs via supervised ADMIXTURE (Alexander et al., 2009) at $K = 3$. References populations from African (YRI; Yoruba in Ibadan, Nigeria), southern Europe (IBS; Iberian Population in Spain) and Native American groups were chosen from the 1KG and selected Native Americans populations from Chacon-Duque et al. (2018). I then removed non-admixed Latin American individuals that I define as having less than 10% or more than 90% Native American genome-wide ancestry. Based on a Principal Component analysis (PCA) I noted that the PEL, MXL and CLM individuals from the 1KG clustered with the Peruvian, Mexican and Colombian individuals from the CANDELA dataset, respectively (Figure B.1) and therefore decided to group these individuals into one population. The final Latin American populations included in the selection analyses consisted of 208 Brazilians (BRA), 1,887 Colombians (COL; composed of CLM and Colombians from CANDELA), 1,770 Chileans (CHL), 1,256 Mexicans (MEX; composed of MXL and Mexicans from CANDELA) and

1,231 Peruvians (PER; composed of PEL and Peruvians from CANDELA).

4.3.4 A new statistical model to detect selection post-admixture

The novel model presented here was written by Dr. Garrett Hellenthal and developed jointly by Dr. Garrett Hellenthal and myself. This model can detect loci under selection post-admixture by using genome-wide SNP data from one admixed population and a set of ancestral ("donor") populations. The model compares the observed allele frequency from the admixed population to those expected under a linear combination of the allele frequencies of ancestral populations and their ancestry contributions. This approach is most closely related to the work of Long (1991) and more recently to that of Mathieson et al. (2015). However and in contrast to previous studies, the model can also simultaneously estimate an additional parameter that controls for the variance of the expected allele frequency in the admixed population using all SNPs across the genome. This parameter can therefore be thought of broadly as a "genetic drift" estimate that is able to capture the amount of drift experienced in the modern admixed population after the admixture event, as well as unmodeled ancestry not captured by the linear combination model. The model then evaluates whether each SNP shows a deviation larger than that expected from the admixture process and genetic drift, which is assumed to be a signal of selection post-admixture.

Formally, assume each bi-allelic SNP $j \in [1, \ldots, S]$ from a modern admixed population k formed from ancestral (donor) populations $d \in [1, \ldots, D]$ draws p_{jk}, its frequency of a chosen allele, from a Beta distribution. The aim is to test whether a modern admixed population has a highly selected (i.e. deviated) allele frequency at each SNP. If SNP j is not selected, the mean of the Beta distribution is α_{jk}, with variance $F_k a_{jk}(1-a_{jk})$. Under this latter setting α_{jk} represents the allele frequency of the chosen allele at SNP j in the ancestral admixed population k formed as a linear combination of allele frequencies in D ancestral (donor) populations. F_k measures the amount by which the modern admixed population has drifted from this ancestral admixed population. Under the former selected setting, p_{jk} follows a uniform distribution.

Let $G_{jk} \in \{0, 1\}$ be an indicator for whether SNP j is selected in the modern admixed population k with prior probability s. β_{dk} is the admixture contribution of ancestral donor population d to the admixed population k and λ the parameter of a symmetric Dirichlet distribution. p_{jd} is the allele frequency at SNP j in ancestral donor population d, n_{jk} the number of non-missing alleles of admixed population k at SNP j, x_{jk} the counts of the chosen allele at SNP j in admixed population k, n_{jd} the number of non-missing alleles of ancestral donor population d at SNP j, and x_{jd} the counts of the chosen allele at SNP j in the ancestral donor population d. Formally, the model assumes:

$$p_{jk}|\alpha_{jk}, F_k \sim \begin{cases} Beta(\alpha_{jk}(1-F_k)/F_k, (1-\alpha_{jk})(1-F_k)/F_k, & \text{if } G_{jk} = 0. \\ Uniform[0,1], & \text{if } G_{jk} = 1. \end{cases} \tag{4.1}$$

$$x_{jk}|p_{jk} \sim Binomial(n_{jk}, p_{jk}) \tag{4.2}$$

$$G_{jk}|s \sim Bernoulli(s) \tag{4.3}$$

$$\alpha_{jk} = \sum_{d=1}^{D} \beta_{dk} p_{jd} \tag{4.4}$$

$$\beta_{dk}|\lambda_1, \ldots, \lambda_D \sim Dirichlet(\lambda_1, ..., \lambda_D) \tag{4.5}$$

$$p_{jd}|x_{j,d}, n_{j_d} \sim Beta(x_{jd} + 1, n_{jd} - x_{jd} + 1) \tag{4.6}$$

$$F_k \sim Beta(c, d) \tag{4.7}$$

Note that $x_{jk}|\alpha_{jk}, F_k, G_{jk}$ follows a Beta-Binomial distribution. I use c=1 and d=500 to define the distribution of F_k, assume s is equal to 0.001 i.e. the prior probability for a SNP to selected and λ_d=1 for $d = [1, \ldots, D]$. I sample posterior probabilities for G_{jk} at each SNP j for each admixed population k, conditional on the data. I use the following Markov Chain Monte Carlo (MCMC) technique to do so, components of which are analogous to Falush et al. (2003).

Start with initial values of $\alpha_{jk}^{(0)}$ equal to the allele frequency of the chosen allele at SNP j using Equation 4.4. β_{dk} is sampled using Equation 4.5 and each p_{jd} is sampled using Equation 4.6

Then for $m = [1, \ldots, M]$:

A. Sample $G_{jk}^{(m)}$ using Gibbs sampling, such that

$$Pr(G_{jk}^{(m)} = g) \propto Pr(x_{jk}|\alpha_{jk}^{(m-1)}, F_k^{(m-1)}, G_{jk}^{(m)} = g)s^{(m-1)} \tag{4.8}$$

for $j = [1, \ldots, S]$ and $k = [1, \ldots, K]$.

B. For $k = [1, \ldots, K]$ update $\beta_{dk}^{(m)}$ using an Metropolis-Hastings (M-H) step. I.e. sample x from a normal distribution with mean zero and standard deviation 0.05. Randomly select one donor population d (i.e. $d = 1$) and set its $\beta_{1k}^{(m)}$ to $\beta_{1k}^{(m-1)} + x$, and then select a second donor population d (i.e. $d = 2$) and set its $\beta_{2k}^{(m)}$ to $\beta_{2k}^{(m-1)} - x$. Accept $\beta_{dk}^{(m)}$ with probability MH_β where $MH_\beta \equiv min(1, \frac{\pi(\beta_{dk}^{(m)})}{\pi(\beta_{dk}^{(m-1)})})$ and

$$\pi(\beta_{dk}^{(m)}) = \prod_j [\sum_g Pr(x_{jk}|\alpha_k^{(m)}, F_k^{(m-1)}, G_{jk}^{(m)} = g) 1_{[G_{jk}^{(m)}]}] Pr(\beta_{dk}^{(m)}|\lambda^{(m-1)}). \tag{4.9}$$

Automatically reject any $\beta_{dk}^{(m)}$ outside (0,1).

D. Update $\alpha_{jk}^{(m)}$ with $\beta_{dk}^{(m)}$ and $p_{jd}^{(m)}$ using Equation 4.4.

E. For $k = [1, \ldots, K]$ update $F_k^{(m)}$ using an M-H step. I.e propose a new $F_k^{(m)}$ by sampling from a normal distribution with mean $F_k^{(m-1)}$ and standard deviation 0.05. Accept $F_k^{(m)}$ with probability MH_F , where $MH_F \equiv min(1, \frac{\pi(F_k^{(m)})}{\pi(F_k^{(m-1)})})$ and

$$\pi(F_k^{(m)}) = \prod_j [\sum_g Pr(x_{jk}|\alpha_k^{(m)}, F_k^{(m-1)}, G_{jk}^{(m)} = g) 1_{[G_{jk}^{(m)}]}] Pr(F_k^{(m)}). \quad (4.10)$$

Automatically reject any $F_k^{(m)}$ outside (0,1).

For large M, this algorithm is guaranteed to converge to the true posterior distribution of the G_{j_k}'s (Gamerman and Lopes, 2006). In practice, for all results presented here I use M=20,000, sampling every 200$^{\text{th}}$ iteration after an initial "burn-in" of 10,000 iterations. To remove an undesirable observed pattern of low minor allele frequency (MAF) SNPs showing higher posterior probabilities, I replaced F_k with $V_k/(\alpha_{jk}(1-\alpha_{jk}))$ and instead infer V_k in step E. by sampling from a normal distribution with mean zero and standard deviation 0.01, but otherwise using the same M-H acceptance procedure.

I defined the ancestral populations for all five Latin American populations using the same Native Americans, Spanish Europeans (IBS) and Western Africans (YRI) used in the ADMIXTURE analysis (described above in Section 4.3.3). This method is run in two steps: for each population I run the algorithm for each chromosome separately, and then averaged the inferred admixture proportions and genetic drift parameters across chromosomes, weighting these parameters by the number of SNPs. I then fixed these parameters in each population separately to estimate the posterior probability for selection post-admixture at each SNP. SNPs with posterior probabilities higher than 0.5 (i.e. those more likely of being selected than being not selected) were considered to be under selection post-admixture.

4.3.5 Local ancestry deviation analysis

To conduct local ancestry assignment I used the discriminative modeling approach implemented in RFMix (Maples et al., 2013). In order to increase the local ancestry assignment accuracy, which is dependent on the number of SNPs (Maples et al., 2013), I used a more lenient QC threshold removing SNPs with more than 5% missingness. A higher number of SNPs is expected to increase the accuracy of local ancestry assigment (Baran et al., 2012; Maples et al., 2013; Pasaniuc et al., 2013). This left a total of 667,674 autosomal SNPs. The phased genotype data needed as input was obtained by using SHAPEIT2 (Delaneau et al., 2013) with default parameter settings. Genetic distances were obtained from the HapMap Phase II genetic map build GRCh37 (International HapMap Consortium, 2003). As reference continental panels I used Native Americans, Spanish Europeans (IBS) and Africans (YRI) individuals, as used in ADMIXTURE and in the new method presented here, but setting the number of individuals per reference population to 100, in order to avoid biases resulting from unbalanced reference panel sizes in the random forest algorithm (Maples et al., 2013). I ran RFMix with default parameters, the phase correction feature

enabled, and performed two rounds of the Expectation-Maximization (EM) algorithm.

To test for selection I followed the rationale that an excess or depletion of a given ancestry on a genomic region can be indicative of selection post-admixture (Tang et al., 2007). I computed the delta-ancestry value (Tang et al., 2007) for each SNP along the genome. The delta-ancestry value is the difference in the ancestry proportion at a SNP compared to the genome-wide average of that particular ancestry (Section 2.5.1). I estimated the average ancestry at each locus as the average of that particular ancestry across all samples. To estimate the genome-wide average I estimated the average of each ancestry across all SNPs and all samples. Formally, for a particular ancestry k at a SNP j, delta ancestry (δ_k^j) is defined as:

$$\delta_k^j = \overline{q}_k^j - \overline{q}_k \tag{4.11}$$

where $\overline{q}_k^j$ is the mean of ancestry k at SNP j averaged over all individuals, and $\overline{q}_k$ is the proportion of ancestry k averaged over all individuals and the entire genome. Since delta-ancestry values are normally distributed (Tang et al., 2007), I used standard deviation (SD) units from the genome-wide average to assess significance. In order to correct for multiple testing, I note that this analysis is similar to that of admixture mapping, where the number of ancestry blocks is the important parameter to account for the number of independent hypotheses and not the total number of SNPs. I therefore used a statistical significance threshold of P-value $= 5.68 \times 10^{-5}$ that has been reported to be a sensitive significance threshold for admixture mapping in Latin American populations (Browning et al., 2016). Under normality, this significance threshold would correspond to 4.03 SD units from the genome-wide average for a particular ancestry.

4.4 Results

4.4.1 Description of the new beta-binomial model to detect selection post-admixture

The new model is based on the principle that allele frequencies in an admixed population can be described as a linear combination of the allele frequencies in the ancestral ("donor") populations and their ancestry contributions. Given that admixture affects all loci equally, the admixture proportions in an admixed population are expected to be the same across all loci, and departures from the expectation are usually assumed to have evolved under some type of selection (Long, 1991; Mathieson et al., 2015). However, since the estimated admixture proportions at a locus can also vary simply due to genetic drift experienced in the admixed population after the admixture event, it is important to account explicitly for this factor. This model detects selection post-admixture by simultaneously estimating admixture proportions and a genetic drift parameter (described in detail in Section 4.3.4) in an admixed population (using joint information across all SNPs). It then aims to identify loci in which the modern admixed population has experienced a larger than expected change in allele frequency compared to the predicted allele frequency based on

the admixture process and genetic drift, which is assumed to be evidence for selection post-admixture (Figure 4.2).

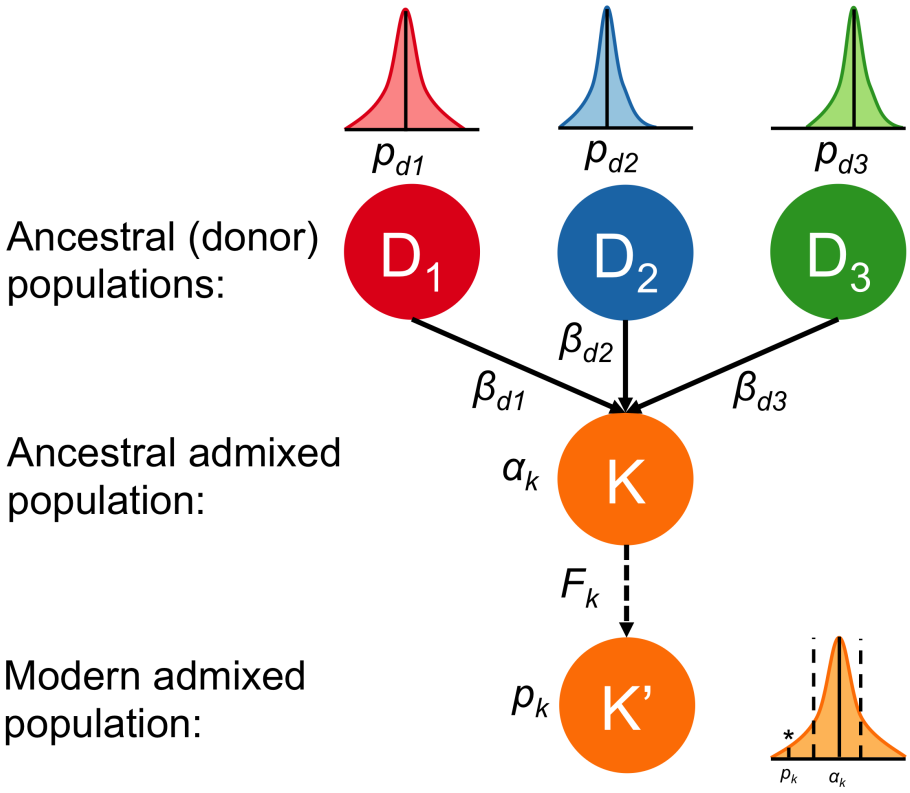

Figure 4.2: Schematic of the new model used to identify variants under selection post-admixture in admixed populations. The model assumes a set of ancestral ("donor") populations (in this case three donor populations: $D1$, $D2$ and $D3$) with allele frequencies $p_{d=1,2,3}$ at a particular SNP, contributing ancestry proportions $\beta_{d=1,2,3}$ (represented as solid lines) to an ancestral admixed population (K) with expected allele frequency α_k. The ancestral admixed population (K) then evolves under genetic drift (F_k; represented as a dashed line) to form the modern (sampled) admixed population (K') with allele frequency p_k. The allele frequencies of the donor populations are drawn from a Beta distribution to reflect uncertainty in ancestral allele frequencies and are therefore represented as distributions on the top of each donor population. Similarly, the observed allele frequency in the modern admixed population is modelled by a Beta-Binomial distribution with variance proportional to the expected allele frequency α_k and genetic drift F_k (represented as a solid and dashed lines in the distribution, respectively) (see Section 4.3.4). In this illustration the observed allele frequency in the modern admixed population p_k has experienced a larger than expected change (i.e. deviation) in allele frequency as expected from the admixture process and genetic drift, and is therefore assumed to have been selected (*) after the admixture event.

4.4.2 Admixture proportions and genetic drift estimates in five admixed Latin American populations

In order to detect selection post-admixture, the new model first estimates admixture proportions in a target admixed population given a set of ancestral donor populations. I first compared the admixture estimates produced by new model to two ancestry estimation softwares: ADMIXTURE (Alexander et al., 2009) and RFMix (Maples et al., 2013). Most modern admixed Latin Americans derive their ancestry from indigenous Native Americans, southern Europeans (mainly from the Iberian Peninsula) and West Africans (Wang et al., 2008; Moreno-Estrada et al., 2013, 2014; Ruiz-Linares et al., 2014; Homburger et al., 2015; Adhikari et al., 2017; Chacon-Duque et al., 2018). I therefore modelled each admixed Latin American population analysed here (Brazil, Chile, Colombia, Mexico and Peru) using modern proxies for these ancestral populations. Encouragingly, the proportions of Native American, European and African ancestry estimated by the new model were similar to those obtained by ADMIXTURE and RFMix (Table 4.1). Differences between the ancestry proportions across methods were low, with slightly closer ancestry estimates obtained by the new model and RFMix, likely due to the fact that these two methods used the whole set of SNPs whereas ADMIXTURE used a set of LD-pruned SNPs. The genetic drift parameters estimated by the new model were low and similar across all populations. Slightly lower genetic drift estimates were also observed in the Brazilian and Peruvian samples.

4.4.3 Candidate regions of selection post-admixture identified by the new beta-binomial model in admixed Latin Americans

The new beta-binomial model identified the strongest signals of selection post-admixture in the Peruvian sample at 10q22 (Figure 4.3 and Table 4.3). SNPs with posterior probability > 0.5 (i.e. those more likely of being under selection than under neutrality) encompass a region of circa 360Kb and are distributed across three genes: Hexokinase Domain Containing 1(*HKDC1*), Storkhead Box 1 (*STOX1*) and VPS26, Retromer Complex Component A (*VPS26A*). *HKDC1* is an hexokinase enzyme, highly conserved across vertebrates and expressed in many tissues (Irwin and Tan, 2008). Variants within *HKDC1* have been associated to 2 hour fasting plasma glucose levels in a sample of pregnant women (including Hispanics) (Hayes et al., 2013), and with gestational mellitus diabetes in a south Indian population (Kanthimathi et al., 2016). The Retromer Complex Component A (*VPS26A*) gene encodes a component of the retromer complex, a multimeric protein involved in transport of proteins from endosomes to the trans-Golgi network (Seaman et al., 1997, 2009), and is expressed in pancreatic, adipose tissues, amongst others (Kim et al., 2008). *VPS26A* has been associated to type 2 diabetes (T2D) in individuals of south Asian ancestry (Kooner et al., 2011). *STOX1* is a protein coding gene. Mutations within this gene have been associated to preeclampsia (van Dijk et al., 2005, 2010), a pathology of pregnancy characterized by high blood pressure and signs of damage to another organ system, that can be lethal for the mother and for the fetus (Sibai, 2005).

In the Brazilian, Chilean, Colombian and Mexican population, no genomic regions

Table 4.1: Ancestry proportions estimates based on ADMIXTURE, the new beta-binomial model and RFMix for five Latin American populations.

	ADMIXTURE		
Population	**Native American**	**European**	**African**
BRA	0.19	0.72	0.09
COL	0.28	0.64	0.08
CHL	0.45	0.54	0.02
MEX	0.55	0.42	0.03
PER	0.65	0.32	0.03

	Beta-binomial model		
Population	**Native American**	**European**	**African**
BRA	0.21	0.69	0.11
COL	0.31	0.59	0.10
CHL	0.46	0.50	0.05
MEX	0.55	0.40	0.06
PER	0.63	0.31	0.06

	RFMix		
Population	**Native American**	**European**	**African**
BRA	0.19	0.71	0.09
COL	0.30	0.61	0.09
CHL	0.45	0.52	0.03
MEX	0.57	0.40	0.03
PER	0.64	0.33	0.04

Abbreviations: BRA, Brazilian sample; COL, Colombian sample; CHL, Chilean sample; MEX, Mexican sample; PER, Peruvian sample.

Table 4.2: Genetic drift estimates based on the new beta-binomial model for five Latin American populations.

Population	**Genetic drift**
BRA	0.00065
COL	0.00097
CHL	0.00104
MEX	0.00105
PER	0.00089

Abbreviations: BRA, Brazilian sample; COL, Colombian sample; CHL, Chilean sample; MEX, Mexican sample; PER, Peruvian sample.

surpassed the posterior probability threshold of 0.5 (Figure 4.3). However, there was a strong signal of selection in the Mexican population at 6p22, that was also shared by all other Latin American populations. Interestingly, this genomic regions harbours the MHC that contains several genes which play an important role in the immune system (Hill, 1998, 2001; Frodsham and Hill, 2004).

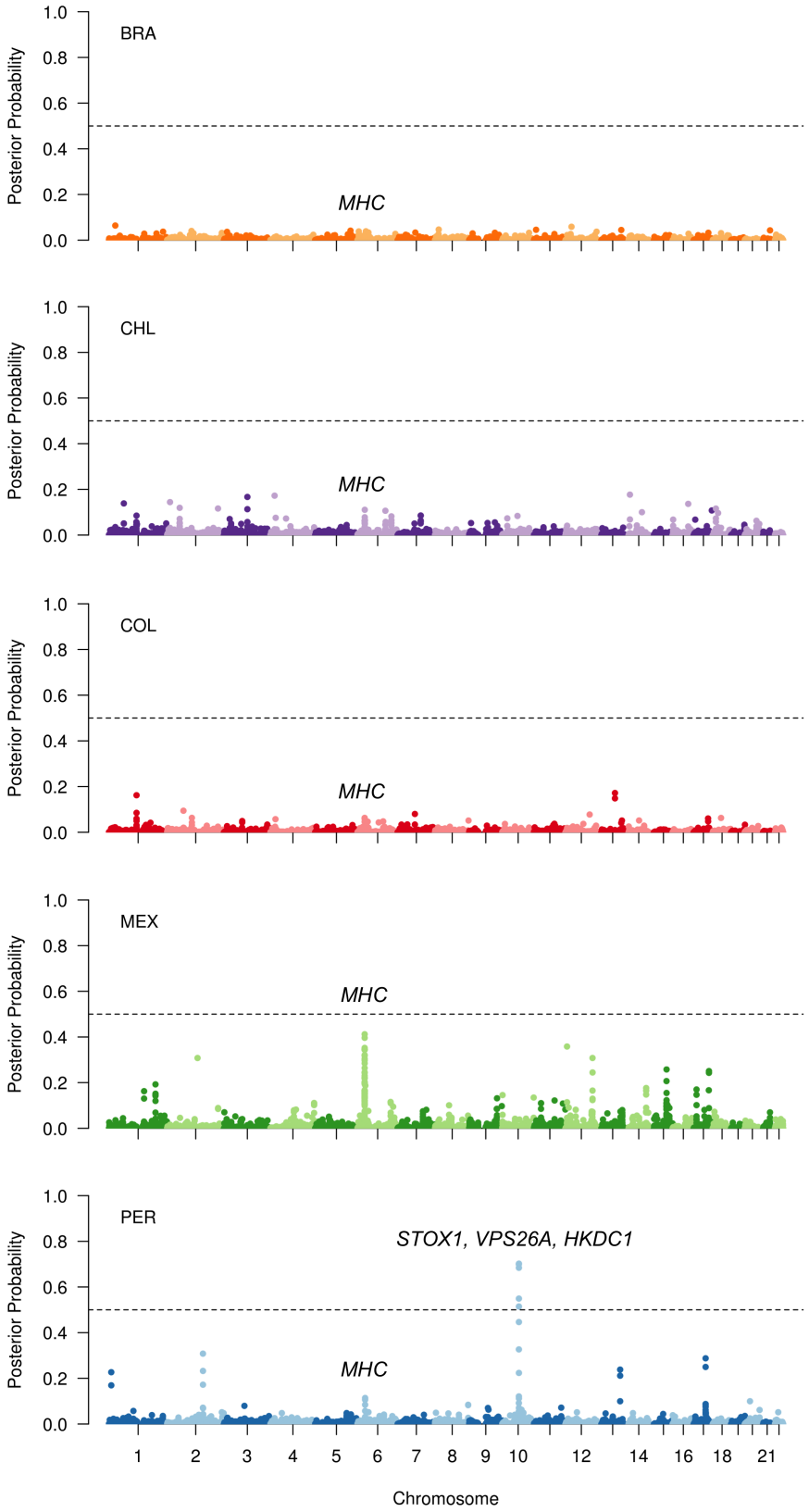

Figure 4.3: Genome-wide scan of selection post-admixture in admixed Latin American populations. Each dot represents the posterior probability estimated by the new beta-binomial model (per-SNP prior probability (s) was set equal to 0.001). The dashed black line represents the 0.5 Posterior Probability. Names of genes discussed in the text are shown. Abbreviations: BRA, Brazilian sample; CHL, Chilean sample; COL, Colombian sample; MEX, Mexican sample; PER, Peruvian sample.

Table 4.3: Top candidate SNPs of selection post-admixture based on the beta-binomial model in the Peruvian sample. SNPs with posterior probability > 0.5 are listed in the table. Genomic position, SNP, annotation, candidate gene and posterior probability, ordered by posterior probability.

Genomic coordinates (hg19)	**SNP**	**Annotation**	**Candidate gene**	**Posterior probability**
Chr10:70975916	rs5030938	5' UTR	*HKDC1*	0.702
Chr10:70613280	rs10998460	Intronic	*STOX1*	0.700
Chr10:70871598	rs2394505	Intergenic	-	0.684
Chr10:70890482	rs6480383	Intronic	*VPS26A*	0.549
Chr10:70906916	rs10823305	Intronic	*VPS26A*	0.514

4.4.4 Candidate regions of selection post-admixture based on local ancestry deviations

I also tested for post-admixture selection by searching for genomic regions with unusually high or low levels of Native America, European or African ancestry in the five Latin American samples. To measure the extent of the contribution of these ancestry across the genome, I used the discriminative modelling approach implemented in RFMix (Maples et al., 2013) to perform local ancestry inference. The distribution of the ancestry inference across each genomic region revealed a highly significant increase in African ancestry (and decrease in Native American ancestry) at 6p22 in the Chilean and Mexican samples (Figure 4.5 and 4.7). The highest amount of inferred African ancestry in this region in the Chilean and Mexican samples are 5.9% and 7.3%, corresponding to 6.9 and 7.2 standard deviation (SD) units from the genome-wide average which, under the normal approximation corresponds to a P-value = 6.3×10^{-13} and P-value = 3.5×10^{-12}, respectively. The increase in African ancestry in the Chilean and Mexican sample (4.03 SD units from the genome-wide average), extended over a large region of ∼12Mb (chr6:23,858,310 — 35,457,396) and ∼14Mb (chr6:22,150,358 — 36,342,181), respectively. Noticeably, this region includes the MHC locus, which plays an important role in the immune system (Hill, 1998, 2001; Frodsham and Hill, 2004). Interestingly, the other Latin American populations also showed some evidence of an increased African ancestry at the MHC region, albeit below the significance threshold set at 4.03 SD units from the genome-wide average (Figure 4.4, 4.6 and 4.8). Compared to the other regions, where fluctuations in ancestry deviation essentially showed a random pattern, the African ancestry in these populations was elevated throughout the MHC region. Additionally, other genomic regions in the Brazilian, Chilean and Colombian samples showed ancestry deviations slightly above the significance threshold (Figure 4.4, 4.5 and 4.6). However, these regions included mostly small genomic segments covered by small number of SNPs and are therefore more likely to represent local ancestry assignment errors than true selection signals.

4.5 Discussion

In this chapter I presented a novel statistical model to detect signals of selection post-admixture. The underlying model is based on the principle that allele frequencies in an admixed population can be modelled as a linear combination of the allele frequencies in the ancestral populations proportional to their admixing contributions, and that deviations from the expectation can be a product of selection after the admixture event. This model is most closely related to the work of Long (1991) and more recently to that of Mathieson et al. (2015), but provides significant improvements for (at least) four reasons. First, the beta-binomial model incorporates the estimation of a parameter that controls the variance in the predicted allele frequencies in the admixed population, given the set of ancestral populations used in the admixture model. This parameter can thus help to control for large deviations in allele frequency arising solely from genetic drift experienced in the admixed population. The role of genetic drift in affecting signatures of selection post-admixture has long been noted and explored in previous and more recent studies.

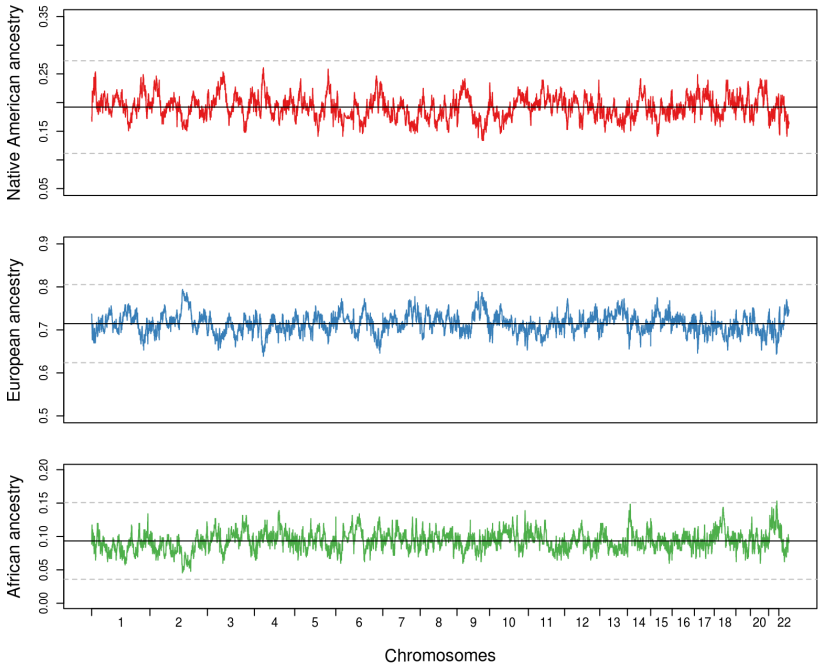

Figure 4.4: Local ancestry deviation in BRA admixed population. Proportion of Native American (upper panel), European (middle panel) and African (bottom panel) ancestry at each SNP along the genome. The black solid line indicates the genome-wide average proportion of each ancestry. Grey dashed line indicates ±4.03 SD from the mean. SD was calculated empirically over all SNPs.

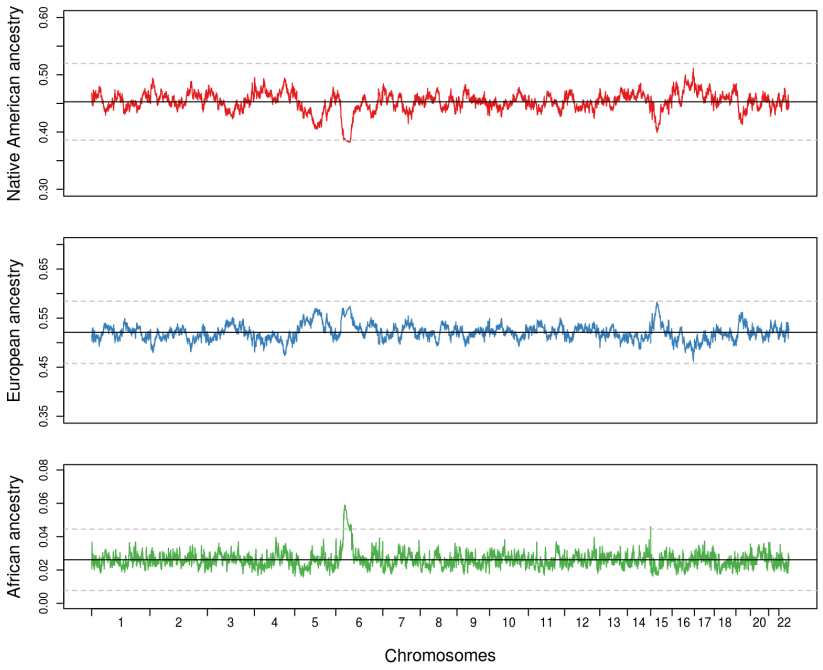

Figure 4.5: Local ancestry deviation in CHL admixed population. Proportion of Native American (upper panel), European (middle panel) and African (bottom panel) ancestry at each SNP along the genome. The black solid line indicates the genome-wide average proportion of each ancestry. Grey dashed line indicates ±4.03 SD from the mean. SD was calculated empirically over all SNPs.

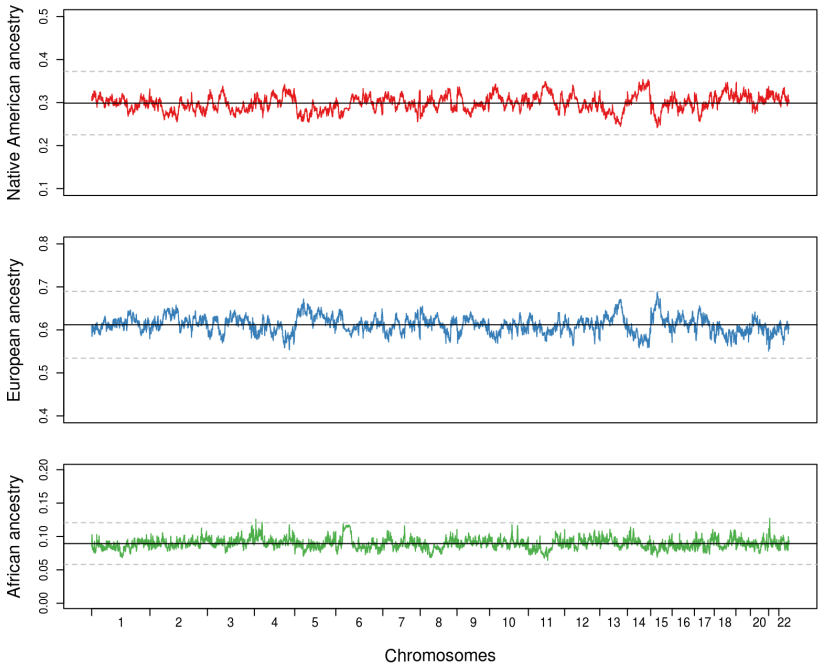

Figure 4.6: Local ancestry deviation in COL admixed population. Proportion of Native American (upper panel), European (middle panel) and African (bottom panel) ancestry at each SNP along the genome. The black solid line indicates the genome-wide average proportion of each ancestry. Grey dashed line indicates ±4.03 SD from the mean. SD was calculated empirically over all SNPs.

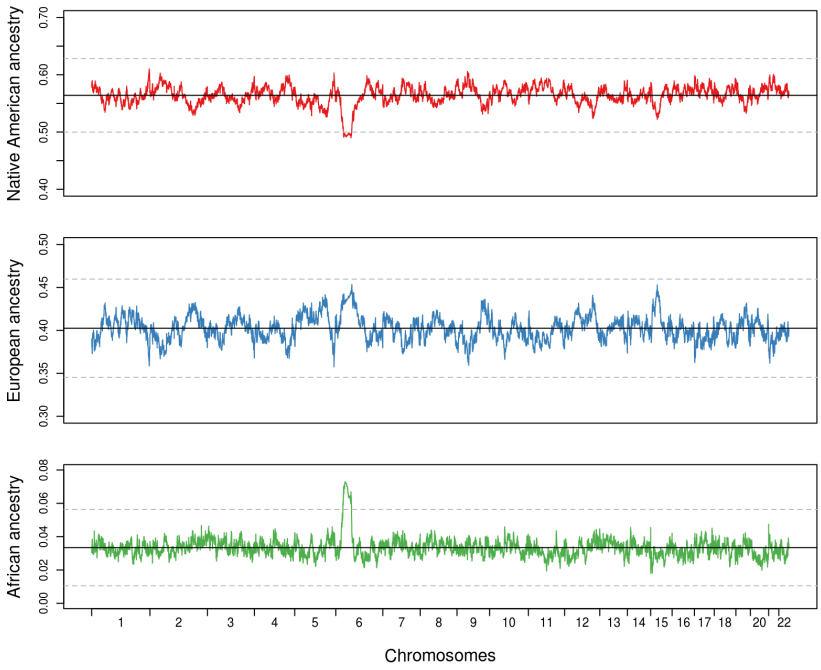

Figure 4.7: Local ancestry deviation in MEX admixed population. Proportion of Native American (upper panel), European (middle panel) and African (bottom panel) ancestry at each SNP along the genome. The black solid line indicates the genome-wide average proportion of each ancestry. Grey dashed line indicates ±4.03 SD from the mean. SD was calculated empirically over all SNPs.

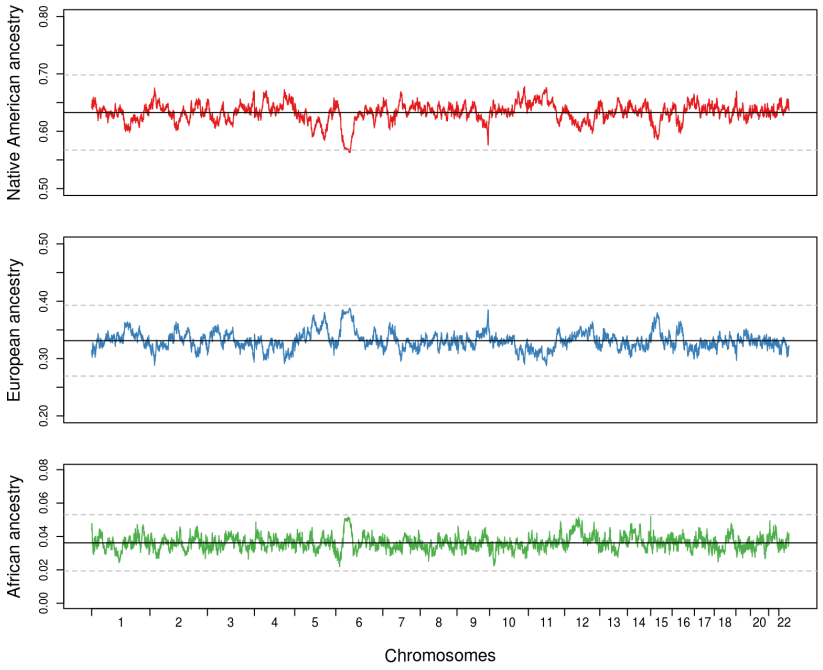

Figure 4.8: Local ancestry deviation in PER admixed population. Proportion of Native American (upper panel), European (middle panel) and African (bottom panel) ancestry at each SNP along the genome. The black solid line indicates the genome-wide average proportion of each ancestry. Grey dashed line indicates ±4.03 SD from the mean. SD was calculated empirically over all SNPs.

For example, based on simple admixture models, Long (1991) showed how genetic drift alone could produce large differences in ancestry proportions along the genome of an admixed population. Similarly, Bhatia et al. (2014) showed via simulations that genetic drift can significantly contribute to variance in average local ancestry, as a function of the effective population size of the admixed population. In a more recent study, Mathieson et al. (2015) identified loci under selection by modelling the observed allele frequencies in modern Europeans as a linear combination of the allele frequencies and admixture proportions of three ancestral populations that contributed most of the ancestry to present day Europeans. Importantly, the authors noted that the test statistic employed showed substantial inflation and suggested that this was most likely due to unmodelled ancestry or additional genetic drift not captured by their model. Second, the beta-binomial model can also account for the uncertainty in ancestral allele frequencies by drawing the ancestral allele frequency estimate from a random Beta distribution. This feature will be of importance when large sample sizes are not available to compute reliable allele frequencies in the ancestral populations, a feature that is common in ancient DNA samples. Third, the Bayesian framework implemented in the new beta-binomial model can also be extended to take into account the differential action of selection across the genome . For instance, a simple extension would be to incorporate different prior probabilities to different genomic contexts. Finally, the beta-binomial model can also simultaneously estimate the ancestry admixture proportions in an admixed population without having to resort to other softwares.

In this chapter I used the new beta-binomial model to test whether admixture contributed to genetic adaptation in Latin American samples from Brazil, Chile, Colombia, Mexico and Peru. Latin Americans represent an ideal setting in which to detect recent selection post-admixture as these populations were formed by the mixing of genetically differentiated ancestral populations that were suddenly exposed to new environmental challenges (Wang et al., 2008; Moreno-Estrada et al., 2013, 2014; Homburger et al., 2015; Chacon-Duque et al., 2018). The beta-binomial model detected strong signals of selection post-admixture only in the Peruvian population, with less evidence of selection post-admixture in the other populations (Figure 4.3). There are two possibilities that might explain this finding. The first possibility is that the recent admixture event in Latin Americans may not have provided enough time for selection to generate a detectable signal (Wang et al., 2008; Moreno-Estrada et al., 2013; Homburger et al., 2015; Chacon-Duque et al., 2018). The second possibility could be due to a combination of factors affecting the statistical power of the new beta-binomial model across Latin American populations. For instance, the different sample sizes, admixture proportions and the current allele frequencies of putatively selected variants are likely to contribute to the statistical power across populations. It is also likely that the fixed prior probability (set at 0.0001) or that the posterior probability threshold (0.5) used here are too stringent and may not be suitable to detect signals of selection in recently admixed populations such as Latin Americans. Future work will be needed to determine the statistical power of this posterior probability threshold under different prior posterior probabilities, different selection coefficients and different starting allele frequencies of the selected variant using simulations that closely resemble the admixture history of these populations.

The genomic region with the strongest selection signal post-admixture was detected in the Peruvian sample at 10q22 (Figure 4.3 and Table 4.3). The SNP with the overall strongest evidence of selection post-admixture was located in the 5' UTR of *HKDC1*, which has an important function in glucose metabolism (Irwin and Tan, 2008; Hayes et al., 2013; Guo et al., 2015). In addition, association studies have shown that variants within this gene might affect glucose levels in pregnant women (Hayes et al., 2013). From an evolutionary perspective, even very small effects on fertility and childhood survival can pose strong selective pressures (Williams, 1957). Interestingly, this same genomic region was also found to possess signals of selection in Andean Native Americans (see Section 3.4.5). It is worthwhile to note another potential caveat in the new beta-binomial model. The five Latin American populations analyzed here were modelled using the same set of Native Americans, Spanish Europeans and West Africans samples, and therefore might not represent the best proxy donor populations for each different Latin American population. For example, Chacon-Duque et al. (2018) showed that different local Native Americans groups contributed to the Native American ancestry of these five Latin American populations, and that the European component in the Brazilian sample is most similar to that of a Portuguese and West Spanish populations, while in Mexico, Colombia, Peru and Chile it is most similar to that of Central and southern Spanish population. It is possible that departures from the expected allele frequencies might be due to large differences between the allele frequencies of the proxies population used in the model and the true ancestral

populations. The fact that the high posterior probabilities at this region was not accompanied by a significant deviation of Native American, European or African ancestry, further suggest this signal might not represent a true selection post-admixture in the Peruvian population (Figure 4.9). Moreover, as modern admixed populations are being modelled using modern reference populations, it is important to note that the new beta-binomial model will not be able to distinguish whether selection happened in the admixed population or one (or more) of the reference populations. For instance, in this case it is more likely that the high posterior probability is detecting a selection signal in the ancestral Native American populations, in agreement with the results presented in Chapter 3 (see Section 3.4.5). Although the most appropriate use for this new model would be to employ ancient DNA samples from the Native Americans, Europeans and Africans populations that contributed to these Latin American populations, there are currently no ancient DNA samples from these specific populations.

As a complementary approach I also searched for signals of selection post-admixture by identifying genomic regions with unusually high or low levels of Native American, European or African ancestry. Applying this approach revealed two genomic regions showing strong ancestry deviations in the Mexican and Chilean populations, both of which were found at the MHC region (Figure 4.5 and 4.7). In the Brazilian, Colombian and Peruvian populations, a higher increase of African ancestry was also observed at MHC, albeit without reaching genome-wide significance (Figure 4.4, 4.6 and4.8). Notably, the new beta-binomial model also showed evidence of selection post-admixture at the MHC regions in all five Latin American populations, with highest signals being observed in the Mexican population (Figure 4.10). MHC harbours various genes with a known function in immune response, including resistance and susceptibility to a broad range of infectious diseases (Hill, 1998, 2001; Frodsham and Hill, 2004). It is well established that many diseases, such as smallpox, measles, mumps and influenza were introduced to the Americas after European contact, and that these caused high mortality that led to many disease outbreaks that lasted even up until colonial times (Duffy, 1972; Crosby, 1976; Dobyns, 1993; Cook and Lovell, 2001; Acuna-Soto et al., 2004; Fields, 2004). Under these adverse conditions, genes related to infectious disease resistance are therefore likely to have been under strong selection post-admixture in Latin Americans. Notably, two recent studies have also reported a strong signal of selection post-admixture at the MHC genomic in Mexicans (Zhou et al., 2016) and Colombians (Rishishwar et al., 2015). I took this opportunity to investigate why I detected the signal in the Mexican population analyzed here, but failed to detected the selection in the Colombian population.

In the first study, Zhou et al. (2016) identified a significant excess of African ancestry at the MHC in a sample of over 2,000 Mexicans. The authors used a sample of Maya and Pima Native Americans from the Human Genome Diversity Project and northern Europeans (CEU) and West Africans (YRI) from the 1KG as reference populations. To test the robustness of the choice of the reference populations used, the authors re-estimated local ancestry inferences using different European and African populations. In addition, they also adjusted their local ancestry assignment statistical model to infer local ancestry us-

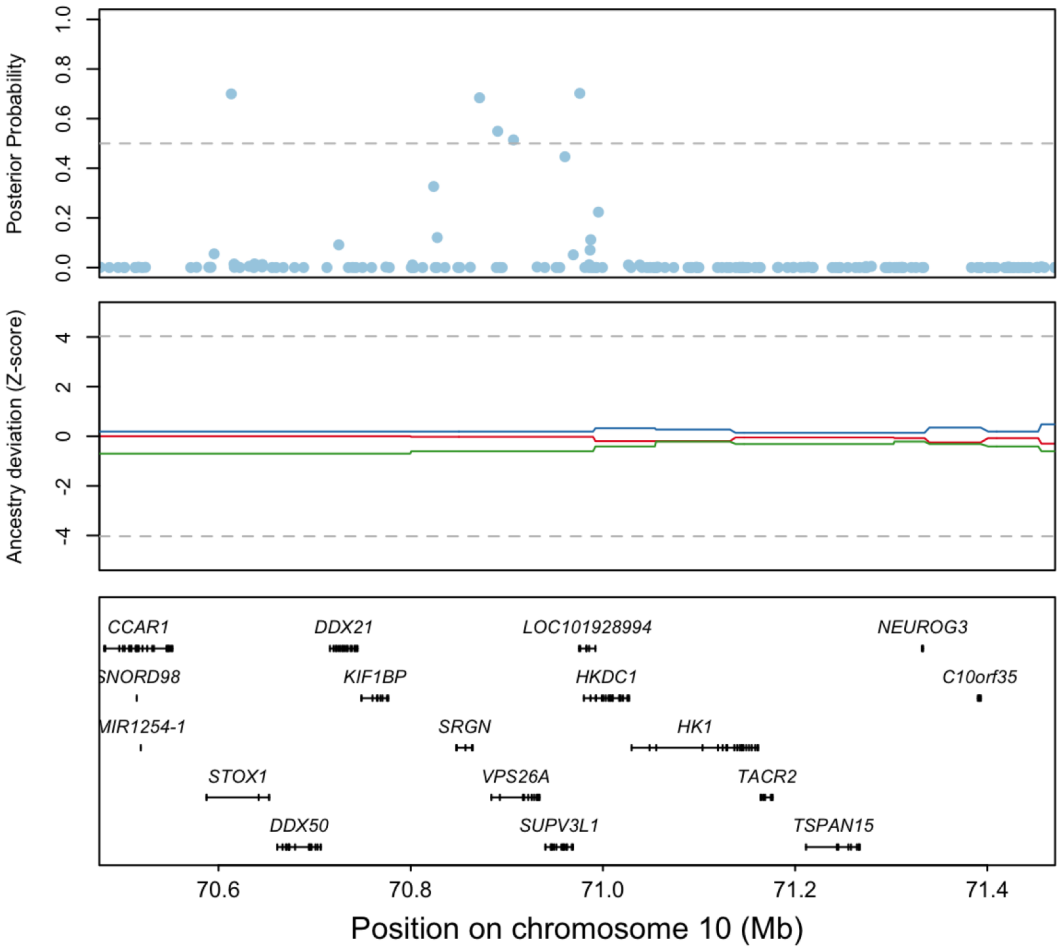

Figure 4.9: Signals of selection post-admixture at 10q22 in the Peruvian population. The Posterior Probabilities estimated by the new beta-binomial model (upper panel). Z-scores representing the excess or deficiency of Native American (red), European (blue) and African (green) ancestries from the genome-wide average (middle panel). UCSC RefSeq genes present at the genomic regions (bottom panel). The grey dashed line in the upper panel indicates the Posterior Probability threshold of 0.5. The dashed grey lines in the middle panel indicates the significant threshold of ±4.03 SD from the genome-wide average.

ing only European and African reference samples and learning Native American segments from the admixed Mexicans. Notably, the selection signal at the MHC region remained significant under these different settings. The results found in the Mexican sample analyzed here further support this finding. In the second study Rishishwar et al. (2015) also detected a significant increase in African ancestry at the MHC region in a sample of over 100 Colombians from the 1KG. However, unlike the study of Zhou et al. (2016), the authors employed East Asians (CHB), northern European (CEU) and West African (YRI) from the 1KG as reference populations. In addition, the authors did not test the robustness of their ancestry assignment using different reference populations. Given that the Colombian sample presented here included a sample of Colombians from the 1KG, contained a larger number of samples, and used a better proxy of reference populations, it is possible that the significant increase in African ancestry reported by Rishishwar et al. (2015) was due to the inaccurate Native American proxy population used. Overall, the results presented here and the findings from the study of Zhou et al. (2016) give further support to a strong selection at MHC in Mexicans since admixture. Nonetheless, it still will be necessary to assess the robustness of the significant African enrichment found in the Mexican and Chilean sample here, ideally using closely related populations for their ancestral donor populations. Finally, it would also be necessary to assess the the accuracy of the genotyping at MHC, given that the exceptionally high diversity and polymorphisms makes this locus hard to map (Szolek et al., 2014; Nariai et al., 2015; Nelson et al., 2015; Duke et al., 2016; Dilthey et al., 2016; Kawaguchi et al., 2017). For example, a simple check such as the one used by Tang et al. (2007) when assessing the reliability of the signal found at MHC in Puerto Rican could be also implemented here. The authors repeated their local ancestry assignment separately with even-numbered and odd-numbered SNPs to eliminate the possibility that the peaks were due to a few outliers (i.e. badly genotyped SNPs). In addition, bioinformatic tools specifically developed to assess the accuracy of the genotyping at MHC could also be employed (Jeanmougin et al., 2017).

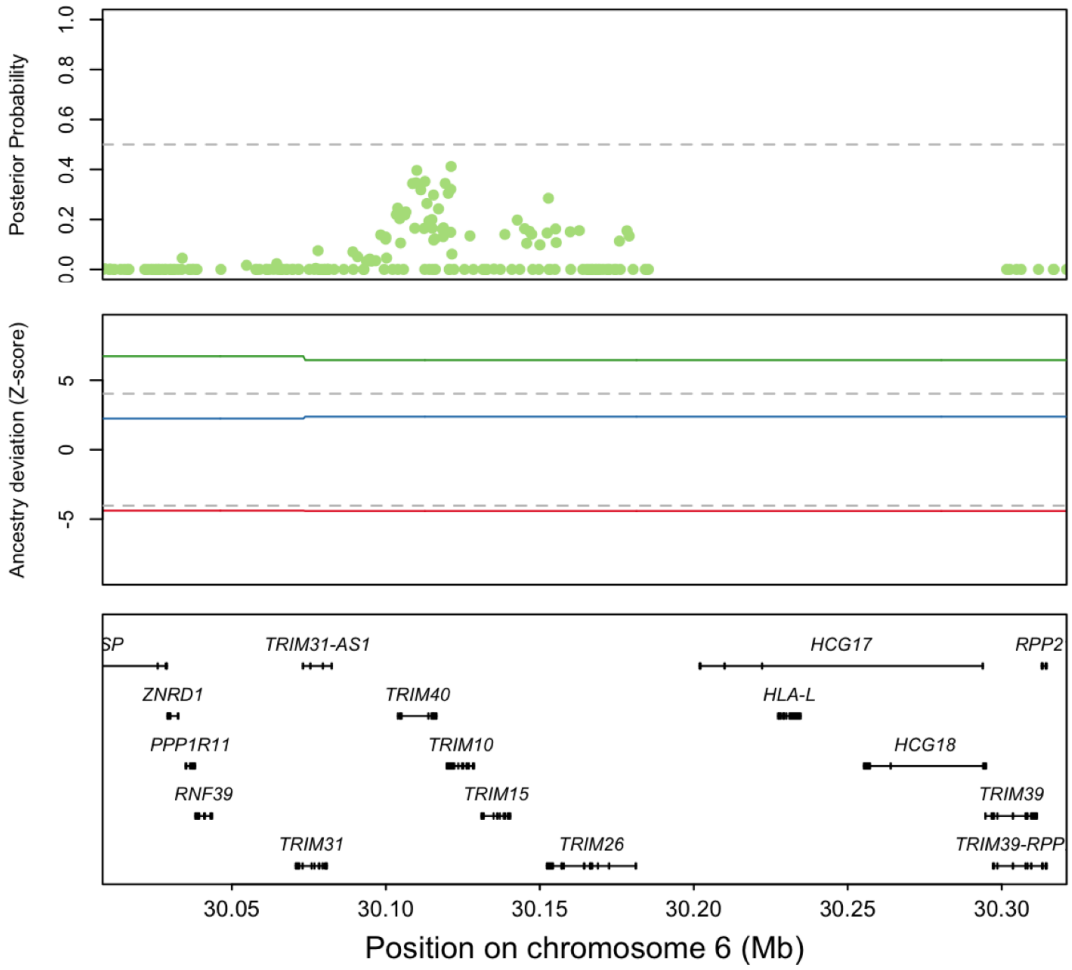

Figure 4.10: Signals of selection post-admixture at the MHC region in the Mexican population. The Posterior Probabilities estimated by the new beta-binomial model (upper panel). Z-scores representing the excess or deficiency of Native American (red), European (blue) and African (green) ancestries from the genome-wide average (middle panel). UCSC RefSeq genes present at the genomic regions (bottom panel). The grey dashed line in the upper panel indicates the Posterior Probability threshold of 0.5. The dashed grey lines in the middle panel indicates the significant threshold of ±4.03 SD from the genome-wide average.

4.6 Summary

In this chapter I presented a novel statistical model aimed to detect selection post-admixture in admixed populations. I apply this new method to five different Latin American populations and showed that there was a strong selective event in the Peruvian sample at a genomic region associated with glucose metabolism. I showed that it is also likely that this signal was driven by the use of inaccurate Native American reference populations and that the selection signal was likely driven by a selection event that occurred in the Native American population that contributed most of the Native American ancestry to the Peruvian sample. Testing for selection via local ancestry deviations in these populations showed a significant enrichment of African ancestry at the MHC in the Chilean and Mexican populations only. The genes at MHC involved in infectious disease resistance might have been selected due to diseases brought from the Old World after European contact.

Chapter 5

Genetic determinants of pigmentation in Latin Americans

5.1 Overview

In this chapter I perform a Genome Wide Association Study (GWAS) on over 6,000 Latin American individuals from the CANDELA sample for skin and eye pigmentation. A GWAS of hair pigmentation measured categorically has been previously reported in these same individuals (Adhikari et al., 2016a). Here, I add to this study by considering the results for all pigmentation traits examined in this sample. The majority of genetic studies of pigmentation have been performed in Northern European populations and thus, a large proportion of the genetic variation that contributes to pigmentation phenotypes in other worldwide populations is yet to be explored. Additionally, and in contrast to most genetic studies of pigmentation variation, the pigmentation phenotypes assessed here (with the exception of hair pigmentation) were evaluated using quantitive measurements, which enhance the power to identify genetic association. In this chapter I report novel associations for both skin and eye pigmentation. I provide evidence that the genomic regions associated with pigmentation phenotypes contain genes that represent important candidates for follow-up functional analyses. The work in this chapter also serves as an example of the necessity and the advantage of conducting GWAS in populations that have been underrepresented in human genetic studies, and how using quantitative pigmentation phenotypes enhances identification of genetic associations.

5.2 Background

Human pigmentation is mainly driven by the type, amount and distribution of melanin in the skin, eye and hair. In recent years, association studies have identified genes affecting common variation in pigmentation in various human populations (Section 1.5.3). The majority of these studies however, have been carried out in European populations (Section 1.7), and there are still substantial gaps in the understanding of the genetic architecture of pigmentation phenotypes in different worldwide groups (recently reviewed in Lasisi and Shriver (2018)). A better understanding of human skin pigmentation genetics is highly relevant for medical studies, for example, due to the shared role between pigmentation loci

and many type of skin cancers (Zhang et al., 2013a; Asgari et al., 2016; Chahal et al., 2016; Ransohoff et al., 2017), evolutionary biology, given the adaptive role of skin pigmentation variation across different latitudes exposed to varying amounts of solar radiation (Jablonski and Chaplin, 2000, 2017), and to forensic applications, such as pigmentation phenotype prediction based on DNA variants (Walsh et al., 2011). Recently, there have also been important advances in the developments of methods based on reflectance and imaging technologies to better characterize pigmentation phenotypes (Liu et al., 2010; Edwards et al., 2010; Andersen et al., 2013; Beleza et al., 2013a; Edwards et al., 2016; Norton et al., 2016; Lloyd-Jones et al., 2017; Rawofi et al., 2017; Wollstein et al., 2017). These studies have shown that using quantitative measurements, compared to using manually defined phenotype categories, can increase both power to detect genetic associations and prediction accuracy of pigmentation phenotypes.

5.2.1 Previous studies

Human pigmentation variation is a trait known to be strongly influenced by genetics (Clark et al., 1981; Rees and Harding, 2012). Candidate genes for pigmentation phenotypes were initially proposed based on model organisms (e.g. the Color Gene database: http://www.espcr.org/micemut/). Functional variants for human pigmentation variation have also been discovered through the study of rare syndromes involving pigmentation anomalies (Hamosh et al., 2005). Previous association studies, including candidate-gene and GWAS have successfully validated previously proposed pigmentation genes as well as discovered novel associations (Section 1.5.3). Established genes involved in normal variation in skin, hair and eye pigmentation in human populations include: *MC1R* (Melanocortin 1 receptor, OMIM: 155555), *OCA2* (Oculocutaneous Albinism II, OMIM: 611409), *HERC2* (HECT and RLD domain Containing E3 ubiquitin protein ligase 2, OMIM: 605837), *ASIP* (Agouti Signalling Protein, OMIM: 600201), *IRF4* (Interferon Regulatory Factor 4, OMIM: 601900), *TYR* (Tyrosinase, OMIM: 606933), *SLC24A4* (Solute Carrier family 24 member 4, OMIM: 609840), *SLC24A5* (Solute carrier family 24 member 5, OMIM: 609802), *KITLG* (KIT ligand, OMIM: 184745), and *TYRP-1* (Tyrosinase related Protein 1, OMIM: 115501).

The vast majority these human pigmentation genetic studies however, have been carried out in European populations, which limits the understanding of the underlying genetic architecture of this phenotype. A recent exception includes a study of skin pigmentation variation in ∼1,500 ethnically diverse Sub-Saharan Africans (Crawford et al., 2017). In this study, the SNPs with the second strongest association to skin pigmentation contained the Major Facilitator Superfamily Domain-containing protein 12 (*MFSD12*) gene. Two of the eight potentially causal SNPs (rs56203814 and rs10424065), where the derived allele was associated with darker pigmentation, are present only in African populations (or those with recent African descent) and are almost absent everywhere else. Additionally, coalescent based analysis estimated the Time to the Most Recent Common Ancestor (TMRCA) of SNP rs10424065 to be 612 kya (95% CI 515-736 kya), thus predating the 300 kya estimate for the origin of modern humans (Richter et al., 2017). This study thus exemplifies

the need for the inclusion of diverse human population in pigmentation genetic studies to further inform about the evolutionary history of pigmentation phenotypes.

Most genetic studies of human pigmentation have assessed these phenotypes categorically or even indirectly for example using questionnaires. Categorizing pigmentation variation in ordinal categories however, represents an oversimplification of the truly continuous nature of human pigmentation variation. It follows that using continuous variables that better capture pigmentation traits, can aid in the identification of novel genes. While many studies have quantified human skin pigmentation variation using continuous variables (most commonly the Melanin Index) (Candille et al., 2012; Abe et al., 2013; Beleza et al., 2013a; Eaton et al., 2015; Liu et al., 2015; Crawford et al., 2017b; Lloyd-Jones et al., 2017; Martin et al., 2017b; Rawofi et al., 2017), the vast majority of genetic studies of eye and hair pigmentation were based on categorical trait information. The first study to quantify continuous eye pigmentation variation was conducted in ∼6,000 northern Europeans (Liu et al., 2010). Liu et al. (2010) showed that eye color varied in more dimensions than the the blue, green and brown categories more commonly used and identified the *LYST* gene and the *DSCR9* gene as promising functional candidates. Similarly, the first GWAS of quantitative skin and eye pigmentation of non-European populations was recently conducted in a small sample (∼300) of East Asians (Rawofi et al., 2017). Although no novel variants were identified for skin pigmentation, one variant within the *ZNF804B* gene was identified as a potential candidate for eye pigmentation variation. These studies show how the use of continuous trait information of pigmentation phenotypes can increase the power to identify novel pigmentation loci.

5.3 Materials and methods

5.3.1 Study subjects

To study genetic determinants of pigmentation in Latin Americans I use a dataset of 6,357 unrelated volunteers from 5 countries (Brazil, Chile, Colombia, Mexico and Peru), part of the CANDELA Consortium sample (Ruiz-Linares et al., 2014; Section 1.8). A detailed description of the main features of the study sample is presented in Table C.1.

5.3.2 DNA genotyping and quality control

DNA samples from participants were genotyped on the Illumina HumanOmniExpress chip at 730,525 SNPs. PLINK v1.9 (Chang et al., 2015) was used to exclude SNPs and individuals with more than 5% missing data, SNPs with minor allele frequency less than 1%, related individuals, and those who failed the X-chromosome sex concordance check (sex estimated from X-chromosome heterozygosity not matching recorded sex information). After applying these filters, 669,462 SNPs and 6,357 individuals were retained for further analysis. Due to the admixed nature of the study sample (Figure C.1) there is an inflation

in Hardy-Weinberg P-values, and SNPs were therefore not excluded based on concordance with Hardy-Weinberg equilibrium.

5.3.3 Description of pigmentation phenotypes

A physical examination of each volunteer was carried out using the same protocol and instruments at all recruitment sites. Eye color was recorded in five ordinal categories (1-blue/grey, 2-honey, 3-green, 4-light brown, 5-dark brown/black). Hair color was recorded in four categories (1-red/reddish, 2-blond, 3-dark blond/light brown or 4-brown/black), as described in Adhikari et al. (2016). Individuals with red hair were excluded prior to the analyses, as it is a rare in the sample (frequency of 0.6%) and this phenotype is known to stem from rare variants in *MC1R* (Valverde et al., 1995). A quantitative measure of constitutive skin pigmentation (the Melanin Index, MI) was obtained using the DermaSpectrometer DSMEII reflectometer (Cortex Technology, Hadsund, Denmark). The MI was recorded from both inner arms and the mean of the two readings used in downstream analyses. The absolute difference of the two measurements was taken as the variability within an individual, and the median variability across all individuals was 1.03 MI units (Figure C.2). For comparison the total variability in the CANDELA sample was 20 to 65 MI units.

In addition to direct assessment of eye color into five ordinal categories, I obtained quantitative variables related to eye color from digital photographs of the volunteers. One of the two eyes was selected based on image quality. Photographs were landmarked manually via a graphical interface tool designed in MATLAB 3.2.5 by Dr. Kaustubh Adhikari. Ten landmarks were used to delimit and extract the visible part of the iris. Additional landmarks were placed to select the whitest part of the sclera. This white reference and the darkest part of the pupil were used to normalize the image, adjusting for variable color casts or illumination levels across images. An adaptive threshold was then used to remove highlights such as reflections on the iris. The resulting images were individually checked for the presence of errors during the digitization steps leading to their exclusion. In total 5,513 iris images were retained for extracting RGB (Red, Green, Blue) pixel color values.

The multivariate median of the RGB values across all pixels was calculated in order to obtain average RGB values for an iris. However, although the RGB color-space is convenient for digital imaging it is not necessarily the most appropriate in terms of human perception or biological relevance. Several other color spaces have therefore been considered in genetic association studies of pigmentation. In particular, the HCL and CIE Lab color spaces have the advantage over RGB of being perception-based (Liu et al., 2010, 2015; Norton et al., 2016; Edwards et al., 2010, 2016; Rawofi et al., 2017). Furthermore, it has been shown that melanosome density and the skin MI are strongly correlated with brightness (L) (Takiwaki et al., 1994). The main difference between the HCL and CIE Lab color spaces is that HCL, being directly derived from RGB, represents the three primary colors (red, green, blue) in opposing corners, while the CIE Lab represents four colors in

different corners (red against green and blue against yellow). Since the HCL values in the CANDELA dataset occupy mainly the opposing red-orange and cyan-blue color hues (Figure 5.2), for this study I considered the HCL color space as more informative than the nearly equivalent CIE Lab color space. H is a circular variable representing color Hue (tone) ranging from 0° to 360°, with red at 0°, green at 120°, blue at 240°. C (Chroma or saturation) ranges from 0 (no color) to 1 (fully saturated color). L (Lightness or brightness) ranges from 0 (black) to 1 (white). HCL values lie approximately on a 2-dimensional plane passing through the vertical central axis (Figure 5.2) at an angle of approximate 20° (obtained from the circular median of H). H values were therefore standardized by subtracting 20°. Furthermore, since H is a circular variable, it was converted to cos(H) prior to its use in the analyses performed here. Cos(H) ranged from -1 (blue-grey eyes) to +1 (olive/brown/dark brown eyes). As the distribution of HCL values was nearly planar, sin(H) showed comparatively little variation (equivalent to taking a projection onto the plane) and was ignored.

5.3.4 Phasing and imputation

The chip genotype data (compromising 669,462 SNPs after QC; Section 5.3.2) was phased using SHAPEIT2 (Delaneau et al., 2013) with default parameters. IMPUTE2 (Howie et al., 2009) was then used to impute genotypes at untyped SNPs using variant positions from the 1000 Genomes Project data (1000 Genomes Project Consortium et al., 2015). The 1000 Genomes Project reference data set included haplotype information for 1,092 individuals across the world at 36,820,992 variant positions. Positions that are monomorphic in 1000 Genomes Latin American samples were excluded, leading to 11,025,002 SNPs being imputed in the dataset. Of these, 48,695 had imputation quality scores <0.4 and were excluded. Median "info" score (imputation certainty score) provided by IMPUTE2 for the remaining imputed SNPs was 0.986. The IMPUTE2 genotype probabilities at each locus were converted into most probable genotypes using PLINK (Chang et al., 2015) (at the default setting of <0.1 uncertainty). Imputed SNPs with >5% uncalled genotypes or minor allele frequency < 1% were excluded. IMPUTE2 provides a "concordance" metric for chip genotyped SNPs, obtained by masking the SNP genotypes and imputing it using nearby chip SNPs. Genotyped SNPs with a low concordance value (<0.7) or a large gap between info and concordance values, suggested poor genotyping quality, and were also removed. The median concordance values of the remaining chip SNPs was 0.994. After these quality control filters, the final imputed dataset contained genotypes for 9,143,600 SNPs.

5.3.5 Narrow sense heritability

Narrow-sense heritability (defined as the additive phenotypic variance explained by a Genetic Relatedness Matrix, GRM, computed from the SNP data) was estimated using the software GCTA (Yang et al., 2011a). GCTA fits an additive linear model with a random effect term whose variance is given by the GRM (with age and sex as covariates). The

GRM was obtained using the LDAK software (Speed et al., 2012), which accounts for LD between SNPs.

5.3.6 ADMIXTURE analysis

Approximate proportions of ancestry for each individual were estimated from a set of LD pruned dataset of 160,858 SNPs via supervised ADMIXTURE (Alexander et al., 2009). Reference populations from African (YRI, Yoruba in Ibadan, Nigeria) European (CEU, Utah Residents with Northern and Western European Ancestry), East Asian (CHB, Han Chinese in Beijing, China) were chosen from the 1000 Genomes Project together with selected Native Americans populations from Ruiz-Linares et al. (2014) and Chacon-Duque et al. (2018). ADMIXTURE was then run with $K = 3$ to $K = 4$.

5.3.7 Association analysis

PLINK 1.9 (Chang et al., 2015) was used to perform the primary GWAS for each phenotype using multiple linear regression with an additive genetic model incorporating age, sex, and 6 genetic Principal Components (PCs) as covariates. PCs were obtained from an LD-pruned dataset of 160,858 SNPs. Individual outliers (including individuals with >20% African or >5% East Asian ancestry, as estimated by ADMIXTURE (Figure C.1)) were removed and PCs recalculated after the removal of these individuals. The number of PCs to be included in the regression was determined by inspecting the proportion of variance explained and by checking scree and PC scatter plots. Based on the proportion of variance explained a total of 6 genetic PCs were used (Figure C.3) as in previous GWAS conducted in the CANDELA sample (Adhikari et al., 2015, 2016a,b). The quantile-quantile (QQ) plots for all association tests showed no evidence of residual population stratification (Figure C.4) , except for skin pigmentation (λ=1.11) (Table C.2). This result however, is expected as highly polygenic traits can show some genomic inflation even in the absence of population structure (Yang et al., 2011b; Bulik-Sullivan et al., 2015b). Polygenicity of the traits were measured using the tail strength (TS) statistic (Taylor and Tibshirani, 2006) (Table C.2), which measures the overall strength of univariate (single-SNP) associations in a chip array dataset. In a GWAS with n SNPs, if the ordered P-values are $p_1 \leq p_2 \leq ... \leq p_n$, the statistic is:

$$TS_{(p_1,...,p_n)} = \frac{1}{n}\sum_{k=1}^{n}(1 - p_k\frac{n+1}{k}). \tag{5.1}$$

Under the null hypothesis of no association between the trait and all SNPs, TS should be equal to zero. A positive value of TS indicates the overall extent of association in the entire dataset and is interpreted as polygenicity, with higher values of TS indicating greater polygenicity. The asymptotic variance of TS can be approximated by $1/n^*$, where n^* is the effective number of independent SNPs. As LD prunning on our dataset yielded 160,858 SNPs, the SD can be estimated as $1/\sqrt{160,858} = 0.0025$, and a confidence interval would be $TS \pm 3 \times SD = TS \pm 0.0075$. The estimated TS statistics obtained in the GWAS

analysis are shown in Table C.2. Association analysis were performed on the imputed dataset using the best-guess imputed genotypes in PLINK (Chang et al., 2015). Upon the initial analysis a set of well-known pigmentation loci were strongly associated with most of the traits: rs16891982 (*SLC45A2*), rs12203592 (*IRF4*), rs10809826 (*TYRP1*), rs1800404 (*OCA2*), rs12913832 (*HERC2*), rs1426654 (*SLC24A5*). Therefore, to increase power, subsequent GWAS analyses were performed conditioned on these SNPs (Yang et al., 2012). P-values reported for other SNPs are taken from the conditioned analysis.

5.3.8 Meta-analysis

A meta-analysis was carried out on the novel index SNPs identified in the primary analyses (Table 5.4) by testing for associations separately in each country sample. Forest plots were produced with MATLAB 3.2.5 combining all regression coefficients and standard errors. Histograms of the traits within each country were compared to the forest plots to examine how trait variability across countries relates to the association signals.

5.3.9 SNP x SNP interaction of genome-wide associated SNPs

Tests of interaction of pairwise index SNPs (i.e. those with genome-wide significant association P-values $< 5 \times 10^{-8}$) were carried out for each phenotype by performing multiple linear regressions of the phenotype value on covariates, index SNPs, and pairwise interactions between index SNPs.

5.4 Results

5.4.1 Distributing of pigmentation phenotypes in Latin Americans

Information on skin, hair and eye (iris) pigmentation (Figure 5.1 A-C) was obtained for 6,357 Latin American individuals. Skin pigmentation was assessed with the Melanin Index (MI, a quantitative variable), measured by reflectometry and showed extensive variation, the MI ranging from 20 to 65 (mean = 34.98 and SD = 5.34). The lightest mean pigmentation was observed in Brazil and the darkest mean pigmentation in Mexico and Peru (5.1A). Hair color was classified into 3 ordinal categories (1-blond, 2-dark blond/light brown, 3-brown/black). The most prevalent hair colors were black and brown, representing $\sim$ 80% of the sample (Figure 5.1B). These were also the most prevalent categories across countries, except in Brazil where $\sim$ 50% of individuals had dark-blond/light-brown or blond hair. Eye color was classified into 5 ordinal categories (1-blue/grey, 2-honey, 3-green, 4-light brown, 5-dark brown/black). The most common categories were dark brown/black and light brown, compromising $\sim$ 85% of the sample (Figure 5.1C). The lighter eye color categories (blue/grey and green) were more common in Brazil ($\sim$ 40%) than in other sampled countries ($\leq$ 10%).

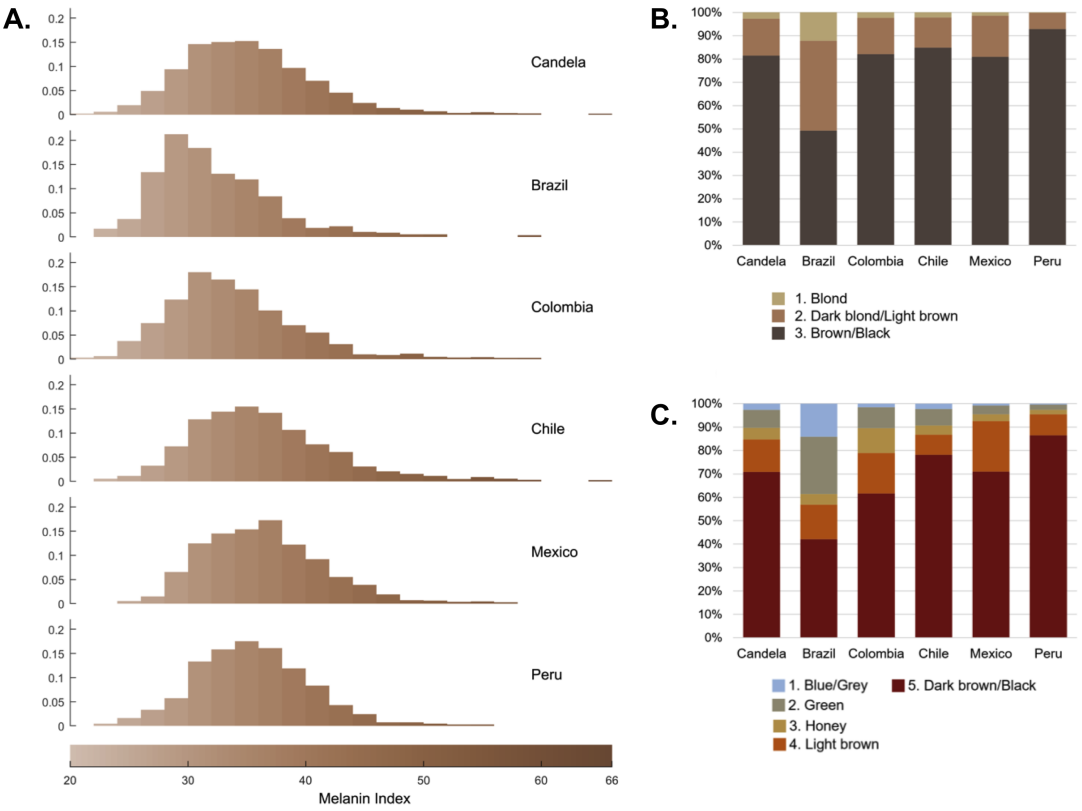

Figure 5.1: Distribution of skin, hair and categorical eye pigmentation phenotypes in the CANDELA sample. A) Frequency distribution of skin Melanin Index (MI). Histograms are shown for the full CANDELA sample and for each country sample separately. To facilitate relating MI values to skin color the MI values (Y-axis) were converted to approximate RGB values. B) Stacked bar plots showing the frequency (percent) of the three hair color categories. C) Stacked bar plots showing the frequency (percent) of eye color categories. Modified from Adhikari & Mendoza-Revilla et al. (2018).

In addition to eye color measured by ordinal categories, I obtained quantitative measurements of eye color from the analysis of digital photographs using the HCL color space (Hue, Chroma, Lightness) (Figure 5.2A-D). Hue (H) measures variation in color tone, whereas Chroma (C) and Lightness (L) measure saturation and brightness, respectively (Figure 5.2A). In contrast to the ordinal categories of eye color, these quantitative color variables capture variation not only in the blue/grey to brown spectrum (mainly captured by L), but also variation within the brown spectrum (mainly captured by C) (Figure 5.2C and D). While individuals with the highest L values exhibited mainly blue/grey eyes, individuals with the highest C values exhibited eye colors with the lightest shades of brown (i.e. light brown or honey, Figure 5.2C). As H is a circular variable it was standardized and converted to cos(H) before testing for association (see Methods Section 5.3.3).

5.5 Correlation between pigmentation phenotypes and covariables.

All pigmentation traits examined are significantly (P-values <0.001) and positively correlated (Table 5.1). The strongest correlation was observed between hair and categorical eye color (r=0.50), while there is lower correlation of these two traits with skin pigmentation (r=0.30 and r=0.31, respectively). Categorical eye color was strongly correlated with the L digital eye color variable (r=-0.78), but only moderately and minimally correlated with cos(H) and C (r of 0.40 and -0.08, respectively), highlighting the considerable amount of variation that is not captured by the eye color categories. Darker pigmentation of hair, skin and eyes was also significantly and negatively (P-values <0.001) correlated with the genetic estimates of European ancestry and significantly and positively (P-values <0.001) correlated with Native American ancestry (Table 5.2). Pigmentation phenotypes were also strongly associated with the first Principal Component (P-values <0.001). Darker pigmentation of skin and eye was also negatively correlated with age (Table 5.2). Darker skin pigmentation was significantly and positively associated with being male (P-value=0.03) and lighter hair and eye pigmentation were significantly and positively associated with being female (P-values <0.001 (Table 5.2).

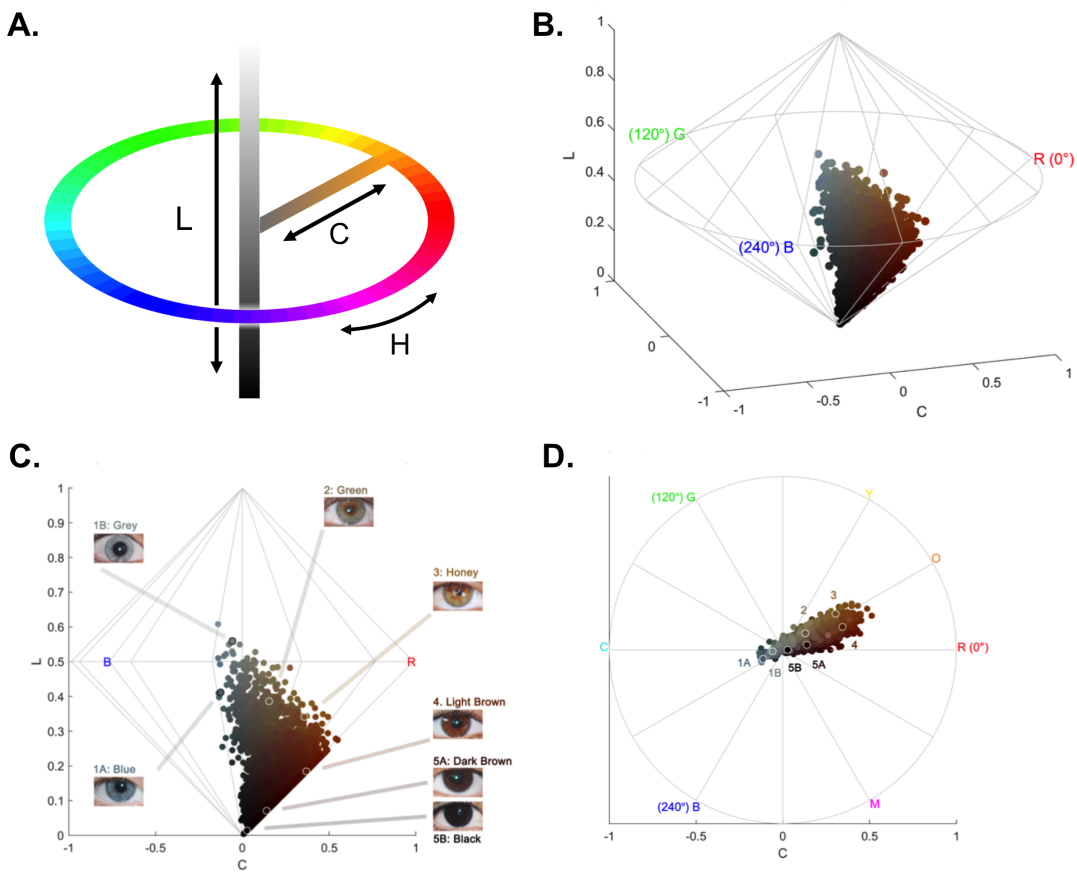

Figure 5.2: Quantitative eye pigmentation phenotypes examined in the CANDELA sample. A) The full range of the HCL color space, showing how the three color components vary in the space. B) 3-dimensional distribution of quantitatively assessed iris colors in the bicone HCL (Hue, Chroma, Lightness) color space. Each dot corresponds to a CANDELA individual and its color represents the average iris color for that person. The color space has a polar coordinate system, where the vertical axis represents L (Lightness/brightness, from dark=0 to light=1), the horizontal distance from the central axis represents C (Chroma/saturation, from desaturated=0 to fully saturated=1), and H (Hue/tone) represents the angle when a vertical plane is rotated along the central axis (the three primary colors red (R), green (G) and blue (B) being situated at angles of 0°, 120° and 240° respectively). C) Side view of the bicone in D showing how the L (Lightness/brightness) and C (Chroma/saturation) of eye colors vary among CANDELA volunteers. The position of the dots corresponding to the average eye colors of the sample images are indicated. D) Top view of the bicone in D showing how H (hue) varies among the eye colors of CANDELA volunteers. The position of the dots, corresponding to the average color of the sample images, are highlighted by white circles. Eye colors cluster around a vertical plane at a 20° angle. In addition to the primary RGB colors, the secondary colors orange (O), yellow (Y), cyan (C) and magenta (M) are shown at their corresponding H angles. Modified from Adhikari & Mendoza-Revilla et al. (2018).

Table 5.1: Correlation between pigmentation phenotypes. Correlation values are presented in the lower left triangle, while corresponding P-values are presented in the upper triangle. The sample size for skin, hair and categorical eye color phenotypes is 6,357 whereas for quantitative eye color is 5,513 individuals.

Trait	**Skin color (MI)**	**Hair color (categorical)**	**Eye color (categorical)**	**L (Brightness)**	**C (Saturation)**	**cos(H) (Hue)**
Skin color (MI)	-	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
Hair color (categorical)	0.30	-	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
Eye color (categorical)	0.31	0.50	-	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
L (Brightness)	-0.35	-0.46	-0.78	-	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
C (Saturation)	-0.20	-0.05	-0.08	0.34	-	$< 2.2 \times 10^{-16}$
cos(H) (Hue)	0.10	0.24	0.40	-0.39	0.23	-

Table 5.2: Correlation between pigmentation phenotypes and covariables. Native American, European and African continental ancestry estimates were obtained using ADMIXTURE. Sex was coded as female = 0 and male = 1.

			Ancestry			**Genetic PCs**					
Trait	**Age**	**Sex**	**European**	**Native American**	**African**	**PC1**	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**
Skin color (MI)	-0.05	0.03	-0.47	0.4	0.19	-0.46	0.13	-0.04	0.15	0.01	-0.03
Hair color (categorical)	-0.01	-0.1	-0.38	0.34	0.08	-0.35	0.21	-0.1	0.12	-0.11	-0.01
Eye color (categorical)	-0.08	0	-0.43	0.39	0.07	-0.39	0.2	-0.08	0.12	-0.06	0
L (Brightness)	0.14	-0.07	0.48	-0.43	-0.13	0.44	-0.29	0.17	0.04	-0.02	-0.01
C (Saturation)	0.07	-0.05	0.24	-0.23	0.01	0.24	-0.03	-0.11	0.08	-0.22	0.01
cos(H) (Hue)	-0.06	0	-0.2	0.18	0.07	-0.18	0.18	-0.05	0.06	-0.14	0.03

	Age	**Sex**	**European**	**Native American**	**African**	**PC1**	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**
Skin color (MI)	0	0.03	0	0	0	0	0	0	0	0.32	0.01
Hair color (categorical)	0.36	0	0	0	0	0	0	0	0	0	0.46
Eye color (categorical)	0	0.85	0	0	0	0	0	0	0	0	0.98
L (Brightness)	0	0	0	0	0	0	0	0	0.01	0.23	0.51
C (Saturation)	0	0	0	0	0.59	0	0.01	0	0	0	0.45
cos(H) (Hue)	0	0.79	0	0	0	0	0	0	0	0	0.05

5.5.1 Heritability of pigmentation phenotypes in Latin Americans

Based on a kinship matrix obtained from the SNP chip data (Speed, 2011), I estimated a narrow-sense heritability for skin pigmentation of 0.85 (S.E. 0.05) and of 1 (S.E. 0.05) for both hair and eye color. Similarly, quantitative eye color variables showed high heritability estimates (between 0.79 and 1.00, S.E 0.06) (Table 5.3). Very high heritabilities for pigmentation traits have also been estimated from family data (Bräuer and Chopra, 1978; Byard and Lees, 1981). Interestingly, lower narrow-sense heritability estimates for skin pigmentation have been estimated for European populations, possibly due to lower within continental skin pigmentation variation due to strong selective pressure (Zaidi et al., 2017).

Table 5.3: Heritability of pigmentation phenotypes

Trait	**Heritability**	**S.E.**	**P-value**
Skin color (MI)	0.85	0.05	$< 2.2 \times 10^{-16}$
Hair color (categorical)	1	0.05	$< 2.2 \times 10^{-16}$
Eye color (categorical)	1	0.06	$< 2.2 \times 10^{-16}$
L (Brightness)	1	0.06	$< 2.2 \times 10^{-16}$
C (Saturation)	0.79	0.06	$< 2.2 \times 10^{-16}$
cosH (Hue)	0.84	0.06	$< 2.2 \times 10^{-16}$

5.5.2 Genomic regions showing signals of association

Genome-wide significant association (P-values $< 5 \times 10^{-8}$) were found for SNPs in twelve genomic regions (Table 5.4 and Figure 5.3). Skin pigmentation showed association with SNPs in eight of these, of which: (i) five have been robustly replicated in previous studies of Europeans or East Asians (Lamason et al., 2005; Soejima and Koda, 2007; Sulem et al., 2007, 2008; Cook et al., 2009; Stokowski et al., 2007; Han et al., 2008; Graf et al., 2005; Eriksson et al., 2010; Zhang et al., 2013a); (ii) one (19p13) has recently been associated with skin pigmentation in Africans (Crawford et al., 2017), but to different SNPs than seen here; and (iii) one (10q26) has not been previously reported. SNPs in four of the skin-pigmentation regions were also found to be significantly associated with eye and hair color (5p13, 6p25, 15q13 and 15q21; Table 5.4). In addition to these, eye pigmentation shows association with SNPs in five other regions (1q32, 9p23, 20q11, 20q13 and 22q12), of which three (1q32, 20q13 and 22q12) have not been reported previously. The genomic regions associated with categorical eye color showed stronger association with the quantitative eye color variables extracted from the individual photographs and all the novel eye-color associations were genome-wide significant only for the quantitative variables (Table 5.4). These observations are consistent with greater statistical power for association testing of quantitative color variables, compared with categorical variables (e.g. see Wollstein et al. (2017)).

Table 5.4: Genome-wide associated SNPs with pigmentation phenotypes in the CANDELA sample. P-values for the associations are presented in a $-log_{10}$ scale. The five novel pigmentation-associated SNPs are shown in bold. Abbreviations: MI, Melanin Index; L, lightness; C, chroma; H, hue.

				Skin	Hair	Eye			
Region	**SNP**	**Gene**	**Annotation**	**MI**	**Categorical**	**Categorical**	**L**	**C**	**cos(H)**
1q32	rs3795556	***DSTYK***	3' UTR	0.8	0.1	0.1	7.7	8.9	0.6
5p13	rs16891982[a,b]	*SLC45A2*	F374L	116.9	65.2	14.9	16.4	7.8	4.3
6p25	rs12203592[a]	*IRF4*	Intronic	9.5	12.7	11.9	13.5	2.8	1.1
9p23	rs10809826[a,b]	*TYRP1*	Intergenic	2.7	1.2	10	15.3	7.7	2.1
10q26	rs11198112	**EMX2**	Intergenic	10.1	0.2	0.4	0.1	0.2	0.2
11q14	rs1042602	*TYR*	S192Y	13	8.6	0.2	0.5	1.7	0
15q13	rs12913832[a]	*HERC2*	Intronic	17	104.1	200	200	6.3	91.9
15q21	rs1426654[a]	*SLC24A5*	T111A	129.8	18	26	49.1	44.2	0.2
16q24	rs885479	*MC1R*	R163Q	8.4	1.3	0.1	0.1	0.1	0.1
19p13	rs2240751	***MFSD12***	Y182H	12.4	0	0.5	0	0.8	0
20q13	rs17422688	***WFDC5***	H97Y	0.1	0.3	0	0.7	0	8.5
22q12	rs5756492	***MPST***	Intronic	2.6	0	1.1	2	7.3	0.7

[a]P-values shown for these SNPs are unconditioned, while P-values for the other index SNPs are conditioned. Results from conditioned and unconditioned analyses are consistent and confirm an increase in power upon conditioning (see Methods). Of the five novel associations shown here (bold), three (rs11198112, rs2240751 and rs17422688) are also genome-wide significant in the unconditioned analyses and two (rs3795556 and rs5756492) are just below the threshold for genome-wide significance in the unconditioned analyses (Tables C.3 and C.4).
[b]These markers were obtained through imputation. Their imputation quality "info" metric was ≥ 0.975, the median value being 0.993. The other markers were obtained from chip genotyping, and their "concordance" metric was > 0.9, the median value being 0.981.

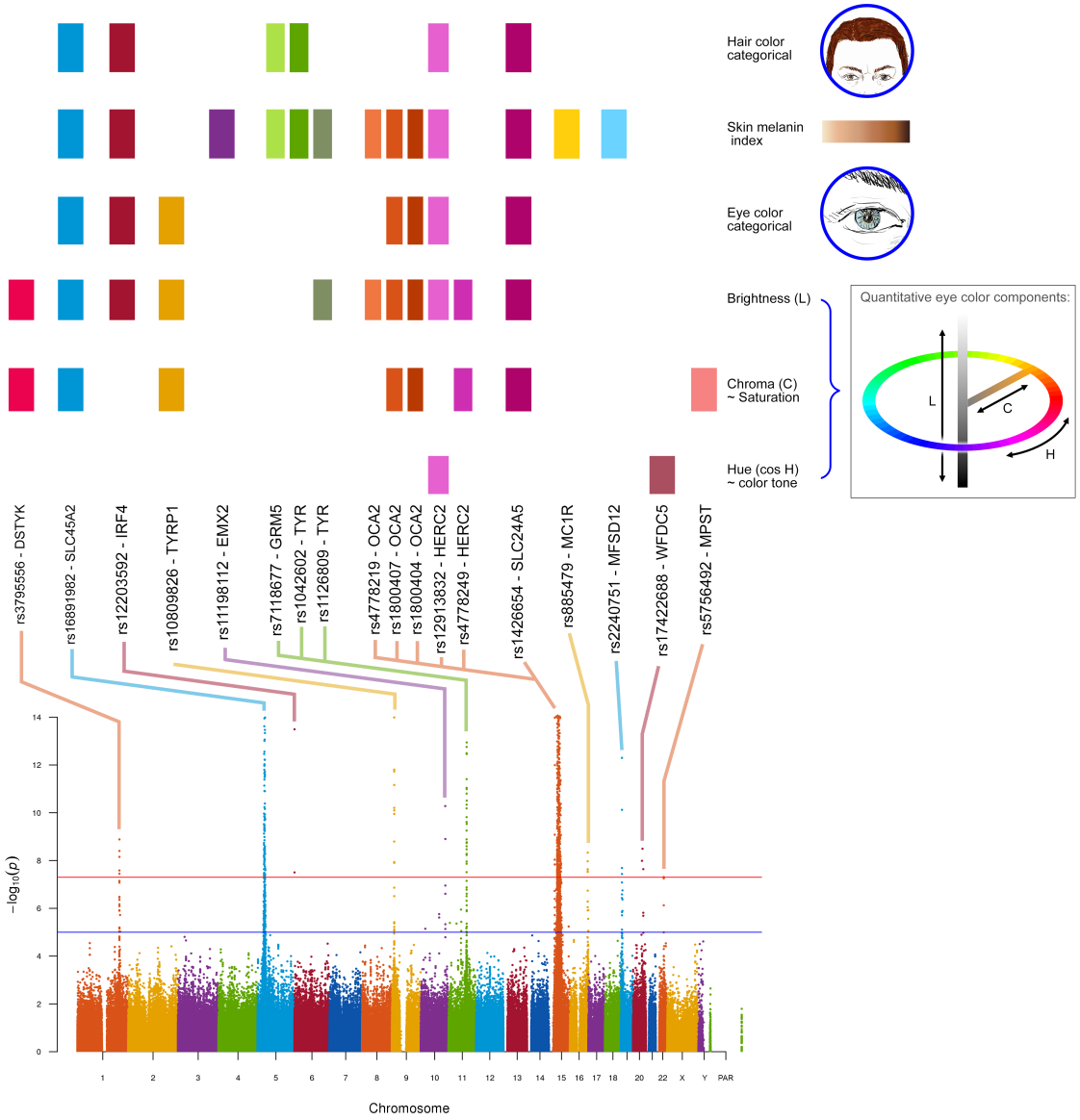

Figure 5.3: Summary of GWAS findings. Results are presented for 6 pigmentation traits: skin melanin index (MI, quantitative), categorical hair color, categorical eye color, and three quantitative eye color variables extracted from digital photographs: L (Lightness/brightness), C (Chroma/saturation), and cos H (cos Hue/tone). These traits are represented on the right. The HCL color-space with the three axes of variation is shown in the inset. To provide a global summary of the results, a composite Manhattan plot is presented at the bottom combining significant signals for all the traits. Horizontal lines indicate the suggestive (blue line, P-value=5×10^{-5}) and significant (red line, P-value = 5×10^{-8}) thresholds. The Y-axis was truncated at -log10(P-value)=14. Index SNPs in each region are listed above the Manhattan plot. The association of these SNPs with the pigmentation traits is indicated in the checkered table at the top: a colored box is shown if a SNP is associated with that trait (Tables 1A and B). Colors correspond to that assigned to each chromosome in the Manhattan plot, with slight variation when multiple independent hits were observed on the same chromosome. From Adhikari & Mendoza-Revilla et al. (2018).

5.5.3 Meta-analysis

I examined association for each genome-wide associated SNPs at newly associated regions (Table 5.4) in all countries sampled separately and combined results as a meta-analysis. Encouragingly, for all associations, significant effects were in the same direction in all countries, with the variability of effect size across countries reflecting sample size (Figure 5.4).

5.5.3.1 Genome-wide association signals at known pigmentation loci

Seven genomic regions associated with pigmentation traits in the CANDELA sample include genes that are well-known to be involved in pigmentation pathways and that have been related with pigmentation phenotypes in a number of previous studies (Table 5.4). The 5p13 region includes the solute carrier family 45 member 2 gene (*SLC45A2*). Variants at this gene have been shown to cause Oculocutaneous Albinism type 4 (OCA4) in humans (Newton et al., 2001) and to impact on pigmentation in mice (Du and Fisher, 2002), horses (Mariat et al., 2003) and tigers (Xu et al., 2013). The *SLC45A2* protein is a transport protein involved in melanogenesis. The strongest association was found for SNP rs16891982 which results in a F374L substitution in *SLC45A2*. In previous association studies this SNP has been associated with skin, hair and eye pigmentation (Stokowski et al., 2007; Han et al., 2008; Eriksson et al., 2010; Liu et al., 2015; Adhikari et al., 2016a; Hernandez-Pacheco et al., 2017). 6p25 shows strongest association with SNP rs12203592 in the second intron of the interferon regulatory factor 4 gene (*IRF4*). This SNP has been associated with skin, hair and eye pigmentation (Han et al., 2008; Eriksson et al., 2010; Zhang et al., 2013; Liu et al., 2015; Adhikari et al., 2016). In-vitro analyses have demonstrated that rs12203592 impacts the function of an enhancer element regulating *IRF4* expression and the induction of tyrosinase (*TYR*), a key enzyme in the melanin synthesis pathway (Visser et al., 2015) (Section 1.5.3). The associated SNP (rs10809826) in 9p23 occurred in a non-coding region, upstream of *TYRP1*. The *TYRP1* gene encodes a melanosomal enzyme with a role in the eumelanin pathway. Rare mutations of *TYRP1* cause Oculocutaneous albinism in humans (Rooryck et al., 2006; Chiang et al., 2009), and coat color alterations in mice (Kobayashi et al., 1998) and cats (Lyons et al., 2005). A private variant in the Solomon Islands (R93C) mutation in exon 2 has been shown to be associated with blond hair (Kenny et al., 2012). Association studies have found variants at *TYRP1* to be associated with variation in skin and eye pigmentation (Frudakis et al., 2003; Sulem et al., 2008; Liu et al., 2010; Zhang et al., 2013a; Martin et al., 2017b). The region in 11q14 shows three independent signals of association impacting on the Tyrosinase (*TYR*) and the Glutamate Metabotropic Receptor 5 (*GRM5*) gene. The tyrosinase enzime, encoded by *TYR* gene, plays a key role in the biosynthesis of melanin by mediating the first steps in melanin formation (Parra, 2007, Liu et al., 2013). Mutations in *TYR* are responsible for Oculocutaneous Albinism type 1 in humans (Kwon et al., 1987) and various pigmentation phenotypes in other organisms (Schmidt-Küntzel et al., 2005; Polanowski et al., 2012). *GRM5* lies upstream of *TYR* and an independent variant has been previously shown to be associated with skin pigmentation in an European admixed population (Beleza et al., 2013). The 15q13 region comprises the oculocutaneous albinism

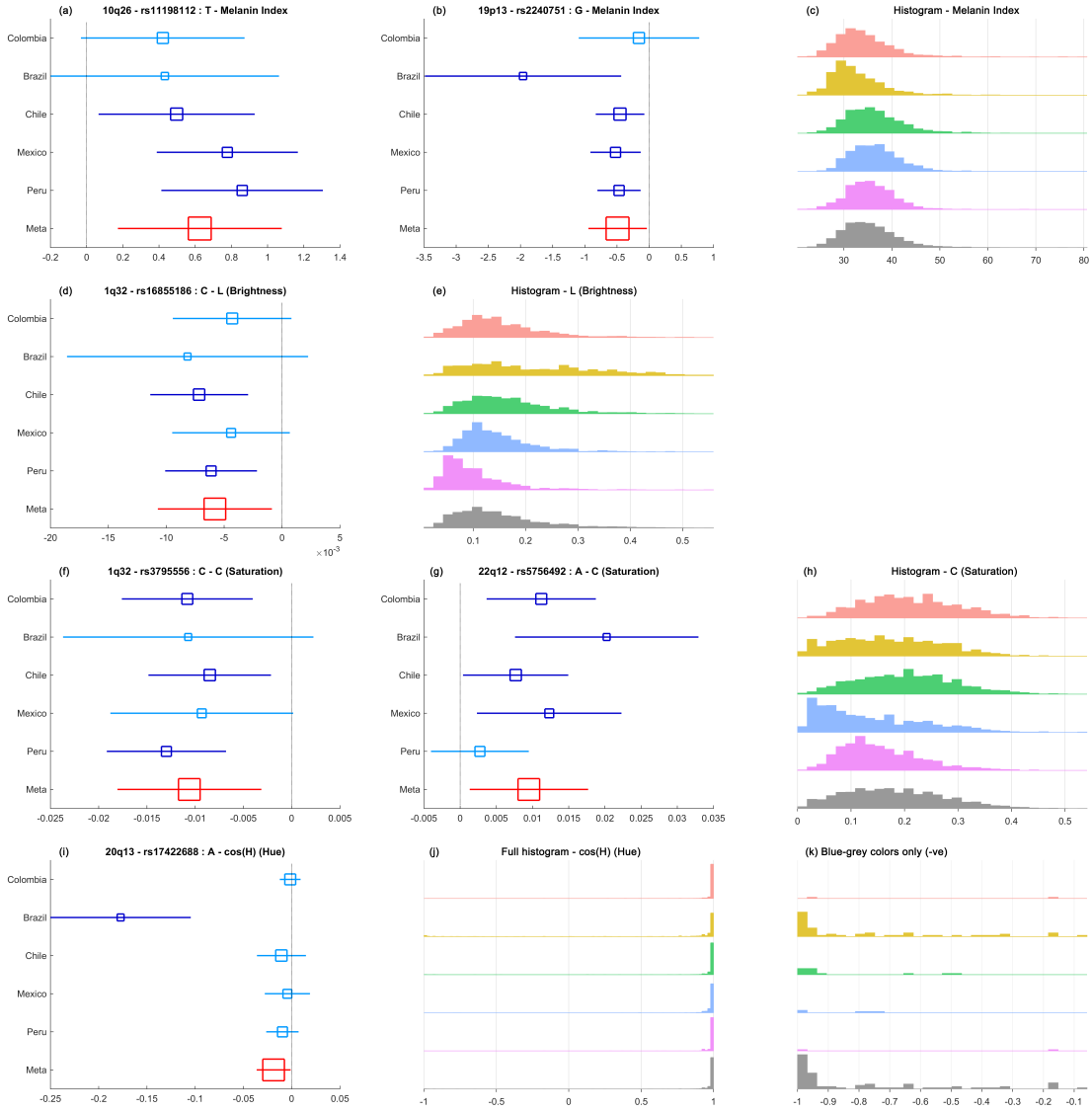

Figure 5.4: Meta-analysis for 6 index SNPs representing novel associations to pigmentation traits. Panels a-b, d, f-g, i show forest plots for the index SNPs in the five novel regions reported in Table 5.4 for four pigmentation traits. Each pigmentation trait is shown in one row. Meta-analysis was performed by combining association results from each country. Estimates obtained in each country are shown as blue boxes. Red boxes indicate estimates obtained in the meta-analysis. Box size is proportional to sample size. Horizontal bars indicate confidence intervals representing 2×standard errors. Intervals that include zero (that is, nonsignificant effects) are shown in light blue. A histogram for the trait in each country and the combined sample is presented at the end of each row (panels c, e, g, h, j). The proportion of negative values (corresponding to blue-grey eyes) is small for cos(H), so a histogram of cos(H) values restricted into the negative values is shown in panel k to show variation across countries. From Adhikari & Mendoza-Revilla et al. (2018).

type 2 (*OCA2/HERC2*) genes. Variants in this region have been shown to result in a range of pigmentation phenotypes in humans and other organisms (Rinchik et al., 1993; Sulem et al., 2007; Sturm, 2009; Liu et al., 2010; Eriksson et al., 2010; Candille et al., 2012; Zhang et al., 2013a; Caduff et al., 2017; Wollstein et al., 2017). The *OCA2* gene encodes the P-protein that assists tyrosinase trafficking and processing, melanosomal pH and glutathionine metabolism (Park et al., 2015). It has also been shown that the encoded protein assists in anion transport increasing chloride conduction from the melanosome (Bellono et al., 2014). Variants located upstream of *OCA2* and within intron 86 of *HERC2* show the strongest association for lighter eye color with the derived allele in European populations (Sturm et al., 2008; Sturm et al., 2009; Eriksson et al., 2010; Liu et al., 2010; Zhang et al., 2013). *HERC2* variants have been shown experimentally to function as an enhancer regulating *OCA2* transcription by modulating chromatin folding (Visser et al., 2012). Specifically, molecular approaches showed that *HERC2* acts as an enhancer to *OCA2* via a long-range chromatin loop that is modulated by several transcriptor factors including the Melanogenesis Associated Transcription Factor (MITF) (Visser et al., 2012). SNPs in the 16q24 region show maximal association for rs885479 leading to a R163Q in the Melanocortin 1 Receptor gene (*MC1R*). Variants of *MC1R* have been previously shown to influence hair pigmentation in humans (notably red hair) (Valverde et al., 1995; Sulem et al., 2007; Han et al., 2008; Eriksson et al., 2010; Lin et al., 2015), as well as skin pigmentation (Liu et al., 2015). SNP rs885479 within *MC1R* has been associated with skin pigmentation variation in East Asians (Yamaguchi et al., 2012).

5.5.3.2 Candidate genes at newly associated regions

Two novel signals of association with skin pigmentation are located in 10q26 and 19p13. The 10q26 region shows SNPs with genome wide significant association spanning ∼100Kb within an intergenic region of ∼400Kb exhibiting relatively low LD (Figure 5.6). Genome annotations indicate that this region overlaps an open chromatin segment, that is highly conserved evolutionarily and includes several transcription factor binding sites (Figure C.5). The derived allele for the index SNP (rs11198112) is segregating at low to moderate frequencies across many populations, but reaches its highest frequency in Native American Amazonians and Melanesians (Figure 5.5).

Interestingly, the derived allele was associated to darker skin pigmentation in contrast to many of the other variants associated to skin pigmentation (Figure C.6). The SNP rs11198112 is also present in the binding site for transcription factor EBF1 (Early B-cell factor) (Figure C.5). If the effect of this SNP is mediated through regulation of nearby genes, of potential interest is the gene encoding for the *EMX2* transcription factor (Empty Spiracles Homeobox 2), which flanks the associated region (Figure 5.6). Mouse experiments have shown that *Emx2* regulates the expression of *Mitf* (a key regulator of melanocyte development and survival) as well as of *Tyr* and *Tyrp-1* (two melanocyte-specific genes responsible for melanin production) (Bordogna et al., 2005). In addition, *EMX2* has also been recently associated to tanning response in Europeans (Visconti et al., 2018).

SNPs showing genome-wide significant association in the 19p13 region span ∼100Kb and show strongest association for SNP rs2240751 located in the Major Facilitator Superfamily Domain Containing 12 (*MFSD12*) gene (Figure 5.6). SNPs in this gene have recently been associated with skin pigmentation variation in Sub-Saharan Africans (Crawford et al., 2017). The index SNP identified in this sample (rs2240751) leads to a tyrosine for histidine substitution at amino-acid 182 of *MFSD12* (Y182H), which is common in East Asians and Native Americans but is rare elsewhere (Figure 5.5) and occurs in a highly conserved sequence (as indicated by GERP and SiPhy metrics). The replacement of a polar for a basic amino acid is likely to affect protein function, as indicated by low SIFT (<0.01) and high PolyPhen2 (>0.99) scores. Animal model studies indicate that *MFSD12* is involved in lysosomal biology (Crawford et al., 2017). A transcriptome analysis has shown that *MFSD12* is amongst the most down-regulated genes in skin biopsies from vitiligo patients (Yu et al., 2012).

Of the three novel regions associated with quantitative digital eye color variables, the one in 1q32 is characterized by substantial LD over a region of ∼300Kb (Figure 5.6) and is associated with the L and C variables (Table 5.4). The strongest association is seen for markers overlapping the *DSTYK* gene (Dual Serine/Threonine and Tyrosine Protein Kinase), the index SNP (rs3795556) being located in the 3' untranslated region of the *DSTYK* transcript. Interestingly, expression studies have shown that *MITF* regulates the expression of *DSTYK* in human melanocytes (Hoek et al., 2008). The 20q13 region associated with the cos(H) variable shows strong LD over a region of ∼200Kb. The strongest association is seen for SNPs overlapping the *WFDC5* gene (WAP Four-Disulfide Core Domain 5, Figure 5.6), with the index SNP (rs17422688) leading to a histidine for tyrosine substitution (H97Y) in a highly conserved region (based on GERP and SiPhy conservation metrics). This amino acid change is predicted to affect protein function, as implied by low SIFT (< 0.03) and high PolyPhen2 (> 0.81) scores. Several WFDC genes have been shown to be expressed in the human iris (Wistow et al., 2002). SNPs in 22q12 associated with the C variable shows LD over a region of ∼100Kb (Figure 5.6). The index SNP (rs5756492) is located in the second intron of the gene encoding Mercaptopyruvate sulfurtransferase (*MPST*) an enzyme playing a role in cyanide detoxification Billaut-Laden et al. (2006) and cellular redox regulation (Nagahara et al., 1998). *MPST* is also expressed in the skin (GTEx Consortium, 2013).

5.5.4 Allelic heterogeneity at OCA2/HERC2 and GRM5/TYR

I evaluated the presence of multiple, independent signals of association at each genomic region highlighted in the primary GWAS by performing step-wise regression (using the same model as in the primary analyses; see Section 5.3.7), conditioning on the index SNP at each region (Table 5.5). Evidence of genome-wide significant association was abolished for all regions except 11q14 and 15q13, where a total of three and five independent signals were detected, respectively (Tables 5.4 and 5.5). These two regions include, respectively,

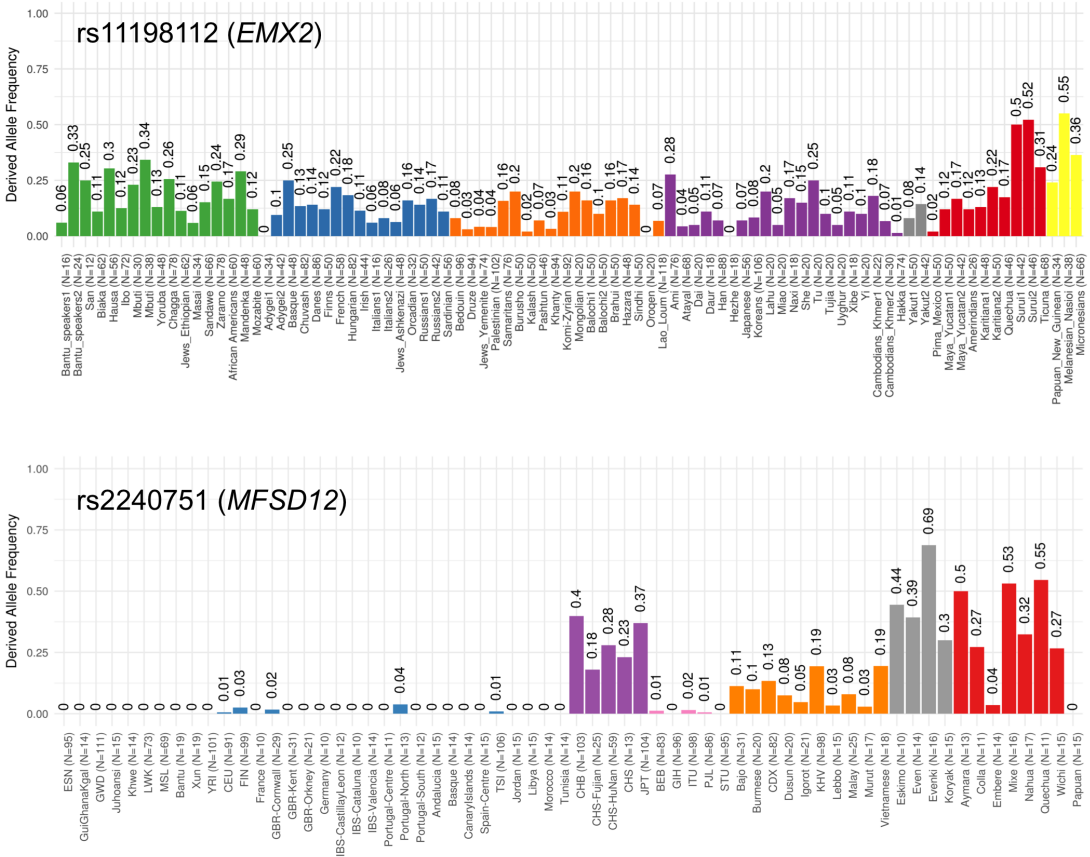

Figure 5.5: Worldwide allele frequencies of novel variants associated to skin pigmentation. The allele frequency for SNP rs11198112 (top) was retrieved from the ALFRED database via https://alfred.med.yale.edu/alfred. This database includes allele frequency information for 84 populations for this SNP. The colors of the bars reflect the major geographic origin of the available populations as categorized by ALFRED: Africa (green), Europe (blue), Middle East and North Africa (orange), East Asia (purple), Siberia (grey), America (red) and Oceania (yellow). The allele frequency for SNP 2240751 (bottom), was estimated from 2391 unrelated individuals from a worldwide dataset including 64 populations (Table D.1). The colors of the bar reflect the geographic origin of the populations: Africa (green), Europe (blue), Middle East and North Africa (brown), East Asia (purple), South Asia (pink), South East Asia (orange), Siberia (grey), America (red) and Oceania (yellow). The numbers of the individuals per population (N) is given next to the population name and the derived allele frequency is displayed on the top of each bar. From Adhikari & Mendoza-Revilla et al. (2018).

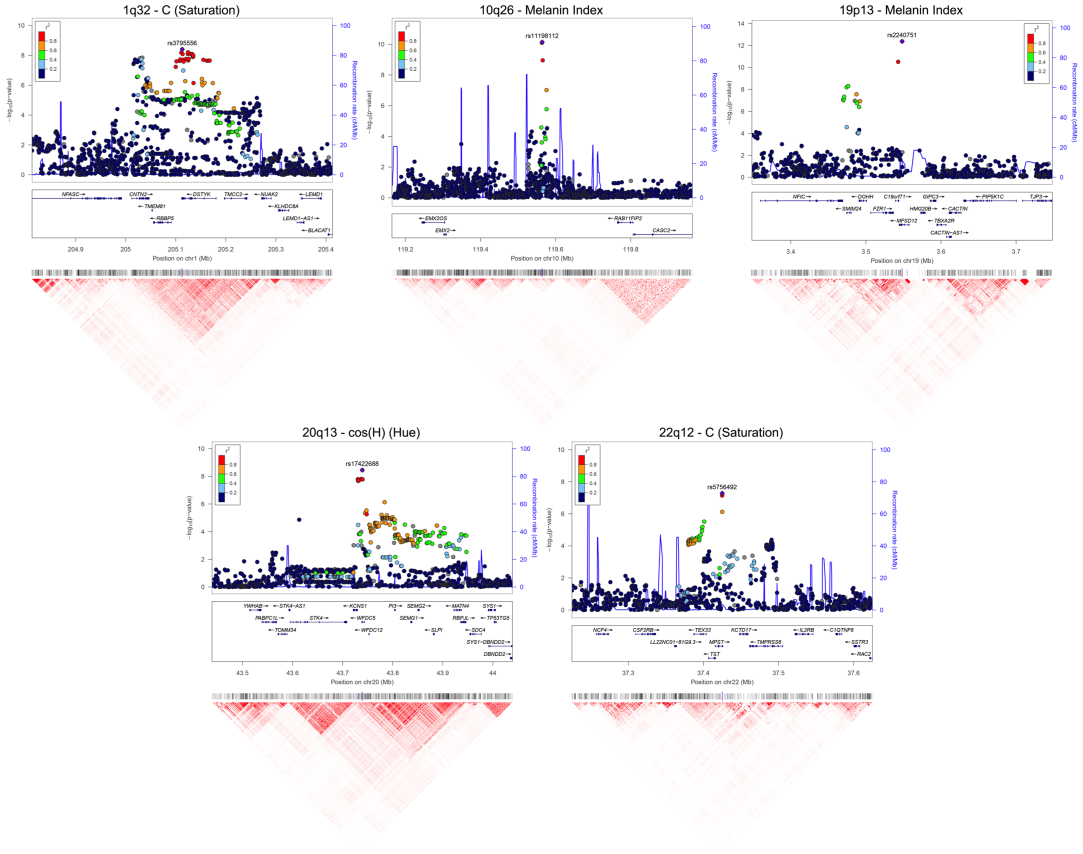

Figure 5.6: Regional association (LocusZoom) plots for SNPs in the five genomic regions showing novel genome-wide significant associations to pigmentation traits. Chromosomal location and trait are specified in the title of each panel. At the top, index SNPs (Table 5.4A) are highlighted with a purple diamond. Colors for other SNPs represent the strength of LD between that SNP and the index SNP (in the 1000 Genomes AMR data). Local recombination rate in the AMR data is shown as a continuous blue line (scale on the right y-axis). Genes in each region, their intron and exon structure, direction of transcription and genomic coordinates (in Mb, using the NCBI human genome sequence, Build 37, as reference) are shown in the middle of each panel. At the bottom of each plot is shown a pair-wise LD heatmap across all SNPs in a region (using r^2, ranging from red indicating $r^2 = 1$ to white indicating $r^2 = 0$). From Adhikari & Mendoza-Revilla et al. (2018).

the *GRM5/TYR* and *OCA2/HERC2* genes, and SNPs in these regions have been consistently associated with pigmentation traits in previous analyses, including several GWAS and candidate gene studies (Section 1.5.3). However, since the SNPs examined in those reports often differ, the independence of their effects has not been systematically evaluated. Consistent with these findings, two independent signals of association in 11q14 have been reported in a GWAS for skin pigmentation in the African/European admixed population of Cabo Verde (Beleza et al., 2013a; Lloyd-Jones et al., 2017). Seven of the eight independent SNPs were associated with skin pigmentation (the exception being rs4778249 in 15q13). In addition to the effect of skin pigmentation for the three associated SNPs in 11q14, two (rs1042602 and rs7118677) were also associated with hair pigmentation, and one (rs1126809) with eye color (Table 5.5). The five independently associated SNPs in 15q13 impact on eye color variation, with one of these SNPs also impacting on hair color (rs12913832). Genome annotations suggest that the eight independently associated SNPs detected here could be functional (Table 5.5). Four occur in exons, of which three result in non-conservative amino-acid substitutions, and one (rs1800404) encodes a synonymous substitution (in exon 10 of *OCA2*) and is located in a conserved binding site for transcription factor YY1 (known to regulate pigmentation in animal models (Li et al., 2012)). The other four independently associated SNPs are located in introns of *GRM5/TYR* or *OCA2/HERC2*. For one of these (rs12913832), intronic within *HERC2*, there is experimental evidence indicating that it regulates transcription of the neighboring *OCA2* gene (Visser et al., 2012).

Table 5.5: Additional index SNPs in the *GRM5/TYR* and *OCA2/HERC2* gene regions showing independent association with pigmentation traits.

				Skin	Hair	Eye			
Region	**SNP**	**Gene**	**Annotation**	**MI**	**Categorical**	**Categorical**	**L**	**C**	**cos(H)**
11q14	rs7118677[b]	GRM5	Intronic	11.2	7.6	0.4	0	0.3	0.2
11q14	rs1126809[b]	TYR	R402Q	9.2	5	4.7	7.5	0.9	3.2
15q13	rs4778219	OCA2	Intronic	8.5	2.4	6.4	10.4	0.2	0.3
15q13	rs1800407	OCA2	R419Q	18.3	3.8	16.2	18.9	7.4	2.1
15q13	rs1800404[a]	OCA2	Synonymous/TFB	10.3	2.4	10.9	18.3	9.5	1.6
15q13	rs4778249[b]	HERC2	Intronic	2.9	0.3	1.5	8.6	13.7	0.8

[a]P-values shown for these SNPs are unconditioned, while P-values for the other index SNPs are conditioned. Results from conditioned and unconditioned analyses are consistent and confirm an increase in power upon conditioning (see Methods).
[b]These markers were obtained through imputation. Their imputation quality "info" metric was ≥ 0.975, the median value being 0.993. The other markers were obtained from chip genotyping, and their "concordance" metric was > 0.9, the median value being 0.981.

5.5.5 Interactions between SNPs independently associated to pigmentation

I examined interactions between the genome-wide significant SNPs (i.e. index SNPs) (Table 5.4 and 5.5) by testing regression models including all possible pairs of index SNPs. A number of significant interactions were detected at a Bonferroni corrected P-value threshold of 3.3×10^{-4} (Figure 5.7). A different pattern of interaction was observed for skin, relative to hair or eye pigmentation. In the case of skin, significant interactions were seen mainly between SNPs that, individually, have strong effects (in *SLC45A2*, *SLC24A5*, *HERC2/OCA2* and *TYR/GRM5*). By contrast, for hair and eye color, SNPs in the regions with strongest individual effects (*SLC45A2*, *SLC24A5* and *HERC2/OCA2*) showed significant interaction with SNPs at most other pigmentation-associated regions. This included regions that individually do not have a significant effect on a particular trait (e.g. *MC1R* and *MFSD12* with hair or eye pigmentation, respectively). This result is in accordance with the common occurrence of epistasis in the determination of pigmentation variation (Pośpiech et al., 2011), particularly in eye pigmentation variation (Wollstein et al., 2017) that is associated with an increase in the prediction accuracy for eye color categorization (Pośpiech et al., 2011; Walsh et al., 2013; Walsh and Kayser, 2016; Wollstein et al., 2017).

5.6 Discussion and limitations

The analyses presented here highlight the complex genetic architecture of pigmentation variation in Latin America. In model organisms, more than 100 genes have been associated to pigmentation variation (Color Gene database: http://www.espcr.org/micemut/), while in humans, less than 20 genes have been robustly associated to pigmentation variation (Section 1.5.3). This observation lead to the assumption that pigmentation is a rather simple phenotype, with only a few variants being highly predictive of pigmentation phenotypes across human populations, as mainly evidenced by genetic forensic analysis (Walsh et al., 2011, 2013; Walsh and Kayser, 2016). However, recent genetic studies of skin pigmentation of underrepresented populations, such as Africans (Crawford et al., 2017; Martin et al., 2017) have shown that previously uncharacterized genes strongly contribute to pigmentation variation and that pigmentation is a rather complex polygenic trait. The results presented here are consistent with these observations. The pigmentation variation in Latin America is affected by multiple gene regions (Table 5.4) as well as multiple independent variations at the *OCA2/HERC2* and *GRM5/TYR* regions (Table 5.5).

Since the history of Latin America involved extensive admixture of Native Americans, Europeans and Africans (Wang et al., 2008; Ruiz-Linares et al., 2014; Chacon-Duque et al., 2018), it is to be expected that variants impacting pigmentation in these populations are segregating in Latin America, and as a consequence, affect pigmentation variation in modern Latin Americans. Further, since Native Americans can trace most of their ancestry to East Asians, it is also expected that variants affecting pigmentation in Native Americans are also shared with Asian populations. Consistent with this scenario, seven variants

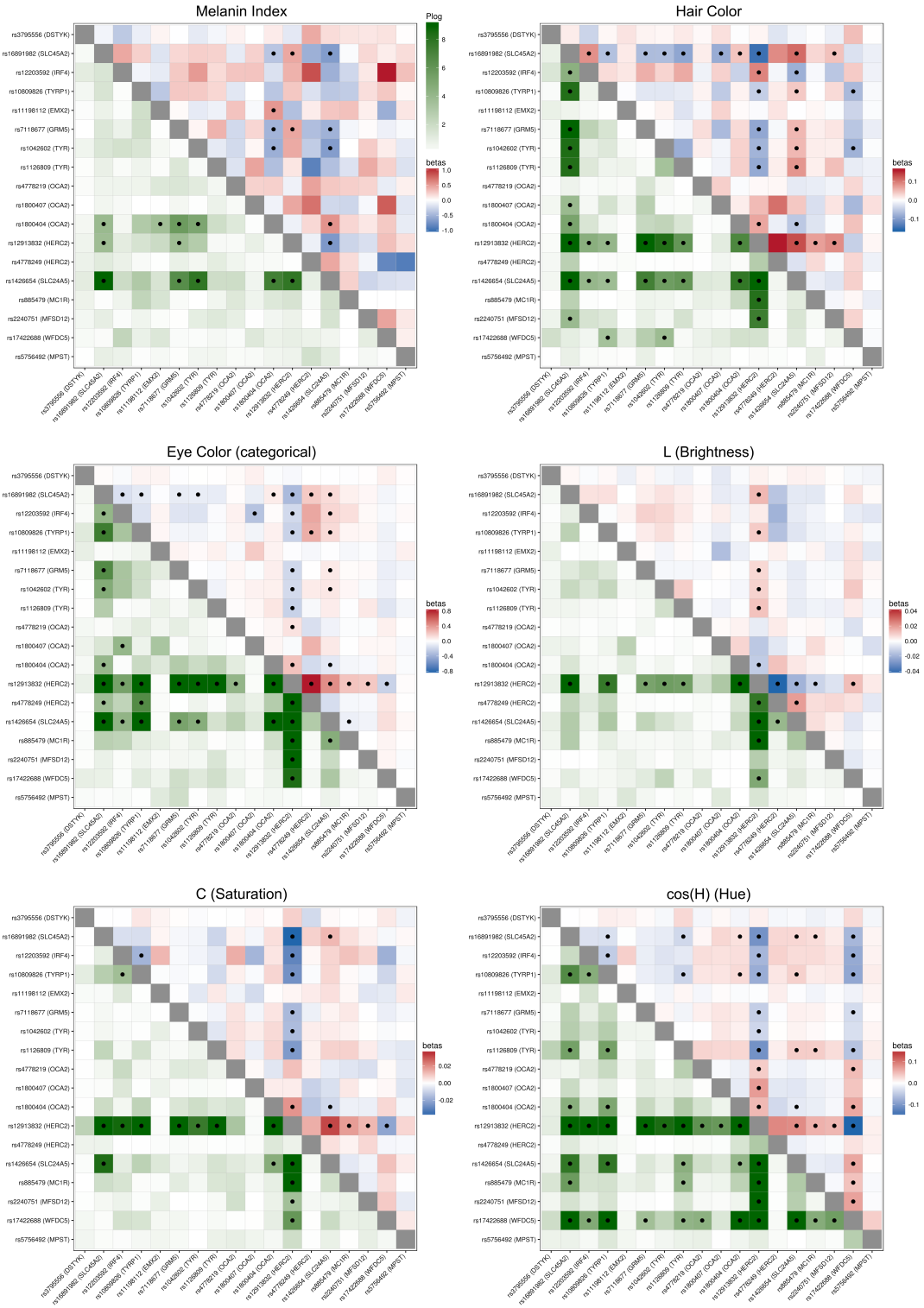

Figure 5.7: Heatmaps of statistical interactions between the 18 SNPs showing genome-wide significant associations to pigmentation phenotypes. Each panel corresponds to a different trait. The lower left triangle represents -log10 P-values for the interaction term included in the regression model (with the color-scale shown at the top). The upper right triangle represents regression beta coefficients for each interaction term, colored from blue (negative effect) to white (no effect) to red (positive effect). As the scale for each trait is different, separate scales for effect sizes are shown next to each panel. Interactions that are significant (after Bonferroni-correction) are marked with a dot. From Adhikari & Mendoza-Revilla et al. (2018).

that have been previously associated in Europeans and one previously associated in East Asians, were replicated in the CANDELA dataset (Table 5.4). It is possible that some of the previously associated variants with pigmentation phenotypes in Eurasian populations were not replicated in this dataset due to a variety of factors affecting power across studies. For example, some of the previously associated variants in Europeans or East Asians could have high frequency in populations that did not contribute to admixture in Latin Americans. The majority of the pigmentation association studies in Europe have been mainly carried out in Northern European populations (Sulem et al., 2007, 2008; Liu et al., 2010; Lin et al., 2015; Jacobs et al., 2015) and pigmentation loci that contribute to pigmentation variation in other European populations (such as Southern Europeans) are yet to be more carefully explored (López et al., 2014). A recent detailed population structure analysis carried out in the CANDELA dataset (Chacon-Duque et al., 2018) showed that the majority of the European component in this sample steemed from Southern European populations, in line with documented historical records. Similarly, variants associated to pigmentation in East Asians populations shows a geographically structured pattern (Figure 6.1), and therefore, it is possible that the Eastern Eurasian ancestors carrying these variants may not have contributed extensive ancestry to the ancestral population of modern Native Americans. Additionally, dissimilarities in phenotype assessment approaches and in phenotype definitions are also likely to explain some of the differences in association results across studies. For example, GWAS carried out in Europeans have mostly focused on variation in the brown to blue color spectrum. By contrast, the C (Saturation) color component examined here, with which two new loci have been associated, captures variation within brown eyes (Table 5.4) and the associated SNPs at these loci have the highest derived allele frequencies in East Asians (Table C.5).

The novel variants associated to the quantitative eye color variables clearly demonstrate the increase in association power when adopting quantitative pigmentation phenotypes compared to the much more common ordinal categorical approach. Although the novel candidate genes represent important candidate genes, as they are related to pigmentation biology, it would be necessary to conduct functional analysis to validate their role in eye pigmentation variation. It is interesting that the highest frequency of the derived allele was present in non-Northern European populations, and as such replication in populations from Southern Europe and Asia could represent important target populations for replication analysis. Similarly, for the novel variant associated to skin pigmentation on the 10q26 region (rs11198112; Table 5.4), highest frequency of the derived allele was found in Native American Amazonians and Melanesians (Figure 5.5). Further, the derived variant was associated to darker skin pigmentation and it would be interesting to assess whether this region shows signals of natural selection in these populations, perhaps associated to their habitation in latitudes with increased solar radiation exposure.

In East Asia, lighter skin pigmentation seems to be due to the effect of derived variants in the genes *OCA2* and *MC1R* (Edwards et al., 2010; Abe et al., 2013; Eaton et al., 2015; Edwards et al., 2016; Norton et al., 2016; Yang et al., 2016; Rawofi et al., 2017).

Here, I also report a new variant (rs2240751) at *MFSD12* as another potential East Asian-specific skin pigmentation association. This same gene has been recently implicated in a study of skin pigmentation variation in a genetically diverse set of sub-Saharan African populations (Crawford et al., 2017). The variants showing the strongest association were present in non-coding SNPs that are segregating mainly in African populations (Crawford et al., 2017). By contrast, the novel variant reported here, is seen for a Y182H amino-acid substitution in *MFSD12* with the highest frequency in East Asian and Native American populations (Figure 5.5). It is therefore likely that this variant was carried by the East Asian populations who entered the Americas. It would also be interesting to assess, whether this variant shows association with skin pigmentation variation in East Asian and Native American populations.

Considering the evidence for solar radiation having shaped the diversity of pigmentation loci in Eurasia populations it is interesting that the GWAS in Latin Americans did not detect any pigmentation variants private to the Americas. The American continent shows extensive variation in solar radiation levels as its territory extends along a North-South axis comprising circumpolar and Equatorial latitudes (Figure 6.4). However, Native Americans do not exhibit a variation in skin pigmentation similar to that seen in Old World populations living at similar latitudes (Jablonski, 2008, 2012). It has been suggested that the difference between continents could relate to cultural adaptations or to environmental factors, as exemplified by the wearing of sewn clothing and the making of shelters, as well as better natural shelter from the sun in the American tropics due to the abundance of high-density canopy (Jablonski, 2008, 2012). Additionally, it has also been suggested that adaptation to solar variation in Native Americans occurred due to a better tanning ability compared to other Old World populations (Jablonski, 2008, 2012). So far, only two genes (*OPRM1* and *EGFR*) have been suggested to contribute to skin pigmentation differences between Native Americans and Europeans (Quillen et al., 2012). However, neither of these SNPs (rs6917661 and rs12668421) showed a significant association with skin pigmentation in the CANDELA sample (P-values > 0.05 for both SNPs). It is also possible that rare variants that contribute to pigmentation variation in Latin American (e.g. inherited from their Native American ancestors) are not well captured by SNPs present in the SNP-array platform or in the imputed dataset used here. Finally, the lack of novel genetic adaptations in relation to solar radiation in the Americas could be related to the relatively recent and rapid settlement of the New World (Tamm et al., 2007; Reich et al., 2012; Raghavan et al., 2015). This settlement history thus limits the timespan for which novel genetic variants could arise and change in frequency in response to regional environmental selection pressures. Thus, it is possible that the majority of genetic adaptations to variable solar radiation levels in the Americas would have involved mainly variants introduced from the Old World. Future research in larger number of diverse Native Americans may reveal additional variants associated with pigmentation phenotypes and will shed light on the evolutionary history and adaptive significance of this phenotype in the Americas.

5.7 Summary

In this chapter I reported novel variants associated to skin and eye pigmentation in Latin Americans. The results here highlight the complex genetic architecture of pigmentation in Latin Americans, as evidenced by independent variants at different genomic regions as well as multiple independent variants within the associated regions. The novel associations using quantitative eye color variables show the greater statistical power obtained by using rich color models. Further, the reported genomic regions associated to eye pigmentation variation represent important candidate genes that should be followed up by functional analyses. Finally, the novel associated variant in the *MFSD12* gene represents a potential East Asian and Native American specific skin pigmentation locus.

Chapter 6

Exploring the convergent evolution of lighter skin pigmentation in Eurasia

6.1 Overview

In this chapter I investigate the convergent evolution of lighter skin pigmentation in Eurasia. I use sequence data from individuals from one North Western European and one East Asian population from the 1000 Genomes Project to conduct genome-wide scans of selection on pigmentation-associated genomic regions, and show that different genomic regions and variants have been selected in these populations. This analysis represents a follow-up investigation of the skin pigmentation-associated loci reported in Chapter 5 and does not represent an extensive survey of selection signals at all known skin pigmentation-associated loci. I also assess the potential selective pressures acting on skin pigmentation, by using genome-wide SNP data from a large set of worldwide human populations, and evaluate the correlation between the allele frequency at pigmentation-associated loci and solar radiation. I provide evidence that the present distribution of a novel candidate SNP at *MFSD12* (associated to skin pigmentation and previously reported in Chapter 5) has probably been affected by exposure to solar radiation in Eastern Eurasia. Finally, I show that the patterns of genomic diversity at this locus are compatible with a scenario of convergent adaptation for lighter skin pigmentation in East Asia and that this selection event probably occurred long after the divergence from Europeans.

6.2 Background

It is well established that the evolution of pigmentation phenotypes, i.e. skin, hair and eye color in human populations have been driven (at least partly) by selection (Section 1.5.2). Whilst hair and eye pigmentation are usually regarded as having evolved through sexual selection, the high correlation between solar radiation exposure and skin pigmentation strongly suggests the role of natural selection in shaping the worldwide distribution of this phenotype (Jablonski and Chaplin, 2000; Jablonski, 2008, 2012; Jablonski and Chaplin, 2014, 2017). The currently accepted hypothesis regarding the evolution on skin pigmentation in human populations is the so called folate/vitamin D hypothesis (Jablonski and

Chaplin, 2000). Under this hypothesis, skin pigmentation is regulated by two opposing forces: one selecting for darker skin tones in populations living in zones with *high* solar radiation exposure to protect against the photolysis of folate, and another selecting for lighter skin tones in populations living in zones with *low* solar radiation exposure to permit the synthesis of adequate levels of Vitamin D. The deficiencies of these two vitamins have been shown to have detrimental effects on the reproductive fitness of individuals and as such, are regarded as good candidates for existing strong selective pressure throughout human history (Jablonski and Chaplin, 2017, 2018). In addition, the similarities in light skin pigmentation in Europeans and East Asians have also been regarded as either the result of a shared genomic origin or independent evolutionary adaptations to low solar radiation environments. Although there is evidence that some loci affecting lighter skin pigmentation have probably evolved prior to their population divergence (McEvoy et al., 2006; Norton et al., 2007), there is also strong evidence supporting convergent evolution in Europeans and East Asians (Edwards et al., 2010; Abe et al., 2013; Eaton et al., 2015; Edwards et al., 2016; Norton et al., 2016; Yang et al., 2016; Rawofi et al., 2017).

6.2.1 Previous studies

Previous studies have shown that convergent evolution of lighter skin pigmentation in Western and Eastern Eurasians occurred through different genetic mechanisms and thus, likely evolved after their population divergence (McEvoy et al., 2006; Norton et al., 2007). In Europe, four genes, namely *SLC24A5*, *OCA2/HERC2*, *GRM5/TYR*, and *SLC45A2*, show signals of selection exclusively in this population. Functionally important derived variants that have been associated to lighter skin pigmentation at these genes are largely restricted to Europeans or their neighbour populations, and consequently only affect skin pigmentation in these populations (or European derived populations such as Latin Americans or African Americans). The onset of selection at these European specific variants in *SLC24A5* and *SLC45A2* has been estimated to between 11,000 to 19,000 ya (Beleza et al., 2013), consistent with their independent evolution long after the divergence of the ancestral population of Europeans and East Asians, which has recently been estimated to have occurred around 42,000 ya (Jouganous et al., 2017). In East Asians, lighter skin pigmentation seems to be due to the effect of derived variants in the genes *OCA2* and *MC1R* (Norton et al., 2007; Edwards et al., 2010; Abe et al., 2013; Eaton et al., 2015; Edwards et al., 2016; Norton et al., 2016; Yang et al., 2016; Rawofi et al., 2017). Interestingly, two non-synonymous SNPs in the *OCA2* gene, rs1800414 and rs74653330, associated to lighter skin pigmentation, show contrasting distributions in Asia (Figure 6.1). As shown on Figure 6.1, SNP rs1800414 is at high frequency across East Asia, whilst rs74653330 is primarily restricted to Northern East Asia, suggesting that these two variants may have been selected independently in different regions of the continent (Murray et al., 2015). The estimated date of the derived allele at both these SNPs has been estimated to be long after the split of European and East Asian populations. SNP rs1800414 is thought to have arisen $\sim 10,000$ ya (Chen et al., 2015) or even more recent $\sim 6,000$ ya (Murray et al., 2015). Similarly, the age of the derived allele at SNP rs74653330 was estimated to $\sim 7,000$ ya (Murray et al., 2015). In addition to these two non-synonymous SNPs in

OCA2, another non-synonymous SNP rs885479 in *MC1R* has also been shown to be associated with lighter skin pigmentation in East Asians (Yamaguchi et al., 2012). Similar to the rs1800414 *OCA2* variant, the derived allele at *MC1R* shows higher frequency broadly across East Asia (Figure 6.2) and strong signals of selection in East Asians (Hider et al., 2013).

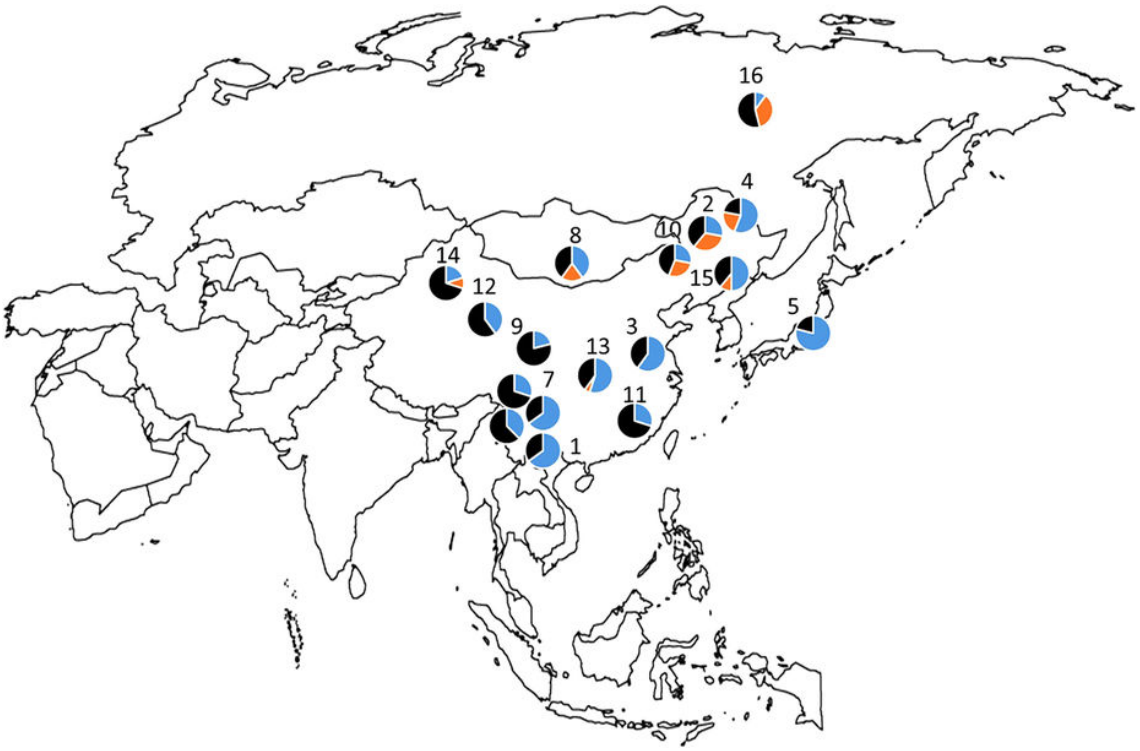

Figure 6.1: Distribution of allele frequencies for SNPs rs1800414 and rs74653330 at *OCA2* in East Asia. Pie-charts show the frequency for SNP rs1800414 in blue and rs74653330 in orange, and the complement of those alleles frequencies in black. The derived allele of SNP rs1800414 has high frequency in a broad East-Asian region, whereas the derived allele of SNP rs74653330 is mainly restricted to northern East Asia. Reproduced from Murray et al. (2015).

6.3 Materials and methods

6.3.1 Description of data

To study the convergent evolution of light skin pigmentation in Eurasia I used two different datasets designed to encapsulate global patterns of human genetic diversity in relation to Eurasia. The first dataset comprised one North Western European (CEU; Utah Residents with Northern and Western European Ancestry), one East Asian (CHB; Han Chinese in Beijing, China) and one West African (YRI; Yoruba in Ibadan, Nigeria) population from the 1000 Genomes Project (1KG) Phase III data release (1000 Genomes Project Consortium et al., 2015). This dataset was used to perform selection scans using three different types of selection statistics (described in Section 6.3.2) on the European and East Asian populations separately. The second dataset was used to explore the correlation between allele frequencies at pigmentation loci with variation in solar radiation levels and contained

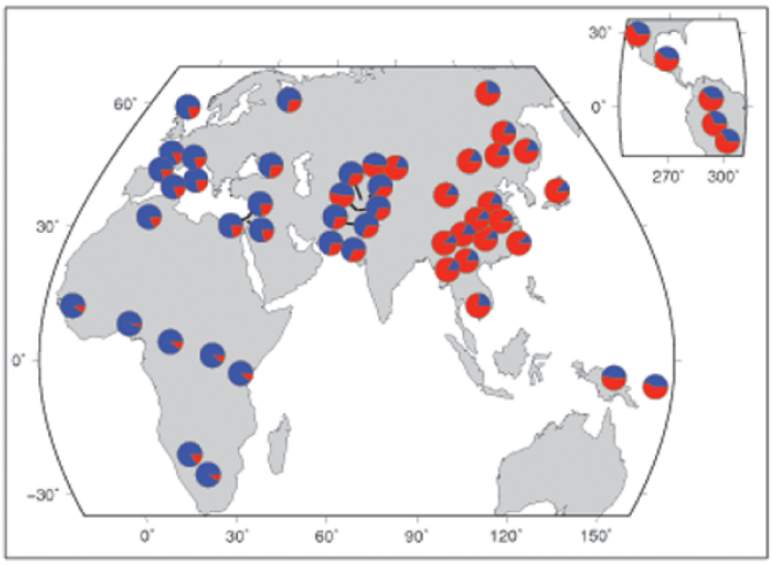

Figure 6.2: Global allele frequency distribution of SNPs rs885479 at *MC1R*. Pie-charts show the frequency of the derived allele in red. Based on the Human Genome Diversity Project (HGDP) browser and adapted from Coop et al. (2009).

populations from several publicly available resources. This dataset included 10 populations from Africa (Schlebusch et al., 2012; 1000 Genomes Project Consortium et al., 2015), 18 populations from Europe (1000 Genomes Project Consortium et al., 2015; Chacon-Duque et al., 2018), 4 populations from North Africa and the Middle East (Chacon-Duque et al., 2018), 20 populations from East, South and South East Asia (1000 Genomes Project Consortium et al., 2015; Mallick et al., 2016; Mörseburg et al., 2016), 4 populations from Siberia (Cardona et al., 2014) and 7 populations from the Americas (Eichstaedt et al., 2014; 1000 Genomes Project Consortium et al., 2015; Chacon-Duque et al., 2018). This dataset is described in detail in Table D.1. I detail the sources of the data, number of individuals per population and geographic coordinates of the sampling locations.

6.3.1.1 Quality control

PLINK v1.9 (Chang et al., 2015) and VCFtools (Danecek et al., 2011) were used to perform quality control (QC) analyses on these datasets. For the first dataset, compromising only populations from the 1KG Project, I filtered duplicated, non-biallelic, and/or SNPs that were not polymorphic between the European (CEU), East Asian (CHB) and West African (YRI) populations. This left a total of 8,304,740 SNPs, which were kept for all subsequent analyses. The number of individuals for CEU, CHB and YRI were 99, 103 and 108, respectively. For the second dataset, SNPs and individuals with >5% missing data were discarded. After performing LD pruning (– *–indep-pairwise 50 5 2*), the PLINK inferred IBD coefficient (*PI-HAT*) was calculated across all pairs of individuals within each population. Individuals with a IBD higher than 0.125 (i.e. third degree relatives) were removed. For the Siberians and Native Americans populations, I used the methodology described in Chacon-Duque et al. (2018), where individuals with more than 10%

from the median IBD value were discarded. This is due to the lower effective population size present in these populations compared to other included populations (Cardona et al., 2014; Chacon-Duque et al., 2018), which can affect IBD estimates based on population allele frequencies (Manichaikul et al., 2010). After these QC filters, a total of 609,005 SNPs were kept for all subsequent analyses. For both datasets only autosomal SNPs were used.

6.3.1.2 ADMIXTURE analysis

To explore correlations between allele frequency and environmental variables it is important to consider only indigenous (i.e. non-recently admixed) populations, in order to obtain allele frequencies that reflect the allele frequency of the indigenous populations inhabiting those areas. I therefore conducted an unsupervised ADMIXTURE analysis (Alexander et al., 2009) applied to the LD pruned data at different K values. All individuals with ancestry components derived from other major geographical region were discarded. Although this ADMIXTURE analysis will not be able to discern sub-continental ancestries within major geographical regions (i.e. migration movements within the same continent), the inclusion of non-indigenous populations is expected to mask the signal of local adaptation, and therefore, should not lead to an increase in the false positive rate. After removing these individuals, only populations with a minimum of 10 individuals were kept. The final total number of individuals passing these criterion were 2,391 from 64 worldwide populations (Table D.1).

6.3.2 Selection signals at skin pigmentation-associated genomic regions

I computed three selection statistics: the Population Branch Statistic (PBS) (Yi et al., 2010), the integrated Haplotype Score (iHS) (Voight et al., 2006) and Tajima's D (Tajima, 1989). PBS scores for CEU were computed using CHB and YRI as references and for CHB using CEU and YRI as reference populations. Pairwise FST were estimated using Reynolds equation (Reynolds et al., 1983) including only SNPs that were polymorphic in at least two populations. The total number of SNPs with PBS scores in CHB and CEU was ∼8,000,000. I calculated iHS using the software selscan (Szpiech and Hernandez, 2014). Ancestral allele states were retrieved from information present in the 1KG VCF files (AA [ancestral allele] field). SNPs with no ancestral allele state were discarded. iHS was computed for SNPs with derived allele frequencies >5% and < 95%. The HapMap (International HapMap Consortium, 2003) GRCh37 genetic map was used to obtain genetic distances between SNPs. SNPs where the extended haplotype homozygosity does not decay below 0.05 beyond 1Mb were also discarded. The final total number of SNPs in CEU and CHB was ∼3,000,000. I calculated Tajima's D using VCFtools on non-overlapping windows of 10kb and discarded windows that contained less than 5 SNPs. The final total number of windows for CEU and CHB was ∼266,000.

6.3.3 Enrichment analysis of selection signals at pigmentation-associated genomic regions

To evaluate whether there is an enrichment of selection signals at gene regions with suggestive evidence of association to pigmentation phenotypes (i.e. SNPs association P-value $< 10^{-5}$ based on the GWAS performed in Chapter 5) I used only PBS scores, as this statistic contained the largest number of SNPs with a selection score (see above). I estimated the maximum PBS score for SNPs in a ±2kb region (to potentially capture nearby regulatory regions) around each gene based on the largest UCSC RefSeq gene transcript retaining genes with at least 5 SNPs. I then contrasted the distribution of maximum PBS at gene regions showing suggestive association with the distribution at gene regions in the rest of the genome. The significance of difference between distributions was then assessed using a one-sided Mann-Whitney U-test. Since the well-established *SLC45A2*, *OCA2*, *HERC2* and *SLC24A5* pigmentation-associated gene regions have strong signals of selection (McEvoy et al., 2006; Lamason et al., 2007; Norton et al., 2007; Sulem et al., 2007; Miller et al., 2007; Coop et al., 2009; Pickrell et al., 2009), and therefore contain very high PBS scores, I also repeated the enrichment analysis after excluding these gene regions.

6.3.4 Using solar radiation data to identify pigmentation loci under selection

To evaluate the possible correlation of allele frequencies at pigmentation genes with solar radiation levels, I examined publicly available data for 64 indigenous population without evidence of recent admixture (Table D.1). Surface solar radiation data was obtained from the NASA Surface meteorology and Solar Energy Web site (https://eosweb.larc.nasa.gov/sse/) in kWh/m^2/day units. These data included annual solar radiation averages from July 1983 to June 2005 on a 1-degree resolution grid over the globe. Annual solar radiation values were obtained for each population based on published coordinates for sampling locations. In case of unpublished sampling location, I obtained this information directly from the authors or used approximate coordinates, for example using the middle of the city or town where the sampling was conducted. I used Bayenv2.0 (Günther and Coop, 2013) to estimate Bayes Factors (BFs) relating solar radiation to allele frequencies at the pigmentation-associated SNPs. These BFs provide a measure of the increase in the fit of allele frequencies to a linear regression model including solar radiation levels over a null model including only population structure as a predictor (see Section 2.3.4). The null model was constructed using a covariance matrix of allele frequencies between populations estimated from 10,000 random SNPs (not in LD) after 100,000 MCMC iterations. These were visually inspected for unexpected low or high correlations and compared across independent runs. As these were not qualitatively distinct from each other, the matrix computed from the first run was used for the Bayenv2 analysis. In addition to BFs, I also estimated Spearman's rank correlation coefficient (ρ) based on normalized allele frequencies as computed by Bayenv2.0. To assess significance, I ranked the SNPs based on their BFs and absolute ρ dividing by the total number of values to obtain empirical P-

values. The allele frequency at a SNP was only considered to be significantly associated to solar radiation if both BF and ρ estimates were significant. As the effect of pigmentation genes could differ between geographic regions, I also conducted separate analyses for Africans, Western Eurasians (including Europeans, North Africans and Middle Easterners), and Eastern Eurasians (including Eastern and Southern Asians, Siberians and Oceanians). Table D.1 lists all the populations included for this analysis restricted by major geographic areas.

6.3.5 Approximate Bayesian Computation (ABC) analysis

To estimate the selection coefficient and the time since the start of selection at the *MFSD12* gene region, I used an Approximate Bayesian Computation (ABC) approach. The software *msms* (Ewing and Hermisson, 2010) was used to perform coalescent-based simulations modelling the demographic history of African, European and East Asian populations (for details of the parameters of the demographic model used, see Jouganous et al. (2017)). I assumed that the minor allele frequency at the time of selection was 1% in Europeans and East Asians and zero in Africans (comparable to the frequency in CEU, CHB and YRI from the 1KG Project). I performed 1,000,000 simulations of a 500kb genome segment with a selected allele in the center, originating in East Asians. I assumed a uniform distribution U[0 — 0.05] for the selection coefficient and a uniform distribution U[5,000 — 42,229 ya] for the starting time of selection. I then only retained simulations where the selected allele was present at the end of the simulation. From the simulations I computed 9 summary statistics in a window of 200kb centered around the selected site: the nucleotide diversity (π), Tajima's D, Fu and Li's D, Fu and Li's F, H1, H2 and H2/H1 as measures of haplotype diversity (Garud et al., 2015), F_{ST} between East Asians and Europeans, F_{ST} between East Asians and Africans, and the derived allele frequency of the selected variant. I used partial least squares (PLS) to identify the most informative statistics based on a subset of 10,000 simulations (prior to PLS analysis, summary statistics were Box-Cox transformed so that their minimum values were between 1 and 2). For parameter inference I used the first 7 PLS components, as they carried the most information for each parameter (estimated by the Root Mean Squared Error [RMSE]; Figure D.14). Estimation of parameters was performed using the *abc* R package (Csilléry et al., 2012). I selected the top 0.5% simulations based on the smallest Euclidean distance between the observed and simulated summary statistics. From these quantities, I obtained the posterior probability distributions for the selection coefficient and the time since selection, and recorded the posterior median and the 95% credible intervals. I examined the accuracy of the ABC parameter estimates using the Predicted Error (PE) (i.e. the mean square error divided by the prior variance of the parameter) based on a leave-one-out cross-validation of 100 observations (Table D.2). Although other indices could have been employed to assess the accuracy of the ABC estimation, such as the relative estimation bias (i.e., the bias expressed a proportion of the true value) and the coverage of the 95% credible interval (i.e., the percent of times where the true value was found within the 95% credible interval), I note that PE estimates found here are similar to that of others obtained using a similar set of summary statistics used to estimate the time since the start of selection and the

selection coefficient (Deschamps et al., 2016).

6.4 Results

6.4.1 Selection has shaped the genetic diversity at pigmentation-associated regions

To explore whether selection had broadly shaped genetic variation at skin pigmentation-associated gene regions, I conducted an enrichment analysis of positive selection based on PBS scores. This enrichment analysis was conducted using only PBS scores as they provided the highest number of selection scores per SNP compared to the other two selection statistics. The enrichment analysis contrasted the distribution of PBS scores at regions showing at least suggestive association with skin pigmentation (i.e. those including SNPs with associated P-values $< 10^{-5}$ based on the GWAS conducted in Chapter 5) against the distribution of PBS scores over the rest of the genome. The enrichment analysis showed significant signals of enrichment for skin pigmentation loci in the European (P-value 6.14×10^{-29}) and East Asian populations (P-value 1.78×10^{-14}). This result is consistent with a strong role of natural selection in shaping skin color in both Western and Eastern Eurasians (Jablonski and Chaplin, 2000; Hider et al., 2013). Interestingly, higher signals of enrichment for positive selection were found in the European population. Although this could suggest higher selective pressure for pigmentation phenotypes in Europeans, this result might also be explained by the slightly higher average European ancestry present in the CANDELA sample (Figure C.1), and consequently higher power to detect pigmentation loci affecting this population. Additionally, to test whether the significant enrichment signals at these skin pigmentation loci were not mainly driven by extremely high PBS scores at only a few gene regions, I repeated the enrichment analysis after removing the *SLC45A2*, *HERC2*, *OCA2* and *SLC24A5* gene regions. Similarly to the previous results, I found strong signals of enrichment of positive selection in Europeans (P-value 4.5×10^{-27}) and East Asians (P-value 8.26×10^{-14}), although with slightly lower significance as expected.

I next tested for signals of selection at the genome-wide associated gene regions (i.e. those with associated P-values $< 5 \times 10^{-8}$ based on the GWAS conducted in Chapter 5) using a suite of three different selection statistics. The first selection statistic computed was iHS, which measures the amount of extended haplotype homozygosity (EHH) at a test SNP along the ancestral allele relative to the derived allele. This selection statistic has strong power to detect recent selective sweeps, especially under a hard-sweep model (Voight et al., 2006). High positive absolute values of iHS scores are indicative of strong recent positive selection. The second selection statistic computed was Tajima's D, a neutrality test that quantifies the reduction in diversity that can be associated with a selective event. An excess of low frequency variants gives a negative Tajima's D score that can be indicative of positive selection. The third statistic, the PBS selection statistic, has strong power to detect signals of positive selection by comparing allele frequency differences in one population relative to two reference populations (Yi et al., 2010). High values of PBS

scores can be indicative of a selective event.

In agreement with previous analyses I found strong evidence of selection at many skin pigmentation-associated gene regions(Figure D.1 — D.13). In the European population strongest selection signals were observed at the *SLC45A2*, *TYR/GRM5*, *OCA2/HERC2* and *SLC24A5* gene regions for at least one selection statistics. All of these gene regions have been previously reported to be under selection in European or European admixed populations (McEvoy et al., 2006; Voight et al., 2006; Sabeti et al., 2007; Coop et al., 2009; Pickrell et al., 2009; Beleza et al., 2013b; Wilde et al., 2014; Mathieson et al., 2015; Field et al., 2016b; Mathieson et al., 2018). In the East Asian population strongest signals of selection were observed at the *OCA2/HERC2*, *MC1R*, and *MFSD12* gene regions, all of which have been previously reported (Norton et al., 2007; Alonso et al., 2008; Hider et al., 2013; Jonnalagadda et al., 2017), with the exception of *MFSD12*. As expected, and based on the convergence of lighter skin pigmentation in Eurasia, different variants within these pigmentation loci seem to have been under selection in Europeans and East Asians (Table 6.1). Five variants (rs16891982 [*SLC45A2*], rs7118677 [*GRM5*], rs4778249 [*HERC2*] and rs1426654 [*SLC24A5*]) show signals of selection only in Europeans and three variants (rs4778219 [*HERC2*], rs885479 [*MC1R*] and rs2240751 [*MFSD12*]) show signals of selection only in East Asians. It is also important to note that in many of the pigmentation-associated regions, the associated SNPs do not show the strongest signals of selection, which suggests that selection may have acted on other nearby SNPs.

Table 6.1: Signals of selection at skin pigmentation genome-wide associated pigmentation SNPs. iHS, Tajima's D and PBS were computed at each index SNPs in CEU and CHB populations from the 1000 Genome Project. P-values were calculated by an outlier approach by ranking all the scores genome-wide and dividing by the number of values in the distribution, taking the upper tail for absolute value of iHS and PBS and the lower tail for Tajima's D. SNPs with empirical P-values lower than 0.05 are shown in bold. Missing values indicate instances where the selection statistic could not be computed at a SNP (see Section 6.3.2).

			CEU			**CHB**		
Region	**Gene**	**SNP**	**abs(iHS) (P-value)**	**Tajima's D (P-value)**	**PBS (P-value)**	**abs(iHS) (P-value)**	**Tajima's D (P-value)**	**PBS (P-value)**
5p13	SLC45A2	rs16891982	-	**-2.47 (1.13E-03)**	**3.64 (3.61E-07)**	-	0.51 (7.47E-01)	0.00 (1.00E+00)
6p25	IRF4	rs12203592	0.37 (6.89E-01)	-0.59 (3.12E-01)	-	-	0.12 (6.26E-01)	-
10q26	EMX2	rs11198112	0.64 (4.90E-01)	0.13 (6.18E-01)	0.00 (1.00E+00)	0.60 (5.21E-01)	0.33 (6.93E-01)	0.00 (6.62E-01)
11q14	GRM5	rs7118677	1.08 (2.50E-01)	0.40 (7.08E-01)	**0.31 (2.68E-02)**	1.11 (2.43E-01)	0.89 (8.39E-01)	0.00 (1.00E+00)
11q14	TYR	rs1042602	1.65 (8.79E-02)	0.25 (6.58E-01)	-	-	0.43 (7.23E-01)	-
11q14	TYR	rs1126809	0.47 (6.10E-01)	-1.19 (1.43E-01)	-	-	-0.77 (2.84E-01)	-
15q13	HERC2	rs4778219	-	0.48 (7.33E-01)	0.10 (1.54E-01)	-	-1.29 (1.46E-01)	**1.27 (1.71E-04)**
15q13	OCA2	rs1800407	0.27 (7.73E-01)	0.13 (6.18E-01)	0.08 (2.01E-01)	-	1.06 (8.72E-01)	0.00 (1.00E+00)
15q13	OCA2	rs1800404	0.75 (4.22E-01)	0.13 (6.18E-01)	**0.71 (1.50E-03)**	**3.45 (3.54E-03)**	1.06 (8.72E-01)	0.00 (1.00E+00)
15q13	HERC2	rs12913832	**3.16 (6.18E-03)**	**-2.20 (8.88E-03)**	-	-	**-1.95 (4.02E-02)**	-
15q13	HERC2	rs4778249	-	**-2.04 (1.87E-02)**	0.02 (4.04E-01)	-	-1.19 (1.71E-01)	0.00 (1.00E+00)
15q21	SLC24A5	rs1426654	-	**-1.94 (2.67E-02)**	**3.87 (2.41E-07)**	-	-1.40 (1.24E-01)	0.00 (1.00E+00)
16q24	MC1R	rs885479	-	-0.03 (5.19E-01)	0.00 (1.00E+00)	0.12 (8.99E-01)	0.24 (6.66E-01)	**0.83 (2.07E-03)**
19p13	MFSD12	rs2240751	-	-0.78 (2.52E-01)	0.00 (1.00E+00)	0.70 (4.55E-01)	-1.71 (7.13E-02)	**0.50 (1.37E-02)**

6.4.2 Loci underlying local adaptation through solar radiation exposure

Selection for variation in skin pigmentation has been proposed to relate to adaptation to patterns of solar radiation (Section 1.5.2). Consistently, a correlation between allele frequencies at certain skin-pigmentation associated SNPs with solar radiation levels has been reported in the Human Genome Diversity Project (HGDP) population panel (Hancock et al., 2008; Coop et al., 2010; Hancock et al., 2011). I re-evaluated this correlation for SNPs associated to skin pigmentation phenotypes in a dataset I compiled including 64 indigenous populations from around the world. Allele frequencies at four SNPs showed a significant correlation with solar radiation (Table 6.2). I replicated three reported associations (rs12913832 and rs1800404 in *OCA2/HERC2* gene region and rs885479 in *MC1R*) (Hancock et al., 2008; Coop et al., 2010; Hancock et al., 2011). The fourth is rs2240751 (*MFSD12*), which showed a strong correlation with solar radiation in Eastern Eurasia (log_{10}(BF)=2.32, P-value = 0.004; ρ =-0.28, P-value = 0.047) (Figure 6.4).

Table 6.2: Correlation between allele frequency at skin pigmentation associated loci and solar radiation. Allele frequency correlations between the derived allele frequency of the associated SNPs and solar radiation was tested using Bayenv2.0. For each SNP I estimated a Bayes Factor (BF) and Spearman's rank correlation coefficient (rho). SNPs not present in the extended worldwide populations dataset (Table D.1) or that were fixed in a geographical region were not included in the analysis. Significant SNPs for both the BF and rho's are in bold.

			Worldwide		**Western Eurasia**		**Eastern Eurasia**	
Region	**Gene**	**SNP**	**log10(BF) (P)**	**rho (P)**	**log10(BF) (P)**	**rho (P)**	**log10(BF) (P)**	**rho (P)**
5p13	*SLC45A2*	rs16891982	-	-	-	-	-	-
6p25	*IRF4*	rs12203592	-0.87 (0.807)	-0.02 (0.834)	-0.30 (0.085)	-0.26 (0.025)	-0.71 (0.826)	0.04 (0.757)
10q26	*EMX2*	rs11198112	-0.80 (0.566)	0.04 (0.642)	-0.69 (0.295)	-0.11 (0.339)	-0.66 (0.703)	0.10 (0.496)
11q14	*GRM5*	rs7118677	-	-	-	-	-	-
11q14	*TYR*	rs1042602	-0.80 (0.592)	-0.05 (0.527)	-0.76 (0.368)	-0.07 (0.568)	-0.54 (0.473)	-0.03 (0.798)
11q14	*TYR*	rs1126809	-	-	-	-	-	-
15q13	*HERC2*	rs4778219	-	-	-	-	-	-
15q13	*OCA2*	rs1800407	-0.49 (0.194)	0.09 (0.279)	-0.90 (0.672)	-0.03 (0.818)	-0.78 (0.968)	0.06 (0.663)
15q13	*OCA2*	rs1800404	**3.12 (0.002)**	**-0.22 (0.019)**	-0.78 (0.406)	0.00 (0.985)	**1.48 (0.017)**	**-0.30 (0.029)**
15q13	*HERC2*	rs12913832	**12.66 (0.001)**	**-0.39 (0.001)**	**9.40 (0.001)**	**-0.36 (0.003)**	-0.40 (0.324)	-0.10 (0.461)
15q13	*HERC2*	rs4778249	-	-	-	-	-	-
15q21	*SLC24A5*	rs1426654	-	-	-	-	-	-
16q24	*MC1R*	rs885479	0.29 (0.029)	-0.14 (0.111)	0.25 (0.017)	-0.19 (0.113)	**1.97 (0.009)**	**-0.29 (0.033)**
19p13	*MFSD12*	rs2240751	0.49 (0.016)	-0.14 (0.116)	-0.53 (0.172)	-0.05 (0.692)	**2.32 (0.004)**	**-0.28 (0.047)**

6.4.3 *MFSD12* is a novel candidate gene for the convergent evolution of lighter skin pigmentation in East Asians

Considering the evidence for selection in the *MFSD12* region, I estimated the time since the start of selection (T) and the selection coefficient (s) for this region using an Approximate Bayesian Computation (ABC) approach in the CHB population from the 1KG Project (Figure 6.3). I obtained a median estimate for the selection coefficient of 1.15% (95% Credible Interval 0.08%—4.4%) and a median age for the start of selection of 10,834 ya (95% Credible Interval of 5,266—33,801 ya). Additionally, I also estimated the joint posterior distributions of the starting time of selection (T) and the selection coefficient (s) (Figure D.15). The joint maximum a posteriori (MAP) for the selection coefficient (s) was equal to 0.139 and for the starting time of selection (T) was equal to 8,508 ya, similar to the marginal posterior distributions obtained for each parameter. These results, thus further suggest that *MFSD12* has been under selection in East Asians, and that the the start of selection probably occurred long after the split from the ancestral population shared with Europeans.

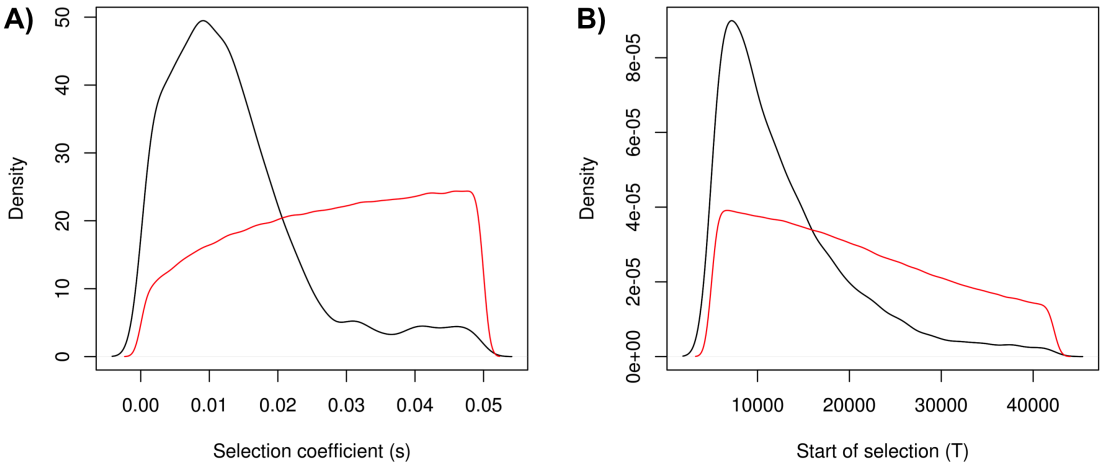

Figure 6.3: Estimation of the start of selection and selection coefficient at *MFSD12* gene region. Prior and posterior distributions are represented as red and black line, respectively. Note that the priors are not uniformly distributed as I only used simulations where the selected allele was still present at the end of the simulation. From Adhikari & Mendoza-Revilla et al. (2018).

6.5 Discussion

Overall the analyses presented in this chapter support the convergent evolution of lighter skin pigmentation in Eurasia. Whilst an initial selection event for lighter skin pigmentation probably happened in the common ancestral population shared by Europeans and East Asians, it seems that the lightening of skin pigmentation also occurred independently in these populations after their divergence. In Europe, functionally important derived variants in the gene regions *SLC45A2*, *TYR/GRM5*, *OCA2/HERC2* and *SLC24A5* seem to have contributed to the evolution of lighter skin pigmentation, whereas in East Asia, func-

tionally variants in the genes regions *OCA2*, *MC1R* and *MFSD12* seem to have contributed to lighter skin pigmentation. It is important to note that the gene regions assessed here are a follow-up investigation of the skin pigmentation associated loci described in Chapter 5 and therefore do not represent all known skin pigmentation associated loci. Nonetheless, many of the associated variants and gene regions analysed here collectively contribute to large variation in skin pigmentation between Europeans and East Asians, and therefore can be used to explore the convergent evolution of this trait.

Consistent with the adaptive importance of skin pigmentation in human populations, the associated genomic regions showed strong signals of selection in European and East Asian populations for at least one of the selection statistics (Figure D.1 — D.13 and Table 6.1). An idealized scenario of positive selection would result in high values of iHS and PBS and significantly negative Tajima's D. Encouragingly, these was observed at the *OCA2/HERC2* gene region (Figure D.1) in both populations. Possible reasons why selection signals were mostly exclusive to one selection statistic however, may be related to the power of each selection statistic, which is affected by the nature of the selection event, such as the age and strength of the selected variant, as well as the pre-existing variation in the genomic region (Sabeti et al., 2006; Oleksyk et al., 2010). To my knowledge, this is the first time that *MFSD12* has been reported to be under positive selection in East Asians and I therefore explored the evolutionary history of this gene region.

The correlation between the derived allele frequency of the associated variant at *MFSD12* and solar radiation suggests that the distribution of this variant has been (at least partly) shaped by solar radiation in Eastern Eurasia (Figure 6.4B and C). This result is in agreement with previous reports showing a correlation between the allele frequencies at certain skin pigmentation associated loci and solar radiation (Coop et al., 2009; Hancock et al., 2011). Although rs2240751 represents a strong candidate locus, I cannot discard that selection targeted another variant. Indeed, the PBS analysis at this region shows that many other variants (possibly in high LD with rs2240751) may also have been targeted by strong selection (Figure 6.4A).

The inference of the start of selection and selection coefficient using ABC contained large confidence intervals and as such, point estimates should be taken with caution. The point estimate for the time of selection at *MFSD12* was estimated to be around 10,834 ya. Interestingly, two other SNPs in *OCA2* that confer lighter skin pigmentation exclusively in East Asia, are thought to have evolved around the same time. Murray et al. (2015) estimated the age of the derived variant of rs74653330 and rs1800414 to have evolved around 7,000 and 6,000 ya, respectively. The estimated date for rs1800414 however, is slightly younger than a previous estimate at around 11,000 ya (Chen et al., 2015). Although, these studies did not estimate the onset of selection of these variants, the younger age of these alleles implies that the selection event at *OCA2* in East Asians probably occurred long after their split from Europeans. It would be interesting to test whether the age and onset of selection of rs885479 variant at *MC1R*, that also confers lighter skin pigmentation

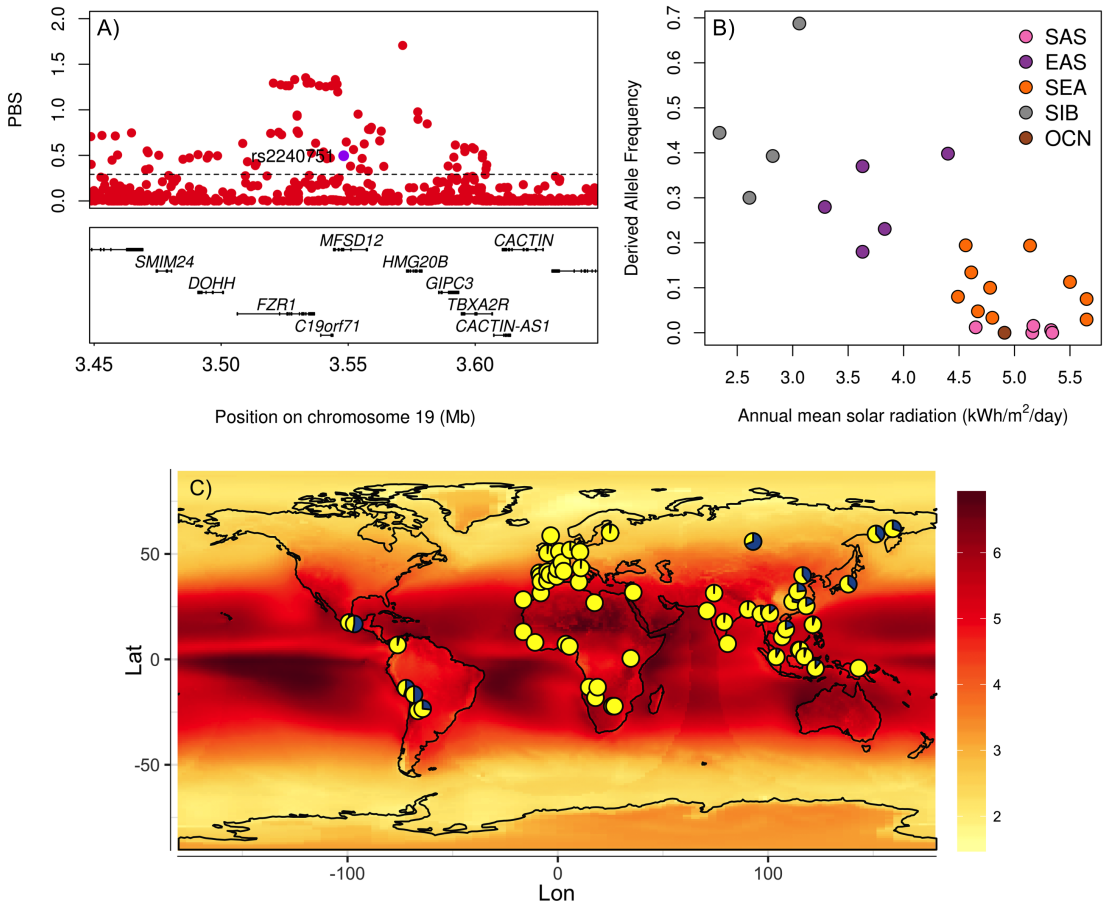

Figure 6.4: Evidence for selection in the *MFSD12* gene region. A) PBS scores in the 1000 Genomes CHB sample for SNPs across the region (index SNP rs2240751 is highlighted in purple and the horizontal line represents the 95th percentile threshold). B) Plot of the derived allele frequency at rs2240751 against mean annual solar radiation in Easter Eurasian populations. C) Allele frequencies at rs2240751 in 64 indigenous populations from across the world, mapped onto solar radiation. Pies charts are centered at the approximate geographic location of each population with the derived allele frequency represented in blue. From Adhikari & Mendoza-Revilla et al. (2018).

in East Asians, is similar to these estimated dates. In European populations, selection of *SLC45A2* and *SLC24A5* have been estimated to have occurred within the last 11,000 to 19,000 ya (Beleza et al., 2013b). These more recent dates have been interpreted as a refinement of the vitamin D hypothesis for lighter skin pigmentation (Mathieson et al., 2015; Stoneking, 2016). It has been suggested that perhaps the varied diet of early hunter-gatherers would have provided a sufficient supply of vitamin D, and that only after the Neolithic revolution, that was coupled with a poorer diet and increased use of clothes and shelter, insufficient vitamin D would have become an important selective pressure (Mathieson et al., 2015; Stoneking, 2016). Interestingly, ancient DNA studies in early Europeans have provided further insight into this hypothesis. While the derived allele at *SLC45A2*, that is associated to lighter skin pigmentation, is absent in European hunter-gatherers, this variant was found in larger frequency in Anatolian farmers from Europe (Mathieson et al., 2015). Given that the estimates for lighter skin pigmentation in East Asia have also probably occurred after the advent of agriculture in East Asia (Higham, 2002), it would be interesting to test whether a similar scenario prompted selection for lighter skin in Eastern Eurasia. Studies on ancient DNA would be needed to address this hypothesis.

The estimate of the selection coefficient for *MFSD12* is best viewed in the context of estimates for other pigmentation loci. Beleza et al. (2013) estimated the selection coefficient of *KITLG* (rs642742 G allele) in Europe and East Asia to be 0.02, whereas the coefficients for *SLC45A2* (rs16891982 G allele) and *SLC24A5* (rs1426654 A allele) were estimated to 0.04 and 0.08, respectively. López et al. (2014) estimated the selection coefficient of *SLC45A2* (rs16891982 G allele) to 0.01 to 0.02 in a Southern European population. Similarly, using an ancient DNA forward simulation approach restricted to European populations, Wilde et al. (2014) estimated the selection coefficient of *SLC45A2* (rs16891982 G allele), *TYR* (rs1042602 A allele) and *HERC2* (rs12913832 G allele) as 0.03, 0.026 and 0.036, respectively. The selection coefficient that I estimated for *MFSD12* thus lies at the lower end of those estimated for other pigmentation genes that appear to have been under selection. This result is in line with the relatively weaker phenotypic effect of *MFSD12*, relative to genes such as *SLC45A2* and *SLC24A5* (Figure C.6).

6.6 Summary

In this chapter I provided evidence that a novel variant at *MFSD12* probably played a role in shaping lighter skin pigmentation in East Asians but not in Europeans. I further showed that the distribution of the derived allele frequency of this variant seemed to have been affected by the solar radiation intensity in East Asia, supporting the role of natural selection in shaping skin pigmentation variation. Finally, I inferred that *MFSD12* was under selection in East Asians probably after their split from Europeans.

Chapter 7

Conclusions

In this thesis I have provided new insights into human adaptive history in the Americas and discovered novel variants associated to pigmentation phenotypes trough a GWAS in a large sample of admixed Latin Americans.

In Chapter 1 I described what is currently known about the demographic and adaptive history Native and admixed Latin American populations. I also described the evolutionary history of pigmentation variation in humans and the key genetic factors influencing this phenotypic trait. Finally, I outlined the rationale and scientific basis of GWAS and highlighted the significance of including underrepresented populations in genetic association studies.

In Chapter 2 I described relevant methods to detect signatures of selection using genome-wide data including allele frequency differentiation and haplotype-based methods. I also described methods that use external data such as environmental data in order to understand the potential adaptive pressures driving the selection signal, as well as methods that use genetic association data to detect instances of polygenic adaptation. In addition, I described the methodological aspects of GWAS and different methods used to account for population structure.

In Chapter 3 I conducted a genome-wide scan of selection in Native Americans and provided important candidate genes that I hypothesized were likely beneficial in the ancentral population of Native Americans in Beringia, prior to their entry into the American continent, particularly in relation to diet. I also showed that some of the top selected variants are found in several Arctic populations, consistent with a shared adaptive event. In this chapter I also explored instances of local adaptation in Native Americans using a large sample of admixed Latin Americans from the CANDELA cohort that derive most of their ancestry from distinct Native American groups. Among the strongest candidate regions of selection are immune-related genes that probably resulted from an adaptation to local pathogens in the Americas or to diseases brought after European contact. I also reported selection signals at genes with an important adaptive interest that have been previously reported in other Native American populations and other human populations highlighting the utility of this approach.

In Chapter 4 I conducted a genome-wide scan of selection post-admixture in five Latin American populations from the CANDELA sample. I presented a novel statistical model aimed at detecting signals of selection post-admixture by identifying larger than expected changes in allele frequency as expected from an admixture event. I showed that there was

a strong selective event in the Peruvian sample at a genomic region associated with glucose metabolism. However, I showed that it is likely that this signal was driven by the use of inaccurate Native American ancestral reference populations and that the selection signal was likely driven by a selection event that occurred in the Native American population that contributed most of the Native American ancestry to the Peruvian sample. In addition, I also detected a strong increase in African ancestry at the MHC locus in the Chilean and Mexican populations. The genes at MHC involved in infectious disease resistance might have been selected due to diseases brought from the Old World after European contact.

In Chapter 5 I reported novel variants associated to skin and eye pigmentation in admixed Latin Americans. The novel reported genomic regions show evidence of being related to various aspects of pigmentation biology and, as such, represent important candidate genes that should be followed up by functional analysis. More generally, these results also highlighted the complex genetic architecture of pigmentation variation in Latin Americans as shown through the independent genetic effect of different genetic variants at different genomic regions as well as within genomic regions. The results also demonstrated the greater statistical power obtained by using sensible continuous color models compared to ordinal categories that represent an oversimplification of the truly continuous nature of human pigmentation variation. Finally, I also reported a novel variant associated to skin pigmentation in the *MFSD12* gene, which represents a potential East Asian and Native American specific skin pigmentation locus.

In Chapter 6 I investigated the convergent evolution of lighter skin pigmentation in western and eastern Eurasians by applying different selection statistics at the skin pigmentation-associated loci reported in Chapter 5. I reported strong signals of selection at most of the pigmentation-associated regions, with one of the strongest signals among the novel loci being observed in the *MFSD12* gene region. I showed that the present geographic distribution of the novel SNP at *MFSD12* has probably been influenced by exposure to solar radiation in eastern Eurasia. In addition, I also showed that the genomic diversity at this locus is compatible with a scenario of convergent adaptation for lighter skin pigmentation in east Asia, and that this selection event probably occurred long after the divergence from Europeans.

7.1 Future directions and significance in studies of human adaptation

Determining how much of the genome is influenced by adaptive selection, and in turn identifying the particular genomic regions targeted by selection has long been and is still of great interest in the field of populations genetics (Haldane, 1957; Kimura et al., 1968; Smith, 1968; Felsenstein, 1971; Kimura and Ohta, 1971; Kreitman, 1983; Tajima, 1989; McDonald and Kreitman, 1991; Wall et al., 2002; Fay et al., 2002; Sabeti et al., 2002b; Bustamante et al., 2005; Sabeti et al., 2007; Pritchard et al., 2010; Hernandez et al., 2011b; Messer and Petrov, 2013; Sheehan and Song, 2016b; Schrider and Kern, 2017b). The explosion of genomic data in the past few decades has allowed the field to swiftly move from a mainly theoretical field (with sparse data) towards a hypothesis-generating

field, driven through empirical observation using large scale genomic data (Schrider and Kern, 2018; Kern and Hahn, 2018). Here, I focus on the future direction of studies aimed at detecting selection, particularly in humans, in the light of the vast volume of genetic data available today. Importantly, establishing a complete picture of human adaptation will remain challenging for (at least) three reasons: i) robustly identifying genomic regions under selection as well as the putative selected alleles, ii) establishing the phenotypes that selection is acting upon and iii) determining the environmental pressure driving inferred selection. I will address each of these issues in turn.

Identifying genomic regions under selection

Given the ever-increasing availability of modern human genomic data, there has been increased focus on using and developing methods that exploit the high-dimensionality of genomic data to detect signals of adaptive selection, notable amongst these, machine learning (ML) approaches (Pavlidis et al., 2010; Lin et al., 2011a; Ronen et al., 2013; Pybus et al., 2015; Schrider and Kern, 2016; Sheehan and Song, 2016a; Flagel et al., 2018; Sugden et al., 2018). Importantly, because the majority of ML applications have been through supervised learning (i.e. when data for the true response value [or label] is known), training sets have been exclusively generated through simulations (Schrider and Kern, 2018). Thus, a detailed description of human demographic history will remain of fundamental importance to identifying loci under selection and pinpointing when and in which populations selection occurred. In this regard, recent methods that can jointly estimate selection and demography (Li and Stephan, 2006; Sheehan and Song, 2016b) are an important way forward, as well as selection methods that are robust to demographic misspecifications (Schrider and Kern, 2016). Additionally, not only the increasing amount of genomic data, but approaches that exploit new types of genomic data — for example, those including ancient DNA (aDNA) — are likely to provide a more accurate view of the mode and tempo of adaptive selection. For the most part, genomic signatures of adaptive selection have been usually obtained from modern genomic data, and as such, represent an indirect approach to detect past selection events. By including past allele frequency estimates, obtained from ancient human populations, new studies have been able to detect adaptive selection by analysing samples from populations before and after an adaptive event, including the putative selected alleles (Mathieson et al., 2015), make direct estimation of selection coefficients for loci involved in a particular adaptive phenotypes of interest, such as pigmentation (Wilde et al., 2014), and to determine the influence of positive selective sweeps in the evolutionary history of different human populations (Key et al., 2016). In addition, recent studies based on comparison of modern humans with archaic hominins have found evidence of adaptive introgression in several genes (Huerta-Sánchez et al., 2014; Vernot and Akey, 2014; Sankararaman et al., 2014; Racimo et al., 2015; Deschamps et al., 2016; Vernot et al., 2016; Dannemann et al., 2017; Dannemann and Kelso, 2017; Racimo et al., 2017; Browning et al., 2018), with future studies likely providing a better view of the origin and evolution of these putative adaptive variants (Wolf and Akey, 2018). Given the ever-increasing amount of aDNA constantly being generated (Callaway, 2018),

coupled with improvements in sequencing technologies to allow for the retrieval of DNA from remains up to several thousand of years old (Orlando et al., 2013), aDNA studies are likely to become ever more important contributors to studies of human adaptation. Finally, there is now a growing realization that other types of mutations besides SNPs, may be underlying adaptive events. Examples include instances of selection at copy-number variants in humans (Perry et al., 2007; Schrider et al., 2013) and insertion of transposable elements (Daborn et al., 2002; Schlenke and Begun, 2004; González et al., 2008), and even large inversions (Kolaczkowski et al., 2011; Cheng et al., 2012; Kirkpatrick and Kern, 2012; Reinhardt et al., 2014) in other organisms. New sequencing technologies that enable long-range haplotype retrieval would constitute a major important advance to elucidating the role of other type of variants in human adaptation.

Establishing the phenotypes that selection has acted upon

Previous studies of adaptive selection have produced an extensive list of putatively selected genomic regions, genes and variants. However, this extensive list is in sharp contrast with the few functionally validated examples of genetic adaptation with a strong candidate genomic region and a convincing explanation for the adaptive phenotype (Pritchard et al., 2010; Hernandez et al., 2011a; Rodríguez et al., 2014; Brown, 2012; Pavlidis et al., 2012; Fan et al., 2016). Recent studies that make use of the large amount of GWAS data for many quantitative phenotypes and diseases are starting to assess whether the genetic variants associated with a phenotype have been under selection (Berg et al., 2017; Guo et al., 2018; Racimo et al., 2018a). Large sampling efforts, such as the UK BioBank, which includes detailed information on many traits including lifestyle, diet and environmental exposures, are likely to greatly contribute to these type of studies of human adaptation. Additionally, novel methods that allow indentification of the putative causal variant (also known as fine-mapping) are of increasing importance as they will likely discover biologically meaningful variation (Akbari et al., 2018; Szpak et al., 2018). Nonetheless, establishing a case for adaptation will ultimately necessitate a combination of genomic and functional evidence. Necessarily, the focus of studies of adaptation will be required to move from candidate variant discovery to fine-mapping of the candidate genomic region, and most importantly, to the biological understanding of their adaptive significance through functional validation. Importantly, further breakthroughs from genomic annotation and genomic manipulation technology (e.g. through CRISPR/Cas9 technology (Jinek et al., 2012)) are likely to become one of the biggest contributors to produce more compelling explanation for adaptation of a relevant phenotype in humans.

Determining the environmental pressure driving the selection

Perhaps the most challenging aspect to establishing a complete picture of human adaptation will be to identifying the environmental pressures driving selection. Approaches that correlate allele-frequency data with environmental variables or other variables known to cause a strong selective pressure (such as pathogens) are likely to greatly contribute

to our understanding of the selective pressures imposed on humans (Coop et al., 2009; Fumagalli et al., 2011; Günther and Coop, 2013). Nonetheless, as vast amounts of aDNA over broad geographic and temporal scales increases, approaches that integrate genomic samples from populations before, during and after adaptive events, are likely to better elucidate key selective pressures in human populations.

7.2 Future directions and significance in GWAS

Over the last years, GWAS have delivered a remarkable set of discoveries in human genetics (Visscher et al., 2017). Undoubtedly GWAS will remain an important experimental design in human genetic studies that will increase our knowledge of complex phenotypes, the biological underpinnings of disease, and help towards translational outcomes for diseases treatment and prevention. In this section I address some of the current and future challenges of GWAS and their future directions. I do not attempt to summarise every challenge, but rather draw out some important themes.

Ensure greater sample diversity in GWAS

A recurrent and worrying challenge is to incorporate a much broader ensemble of human populations in GWAS. The majority of GWAS so far have been carried out in European derived populations, and the only significant growth in non-European populations in recent years has been due to an increase in the number of samples of Asian origins (Need and Goldstein, 2009; Bustamante et al., 2011; Petrovski and Goldstein, 2016; Popejoy and Fullerton, 2016). Poor representation of different populations will fail to capture the full genetic diversity of human populations and will therefore provide only a biased view of which variants are relevant for future genomic medicine studies. Studying underrepresented populations are likely to ensure novel findings and provide new insights into the genetic architecture of complex phenotypes, and thus will represent a cost-effective approach. This will be especially relevant in Africans, as these populations have more genetic and often more phenotypic diversity that increases statistical power. Moreover, lower LD genomic blocks in Africans means there is a greater chance for better fine-mapping resolution to pinpoint casual variants (Martin et al., 2018). In addition, recent studies have also questioned the transferability of existing study results to other human populations, particularly regarding the construction of Polygenic Risk Scores (PRS) (Vilhjálmsson et al., 2015; Martin et al., 2017a; Akiyama et al., 2017; Li et al., 2017; Ware et al., 2017). Importantly, this bias may hinder the development of precision medicine and thus could not benefit the populations which currently have the greatest health disparities (Bustamante et al., 2011). While developing new methods that will help even the transferability of PRS, for example by considering LD within and between populations as shown by Vilhjálmsson et al. (2015) represents an important way forward, the best way to even out genetic prediction power in human populations will be by producing similar-sized GWAS of relevant phenotypes in these populations.

Biobanking

Large genomic biobanks are also, and will become, an ever more important component for GWAS. By amassing detailed information on many traits including lifestyle, diet and environmental exposure, together with genomic data, biobanks look set to better establish the heritable component of common and complex phenotypes and diseases. Importantly, for diseases with very low prevalence, the discovery of variants is limited by the sample size and thus, large-scale biobanks will likely play a crucial role in detecting these type of variants, given that it takes many years to accumulate sample sizes big enough to ensure high enough statistical power. Further, as sample sizes increase, it is expected that variants with smaller effects will also be discovered. Additionally, given the amount of detailed information on environmental data, new studies based on biobank data are likely to shed further light on the interactions between genetic variants and many different environmental risk factors. Importantly, the wide range of ancestries of the individuals in these large cohorts will likely emerge as a recurrent challenge, as well as as optimizing statistical power and computational efficiency on biobank-based GWAS (Bycroft et al., 2017; Loh et al., 2018).

The value of sequencing data

The majority of current GWAS are still largely based on SNP arrays and complemented by genomic imputation. Although this methodology has proved successful, it is becoming clear that many individual populations carry their own rare variants that can only be reliable imputed by combining reference samples with sequence data from these local populations. This has been recently exemplified by using population-specific reference in an association study of Anabaptist (Hou et al., 2017) and Ashkenazi Jewish populations (Rivas et al., 2018). Nonetheless, it is clear that ultimately Whole Exome Sequencing (WES) and Whole Genome Sequencing (WGS) techonologies of large cohorts will be routinely created and used for GWAS. WES association studies will likely become a popular approach in human genetics, especially for testing for association of rare coding variants that are not queried in SNP arrays. This will also be especially important for Mendelian disorders as these are underpinned by variation at the coding level. The proteins affected by these rare mutations could provide important potential drug targets. Additionally, WES and WGS will not only be important for the discovery of novel loci, but also to fine-map variants at novel and previously reported loci. An important additional challenge for sequence-based GWAS will be to step up computational resources.

Integration of data

Finally, perhaps the greatest efforts will have to be into discovering the biological function of the many thousands of variants identified by past and new GWAS. Integrating the results from the associated variants with functional genomic data from relevant tissues and cell types at multi-omics levels (i.e. at the transcriptome, proteome and epigenome

levels), has the potential to lead to a refined understanding of the biological mechanisms underpinning complex phenotypes including disease etiology. A transition from initial genomic region discovery, to fine-mapping of casual variants and experimental functional validation, will be needed using a combination of in silico, in vitro and in vivo methods to ultimately further the translational path, and in the case of diseases, to potentially novel therapeutics and prevention strategies.

Bibliography

1000 Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A. et al. A global reference for human genetic variation. *Nature*, 526(7571):68–74, Oct 2015. doi: 10.1038/nature15393.

Abe, Y., Tamiya, G., Nakamura, T., Hozumi, Y. and Suzuki, T. Association of melanogenesis genes with skin color variation among japanese females. *J Dermatol Sci*, 69(2): 167–72, Feb 2013. doi: 10.1016/j.jdermsci.2012.10.016.

Abernathy, C.O., Thomas, D.J. and Calderon, R.L. Health effects and risk assessment of arsenic. *The Journal of nutrition*, 133(5):1536S–1538S, 2003.

Acuña-Alonzo, V., Flores-Dorantes, T., Kruit, J.K., Villarreal-Molina, T., Arellano-Campos, O., Hünemeier, T., Moreno-Estrada, A., Ortiz-López, M.G., Villamil-Ramírez, H., León-Mimila, P. et al. A functional abca1 gene variant is associated with low hdl-cholesterol levels and shows evidence of positive selection in native americans. *Hum Mol Genet*, 19(14):2877–85, Jul 2010. doi: 10.1093/hmg/ddq173.

Acuna-Soto, R., Stahle, D.W., Therrell, M.D., Griffin, R.D. and Cleaveland, M.K. When half of the population died: the epidemic of hemorrhagic fevers of 1576 in mexico. *FEMS microbiology letters*, 240(1):1–5, 2004.

Adhikari, K., Reales, G., Smith, A.J.P., Konka, E., Palmen, J., Quinto-Sanchez, M., Acuña-Alonzo, V., Jaramillo, C., Arias, W., Fuentes, M. et al. A genome-wide association study identifies multiple loci for variation in human ear morphology. *Nat Commun*, 6:7500, Jun 2015. doi: 10.1038/ncomms8500.

Adhikari, K., Fontanil, T., Cal, S., Mendoza-Revilla, J., Fuentes-Guajardo, M., Chacón-Duque, J.C., Al-Saadi, F., Johansson, J.A., Quinto-Sanchez, M., Acuña-Alonzo, V. et al. A genome-wide association scan in admixed latin americans identifies loci influencing facial and scalp hair features. *Nat Commun*, 7:10815, Mar 2016a. doi: 10.1038/ncomms10815.

Adhikari, K., Fuentes-Guajardo, M., Quinto-Sánchez, M., Mendoza-Revilla, J., Camilo Chacón-Duque, J., Acuña-Alonzo, V., Jaramillo, C., Arias, W., Lozano, R.B., Pérez, G.M. et al. A genome-wide association scan implicates dchs2, runx2, gli3, pax1 and edar in human facial variation. *Nat Commun*, 7:11616, May 2016b. doi: 10.1038/ncomms11616.

Adhikari, K., Chacón-Duque, J.C., Mendoza-Revilla, J., Fuentes-Guajardo, M. and Ruiz-Linares, A. The genetic diversity of the americas. *Annu Rev Genomics Hum Genet*, 18: 277–296, Aug 2017. doi: 10.1146/annurev-genom-083115-022331.

Agius, A., Sultana, R., Camenzuli, C., Calleja-Agius, J. and Balzan, R. An update on the genetics of pre-eclampsia. *Minerva Ginecol*, Oct 2017. doi: 10.23736/S0026-4784.17.04150-8.

Aguilar Salinas, C.A., Cruz-Bautista, I., Mehta, R., Villarreal-Molina, M.T., Pérez, F.J., Tusié-Luna, M.T. and Canizales-Quinteros, S. The atp-binding cassette transporter subfamily a member 1 (abc-a1) and type 2 diabetes: an association beyond hdl cholesterol. *Current diabetes reviews*, 3(4):264–267, 2007.

Ainger, S.A., Jagirdar, K., Lee, K.J., Soyer, H.P. and Sturm, R.A. Skin pigmentation genetics for the clinic. *Dermatology*, 233(1):1–15, 2017. doi: 10.1159/000468538.

Akbari, A., Vitti, J.J., Iranmehr, A., Bakhtiari, M., Sabeti, P.C., Mirarab, S. and Bafna, V. Identifying the favored mutation in a positive selective sweep. *Nature methods*, 15 (4):279, 2018.

Akiyama, M., Okada, Y., Kanai, M., Takahashi, A., Momozawa, Y., Ikeda, M., Iwata, N., Ikegawa, S., Hirata, M., Matsuda, K. et al. Genome-wide association study identifies 112 new loci for body mass index in the japanese population. *Nat Genet*, 49(10):1458–1467, Oct 2017. doi: 10.1038/ng.3951.

Alaluf, S., Atkins, D., Barrett, K., Blount, M., Carter, N. and Heath, A. The impact of epidermal melanin on objective measurements of human skin colour. *Pigment Cell & Melanoma Research*, 15(2):119–126, 2002.

Alarcón-Riquelme, M.E., Ziegler, J.T., Molineros, J., Howard, T.D., Moreno-Estrada, A., Sánchez-Rodríguez, E., Ainsworth, H.C., Ortiz-Tello, P., Comeau, M.E., Rasmussen, A. et al. Genome-wide association study in an amerindian ancestry population reveals novel systemic lupus erythematosus risk loci and the role of european admixture. *Arthritis & rheumatology*, 68(4):932–943, 2016.

Alexander, D.H., Novembre, J. and Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res*, 19(9):1655–64, Sep 2009. doi: 10.1101/gr.094052.109.

Allentoft, M.E., Sikora, M., Sjögren, K.G., Rasmussen, S., Rasmussen, M., Stenderup, J., Damgaard, P.B., Schroeder, H., Ahlström, T., Vinner, L. et al. Population genomics of bronze age eurasia. *Nature*, 522(7555):167–72, Jun 2015. doi: 10.1038/nature14507.

Alonso, S., Izagirre, N., Smith-Zubiaga, I., Gardeazabal, J., Díaz-Ramón, J.L., Díaz-Pérez, J.L., Zelenika, D., Boyano, M.D., Smit, N. and De la Rúa, C. Complex signatures of selection for the melanogenic loci tyr, tyrp1 and dct in humans. *BMC evolutionary biology*, 8(1):74, 2008.

Amorim, C.E., Nunes, K., Meyer, D., Comas, D., Bortolini, M.C., Salzano, F.M. and Hünemeier, T. Genetic signature of natural selection in first americans. *Proc Natl Acad Sci U S A*, 114(9):2195–2199, 02 2017. doi: 10.1073/pnas.1620541114.

Amorim, C.E.G., Daub, J.T., Salzano, F.M., Foll, M. and Excoffier, L. Detection of convergent genome-wide signals of adaptation to tropical forests in humans. *PLoS One*, 10(4):e0121557, 2015. doi: 10.1371/journal.pone.0121557.

Amundsen, T. and Forsgren, E. Male mate choice selects for female coloration in a fish. *Proceedings of the National Academy of Sciences*, 98(23):13155–13160, 2001.

Amyere, M., Vogt, T., Hoo, J., Brandrup, F., Bygum, A., Boon, L. and Vikkula, M. Kitlg mutations cause familial progressive hyper- and hypopigmentation. *J Invest Dermatol*, 131(6):1234–9, Jun 2011. doi: 10.1038/jid.2011.29.

Andersen, J.D., Johansen, P., Harder, S., Christoffersen, S.R., Delgado, M.C., Henriksen, S.T., Nielsen, M.M., Sørensen, E., Ullum, H., Hansen, T. et al. Genetic analyses of the human eye colours using a novel objective method for eye colour classification. *Forensic Sci Int Genet*, 7(5):508–15, Sep 2013. doi: 10.1016/j.fsigen.2013.05.003.

Apata, M., Arriaza, B., Llop, E. and Moraga, M. Human adaptation to arsenic in andean populations of the atacama desert. *Am J Phys Anthropol*, 163(1):192–199, 05 2017. doi: 10.1002/ajpa.23193.

Asgari, M.M., Wang, W., Ioannidis, N.M., Itnyre, J., Hoffmann, T., Jorgenson, E. and Whittemore, A.S. Identification of susceptibility loci for cutaneous squamous cell carcinoma. *J Invest Dermatol*, 136(5):930–7, 05 2016. doi: 10.1016/j.jid.2016.01.013.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T. et al. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, 25(1):25–9, May 2000. doi: 10.1038/75556.

Astle, W., Balding, D.J. et al. Population structure and cryptic relatedness in genetic association studies. *Statistical Science*, 24(4):451–471, 2009.

Augusto, D.G., Hollenbach, J.A. and Petzl-Erler, M.L. A deep look at kir–hla in amerindians: comprehensive meta-analysis reveals limited diversity of kir haplotypes. *Human immunology*, 76(4):272–280, 2015.

Azzedine, H., Zavadakova, P., Planté-Bordeneuve, V., Vaz Pato, M., Pinto, N., Bartesaghi, L., Zenker, J., Poirot, O., Bernard-Marissal, N., Arnaud Gouttenoire, E. et al. Plekhg5 deficiency leads to an intermediate form of autosomal-recessive charcot-marie-tooth disease. *Hum Mol Genet*, 22(20):4224–32, Oct 2013. doi: 10.1093/hmg/ddt274.

Baik, I., Cho, N.H., Kim, S.H., Han, B.G. and Shin, C. Genome-wide association studies identify genetic loci related to alcohol consumption in korean men. *Am J Clin Nutr*, 93(4):809–16, Apr 2011. doi: 10.3945/ajcn.110.001776.

Baily, S.L. and Miguez, E.J. *Mass migration to modern Latin America*. Rowman & Littlefield Publishers, 2003.

Balding, D.J. A tutorial on statistical methods for population association studies. *Nature Reviews Genetics*, 7(10):781, 2006.

Balloux, F., Handley, L.J.L., Jombart, T., Liu, H. and Manica, A. Climate shaped the worldwide distribution of human mitochondrial dna sequence variation. *Proceedings of the Royal Society of London B: Biological Sciences*, 276(1672):3447–3455, 2009.

Banerjee, M., Banerjee, N., Bhattacharjee, P., Mondal, D., Lythgoe, P.R., Martínez, M., Pan, J., Polya, D.A. and Giri, A.K. High arsenic in rice is associated with elevated genotoxic effects in humans. *Scientific reports*, 3:2195, 2013.

Baran, Y., Pasaniuc, B., Sankararaman, S., Torgerson, D.G., Gignoux, C., Eng, C., Rodriguez-Cintron, W., Chapela, R., Ford, J.G., Avila, P.C. et al. Fast and accurate inference of local ancestry in latino populations. *Bioinformatics*, 28(10):1359–67, May 2012. doi: 10.1093/bioinformatics/bts144.

Barrett, J.H., Iles, M.M., Harland, M., Taylor, J.C., Aitken, J.F., Andresen, P.A., Akslen, L.A., Armstrong, B.K., Avril, M.F., Azizi, E. et al. Genome-wide association study identifies three new melanoma susceptibility loci. *Nat Genet*, 43(11):1108–13, Oct 2011. doi: 10.1038/ng.959.

Barsh, G.S. What controls variation in human skin color? *PLoS biology*, 1(1):e27, 2003.

Bastiaens, M.T., ter Huurne, J.A., Kielich, C., Gruis, N.A., Westendorp, R.G., Vermeer, B.J., Bavinck, J.N. and Leiden Skin Cancer Study Team. Melanocortin-1 receptor gene variants determine the risk of nonmelanoma skin cancer independently of fair skin and red hair. *Am J Hum Genet*, 68(4):884–94, Apr 2001.

Basu, A., Tang, H., Zhu, X., Gu, C.C., Hanis, C., Boerwinkle, E. and Risch, N. Genome-wide distribution of ancestry in mexican americans. *Hum Genet*, 124(3):207–14, Oct 2008. doi: 10.1007/s00439-008-0541-5.

Battaglia, V., Grugni, V., Perego, U.A., Angerhofer, N., Gomez-Palmieri, J.E., Woodward, S.R., Achilli, A., Myres, N., Torroni, A. and Semino, O. The first peopling of south america: new evidence from y-chromosome haplogroup q. *PLoS One*, 8(8):e71390, 2013. doi: 10.1371/journal.pone.0071390.

Beall, C.M., Strohl, K.P., Blangero, J., Williams-Blangero, S., Almasy, L.A., Decker, M.J., Worthman, C.M., Goldstein, M.C., Vargas, E., Villena, M. et al. Ventilation and hypoxic ventilatory response of tibetan and aymara high altitude natives. *Am J Phys Anthropol*, 104(4):427–47, Dec 1997. doi: 10.1002/(SICI)1096-8644(199712)104:4⟨427::AID-AJPA1⟩3.0.CO;2-P.

Beaumont, M.A., Zhang, W. and Balding, D.J. Approximate bayesian computation in population genetics. *Genetics*, 162(4):2025–2035, 2002.

Bedoya, G., Montoya, P., García, J., Soto, I., Bourgeois, S., Carvajal, L., Labuda, D., Alvarez, V., Ospina, J., Hedrick, P.W. et al. Admixture dynamics in hispanics: a shift in the nuclear genetic ancestry of a south american population isolate. *Proceedings of the National Academy of Sciences*, 103(19):7234–7239, 2006.

Beleza, S., Johnson, N.A., Candille, S.I., Absher, D.M., Coram, M.A., Lopes, J., Campos, J., Araújo, I.I., Anderson, T.M., Vilhjálmsson, B.J. et al. Genetic architecture of skin and eye color in an african-european admixed population. *PLoS Genet*, 9(3):e1003372, Mar 2013a. doi: 10.1371/journal.pgen.1003372.

Beleza, S., Santos, A.M., McEvoy, B., Alves, I., Martinho, C., Cameron, E., Shriver, M.D., Parra, E.J. and Rocha, J. The timing of pigmentation lightening in europeans. *Mol Biol Evol*, 30(1):24–35, Jan 2013b. doi: 10.1093/molbev/mss207.

Bellono, N.W., Escobar, I.E., Lefkovith, A.J., Marks, M.S. and Oancea, E. An intracellular anion channel critical for pigmentation. *Elife*, 3:e04543, Dec 2014. doi: 10.7554/eLife.04543.

Bellwood, P., Gamble, C., Le Blanc, S.A., Pluciennik, M., Richards, M. and Terrell, J.E. First farmers: the origins of agricultural societies, by peter bellwood. malden (ma): Blackwell, 2005; isbn 0-631-20565-9 hardback£ 60; isbn 0-631-20566-7 paperback£ 17.99, xix+ 360 pp., 59 figs., 3 tables. *Cambridge archaeological journal*, 17(1): 87–109, 2007.

Bentham, J., Morris, D.L., Graham, D.S.C., Pinder, C.L., Tombleson, P., Behrens, T.W., Martín, J., Fairfax, B.P., Knight, J.C., Chen, L. et al. Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. *Nature genetics*, 47(12):1457, 2015.

Berg, J.J. and Coop, G. A population genetic signal of polygenic adaptation. *PLoS genetics*, 10(8):e1004412, 2014.

Berg, J.J., Zhang, X. and Coop, G. Polygenic adaptation has impacted multiple anthropometric traits. *bioRxiv*, 2017. doi: 10.1101/167551. URL `https://www.biorxiv.org/content/early/2017/08/02/167551`.

Berg, J.J., Harpak, A., Sinnott-Armstrong, N., Joergensen, A.M., Mostafavi, H., Field, Y., Boyle, E.A., Zhang, X., Racimo, F., Pritchard, J.K. et al. Reduced signal for polygenic adaptation of height in uk biobank. *bioRxiv*, 2018. doi: 10.1101/354951. URL `https://www.biorxiv.org/content/early/2018/06/27/354951`.

Bergey, C.M., Lopez, M., Harrison, G.F., Patin, E., Cohen, J., Quintana-Murci, L., Barreiro, L.B. and Perry, G.H. Polygenic adaptation and convergent evolution across both growth and cardiac genetic pathways in african and asian rainforest hunter-gatherers. *bioRxiv*, 2018. doi: 10.1101/300574. URL `https://www.biorxiv.org/content/early/2018/04/18/300574`.

Bersaglieri, T., Sabeti, P.C., Patterson, N., Vanderploeg, T., Schaffner, S.F., Drake, J.A., Rhodes, M., Reich, D.E. and Hirschhorn, J.N. Genetic signatures of strong recent positive selection at the lactase gene. *Am J Hum Genet*, 74(6):1111–20, Jun 2004. doi: 10.1086/421051.

Bhatia, G., Patterson, N., Pasaniuc, B., Zaitlen, N., Genovese, G., Pollack, S., Mallick, S., Myers, S., Tandon, A., Spencer, C. et al. Genome-wide comparison of african-ancestry populations from care and other cohorts reveals signals of natural selection. *Am J Hum Genet*, 89(3):368–81, Sep 2011. doi: 10.1016/j.ajhg.2011.07.025.

Bhatia, G., Patterson, N., Sankararaman, S. and Price, A.L. Estimating and interpreting fst: the impact of rare variants. *Genome Res*, 23(9):1514–21, Sep 2013. doi: 10.1101/gr.154831.113.

Bhatia, G., Tandon, A., Patterson, N., Aldrich, M.C., Ambrosone, C.B., Amos, C., Bandera, E.V., Berndt, S.I., Bernstein, L., Blot, W.J. et al. Genome-wide scan of 29,141 african americans finds no evidence of directional selection since admixture. *Am J Hum Genet*, 95(4):437–44, Oct 2014. doi: 10.1016/j.ajhg.2014.08.011.

Bhatia, K.K., Black, F.L., Smith, T.A., Prasad, M.L. and Koki, G.N. Class i hla antigens in two long-separated populations: Melanesians and south amerinds. *American journal of physical anthropology*, 97(3):291–305, 1995.

Bianchi, N.O., Catanesi, C.I., Bailliet, G., Martinez-Marignac, V.L., Bravi, C.M., Vidal-Rioja, L.B., Herrera, R.J. and López-Camelo, J.S. Characterization of ancestral and derived y-chromosome haplotypes of new world native populations. *Am J Hum Genet*, 63(6):1862–71, Dec 1998. doi: 10.1086/302141.

Bigham, A., Bauchet, M., Pinto, D., Mao, X., Akey, J.M., Mei, R., Scherer, S.W., Julian, C.G., Wilson, M.J., López Herráez, D. et al. Identifying signatures of natural selection in tibetan and andean populations using dense genome scan data. *PLoS Genet*, 6(9): e1001116, Sep 2010. doi: 10.1371/journal.pgen.1001116.

Bigham, A.W. Genetics of human origin and evolution: high-altitude adaptations. *Curr Opin Genet Dev*, 41:8–13, Dec 2016. doi: 10.1016/j.gde.2016.06.018.

Bigham, A.W., Mao, X., Mei, R., Brutsaert, T., Wilson, M.J., Julian, C.G., Parra, E.J., Akey, J.M., Moore, L.G. and Shriver, M.D. Identifying positive selection candidate loci for high-altitude adaptation in andean populations. *Hum Genomics*, 4(2):79–90, Dec 2009.

Bigham, A.W., Julian, C.G., Wilson, M.J., Vargas, E., Browne, V.A., Shriver, M.D. and Moore, L.G. Maternal prkaa1 and ednra genotypes are associated with birth weight, and prkaa1 with uterine artery diameter and metabolic homeostasis at high altitude. *Physiol Genomics*, 46(18):687–97, Sep 2014. doi: 10.1152/physiolgenomics.00063.2014.

Billaut-Laden, I., Rat, E., Allorge, D., Crunelle-Thibaut, A., Cauffiez, C., Chevalier, D., Lo-Guidice, J.M. and Broly, F. Evidence for a functional genetic polymorphism of

the human mercaptopyruvate sulfurtransferase (mpst), a cyanide detoxification enzyme. *Toxicol Lett*, 165(2):101–11, Aug 2006. doi: 10.1016/j.toxlet.2006.02.002.

Blum, M.G. and François, O. Non-linear regression models for approximate bayesian computation. *Statistics and Computing*, 20(1):63–73, 2010.

Bocquet-Appel, J.P. When the world's population took off: the springboard of the neolithic demographic transition. *Science*, 333(6042):560–1, Jul 2011. doi: 10.1126/science.1208880.

Bolognia, J.L. and Orlow, S.J. Melanocyte biology. *Dermatology*, 1:935–945, 2003.

Bonàs-Guarch, S., Guindo-Martínez, M., Miguel-Escalada, I., Grarup, N., Sebastian, D., Rodriguez-Fos, E., Sánchez, F., Planas-Fèlix, M., Cortes-Sánchez, P., González, S. et al. Re-analysis of public genetic data reveals a rare x-chromosomal variant associated with type 2 diabetes. *Nat Commun*, 9(1):321, 01 2018. doi: 10.1038/s41467-017-02380-9.

Bordogna, W., Hudson, J.D., Buddle, J., Bennett, D.C., Beach, D.H. and Carnero, A. Emx homeobox genes regulate microphthalmia and alter melanocyte biology. *Exp Cell Res*, 311(1):27–38, Nov 2005. doi: 10.1016/j.yexcr.2005.08.013.

Bortolini, M.C., Salzano, F.M., Thomas, M.G., Stuart, S., Nasanen, S.P.K., Bau, C.H.D., Hutz, M.H., Layrisse, Z., Petzl-Erler, M.L., Tsuneto, L.T. et al. Y-chromosome evidence for differing ancient demographic histories in the americas. *Am J Hum Genet*, 73(3):524–39, Sep 2003. doi: 10.1086/377588.

Bortolini, M.C., González-José, R., Bonatto, S.L. and Santos, F.R. Reconciling pre-columbian settlement hypotheses requires integrative, multidisciplinary, and model-bound approaches. *Proc Natl Acad Sci U S A*, 111(2):E213–4, Jan 2014. doi: 10.1073/pnas.1321197111.

Botstein, D. and Risch, N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat Genet*, 33 Suppl:228–37, Mar 2003. doi: 10.1038/ng1090.

Boulesteix, A.L. and Strimmer, K. Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Briefings in bioinformatics*, 8(1):32–44, 2006.

Boyle, E.A., Li, Y.I. and Pritchard, J.K. An expanded view of complex traits: From polygenic to omnigenic. *Cell*, 169(7):1177–1186, Jun 2017. doi: 10.1016/j.cell.2017.05.038.

Brace, S., Diekmann, Y., Booth, T.J., Faltyskova, Z., Rohland, N., Mallick, S., Ferry, M., Michel, M., Oppenheimer, J., Broomandkhoshbacht, N. et al. Population replacement in early neolithic britain. *bioRxiv*, 2018. doi: 10.1101/267443. URL `https://www.biorxiv.org/content/early/2018/02/18/267443`.

Bramble, D.M. and Lieberman, D.E. Endurance running and the evolution of homo. *Nature*, 432(7015):345, 2004.

Brandini, S., Bergamaschi, P., Cerna, M.F., Gandini, F., Bastaroli, F., Bertolini, E., Cereda, C., Ferretti, L., Gómez-Carballa, A., Battaglia, V. et al. The paleo-indian entry into south america according to mitogenomes. *Mol Biol Evol*, 35(2):299–311, Feb 2018. doi: 10.1093/molbev/msx267.

Branicki, W., Brudnik, U. and Wojas-Pelc, A. Interactions between herc2, oca2 and mc1r may influence human pigmentation phenotype. *Ann Hum Genet*, 73(2):160–70, Mar 2009. doi: 10.1111/j.1469-1809.2009.00504.x.

Bräuer, G. and Chopra, V. Estimation of the heritability of hair and eye color. *Anthropologischer Anzeiger; Bericht uber die biologisch-anthropologische Literatur*, 36(2): 109–120, 1978.

Brocker, C.N., Vasiliou, V. and Nebert, D.W. Evolutionary divergence and functions of the adam and adamts gene families. *Hum Genomics*, 4(1):43–55, Oct 2009.

Brown, E.A. Genetic explorations of recent human metabolic adaptations: hypotheses and evidence. *Biol Rev Camb Philos Soc*, 87(4):838–55, Nov 2012. doi: 10.1111/j.1469-185X.2012.00227.x.

Browning, S.R., Grinde, K., Plantinga, A., Gogarten, S.M., Stilp, A.M., Kaplan, R.C., Avilés-Santa, M.L., Browning, B.L. and Laurie, C.C. Local ancestry inference in a large us-based hispanic/latino study: Hispanic community health study/study of latinos (hchs/sol). *G3 (Bethesda)*, 6(6):1525–34, 06 2016. doi: 10.1534/g3.116.028779.

Browning, S.R., Browning, B.L., Zhou, Y., Tucci, S. and Akey, J.M. Analysis of human sequence data reveals two pulses of archaic denisovan admixture. *Cell*, 173(1):53–61.e9, Mar 2018. doi: 10.1016/j.cell.2018.02.031.

Bulik-Sullivan, B., Finucane, H.K., Anttila, V., Gusev, A., Day, F.R., Loh, P.R., ReproGen Consortium, Psychiatric Genomics Consortium, Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control Consortium 3, Duncan, L. et al. An atlas of genetic correlations across human diseases and traits. *Nat Genet*, 47(11):1236–41, Nov 2015a. doi: 10.1038/ng.3406.

Bulik-Sullivan, B.K., Loh, P.R., Finucane, H.K., Ripke, S., Yang, J., Schizophrenia Working Group of the Psychiatric Genomics Consortium, Patterson, N., Daly, M.J., Price, A.L. and Neale, B.M. Ld score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet*, 47(3):291–5, Mar 2015b. doi: 10.1038/ng.3211.

Bunger, M.K., Wilsbacher, L.D., Moran, S.M., Clendenin, C., Radcliffe, L.A., Hogenesch, J.B., Simon, M.C., Takahashi, J.S. and Bradfield, C.A. Mop3 is an essential component of the master circadian pacemaker in mammals. *Cell*, 103(7):1009–17, Dec 2000.

Busby, G., Christ, R., Band, G., Leffler, E., Si Le, Q., Rockett, K., Kwiatkowski, D. and Spencer, C. Inferring adaptive gene-flow in recent african history. *bioRxiv*, 2017. doi: 10.1101/205252. URL `https://www.biorxiv.org/content/early/2017/11/03/205252`.

Bustamante, C.D., Fledel-Alon, A., Williamson, S., Nielsen, R., Hubisz, M.T., Glanowski, S., Tanenbaum, D.M., White, T.J., Sninsky, J.J., Hernandez, R.D. et al. Natural selection on protein-coding genes in the human genome. *Nature*, 437(7062):1153, 2005.

Bustamante, C.D., Francisco, M. and Burchard, E.G. Genomics for the world. *Nature*, 475(7355):163, 2011.

Byard, P. and Lees, F. Estimating the number of loci determining skin colour in a hybrid population. *Annals of human biology*, 8(1):49–58, 1981.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J. et al. Genome-wide genetic data on 500,000 uk biobank participants. *bioRxiv*, 2017. doi: 10.1101/166298. URL `https://www.biorxiv.org/content/early/2017/07/20/166298`.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J. et al. The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, Oct 2018. doi: 10.1038/s41586-018-0579-z.

Caduff, M., Bauer, A., Jagannathan, V. and Leeb, T. Oca2 splice site variant in german spitz dogs with oculocutaneous albinism. *PLoS One*, 12(10):e0185944, 2017. doi: 10.1371/journal.pone.0185944.

Callaway, E. Divided by dna: The uneasy relationship between archaeology and ancient genomics. *Nature*, 555(7698):573–576, Mar 2018. doi: 10.1038/d41586-018-03773-6.

Candille, S.I., Absher, D.M., Beleza, S., Bauchet, M., McEvoy, B., Garrison, N.A., Li, J.Z., Myers, R.M., Barsh, G.S., Tang, H. et al. Genome-wide association studies of quantitatively measured skin, hair, and eye pigmentation in four european populations. *PLoS One*, 7(10):e48294, 2012. doi: 10.1371/journal.pone.0048294.

Cardona, A., Pagani, L., Antao, T., Lawson, D.J., Eichstaedt, C.A., Yngvadottir, B., Shwe, M.T.T., Wee, J., Romero, I.G., Raj, S. et al. Genome-wide analysis of cold adaptation in indigenous siberian populations. *PLoS One*, 9(5):e98076, 2014. doi: 10.1371/journal.pone.0098076.

Cario-André, M., Pain, C., Gauthier, Y., Casoli, V. and Taieb, A. In vivo and in vitro evidence of dermal fibroblasts influence on human epidermal pigmentation. *Pigment Cell Res*, 19(5):434–42, Oct 2006. doi: 10.1111/j.1600-0749.2006.00326.x.

Carr, I.M. and Markham, A.F. Molecular genetic analysis of the human sorbitol dehydrogenase gene. *Mamm Genome*, 6(9):645–52, Sep 1995.

Carvajal-Carmona, L.G., Soto, I.D., Pineda, N., Ortíz-Barrientos, D., Duque, C., Ospina-Duque, J., McCarthy, M., Montoya, P., Alvarez, V.M., Bedoya, G. et al. Strong amerind/white sex bias and a possible sephardic contribution among the founders of a population in northwest colombia. *The American Journal of Human Genetics*, 67(5):1287–1295, 2000.

Carvajal-Carmona, L.G., Ophoff, R., Hartiala, J., Molina, J., Leon, P., Ospina, J., Bedoya, G., Freimer, N., Ruiz-Linares, A. et al. Genetic demography of antioquia (colombia) and the central valley of costa rica. *Human genetics*, 112(5-6):534–541, 2003.

Cavalli-Sforza, L.L. Human diversity. In *Proc. 12th Int. Congr. Genet*, volume 2, pages 405–416, 1969.

Cavalli-Sforza, L.L. Population structure and human evolution. *Proc. R. Soc. Lond. B*, 164(995):362–379, 1966.

Chacon-Duque, J.C., Adhikari, K., Fuentes-Guajardo, M., Mendoza-Revilla, J., Acuna-Alonzo, V., Barquera Lozano, R., Quinto-Sanchez, M., Gomez-Valdes, J., Everardo Martinez, P., Villamil-Ramirez, H. et al. Latin americans show wide-spread converso ancestry and the imprint of local native ancestry on physical appearance. *bioRxiv*, 2018. doi: 10.1101/252155. URL `https://www.biorxiv.org/content/early/2018/01/23/252155`.

Chahal, H.S., Wu, W., Ransohoff, K.J., Yang, L., Hedlin, H., Desai, M., Lin, Y., Dai, H.J., Qureshi, A.A., Li, W.Q. et al. Genome-wide association study identifies 14 novel risk alleles associated with basal cell carcinoma. *Nat Commun*, 7:12510, Aug 2016. doi: 10.1038/ncomms12510.

Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M. and Lee, J.J. Second-generation plink: rising to the challenge of larger and richer datasets. *Gigascience*, 4:7, 2015. doi: 10.1186/s13742-015-0047-8.

Chang, Y.C., Chang, L.Y., Chang, T.J., Jiang, Y.D., Lee, K.C., Kuo, S.S., Lee, W.J. and Chuang, L.M. The associations of lpin1 gene expression in adipose tissue with metabolic phenotypes in the chinese population. *Obesity (Silver Spring)*, 18(1):7–12, Jan 2010. doi: 10.1038/oby.2009.198.

Chaplin, G. Geographic distribution of environmental factors influencing human skin coloration. *Am J Phys Anthropol*, 125(3):292–302, Nov 2004. doi: 10.1002/ajpa.10263.

Chapman, J.M., Cooper, J.D., Todd, J.A. and Clayton, D.G. Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Hum Hered*, 56(1-3):18–31, 2003. doi: 10.1159/000073729.

Charlesworth, D., Charlesworth, B. and Morgan, M.T. The pattern of neutral molecular variation under the background selection model. *Genetics*, 141(4):1619–32, Dec 1995.

Chen, H., Hey, J. and Slatkin, M. A hidden markov model for investigating recent positive selection through haplotype structure. *Theor Popul Biol*, 99:18–30, Feb 2015. doi: 10.1016/j.tpb.2014.11.001.

Cheng, C., White, B.J., Kamdem, C., Mockaitis, K., Costantini, C., Hahn, M.W. and Besansky, N.J. Ecological genomics of anopheles gambiae along a latitudinal cline: a population-resequencing approach. *Genetics*, 190(4):1417–1432, 2012.

Cheng, C.J., Lin, Y.C., Tsai, M.T., Chen, C.S., Hsieh, M.C., Chen, C.L. and Yang, R.B. Scube2 suppresses breast tumor cell proliferation and confers a favorable prognosis in invasive breast cancer. *Cancer Res*, 69(8):3634–41, Apr 2009. doi: 10.1158/0008-5472.CAN-08-3615.

Chiang, P.W., Spector, E. and Scheuerle, A. A case of asian indian oca3 patient. *American Journal of Medical Genetics Part A*, 149(7):1578–1580, 2009.

Clark, P., Stark, A.E., Walsh, R.J., Jardine, R. and Martin, N.G. A twin study of skin reflectance. *Ann Hum Biol*, 8(6):529–41, 1981.

Concha, G., Nermell, B. and Vahter, M. Spatial and temporal variations in arsenic exposure via drinking-water in northern argentina. *J Health Popul Nutr*, 24(3):317–26, Sep 2006.

Cook, A.L., Chen, W., Thurber, A.E., Smit, D.J., Smith, A.G., Bladen, T.G., Brown, D.L., Duffy, D.L., Pastorino, L., Bianchi-Scarra, G. et al. Analysis of cultured human melanocytes based on polymorphisms within the slc45a2/matp, slc24a5/nckx5, and oca2/p loci. *J Invest Dermatol*, 129(2):392–405, Feb 2009. doi: 10.1038/jid.2008.211.

Cook, N.D. and Lovell, W.G. *Secret judgments of God: Old world disease in colonial Spanish America*, volume 205. University of Oklahoma Press, 2001.

Coop, G., Pickrell, J.K., Novembre, J., Kudaravalli, S., Li, J., Absher, D., Myers, R.M., Cavalli-Sforza, L.L., Feldman, M.W. and Pritchard, J.K. The role of geography in human adaptation. *PLoS genetics*, 5(6):e1000500, 2009.

Coop, G., Witonsky, D., Di Rienzo, A. and Pritchard, J.K. Using environmental correlations to identify loci underlying local adaptation. *Genetics*, 185(4):1411–23, Aug 2010. doi: 10.1534/genetics.110.114819.

Corona, E., Chen, R., Sikora, M., Morgan, A.A., Patel, C.J., Ramesh, A., Bustamante, C.D. and Butte, A.J. Analysis of the genetic basis of disease in the context of worldwide human relationships and migration. *PLoS Genet*, 9(5):e1003447, May 2013. doi: 10.1371/journal.pgen.1003447.

Cossrow, N. and Falkner, B. Race/ethnic issues in obesity and obesity-related comorbidities. *The Journal of Clinical Endocrinology & Metabolism*, 89(6):2590–2594, 2004.

Crawford, J.E., Amaru, R., Song, J., Julian, C.G., Racimo, F., Cheng, J.Y., Guo, X., Yao, J., Ambale-Venkatesh, B., Lima, J.A. et al. Natural selection on genes related to cardiovascular health in high-altitude adapted andeans. *Am J Hum Genet*, 101(5):752–767, Nov 2017a. doi: 10.1016/j.ajhg.2017.09.023.

Crawford, N.G., Kelly, D.E., Hansen, M.E.B., Beltrame, M.H., Fan, S., Bowman, S.L., Jewett, E., Ranciaro, A., Thompson, S., Lo, Y. et al. Loci associated with skin pigmentation identified in african populations. *Science*, 358(6365), 11 2017b. doi: 10.1126/science.aan8433.

Crosby, A.W. Virgin soil epidemics as a factor in the aboriginal depopulation in america. *The William and Mary Quarterly: A Magazine of Early American History*, pages 289–299, 1976.

Csilléry, K., Blum, M.G., Gaggiotti, O.E. and François, O. Approximate bayesian computation (abc) in practice. *Trends in ecology & evolution*, 25(7):410–418, 2010.

Csilléry, K., François, O. and Blum, M.G. abc: an r package for approximate bayesian computation (abc). *Methods in ecology and evolution*, 3(3):475–479, 2012.

Cui, R., Kamatani, Y., Takahashi, A., Usami, M., Hosono, N., Kawaguchi, T., Tsunoda, T., Kamatani, N., Kubo, M., Nakamura, Y. et al. Functional variants in adh1b and aldh2 coupled with alcohol and smoking synergistically enhance esophageal cancer risk. *Gastroenterology*, 137(5):1768–75, Nov 2009. doi: 10.1053/j.gastro.2009.07.070.

Curtin, P. The at/antic slave trade: A census. madison, 1969.

Daborn, P., Yen, J., Bogwitz, M., Le Goff, G., Feil, E., Jeffers, S., Tijet, N., Perry, T., Heckel, D., Batterham, P. et al. A single p450 allele associated with insecticide resistance in drosophila. *Science*, 297(5590):2253–2256, 2002.

Dall'Ara, I., Ghirotto, S., Ingusci, S., Bagarolo, G., Bertolucci, C. and Barbujani, G. Demographic history and adaptation account for clock gene diversity in humans. *Heredity (Edinb)*, 117(3):165–72, 09 2016. doi: 10.1038/hdy.2016.39.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T. et al. The variant call format and vcftools. *Bioinformatics*, 27(15):2156–8, Aug 2011. doi: 10.1093/bioinformatics/btr330.

Dannemann, M. and Kelso, J. The contribution of neanderthals to phenotypic variation in modern humans. *Am J Hum Genet*, 101(4):578–589, Oct 2017. doi: 10.1016/j.ajhg.2017.09.010.

Dannemann, M., Prüfer, K. and Kelso, J. Functional implications of neandertal introgression in modern humans. *Genome Biol*, 18(1):61, 04 2017. doi: 10.1186/s13059-017-1181-7.

Darwin, C. *The descent of man and selection in relation to sex*, volume 1. Murray, 1871.

Daub, J.T., Moretti, S., Davydov, I.I., Excoffier, L. and Robinson-Rechavi, M. Detection of pathways affected by positive selection in primate lineages ancestral to humans. *Mol Biol Evol*, 34(6):1391–1402, Jun 2017. doi: 10.1093/molbev/msx083.

Daub, J.T., Hofer, T., Cutivet, E., Dupanloup, I., Quintana-Murci, L., Robinson-Rechavi, M. and Excoffier, L. Evidence for polygenic adaptation to pathogens in the human genome. *Mol Biol Evol*, 30(7):1544–58, Jul 2013. doi: 10.1093/molbev/mst080.

Daub, J.T., Dupanloup, I., Robinson-Rechavi, M. and Excoffier, L. Inference of evolutionary forces acting on human biological pathways. *Genome Biol Evol*, 7(6):1546–58, May 2015. doi: 10.1093/gbe/evv083.

de la Fuente, C., Ávila-Arcos, M.C., Galimany, J., Carpenter, M.L., Homburger, J.R., Blanco, A., Contreras, P., Cruz Dávalos, D., Reyes, O., San Roman, M. et al. Genomic insights into the origin and diversification of late maritime hunter-gatherers from the chilean patagonia. *Proc Natl Acad Sci U S A*, 115(17):E4006–E4012, Apr 2018. doi: 10.1073/pnas.1715688115.

De Mita, S., Thuillet, A.C., Gay, L., Ahmadi, N., Manel, S., Ronfort, J. and Vigouroux, Y. Detecting selection along environmental gradients: analysis of eight methods and their effectiveness for outbreeding and selfing populations. *Mol Ecol*, 22(5):1383–99, Mar 2013. doi: 10.1111/mec.12182.

DeBruyne, J.P., Weaver, D.R. and Reppert, S.M. Clock and npas2 have overlapping roles in the suprachiasmatic circadian clock. *Nat Neurosci*, 10(5):543–5, May 2007. doi: 10.1038/nn1884.

Del Marmol, V. and Beermann, F. Tyrosinase and related proteins in mammalian pigmentation. *FEBS letters*, 381(3):165–168, 1996.

Delaneau, O., Zagury, J.F. and Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat Methods*, 10(1):5–6, Jan 2013. doi: 10.1038/nmeth.2307.

Denevan, W.M. *The native population of the Americas in 1492*. Univ of Wisconsin Press, 1992.

Deng, L., Ruiz-Linares, A., Xu, S. and Wang, S. Ancestry variation and footprints of natural selection along the genome in latin american populations. *Sci Rep*, 6:21766, Feb 2016. doi: 10.1038/srep21766.

Deschamps, M., Laval, G., Fagny, M., Itan, Y., Abel, L., Casanova, J.L., Patin, E. and Quintana-Murci, L. Genomic signatures of selective pressures and introgression from archaic hominins at human innate immunity genes. *Am J Hum Genet*, 98(1):5–21, Jan 2016. doi: 10.1016/j.ajhg.2015.11.014.

Devlin, B. and Roeder, K. Genomic control for association studies. *Biometrics*, 55(4):997–1004, 1999.

Dillehay, T.D. and Collins, M.B. Early cultural evidence from monte verde in chile. *Nature*, 332(6160):150, 1988.

Dilthey, A.T., Gourraud, P.A., Mentzer, A.J., Cereb, N., Iqbal, Z. and McVean, G. High-accuracy hla type inference from whole-genome sequencing data using population reference graphs. *PLoS Comput Biol*, 12(10):e1005151, Oct 2016. doi: 10.1371/journal.pcbi.1005151.

Dobyns, H.F. Disease transfer at contact. *Annual Review of Anthropology*, 22(1):273–291, 1993.

D'Orazio, J.A., Nobuhisa, T., Cui, R., Arya, M., Spry, M., Wakamatsu, K., Igras, V., Kunisada, T., Granter, S.R., Nishimura, E.K. et al. Topical drug rescue strategy and skin protection based on the role of mc1r in uv-induced tanning. *Nature*, 443(7109): 340–4, Sep 2006. doi: 10.1038/nature05098.

Du, J. and Fisher, D.E. Identification of aim-1 as the underwhite mouse mutant and its transcriptional regulation by mitf. *J Biol Chem*, 277(1):402–6, Jan 2002. doi: 10.1074/jbc.M110229200.

Duffy, D.L., Box, N.F., Chen, W., Palmer, J.S., Montgomery, G.W., James, M.R., Hayward, N.K., Martin, N.G. and Sturm, R.A. Interactive effects of mc1r and oca2 on melanoma risk phenotypes. *Hum Mol Genet*, 13(4):447–61, Feb 2004. doi: 10.1093/hmg/ddh043.

Duffy, J. *Epidemics in Colonial America*. Kennikat Press, 1972. ISBN 9780804616645. URL `https://books.google.fr/books?id=Px30RAAACAAJ`.

Duke, J.L., Lind, C., Mackiewicz, K., Ferriola, D., Papazoglou, A., Gasiewski, A., Heron, S., Huynh, A., McLaughlin, L., Rogers, M. et al. Determining performance characteristics of an ngs-based hla typing method for clinical applications. *HLA*, 87(3):141–52, Mar 2016. doi: 10.1111/tan.12736.

Dunlap, J.C., Loros, J.J. and DeCoursey, P.J. *Chronobiology: biological timekeeping.* Sinauer Associates, 2004.

Eaton, K., Edwards, M., Krithika, S., Cook, G., Norton, H. and Parra, E.J. Association study confirms the role of two oca2 polymorphisms in normal skin pigmentation variation in east asian populations. *Am J Hum Biol*, 27(4):520–5, 2015. doi: 10.1002/ajhb.22678.

Eden, E., Navon, R., Steinfeld, I., Lipson, D. and Yakhini, Z. Gorilla: a tool for discovery and visualization of enriched go terms in ranked gene lists. *BMC Bioinformatics*, 10: 48, Feb 2009. doi: 10.1186/1471-2105-10-48.

Edwards, M., Bigham, A., Tan, J., Li, S., Gozdzik, A., Ross, K., Jin, L. and Parra, E.J. Association of the oca2 polymorphism his615arg with melanin content in east asian populations: further evidence of convergent evolution of skin pigmentation. *PLoS Genet*, 6(3):e1000867, Mar 2010. doi: 10.1371/journal.pgen.1000867.

Edwards, M., Cha, D., Krithika, S., Johnson, M., Cook, G. and Parra, E.J. Iris pigmentation as a quantitative trait: variation in populations of european, east asian and south asian ancestry and association with candidate gene polymorphisms. *Pigment Cell Melanoma Res*, 29(2):141–62, Mar 2016. doi: 10.1111/pcmr.12435.

Eichstaedt, C.A., Antão, T., Pagani, L., Cardona, A., Kivisild, T. and Mormina, M. The andean adaptive toolkit to counteract high altitude maladaptation: genome-wide and phenotypic analysis of the collas. *PLoS One*, 9(3):e93314, 2014. doi: 10.1371/journal.pone.0093314.

Eichstaedt, C.A., Antão, T., Cardona, A., Pagani, L., Kivisild, T. and Mormina, M. Genetic and phenotypic differentiation of an andean intermediate altitude population. *Physiol Rep*, 3(5), May 2015a. doi: 10.14814/phy2.12376.

Eichstaedt, C.A., Antao, T., Cardona, A., Pagani, L., Kivisild, T. and Mormina, M. Positive selection of as3mt to arsenic water in andean populations. *Mutat Res*, 780: 97–102, Oct 2015b. doi: 10.1016/j.mrfmmm.2015.07.007.

Elks, C.E., Perry, J.R.B., Sulem, P., Chasman, D.I., Franceschini, N., He, C., Lunetta, K.L., Visser, J.A., Byrne, E.M., Cousminer, D.L. et al. Thirty new loci for age at menarche identified by a meta-analysis of genome-wide association studies. *Nat Genet*, 42(12):1077–85, Dec 2010. doi: 10.1038/ng.714.

Endler, J.A. Natural and sexual selection on color patterns in poeciliid fishes. *Environmental biology of Fishes*, 9(2):173–190, 1983.

Eng, M.Y., Luczak, S.E. and Wall, T.L. Aldh2, adh1b, and adh1c genotypes in asians: a literature review. *Alcohol Res Health*, 30(1):22–7, 2007.

Engström, K.S., Hossain, M.B., Lauss, M., Ahmed, S., Raqib, R., Vahter, M. and Broberg, K. Efficient arsenic metabolism–the as3mt haplotype is associated with dna methylation and expression of multiple genes around as3mt. *PLoS One*, 8(1):e53732, 2013. doi: 10.1371/journal.pone.0053732.

Eriksson, N., Macpherson, J.M., Tung, J.Y., Hon, L.S., Naughton, B., Saxonov, S., Avey, L., Wojcicki, A., Pe'er, I. and Mountain, J. Web-based, participant-driven studies yield novel genetic associations for common traits. *PLoS Genet*, 6(6):e1000993, Jun 2010. doi: 10.1371/journal.pgen.1000993.

Ettinger, N.A., Duggal, P., Braz, R.F.S., Nascimento, E.T., Beaty, T.H., Jeronimo, S.M.B., Pearson, R.D., Blackwell, J.M., Moreno, L. and Wilson, M.E. Genetic admixture in brazilians exposed to infection with leishmania chagasi. *Ann Hum Genet*, 73 (Pt 3):304–13, May 2009. doi: 10.1111/j.1469-1809.2009.00510.x.

Ewing, G. and Hermisson, J. Msms: a coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics*, 26(16): 2064–5, Aug 2010. doi: 10.1093/bioinformatics/btq322.

Fagundes, N.J., Ray, N., Beaumont, M., Neuenschwander, S., Salzano, F.M., Bonatto, S.L. and Excoffier, L. Statistical evaluation of alternative models of human evolution. *Proceedings of the National Academy of Sciences*, 104(45):17614–17619, 2007.

Falconer, D.S. *Introduction to quantitative genetics*. Oliver And Boyd; Edinburgh; London, 1960.

Falush, D., Stephens, M. and Pritchard, J.K. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 164 (4):1567–87, Aug 2003.

Fan, S., Hansen, M.E.B., Lo, Y. and Tishkoff, S.A. Going global by adapting local: A review of recent human adaptation. *Science*, 354(6308):54–59, 10 2016. doi: 10.1126/science.aaf5098.

Farh, K.K.H., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W.J., Beik, S., Shoresh, N., Whitton, H., Ryan, R.J.H., Shishkin, A.A. et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*, 518(7539):337–43, Feb 2015. doi: 10.1038/nature13835.

Fay, J.C., Wyckoff, G.J. and Wu, C.I. Testing the neutral theory of molecular evolution with genomic data from drosophila. *Nature*, 415(6875):1024, 2002.

Fehren-Schmitz, L. and Georges, L. Ancient dna reveals selection acting on genes associated with hypoxia response in pre-columbian peruvian highlanders in the last 8500 years. *Sci Rep*, 6:23485, Mar 2016. doi: 10.1038/srep23485.

Felsenstein, J. On the biological significance of the cost of gene substitution. *The American Naturalist*, 105(941):1–11, 1971.

Ferrer-Admetlla, A., Liang, M., Korneliussen, T. and Nielsen, R. On detecting incomplete soft or hard selective sweeps using haplotype structure. *Mol Biol Evol*, 31(5):1275–91, May 2014. doi: 10.1093/molbev/msu077.

Field, Y., Boyle, E.A., Telis, N., Gao, Z., Gaulton, K.J., Golan, D., Yengo, L., Rocheleau, G., Froguel, P., McCarthy, M.I. et al. Detection of human adaptation during the past 2000 years. *Science*, 354(6313):760–764, 11 2016a. doi: 10.1126/science.aag0776.

Field, Y., Boyle, E.A., Telis, N., Gao, Z., Gaulton, K.J., Golan, D., Yengo, L., Rocheleau, G., Froguel, P., McCarthy, M.I. et al. Detection of human adaptation during the past 2000 years. *Science*, page aag0776, 2016b.

Fields, S.L. *Pestilence and headcolds: Encountering illness in colonial Mexico.* University of California, Davis, 2004.

Flagel, L., Brandvain, Y.J. and Schrider, D.R. The unreasonable effectiveness of convolutional neural networks in population genetic inference. *bioRxiv*, 2018. doi: 10.1101/336073. URL `https://www.biorxiv.org/content/early/2018/05/31/336073`.

Fleming, A. and Copp, A.J. Embryonic folate metabolism and mouse neural tube defects. *Science*, 280(5372):2107–2109, 1998.

Foll, M., Gaggiotti, O.E., Daub, J.T., Vatsiou, A. and Excoffier, L. Widespread signals of convergent adaptation to high altitude in asia and america. *Am J Hum Genet*, 95(4):394–407, Oct 2014. doi: 10.1016/j.ajhg.2014.09.002.

Fox, C.S., Liu, Y., White, C.C., Feitosa, M., Smith, A.V., Heard-Costa, N., Lohman, K., GIANT Consortium, MAGIC Consortium, GLGC Consortium et al. Genome-wide association for abdominal subcutaneous and visceral adipose reveals a novel locus for visceral fat in women. *PLoS Genet*, 8(5):e1002695, 2012. doi: 10.1371/journal.pgen.1002695.

Fraser, H.B. Gene expression drives local adaptation in humans. *Genome Res*, 23(7): 1089–96, Jul 2013. doi: 10.1101/gr.152710.112.

Frayn, K.N. Adipose tissue as a buffer for daily lipid flux. *Diabetologia*, 45(9):1201–10, Sep 2002. doi: 10.1007/s00125-002-0873-y.

Frichot, E., Schoville, S.D., Bouchard, G. and François, O. Testing for associations between loci and environmental gradients using latent factor mixed models. *Molecular biology and evolution*, 30(7):1687–1699, 2013.

Frodsham, A.J. and Hill, A.V.S. Genetics of infectious diseases. *Hum Mol Genet*, 13 Spec No 2:R187–94, Oct 2004. doi: 10.1093/hmg/ddh225.

Frost, P. European hair and eye color: a case of frequency-dependent sexual selection? *Evolution and Human Behavior*, 27(2):85–103, 2006.

Frost, P. Human skin-color sexual dimorphism: a test of the sexual selection hypothesis. *American journal of physical anthropology*, 133(1):779–780, 2007.

Frudakis, T., Thomas, M., Gaskin, Z., Venkateswarlu, K., Chandra, K.S., Ginjupalli, S., Gunturi, S., Natrajan, S., Ponnuswamy, V.K. and Ponnuswamy, K.N. Sequences associated with human iris pigmentation. *Genetics*, 165(4):2071–83, Dec 2003.

Fry, R.C., Navasumrit, P., Valiathan, C., Svensson, J.P., Hogan, B.J., Luo, M., Bhattacharya, S., Kandjanapa, K., Soontararuks, S., Nookabkaew, S. et al. Activation of inflammation/nf-kappab signaling in infants born to arsenic-exposed mothers. *PLoS Genet*, 3(11):e207, Nov 2007. doi: 10.1371/journal.pgen.0030207.

Fujimoto, A., Ohashi, J., Nishida, N., Miyagawa, T., Morishita, Y., Tsunoda, T., Kimura, R. and Tokunaga, K. A replication study confirmed the edar gene to be a major contributor to population differentiation regarding head hair thickness in asia. *Hum Genet*, 124(2):179–85, Sep 2008. doi: 10.1007/s00439-008-0537-1.

Fumagalli, M., Sironi, M., Pozzoli, U., Ferrer-Admetlla, A., Ferrer-Admettla, A., Pattini, L. and Nielsen, R. Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. *PLoS Genet*, 7(11):e1002355, Nov 2011. doi: 10.1371/journal.pgen.1002355.

Fumagalli, M., Moltke, I., Grarup, N., Racimo, F., Bjerregaard, P., Jørgensen, M.E., Korneliussen, T.S., Gerbault, P., Skotte, L., Linneberg, A. et al. Greenlandic inuit show genetic signatures of diet and climate adaptation. *Science*, 349(6254):1343–7, Sep 2015. doi: 10.1126/science.aab2319.

Gallego Romero, I., Basu Mallick, C., Liebert, A., Crivellaro, F., Chaubey, G., Itan, Y., Metspalu, M., Eaaswarkhanth, M., Pitchappan, R., Villems, R. et al. Herders of indian and european cattle share their predominant allele for lactase persistence. *Mol Biol Evol*, 29(1):249–60, Jan 2012. doi: 10.1093/molbev/msr190.

Gamerman, D. and Lopes, H.F. *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. Chapman and Hall/CRC, 2006.

García-Borrón, J.C., Sánchez-Laorden, B.L. and Jiménez-Cervantes, C. Melanocortin-1 receptor structure and functional regulation. *Pigment Cell Res*, 18(6):393–410, Dec 2005. doi: 10.1111/j.1600-0749.2005.00278.x.

Garud, N.R., Messer, P.W., Buzbas, E.O. and Petrov, D.A. Recent selective sweeps in north american drosophila melanogaster show signatures of soft sweeps. *PLoS Genet*, 11(2):e1005004, Feb 2015. doi: 10.1371/journal.pgen.1005004.

Gautam, P., Chaurasia, A., Bhattacharya, A., Grover, R., Consortium, I.G.V., Mukerji, M. and Natarajan, V.T. Population diversity and adaptive evolution in keratinization genes: impact of environment in shaping skin phenotypes. *Molecular biology and evolution*, 32 (3):555–573, 2014.

Gelernter, J., Zhou, H., Nuñez, Y.Z., Mutirangura, A., Malison, R.T. and Kalayasiri, R. Genomewide association study of alcohol dependence and related traits in a thai population. *Alcohol Clin Exp Res*, 42(5):861–868, May 2018. doi: 10.1111/acer.13614.

Gerbault, P., Moret, C., Currat, M. and Sanchez-Mazas, A. Impact of selection and demography on the diffusion of lactase persistence. *PLoS One*, 4(7):e6369, Jul 2009. doi: 10.1371/journal.pone.0006369.

González, J., Lenkov, K., Lipatov, M., Macpherson, J.M. and Petrov, D.A. High rate of recent transposable element–induced adaptation in drosophila melanogaster. *PLoS biology*, 6(10):e251, 2008.

González-José, R., Bortolini, M.C., Santos, F.R. and Bonatto, S.L. The peopling of america: craniofacial shape variation on a continental scale and its interpretation from an interdisciplinary view. *Am J Phys Anthropol*, 137(2):175–87, Oct 2008. doi: 10.1002/ajpa.20854.

Graf, J., Hodgson, R. and van Daal, A. Single nucleotide polymorphisms in the matp gene are associated with normal human pigmentation variation. *Hum Mutat*, 25(3):278–84, Mar 2005. doi: 10.1002/humu.20143.

Graf, J., Voisey, J., Hughes, I. and van Daal, A. Promoter polymorphisms in the matp (slc45a2) gene are associated with normal human skin color variation. *Hum Mutat*, 28 (7):710–7, Jul 2007. doi: 10.1002/humu.20504.

Grafen, A. The phylogenetic regression. *Phil. Trans. R. Soc. Lond. B*, 326(1233):119–157, 1989.

Graff, M., Scott, R.A., Justice, A.E., Young, K.L., Feitosa, M.F., Barata, L., Winkler, T.W., Chu, A.Y., Mahajan, A., Hadley, D. et al. Genome-wide physical activity interactions in adiposity - a meta-analysis of 200,452 adults. *PLoS Genet*, 13(4):e1006528, Apr 2017. doi: 10.1371/journal.pgen.1006528.

Greaves, M. Was skin cancer a selective force for black pigmentation in early hominin evolution? *Proc. R. Soc. B*, 281(1781):20132955, 2014.

Green, R.E., Krause, J., Briggs, A.W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M.H.Y. et al. A draft sequence of the neandertal genome. *Science*, 328(5979):710–722, May 2010. doi: 10.1126/science.1188021.

Grossman, S.R., Shlyakhter, I., Shylakhter, I., Karlsson, E.K., Byrne, E.H., Morales, S., Frieden, G., Hostetter, E., Angelino, E., Garber, M. et al. A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science*, 327(5967): 883–6, Feb 2010. doi: 10.1126/science.1183863.

GTEx Consortium. The genotype-tissue expression (gtex) project. *Nat Genet*, 45(6): 580–5, Jun 2013. doi: 10.1038/ng.2653.

Guan, Y. Detecting structure of haplotypes and local ancestry. *Genetics*, 196(3):625–42, Mar 2014. doi: 10.1534/genetics.113.160697.

Guidon, N. and Delibrias, G. Carbon-14 dates point to man in the americas 32,000 years ago. *Nature*, 321(6072):769, 1986.

Guillot, G., Vitalis, R., le Rouzic, A. and Gautier, M. Detecting correlation between allele frequencies and environmental variables as a signature of selection. a fast computational approach for genome-wide studies. *Spatial Statistics*, 8:145–155, 2014.

Günther, T. and Coop, G. Robust identification of local adaptation from allele frequencies. *Genetics*, 195(1):205–20, Sep 2013. doi: 10.1534/genetics.113.152462.

Guo, C., Ludvik, A.E., Arlotto, M.E., Hayes, M.G., Armstrong, L.L., Scholtens, D.M., Brown, C.D., Newgard, C.B., Becker, T.C., Layden, B.T. et al. Coordinated regulatory variation associated with gestational hyperglycaemia regulates expression of the novel hexokinase hkdc1. *Nat Commun*, 6:6069, Feb 2015. doi: 10.1038/ncomms7069.

Guo, J., Wu, Y., Zhu, Z., Zheng, Z., Trzaskowski, M., Zeng, J., Robinson, M.R., Visscher, P.M. and Yang, J. Global genetic differentiation of complex traits shaped by natural selection in humans. *Nat Commun*, 9(1):1865, May 2018. doi: 10.1038/s41467-018-04191-y.

Ha, T., Naysmith, L., Waterston, K., Oh, C., Weller, R. and Rees, J.L. Defining the quantitative contribution of the melanocortin 1 receptor (mc1r) to variation in pigmentary phenotype. *Ann N Y Acad Sci*, 994:339–47, Jun 2003.

Haitina, T., Ringholm, A., Kelly, J., Mundy, N.I. and Schiöth, H.B. High diversity in functional properties of melanocortin 1 receptor (mc1r) in divergent primate species is more strongly associated with phylogeny than coat color. *Mol Biol Evol*, 24(9):2001–8, Sep 2007. doi: 10.1093/molbev/msm134.

Haldane, J. The theory of a cline. *Journal of genetics*, 48(3):277–284, 1948.

Haldane, J.B.S. The cost of natural selection. *Journal of Genetics*, 55(3):511, 1957.

Hallam, S.J., Goncharov, A., McEwen, J., Baran, R. and Jin, Y. Syd-1, a presynaptic protein with pdz, c2 and rhogap-like domains, specifies axon identity in c. elegans. *Nat Neurosci*, 5(11):1137–46, Nov 2002. doi: 10.1038/nn959.

Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A. and McKusick, V.A. Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res*, 33(Database issue):D514–7, Jan 2005. doi: 10.1093/nar/gki033.

Han, J., Kraft, P., Nan, H., Guo, Q., Chen, C., Qureshi, A., Hankinson, S.E., Hu, F.B., Duffy, D.L., Zhao, Z.Z. et al. A genome-wide association study identifies novel alleles associated with hair color and skin pigmentation. *PLoS Genet*, 4(5):e1000074, May 2008. doi: 10.1371/journal.pgen.1000074.

Hancock, A.M., Witonsky, D.B., Gordon, A.S., Eshel, G., Pritchard, J.K., Coop, G. and Di Rienzo, A. Adaptations to climate in candidate genes for common metabolic disorders. *PLoS Genet*, 4(2):e32, Feb 2008. doi: 10.1371/journal.pgen.0040032.

Hancock, A.M., Witonsky, D.B., Alkorta-Aranburu, G., Beall, C.M., Gebremedhin, A., Sukernik, R., Utermann, G., Pritchard, J.K., Coop, G. and Di Rienzo, A. Adaptations to climate-mediated selective pressures in humans. *PLoS Genet*, 7(4):e1001375, Apr 2011. doi: 10.1371/journal.pgen.1001375.

Harada, M., Li, Y.F., El-Gamil, M., Rosenberg, S.A. and Robbins, P.F. Use of an in vitro immunoselected tumor line to identify shared melanoma antigens recognized by hla-a*0201-restricted t cells. *Cancer Res*, 61(3):1089–94, Feb 2001.

Harada, S., Agarwal, D.P. and Goedde, H.W. Aldehyde dehydrogenase deficiency as cause of facial flushing reaction to alcohol in japanese. *Lancet*, 2(8253):982, Oct 1981.

Harding, R.M., Healy, E., Ray, A.J., Ellis, N.S., Flanagan, N., Todd, C., Dixon, C., Sajantila, A., Jackson, I.J., Birch-Machin, M.A. et al. Evidence for variable selective pressures at mc1r. *Am J Hum Genet*, 66(4):1351–61, Apr 2000a. doi: 10.1086/302863.

Harding, R.M., Healy, E., Ray, A.J., Ellis, N.S., Flanagan, N., Todd, C., Dixon, C., Sajantila, A., Jackson, I.J., Birch-Machin, M.A. et al. Evidence for variable selective pressures at mc1r. *The American Journal of Human Genetics*, 66(4):1351–1361, 2000b.

Hartl, D.L., Clark, A.G. and Clark, A.G. *Principles of population genetics*, volume 116. Sinauer associates Sunderland, 1997.

Hayes, M.G., Urbanek, M., Hivert, M.F., Armstrong, L.L., Morrison, J., Guo, C., Lowe, L.P., Scheftner, D.A., Pluzhnikov, A., Levine, D.M. et al. Identification of hkdc1 and bace2 as genes influencing glycemic traits during pregnancy through genome-wide association studies. *Diabetes*, 62(9):3282–91, Sep 2013. doi: 10.2337/db12-1692.

Heid, I.M., Jackson, A.U., Randall, J.C., Winkler, T.W., Qi, L., Steinthorsdottir, V., Thorleifsson, G., Zillikens, M.C., Speliotes, E.K., Mägi, R. et al. Meta-analysis identifies

13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution. *Nat Genet*, 42(11):949–60, Nov 2010. doi: 10.1038/ng.685.

Heintzman, P.D., Froese, D., Ives, J.W., Soares, A.E.R., Zazula, G.D., Letts, B., Andrews, T.D., Driver, J.C., Hall, E., Hare, P.G. et al. Bison phylogeography constrains dispersal and viability of the ice free corridor in western canada. *Proc Natl Acad Sci U S A*, 113 (29):8057–63, 07 2016. doi: 10.1073/pnas.1601077113.

Held, L.I. The evo-devo puzzle of human hair patterning. *Evolutionary biology*, 37(2-3): 113–122, 2010.

Hellenthal, G., Busby, G.B.J., Band, G., Wilson, J.F., Capelli, C., Falush, D. and Myers, S. A genetic atlas of human admixture history. *Science*, 343(6172):747–751, Feb 2014. doi: 10.1126/science.1243518.

Hermisson, J. and Pennings, P.S. Soft sweeps: molecular population genetics of adaptation from standing genetic variation. *Genetics*, 169(4):2335–52, Apr 2005. doi: 10.1534/genetics.104.036947.

Hernandez, R.D., Kelley, J.L., Elyashiv, E., Melton, S.C., Auton, A., McVean, G., 1000 Genomes Project, Sella, G. and Przeworski, M. Classic selective sweeps were rare in recent human evolution. *Science*, 331(6019):920–4, Feb 2011a. doi: 10.1126/science.1198878.

Hernandez, R.D., Kelley, J.L., Elyashiv, E., Melton, S.C., Auton, A., McVean, G., Sella, G., Przeworski, M. et al. Classic selective sweeps were rare in recent human evolution. *science*, 331(6019):920–924, 2011b.

Hernandez-Pacheco, N., Flores, C., Alonso, S., Eng, C., Mak, A.C.Y., Hunstman, S., Hu, D., White, M.J., Oh, S.S., Meade, K. et al. Identification of a novel locus associated with skin colour in african-admixed populations. *Sci Rep*, 7:44548, Mar 2017. doi: 10.1038/srep44548.

Hernando, B., Ibarrola-Villava, M., Peña-Chilet, M., Alonso, S., Ribas, G. and Martinez-Cadenas, C. Sex and mc1r variants in human pigmentation: Differences in tanning ability and sensitivity to sunlight between sexes. *J Dermatol Sci*, 84(3):346–348, Dec 2016. doi: 10.1016/j.jdermsci.2016.09.004.

Heykants, L., Schollen, E., Grünewald, S. and Matthijs, G. Identification and localization of two mouse phosphomannomutase genes, pmm1 and pmm2. *Gene*, 270(1-2):53–9, May 2001.

Hider, J.L., Gittelman, R.M., Shah, T., Edwards, M., Rosenbloom, A., Akey, J.M. and Parra, E.J. Exploring signatures of positive selection in pigmentation candidate genes in populations of east asian ancestry. *BMC Evol Biol*, 13:150, Jul 2013. doi: 10.1186/1471-2148-13-150.

Higham, C. *Early cultures of mainland Southeast Asia*. River Books Bangkok, 2002.

Higuchi, S., Motohashi, Y., Ishibashi, K. and Maeda, T. Influence of eye colors of caucasians and asians on suppression of melatonin secretion by light. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, 292(6):R2352–R2356, 2007.

Hill, A.V. The immunogenetics of human infectious diseases. *Annu Rev Immunol*, 16: 593–617, 1998. doi: 10.1146/annurev.immunol.16.1.593.

Hill, A.V. Immunogenetics and genomics. *Lancet*, 357(9273):2037–41, Jun 2001. doi: 10.1016/S0140-6736(00)05117-5.

Hinch, A.G., Tandon, A., Patterson, N., Song, Y., Rohland, N., Palmer, C.D., Chen, G.K., Wang, K., Buxbaum, S.G., Akylbekova, E.L. et al. The landscape of recombination in african americans. *Nature*, 476(7359):170–5, Jul 2011. doi: 10.1038/nature10336.

Hlusko, L.J., Carlson, J.P., Chaplin, G., Elias, S.A., Hoffecker, J.F., Huffman, M., Jablonski, N.G., Monson, T.A., O'Rourke, D.H., Pilloud, M.A. et al. Environmental selection during the last ice age on the mother-to-infant transmission of vitamin d and fatty acids through breast milk. *Proc Natl Acad Sci U S A*, Apr 2018. doi: 10.1073/pnas.1711788115.

Hodgson, J.A., Pickrell, J.K., Pearson, L.N., Quillen, E.E., Prista, A., Rocha, J., Soodyall, H., Shriver, M.D. and Perry, G.H. Natural selection for the duffy-null allele in the recently admixed people of madagascar. *Proc Biol Sci*, 281(1789):20140930, Aug 2014. doi: 10.1098/rspb.2014.0930.

Hoek, K.S., Schlegel, N.C., Eichhoff, O.M., Widmer, D.S., Praetorius, C., Einarsson, S.O., Valgeirsdottir, S., Bergsteinsdottir, K., Schepsky, A., Dummer, R. et al. Novel mitf targets identified using a two-step dna microarray strategy. *Pigment Cell Melanoma Res*, 21(6):665–76, Dec 2008. doi: 10.1111/j.1755-148X.2008.00505.x.

Hoffecker, J.F., Elias, S.A., O'Rourke, D.H., Scott, G.R. and Bigelow, N.H. Beringia and the global dispersal of modern humans. *Evol Anthropol*, 25(2):64–78, 2016. doi: 10.1002/evan.21478.

Holden, M., Deng, S., Wojnowski, L. and Kulle, B. Gsea-snp: applying gene set enrichment analysis to snp data from genome-wide association studies. *Bioinformatics*, 24(23):2784–5, Dec 2008. doi: 10.1093/bioinformatics/btn516.

Homburger, J.R., Moreno-Estrada, A., Gignoux, C.R., Nelson, D., Sanchez, E., Ortiz-Tello, P., Pons-Estel, B.A., Acevedo-Vasquez, E., Miranda, P., Langefeld, C.D. et al. Genomic insights into the ancestry and demographic history of south america. *PLoS Genet*, 11(12):e1005602, Dec 2015. doi: 10.1371/journal.pgen.1005602.

Horai, S., Kondo, R., Nakagawa-Hattori, Y., Hayashi, S., Sonoda, S. and Tajima, K. Peopling of the americas, founded by four major lineages of mitochondrial dna. *Mol Biol Evol*, 10(1):23–47, Jan 1993. doi: 10.1093/oxfordjournals.molbev.a039987.

Hou, L., Kember, R.L., Roach, J.C., O'Connell, J.R., Craig, D.W., Bucan, M., Scott, W.K., Pericak-Vance, M., Haines, J.L., Crawford, M.H. et al. A population-specific reference panel empowers genetic studies of anabaptist populations. *Sci Rep*, 7(1):6079, Jul 2017. doi: 10.1038/s41598-017-05445-3.

Howie, B.N., Donnelly, P. and Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet*, 5(6): e1000529, Jun 2009. doi: 10.1371/journal.pgen.1000529.

Hsieh, P., Veeramah, K.R., Lachance, J., Tishkoff, S.A., Wall, J.D., Hammer, M.F. and Gutenkunst, R.N. Whole-genome sequence analyses of western central african pygmy hunter-gatherers reveal a complex demographic history and identify candidate genes under positive natural selection. *Genome Res*, 26(3):279–90, Mar 2016. doi: 10.1101/gr.192971.115.

Huang, J., Xiao, L., Gong, X., Shao, W., Yin, Y., Liao, Q., Meng, Y., Zhang, Y., Ma, D. and Qiu, X. Cytokine-like molecule ccdc134 contributes to cd8 t-cell effector functions in cancer immunotherapy. *Cancer Res*, 74(20):5734–45, Oct 2014. doi: 10.1158/0008-5472.CAN-13-3132.

Hudson, R.R. and Kaplan, N.L. Deleterious background selection with recombination. *Genetics*, 141(4):1605–17, Dec 1995.

Huerta-Sánchez, E., DeGiorgio, M., Pagani, L., Tarekegn, A., Ekong, R., Antao, T., Cardona, A., Montgomery, H.E., Cavalleri, G.L., Robbins, P.A. et al. Genetic signatures reveal high-altitude adaptation in a set of ethiopian populations. *Molecular Biology and Evolution*, 30(8):1877–1888, 2013. doi: 10.1093/molbev/mst089. URL `http://dx.doi.org/10.1093/molbev/mst089`.

Huerta-Sánchez, E., Jin, X., Asan, Bianba, Z., Peter, B.M., Vinckenbosch, N., Liang, Y., Yi, X., He, M., Somel, M. et al. Altitude adaptation in tibetans caused by introgression of denisovan-like dna. *Nature*, 512(7513):194–7, Aug 2014. doi: 10.1038/nature13408.

Hysi, P.G., Valdes, A.M., Liu, F., Furlotte, N.A., Evans, D.M., Bataille, V., Visconti, A., Hemani, G., McMahon, G., Ring, S.M. et al. Genome-wide association meta-analysis of individuals of european ancestry identifies new loci explaining a substantial fraction of hair color variation and heritability. *Nat Genet*, 50(5):652–656, May 2018. doi: 10.1038/s41588-018-0100-5.

Ilardo, M.A., Moltke, I., Korneliussen, T.S., Cheng, J., Stern, A.J., Racimo, F., de Barros Damgaard, P., Sikora, M., Seguin-Orlando, A., Rasmussen, S. et al. Physiological and genetic adaptations to diving in sea nomads. *Cell*, 173(3):569–580.e15, Apr 2018. doi: 10.1016/j.cell.2018.03.054.

Inagaki, K., Suzuki, T., Shimizu, H., Ishii, N., Umezawa, Y., Tada, J., Kikuchi, N., Takata, M., Takamori, K., Kishibe, M. et al. Oculocutaneous albinism type 4 is one of the most common types of albinism in japan. *Am J Hum Genet*, 74(3):466–71, Mar 2004. doi: 10.1086/382195.

International HapMap Consortium. The international hapmap project. *Nature*, 426(6968): 789–96, Dec 2003. doi: 10.1038/nature02168.

Irwin, D.M. and Tan, H. Molecular evolution of the vertebrate hexokinase gene family: Identification of a conserved fifth vertebrate hexokinase gene. *Comp Biochem Physiol Part D Genomics Proteomics*, 3(1):96–107, Mar 2008. doi: 10.1016/j.cbd.2007.11.002.

Jablonski, N.G. *Skin: A natural history*. Univ of California Press, 2008.

Jablonski, N.G. *Living color: the biological and social meaning of skin color*. Univ of California Press, 2012.

Jablonski, N.G. and Chaplin, G. The evolution of human skin coloration. *Journal of human evolution*, 39(1):57–106, 2000.

Jablonski, N.G. and Chaplin, G. Skin cancer was not a potent selective force in the evolution of protective pigmentation in early hominins. *Proceedings of the Royal Society of London B: Biological Sciences*, 281(1789):20140517, 2014.

Jablonski, N.G. and Chaplin, G. The colours of humanity: the evolution of pigmentation in the human lineage. *Phil. Trans. R. Soc. B*, 372(1724):20160349, 2017.

Jablonski, N.G. and Chaplin, G. The roles of vitamin d and cutaneous vitamin d production in human evolution and health. *Int J Paleopathol*, Mar 2018. doi: 10.1016/j.ijpp.2018.01.005.

Jackson, I.J. Homologous pigmentation mutations in human, mouse and other model organisms. *Hum Mol Genet*, 6(10):1613–24, 1997.

Jacobs, L.C., Hamer, M.A., Gunn, D.A., Deelen, J., Lall, J.S., van Heemst, D., Uh, H.W., Hofman, A., Uitterlinden, A.G., Griffiths, C.E.M. et al. A genome-wide association study identifies the skin color genes irf4, mc1r, asip, and bnc2 influencing facial pigmented spots. *J Invest Dermatol*, 135(7):1735–1742, Jul 2015. doi: 10.1038/jid.2015.62.

Jarrett, S.G., Wolf Horrell, E.M., Boulanger, M.C. and D'Orazio, J.A. Defining the contribution of mc1r physiological ligands to atr phosphorylation at ser435, a predictor of dna repair in melanocytes. *J Invest Dermatol*, 135(12):3086–3095, Dec 2015. doi: 10.1038/jid.2015.280.

Jeanmougin, M., Noirel, J., Coulonges, C. and Zagury, J.F. Hla-check: evaluating hla data from snp information. *BMC Bioinformatics*, 18(1):334, Jul 2017. doi: 10.1186/s12859-017-1746-1.

Jin, Y., Andersen, G., Yorgov, D., Ferrara, T.M., Ben, S., Brownson, K.M., Holland, P.J., Birlea, S.A., Siebert, J., Hartmann, A. et al. Genome-wide association studies of autoimmune vitiligo identify 23 new risk loci and highlight key pathways and regulatory variants. *Nat Genet*, 48(11):1418–1424, 11 2016. doi: 10.1038/ng.3680.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J.A. and Charpentier, E. A programmable dual-rna-guided dna endonuclease in adaptive bacterial immunity. *Science*, 337(6096):816–21, Aug 2012. doi: 10.1126/science.1225829.

Jobling, M., Hurles, M. and Tyler-Smith, C. *Human evolutionary genetics: origins, peoples & disease*. Garland Science, 2013.

Jonnalagadda, M., Bharti, N., Patil, Y., Ozarkar, S., K, S.M., Joshi, R. and Norton, H. Identifying signatures of positive selection in pigmentation genes in two south asian populations. *Am J Hum Biol*, 29(5), Sep 2017. doi: 10.1002/ajhb.23012.

Joost, S., Bonin, A., Bruford, M.W., Després, L., Conord, C., Erhardt, G. and Taberlet, P. A spatial analysis method (sam) to detect candidate loci for selection: towards a landscape genomics approach to adaptation. *Molecular ecology*, 16(18):3955–3969, 2007.

Jouganous, J., Long, W., Ragsdale, A.P. and Gravel, S. Inferring the joint demographic history of multiple populations: Beyond the diffusion approximation. *Genetics*, 206(3): 1549–1567, 07 2017. doi: 10.1534/genetics.117.200493.

Joyce, P. and Marjoram, P. Approximately sufficient statistics and bayesian computation. *Statistical applications in genetics and molecular biology*, 7(1), 2008.

Julian, C.G. High altitude during pregnancy. *Clinics in chest medicine*, 32(1):21–31, 2011.

Justice, A.E., Winkler, T.W., Feitosa, M.F., Graff, M., Fisher, V.A., Young, K., Barata, L., Deng, X., Czajkowski, J., Hadley, D. et al. Genome-wide meta-analysis of 241,258 adults accounting for smoking behaviour identifies novel loci for obesity traits. *Nat Commun*, 8:14977, Apr 2017. doi: 10.1038/ncomms14977.

Kamberov, Y.G., Wang, S., Tan, J., Gerbault, P., Wark, A., Tan, L., Yang, Y., Li, S., Tang, K., Chen, H. et al. Modeling recent human evolution in mice by expression of a selected edar variant. *Cell*, 152(4):691–702, Feb 2013. doi: 10.1016/j.cell.2013.01.016.

Kanehisa, M. and Goto, S. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 28(1):27–30, Jan 2000.

Kang, H.M., Sul, J.H., Zaitlen, N.A., Kong, S.y., Freimer, N.B., Sabatti, C., Eskin, E. et al. Variance component model to account for sample structure in genome-wide association studies. *Nature genetics*, 42(4):348, 2010.

Kanthimathi, S., Liju, S., Laasya, D., Anjana, R.M., Mohan, V. and Radha, V. Hexokinase domain containing 1 (hkdc1) gene variants and their association with gestational diabetes mellitus in a south indian population. *Ann Hum Genet*, 80(4):241–5, Jul 2016. doi: 10.1111/ahg.12155.

Karafet, T., Zegura, S.L., Vuturo-Brady, J., Posukh, O., Osipova, L., Wiebe, V., Romero, F., Long, J.C., Harihara, S., Jin, F. et al. Y chromosome markers and trans-bering strait dispersals. *Am J Phys Anthropol*, 102(3):301–14, Mar 1997. doi: 10.1002/(SICI)1096-8644(199703)102:3⟨301::AID-AJPA1⟩3.0.CO;2-Y.

Karafet, T.M., Zegura, S.L., Posukh, O., Osipova, L., Bergen, A., Long, J., Goldman, D., Klitz, W., Harihara, S., de Knijff, P. et al. Ancestral asian source(s) of new world y-chromosome founder haplotypes. *Am J Hum Genet*, 64(3):817–31, Mar 1999. doi: 10.1086/302282.

Karafet, T.M., Osipova, L.P., Gubina, M.A., Posukh, O.L., Zegura, S.L. and Hammer, M.F. High levels of y-chromosome differentiation among native siberian populations and the genetic signature of a boreal hunter-gatherer way of life. *Hum Biol*, 74(6):761–89, Dec 2002.

Karlsson, E.K., Kwiatkowski, D.P. and Sabeti, P.C. Natural selection and infectious disease in human populations. *Nat Rev Genet*, 15(6):379–93, Jun 2014. doi: 10.1038/nrg3734.

Kawaguchi, S., Higasa, K., Shimizu, M., Yamada, R. and Matsuda, F. Hla-hd: An accurate hla typing algorithm for next-generation sequencing data. *Hum Mutat*, 38(7):788–797, 07 2017. doi: 10.1002/humu.23230.

Kehdy, F.S.G., Gouveia, M.H., Machado, M., Magalhães, W.C.S., Horimoto, A.R., Horta, B.L., Moreira, R.G., Leal, T.P., Scliar, M.O., Soares-Souza, G.B. et al. Origin and dynamics of admixture in brazilians and its effect on the pattern of deleterious mutations. *Proc Natl Acad Sci U S A*, 112(28):8696–701, Jul 2015. doi: 10.1073/pnas.1504447112.

Kenny, E.E., Timpson, N.J., Sikora, M., Yee, M.C., Moreno-Estrada, A., Eng, C., Huntsman, S., Burchard, E.G., Stoneking, M., Bustamante, C.D. et al. Melanesian blond hair is caused by an amino acid change in tyrp1. *Science*, 336(6081):554, May 2012. doi: 10.1126/science.1217849.

Kent, R.B. *Latin America: regions and people*. Guilford Publications, 2016.

Kern, A.D. and Hahn, M.W. The neutral theory in light of natural selection. *Mol Biol Evol*, 35(6):1366–1371, Jun 2018. doi: 10.1093/molbev/msy092.

Key, F.M., Fu, Q., Romagné, F., Lachmann, M. and Andrés, A.M. Human adaptation and population differentiation in the light of ancient genomes. *Nat Commun*, 7:10775, Mar 2016. doi: 10.1038/ncomms10775.

Key, F.M., Abdul-Aziz, M.A., Mundry, R., Peter, B.M., Sekar, A., D'Amato, M., Dennis, M.Y., Schmidt, J.M. and Andrés, A.M. Human local adaptation of the trpm8 cold receptor along a latitudinal cline. *PLoS Genet*, 14(5):e1007298, May 2018. doi: 10.1371/journal.pgen.1007298.

Kim, E., Lee, J.W., Baek, D.C., Lee, S.R., Kim, M.S., Kim, S.H., Imakawa, K. and Chang, K.T. Identification of novel retromer complexes in the mouse testis. *Biochem Biophys Res Commun*, 375(1):16–21, Oct 2008. doi: 10.1016/j.bbrc.2008.07.067.

Kimura, M. and Ohta, T. Protein polymorphism as a phase of molecular evolution. *Nature*, 229:467–469, 1971.

Kimura, M. et al. Evolutionary rate at the molecular level. *Nature*, 217(5129):624–626, 1968.

Kimura, R., Yamaguchi, T., Takeda, M., Kondo, O., Toma, T., Haneji, K., Hanihara, T., Matsukusa, H., Kawamura, S., Maki, K. et al. A common variation in edar is a genetic determinant of shovel-shaped incisors. *Am J Hum Genet*, 85(4):528–35, Oct 2009. doi: 10.1016/j.ajhg.2009.09.006.

Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M. and Shendure, J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet*, 46(3):310–5, Mar 2014. doi: 10.1038/ng.2892.

Kirkpatrick, M. and Kern, A. Where's the money? inversions, genes, and the hunt for genomic targets of selection. *Genetics*, 190(4):1153–1155, 2012.

Kita, R. and Fraser, H.B. Local adaptation of sun-exposure-dependent gene expression regulation in human skin. *PLoS Genet*, 12(10):e1006382, Oct 2016. doi: 10.1371/journal.pgen.1006382.

Kobayashi, T. and Hearing, V.J. Direct interaction of tyrosinase with tyrp1 to form heterodimeric complexes in vivo. *Journal of cell science*, 120(24):4261–4268, 2007.

Kobayashi, T., Imokawa, G., Bennett, D.C. and Hearing, V.J. Tyrosinase stabilization by tyrp1 (the brown locus protein). *Journal of Biological Chemistry*, 273(48):31801–31805, 1998.

Kofler, R. and Schlötterer, C. Gowinda: unbiased analysis of gene set enrichment for genome-wide association studies. *Bioinformatics*, 28(15):2084–5, Aug 2012. doi: 10.1093/bioinformatics/bts315.

Kolaczkowski, B., Kern, A.D., Holloway, A.K. and Begun, D.J. Genomic differentiation between temperate and tropical australian populations of drosophila melanogaster. *Genetics*, 187(1):245–260, 2011.

Kooner, J.S., Saleheen, D., Sim, X., Sehmi, J., Zhang, W., Frossard, P., Been, L.F., Chia, K.S., Dimas, A.S., Hassanali, N. et al. Genome-wide association study in individuals of south asian ancestry identifies six new type 2 diabetes susceptibility loci. *Nat Genet*, 43(10):984–9, Aug 2011. doi: 10.1038/ng.921.

Kreitman, M. Nucleotide polymorphism at the alcohol dehydrogenase locus of drosophila melanogaster. *Nature*, 304(5925):412, 1983.

Kwon, B.S., Haq, A.K., Pomerantz, S.H. and Halaban, R. Isolation and sequence of a cdna clone for human tyrosinase that maps at the mouse c-albino locus. *Proc Natl Acad Sci U S A*, 84(21):7473–7, Nov 1987.

Lachance, J., Vernot, B., Elbers, C.C., Ferwerda, B., Froment, A., Bodo, J.M., Lema, G., Fu, W., Nyambo, T.B., Rebbeck, T.R. et al. Evolutionary history and adaptation from high-coverage whole-genome sequences of diverse african hunter-gatherers. *Cell*, 150(3):457–69, Aug 2012. doi: 10.1016/j.cell.2012.07.009.

Laeng, B., Mathisen, R. and Johnsen, J.A. Why do blue-eyed men prefer women with the same eye color? *Behavioral Ecology and Sociobiology*, 61(3):371–384, 2007.

Lamason, R.L., Mohideen, M.A.P.K., Mest, J.R., Wong, A.C., Norton, H.L., Aros, M.C., Jurynec, M.J., Mao, X., Humphreville, V.R., Humbert, J.E. et al. Slc24a5, a putative cation exchanger, affects pigmentation in zebrafish and humans. *Science*, 310(5755): 1782–6, Dec 2005. doi: 10.1126/science.1116238.

Langner, C.A., Birkenmeier, E.H., Ben-Zeev, O., Schotz, M.C., Sweet, H.O., Davisson, M.T. and Gordon, J.I. The fatty liver dystrophy (fld) mutation. a new mutant mouse with a developmental abnormality in triglyceride metabolism and associated tissue-specific defects in lipoprotein lipase and hepatic lipase activities. *J Biol Chem*, 264(14): 7994–8003, May 1989.

Lango Allen, H., Estrada, K., Lettre, G., Berndt, S.I., Weedon, M.N., Rivadeneira, F., Willer, C.J., Jackson, A.U., Vedantam, S., Raychaudhuri, S. et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, 467 (7317):832–8, Oct 2010. doi: 10.1038/nature09410.

Lasisi, T. and Shriver, M.D. Focus on african diversity confirms complexity of skin pigmentation genetics. *Genome Biol*, 19(1):13, 01 2018. doi: 10.1186/s13059-018-1395-3.

Latta, R.G. Differentiation of allelic frequencies at quantitative trait loci affecting locally adaptive traits. *Am Nat*, 151(3):283–92, Mar 1998. doi: 10.1086/286119.

Latta, R.G. Gene flow, adaptive population divergence and comparative population structure across loci. *New Phytologist*, 161(1):51–58, 2004.

Lattka, E., Illig, T., Koletzko, B. and Heinrich, J. Genetic variants of the fads1 fads2 gene cluster as related to essential fatty acid metabolism. *Curr Opin Lipidol*, 21(1):64–9, Feb 2010. doi: 10.1097/MOL.0b013e3283327ca8.

Law, M.H., Bishop, D.T., Lee, J.E., Brossard, M., Martin, N.G., Moses, E.K., Song, F., Barrett, J.H., Kumar, R., Easton, D.F. et al. Genome-wide meta-analysis identifies five new susceptibility loci for cutaneous malignant melanoma. *Nat Genet*, 47(9):987–995, Sep 2015. doi: 10.1038/ng.3373.

Law, M.H., Medland, S.E., Zhu, G., Yazar, S., Viñuela, A., Wallace, L., Shekar, S.N., Duffy, D.L., Bataille, V., Glass, D. et al. Genome-wide association shows that pigmentation genes play a role in skin aging. *J Invest Dermatol*, 137(9):1887–1894, Sep 2017. doi: 10.1016/j.jid.2017.04.026.

Lawson, D.J., Hellenthal, G., Myers, S. and Falush, D. Inference of population structure using dense haplotype data. *PLoS Genet*, 8(1):e1002453, Jan 2012. doi: 10.1371/journal.pgen.1002453.

Lazaridis, I., Patterson, N., Mittnik, A., Renaud, G., Mallick, S., Kirsanow, K., Sudmant, P.H., Schraiber, J.G., Castellano, S., Lipson, M. et al. Ancient human genomes suggest

three ancestral populations for present-day europeans. *Nature*, 513(7518):409–13, Sep 2014. doi: 10.1038/nature13673.

Le Corre, V. and Kremer, A. The genetic differentiation at quantitative trait loci under local adaptation. *Molecular ecology*, 21(7):1548–1566, 2012.

Lee, S., Zhang, C., Kilicarslan, M., Piening, B.D., Bjornson, E., Hallström, B.M., Groen, A.K., Ferrannini, E., Laakso, M., Snyder, M. et al. Integrated network analysis reveals an association between plasma mannose levels and insulin resistance. *Cell Metab*, 24 (1):172–84, Jul 2016. doi: 10.1016/j.cmet.2016.05.026.

Lell, J.T., Brown, M.D., Schurr, T.G., Sukernik, R.I., Starikovskaya, Y.B., Torroni, A., Moore, L.G., Troup, G.M. and Wallace, D.C. Y chromosome polymorphisms in native american and siberian populations: identification of native american y chromosome haplotypes. *Hum Genet*, 100(5-6):536–43, Oct 1997.

Lewontin, R.C. and Krakauer, J. Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics*, 74(1):175–95, May 1973.

Lewontin, R.C. The apportionment of human diversity. In *Evolutionary biology*, pages 381–398. Springer, 1972.

Li, H. and Stephan, W. Inferring the demographic history and rate of adaptive substitution in drosophila. *PLoS Genet*, 2(10):e166, Oct 2006. doi: 10.1371/journal.pgen.0020166.

Li, J., Song, J.S., Bell, R.J.A., Tran, T.N.T., Haq, R., Liu, H., Love, K.T., Langer, R., Anderson, D.G., Larue, L. et al. Yy1 regulates melanocyte development and function by cooperating with mitf. *PLoS Genet*, 8(5):e1002688, 2012. doi: 10.1371/journal.pgen.1002688.

Li, Y.I., van de Geijn, B., Raj, A., Knowles, D.A., Petti, A.A., Golan, D., Gilad, Y. and Pritchard, J.K. Rna splicing is a primary link between genetic variation and disease. *Science*, 352(6285):600–4, Apr 2016. doi: 10.1126/science.aad9417.

Li, Z., Chen, J., Yu, H., He, L., Xu, Y., Zhang, D., Yi, Q., Li, C., Li, X., Shen, J. et al. Genome-wide association analysis identifies 30 new susceptibility loci for schizophrenia. *Nat Genet*, 49(11):1576–1583, Nov 2017. doi: 10.1038/ng.3973.

Lieberman, D.E. and Bramble, D.M. The evolution of marathon running. *Sports Medicine*, 37(4-5):288–290, 2007.

Lin, B.D., Mbarek, H., Willemsen, G., Dolan, C.V., Fedko, I.O., Abdellaoui, A., de Geus, E.J., Boomsma, D.I. and Hottenga, J.J. Heritability and genome-wide association studies for hair color in a dutch twin family based sample. *Genes*, 6(3):559–576, 2015.

Lin, K., Li, H., Schlötterer, C. and Futschik, A. Distinguishing positive selection from neutral evolution: boosting the performance of summary statistics. *Genetics*, 187(1): 229–44, Jan 2011a. doi: 10.1534/genetics.110.122614.

Lin, Y.C., Chen, C.C., Cheng, C.J. and Yang, R.B. Domain and functional analysis of a novel breast tumor suppressor protein, scube2. *J Biol Chem*, 286(30):27039–47, Jul 2011b. doi: 10.1074/jbc.M111.244418.

Lindenau, J.D., Guimarães, L.S.P., Hurtado, A.M., Hill, K.R., Tsuneto, L.T., Salzano, F.M., Petzl-Erler, M.L. and Hutz, M.H. Association between hla-dr4 haplotypes and tuberculin skin test response in the aché population. *Tissue Antigens*, 84(5):479–83, Nov 2014a. doi: 10.1111/tan.12451.

Lindenau, J., Guimarães, L., Hurtado, A., Hill, K., Tsuneto, L., Salzano, F., Petzl-Erler, M. and Hutz, M. Association between hla-dr4 haplotypes and tuberculin skin test response in the aché population. *HLA*, 84(5):479–483, 2014b.

Lindenau, J.D.R., Salzano, F.M., Hurtado, A.M., Hill, K.R., Petzl-Erler, M.L., Tsuneto, L.T. and Hutz, M.H. Variability of innate immune system genes in native american populations—relationship with history and epidemiology. *American journal of physical anthropology*, 159(4):722–728, 2016.

Lindo, J., Huerta-Sánchez, E., Nakagome, S., Rasmussen, M., Petzelt, B., Mitchell, J., Cybulski, J.S., Willerslev, E., DeGiorgio, M. and Malhi, R.S. A time transect of exomes from a native american population before and after european contact. *Nat Commun*, 7: 13175, Nov 2016. doi: 10.1038/ncomms13175.

Lippert, C., Listgarten, J., Liu, Y., Kadie, C.M., Davidson, R.I. and Heckerman, D. Fast linear mixed models for genome-wide association studies. *Nature methods*, 8(10):833, 2011.

Liu, C.T., Monda, K.L., Taylor, K.C., Lange, L., Demerath, E.W., Palmas, W., Wojczynski, M.K., Ellis, J.C., Vitolins, M.Z., Liu, S. et al. Genome-wide association of body fat distribution in african ancestry populations suggests new loci. *PLoS Genet*, 9(8): e1003681, 2013a. doi: 10.1371/journal.pgen.1003681.

Liu, F., van Duijn, K., Vingerling, J.R., Hofman, A., Uitterlinden, A.G., Janssens, A.C.J.W. and Kayser, M. Eye color and the prediction of complex phenotypes from genotypes. *Curr Biol*, 19(5):R192–3, Mar 2009. doi: 10.1016/j.cub.2009.01.027.

Liu, F., Wollstein, A., Hysi, P.G., Ankra-Badu, G.A., Spector, T.D., Park, D., Zhu, G., Larsson, M., Duffy, D.L., Montgomery, G.W. et al. Digital quantification of human eye color highlights genetic association of three new loci. *PLoS Genet*, 6(5):e1000934, May 2010. doi: 10.1371/journal.pgen.1000934.

Liu, F., Wen, B. and Kayser, M. Colorful dna polymorphisms in humans. *Semin Cell Dev Biol*, 24(6-7):562–75, 2013b. doi: 10.1016/j.semcdb.2013.03.013.

Liu, F., Visser, M., Duffy, D.L., Hysi, P.G., Jacobs, L.C., Lao, O., Zhong, K., Walsh, S., Chaitanya, L., Wollstein, A. et al. Genetics of skin color variation in europeans: genome-wide association studies with functional follow-up. *Hum Genet*, 134(8):823–35, Aug 2015. doi: 10.1007/s00439-015-1559-0.

Liu, F., Hamer, M.A., Deelen, J., Lall, J.S., Jacobs, L., van Heemst, D., Murray, P.G., Wollstein, A., de Craen, A.J.M., Uh, H.W. et al. The mc1r gene and youthful looks. *Curr Biol*, 26(9):1213–20, 05 2016. doi: 10.1016/j.cub.2016.03.008.

Llamas, B., Fehren-Schmitz, L., Valverde, G., Soubrier, J., Mallick, S., Rohland, N., Nordenfelt, S., Valdiosera, C., Richards, S.M., Rohrlach, A. et al. Ancient mitochondrial dna provides high-resolution time scale of the peopling of the americas. *Sci Adv*, 2(4): e1501385, Apr 2016. doi: 10.1126/sciadv.1501385.

Lloyd-Jones, L.R., Robinson, M.R., Moser, G., Zeng, J., Beleza, S., Barsh, G.S., Tang, H. and Visscher, P.M. Inference on the genetic basis of eye and skin color in an admixed population via bayesian linear mixed models. *Genetics*, 206(2):1113–1126, 06 2017. doi: 10.1534/genetics.116.193383.

Locke, A.E., Kahali, B., Berndt, S.I., Justice, A.E., Pers, T.H., Day, F.R., Powell, C., Vedantam, S., Buchkovich, M.L., Yang, J. et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature*, 518(7538):197–206, Feb 2015. doi: 10.1038/nature14177.

Loh, P.R., Kichaev, G., Gazal, S., Schoech, A.P. and Price, A.L. Mixed-model association for biobank-scale datasets. *Nat Genet*, Jun 2018. doi: 10.1038/s41588-018-0144-6.

Long, J.C. The genetic structure of admixed populations. *Genetics*, 127(2):417–28, Feb 1991.

Long, J.C., Knowler, W.C., Hanson, R.L., Robin, R.W., Urbanek, M., Moore, E., Bennett, P.H., Goldman, D. et al. Evidence for genetic linkage to alcohol dependence on chromosomes 4 and 11 from an autosome-wide scan in an american indian population. *American journal of medical genetics*, 81(3):216–221, 1998.

Loomis, W.F. Skin-pigment regulation of vitamin-d biosynthesis in man: Variation in solar ultraviolet at different latitudes may have caused racial differentiation in man. *Science*, 157(3788):501–506, 1967.

Loos, R.J.F., Rankinen, T., Pérusse, L., Tremblay, A., Després, J.P. and Bouchard, C. Association of lipin 1 gene polymorphisms with measures of energy and glucose metabolism. *Obesity (Silver Spring)*, 15(11):2723–32, Nov 2007. doi: 10.1038/oby.2007.324.

López, D.L., Bundschuh, J., Birkle, P., Armienta, M.A., Cumbal, L., Sracek, O., Cornejo, L. and Ormachea, M. Arsenic in volcanic geothermal fluids of latin america. *Sci Total Environ*, 429:57–75, Jul 2012. doi: 10.1016/j.scitotenv.2011.08.043.

López, S., García, Ó., Yurrebaso, I., Flores, C., Acosta-Herrera, M., Chen, H., Gardeazabal, J., Careaga, J.M., Boyano, M.D., Sánchez, A. et al. The interplay between natural selection and susceptibility to melanoma on allele 374f of slc45a2 gene in a south european population. *PloS one*, 9(8):e104367, 2014.

Lucock, M. Folic acid: nutritional biochemistry, molecular biology, and role in disease processes. *Molecular genetics and metabolism*, 71(1):121–138, 2000.

Luczak, S.E., Glatt, S.J. and Wall, T.L. Meta-analyses of aldh2 and adh1b with alcohol dependence in asians. *Psychol Bull*, 132(4):607–21, Jul 2006. doi: 10.1037/0033-2909.132.4.607.

Lyons, L.A., Foe, I.T., Rah, H.C. and Grahn, R.A. Chocolate coated cats: Tyrp1 mutations for brown color in domestic cats. *Mamm Genome*, 16(5):356–66, May 2005.

Makova, K. and Norton, H. Worldwide polymorphism at the mc1r locus and normal pigmentation variation in humans. *Peptides*, 26(10):1901–8, Oct 2005. doi: 10.1016/j.peptides.2004.12.032.

Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., Zhao, M., Chennagiri, N., Nordenfelt, S., Tandon, A. et al. The simons genome diversity project: 300 genomes from 142 diverse populations. *Nature*, 538(7624):201–206, Oct 2016. doi: 10.1038/nature18964.

Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M. and Chen, W.M. Robust relationship inference in genome-wide association studies. *Bioinformatics*, 26(22):2867–73, Nov 2010. doi: 10.1093/bioinformatics/btq559.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A. et al. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747, 2009.

Maples, B.K., Gravel, S., Kenny, E.E. and Bustamante, C.D. Rfmix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am J Hum Genet*, 93(2):278–88, Aug 2013. doi: 10.1016/j.ajhg.2013.06.020.

Marchini, J., Cardon, L.R., Phillips, M.S. and Donnelly, P. The effects of human population structure on large genetic association studies. *Nature genetics*, 36(5):512, 2004.

Mariat, D., Taourit, S. and Guérin, G. A mutation in the matp gene causes the cream coat colour in the horse. *Genet Sel Evol*, 35(1):119–33, 2003. doi: 10.1051/gse:2002039.

Marigorta, U.M. and Navarro, A. High trans-ethnic replicability of gwas results implies common causal variants. *PLoS genetics*, 9(6):e1003566, 2013.

Marouli, E., Graff, M., Medina-Gomez, C., Lo, K.S., Wood, A.R., Kjaer, T.R., Fine, R.S., Lu, Y., Schurmann, C., Highland, H.M. et al. Rare and low-frequency coding variants alter human adult height. *Nature*, 542(7640):186–190, 02 2017. doi: 10.1038/nature21039.

Márquez, A., Vidal-Bralo, L., Rodríguez-Rodríguez, L., González-Gay, M.A., Balsa, A., González-Álvaro, I., Carreira, P., Ortego-Centeno, N., Ayala-Gutiérrez, M.M., García-Hernández, F.J. et al. A combined large-scale meta-analysis identifies cog6 as a novel shared risk locus for rheumatoid arthritis and systemic lupus erythematosus. *Ann Rheum Dis*, 76(1):286–294, Jan 2017. doi: 10.1136/annrheumdis-2016-209436.

Martin, A.R., Gignoux, C.R., Walters, R.K., Wojcik, G.L., Neale, B.M., Gravel, S., Daly, M.J., Bustamante, C.D. and Kenny, E.E. Human demographic history impacts genetic risk prediction across diverse populations. *Am J Hum Genet*, 100(4):635–649, Apr 2017a. doi: 10.1016/j.ajhg.2017.03.004.

Martin, A.R., Lin, M., Granka, J.M., Myrick, J.W., Liu, X., Sockell, A., Atkinson, E.G., Werely, C.J., Möller, M., Sandhu, M.S. et al. An unexpectedly complex architecture for skin pigmentation in africans. *Cell*, 171(6):1340–1353.e14, Nov 2017b. doi: 10.1016/j.cell.2017.11.015.

Martin, A.R., Solomon, T., Marlo, M., Eileen G., H. and Mark J., D. The critical needs and challenges for genetic architecture studies in africa. *Preprints*, 2018. doi: 10.20944/preprints201806.0201.v1. URL `https://www.preprints.org/manuscript/201806.0201/v1`.

Mathieson, I., Lazaridis, I., Rohland, N., Mallick, S., Patterson, N., Roodenberg, S.A., Harney, E., Stewardson, K., Fernandes, D., Novak, M. et al. Genome-wide patterns of selection in 230 ancient eurasians. *Nature*, 528(7583):499–503, Dec 2015. doi: 10.1038/nature16152.

Mathieson, I., Alpaslan-Roodenberg, S., Posth, C., Szécsényi-Nagy, A., Rohland, N., Mallick, S., Olalde, I., Broomandkhoshbacht, N., Candilio, F., Cheronet, O. et al. The genomic history of southeastern europe. *Nature*, 555(7695):197–203, Mar 2018. doi: 10.1038/nature25778.

Matsuda, A., Suzuki, Y., Honda, G., Muramatsu, S., Matsuzaki, O., Nagano, Y., Doi, T., Shimotohno, K., Harada, T., Nishida, E. et al. Large-scale identification and characterization of human genes that activate nf-kappab and mapk signaling pathways. *Oncogene*, 22(21):3307–18, May 2003. doi: 10.1038/sj.onc.1206406.

Matsushita, S. and Higuchi, S. Review: Use of asian samples in genetic research of alcohol use disorders: Genetic variation of alcohol metabolizing enzymes and the effects of acetaldehyde. *Am J Addict*, 26(5):469–476, Aug 2017. doi: 10.1111/ajad.12477.

Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J. et al. Systematic localization of common disease-associated variation in regulatory dna. *Science*, 337(6099):1190–5, Sep 2012. doi: 10.1126/science.1222794.

Maystadt, I., Rezsöhazy, R., Barkats, M., Duque, S., Vannuffel, P., Remacle, S., Lambert, B., Najimi, M., Sokal, E., Munnich, A. et al. The nuclear factor kappab-activator gene plekhg5 is mutated in a form of autosomal recessive lower motor neuron disease with childhood onset. *Am J Hum Genet*, 81(1):67–76, Jul 2007. doi: 10.1086/518900.

McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A.R., Teumer, A., Kang, H.M., Fuchsberger, C., Danecek, P., Sharp, K. et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet*, 48(10):1279–83, 10 2016. doi: 10.1038/ng.3643.

McDonald, J.H. and Kreitman, M. Adaptive protein evolution at the adh locus in drosophila. *Nature*, 351(6328):652, 1991.

McEvoy, B., Beleza, S. and Shriver, M.D. The genetic architecture of normal variation in human pigmentation: an evolutionary perspective and model. *Hum Mol Genet*, 15 Spec No 2:R176–81, Oct 2006. doi: 10.1093/hmg/ddl217.

Meier, U.T. and Blobel, G.n. A nuclear localization signal binding protein in the nucleolus. *The Journal of cell biology*, 111(6):2235–2245, 1990.

Meier, U.T. and Blobel, G. Nopp 140 shuttles on tracks between nucleolus and cytoplasm. *Cell*, 70(1):127–138, 1992.

Meredith, P. and Sarna, T. The physical and chemical properties of eumelanin. *Pigment Cell & Melanoma Research*, 19(6):572–594, 2006.

Messer, P.W. and Petrov, D.A. Population genomics of rapid adaptation by soft selective sweeps. *Trends in ecology & evolution*, 28(11):659–669, 2013.

Meyer, M., Kircher, M., Gansauge, M.T., Li, H., Racimo, F., Mallick, S., Schraiber, J.G., Jay, F., Prüfer, K., de Filippo, C. et al. A high-coverage genome sequence from an archaic denisovan individual. *Science*, 338(6104):222–6, Oct 2012. doi: 10.1126/science.1224344.

Miller, C.T., Beleza, S., Pollen, A.A., Schluter, D., Kittles, R.A., Shriver, M.D. and Kingsley, D.M. cis-regulatory changes in kit ligand expression and parallel evolution of pigmentation in sticklebacks and humans. *Cell*, 131(6):1179–89, Dec 2007. doi: 10.1016/j.cell.2007.10.055.

Mitchell-Olds, T., Willis, J.H. and Goldstein, D.B. Which evolutionary processes influence natural genetic variation for phenotypic traits? *Nat Rev Genet*, 8(11):845–56, Nov 2007. doi: 10.1038/nrg2207.

Moltke, I., Fumagalli, M., Korneliussen, T.S., Crawford, J.E., Bjerregaard, P., Jørgensen, M.E., Grarup, N., Gulløv, H.C., Linneberg, A., Pedersen, O. et al. Uncovering the genetic history of the present-day greenlandic population. *Am J Hum Genet*, 96(1):54–69, Jan 2015. doi: 10.1016/j.ajhg.2014.11.012.

Moore, L.G., Hershey, D.W., Jahnigen, D. and Bowes, Jr, W. The incidence of pregnancy-induced hypertension is increased among colorado residents at high altitude. *Am J Obstet Gynecol*, 144(4):423–9, Oct 1982.

Moore, L.G., Shriver, M., Bemis, L., Hickler, B., Wilson, M., Brutsaert, T., Parra, E. and Vargas, E. Maternal adaptation to high-altitude pregnancy: an experiment of nature–a review. *Placenta*, 25 Suppl A:S60–71, Apr 2004. doi: 10.1016/j.placenta.2004.01.008.

Moreno-Estrada, A., Gravel, S., Zakharia, F., McCauley, J.L., Byrnes, J.K., Gignoux, C.R., Ortiz-Tello, P.A., Martínez, R.J., Hedges, D.J., Morris, R.W. et al. Reconstructing the population genetic history of the caribbean. *PLoS Genet*, 9(11):e1003925, Nov 2013. doi: 10.1371/journal.pgen.1003925.

Moreno-Estrada, A., Gignoux, C.R., Fernández-López, J.C., Zakharia, F., Sikora, M., Contreras, A.V., Acuña-Alonzo, V., Sandoval, K., Eng, C., Romero-Hidalgo, S. et al. Human genetics. the genetics of mexico recapitulates native american substructure and affects biomedical traits. *Science*, 344(6189):1280–5, Jun 2014. doi: 10.1126/science.1251688.

Moreno-Mayar, J.V., Potter, B.A., Vinner, L., Steinrücken, M., Rasmussen, S., Terhorst, J., Kamm, J.A., Albrechtsen, A., Malaspinas, A.S., Sikora, M. et al. Terminal pleistocene alaskan genome reveals first founding population of native americans. *Nature*, 553(7687):203–207, Jan 2018. doi: 10.1038/nature25173.

Mori, S., Kou, I., Sato, H., Emi, M., Ito, H., Hosoi, T. and Ikegawa, S. Nucleotide variations in genes encoding carbonic anhydrase 8 and 10 associated with femoral bone mineral density in japanese female with osteoporosis. *J Bone Miner Metab*, 27(2):213–6, 2009. doi: 10.1007/s00774-008-0031-9.

Morris, A.P., Voight, B.F., Teslovich, T.M., Ferreira, T., Segrè, A.V., Steinthorsdottir, V., Strawbridge, R.J., Khan, H., Grallert, H., Mahajan, A. et al. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat Genet*, 44(9):981–90, Sep 2012. doi: 10.1038/ng.2383.

Morris, D.L., Sheng, Y., Zhang, Y., Wang, Y.F., Zhu, Z., Tombleson, P., Chen, L., Graham, D.S.C., Bentham, J., Roberts, A.L. et al. Genome-wide association meta-analysis in chinese and european individuals identifies ten new loci associated with systemic lupus erythematosus. *Nature genetics*, 48(8):940, 2016.

Mörseburg, A., Pagani, L., Ricaut, F.X., Yngvadottir, B., Harney, E., Castillo, C., Hoogervorst, T., Antao, T., Kusuma, P., Brucato, N. et al. Multi-layered population structure in island southeast asians. *Eur J Hum Genet*, 24(11):1605–1611, 11 2016. doi: 10.1038/ejhg.2016.60.

Mosaad, Y.M. Clinical role of human leukocyte antigen in health and disease. *Scand J Immunol*, 82(4):283–306, Oct 2015. doi: 10.1111/sji.12329.

Moskvina, V., Smith, M., Ivanov, D., Blackwood, D., StClair, D., Hultman, C., Toncheva, D., Gill, M., Corvin, A., O'Dushlaine, C. et al. Genetic differences between five european populations. *Human heredity*, 70(2):141–149, 2010.

Mou, C., Thomason, H.A., Willan, P.M., Clowes, C., Harris, W.E., Drew, C.F., Dixon, J., Dixon, M.J. and Headon, D.J. Enhanced ectodysplasin-a receptor (edar) signaling alters multiple fiber characteristics to produce the east asian hair form. *Hum Mutat*, 29 (12):1405–11, Dec 2008. doi: 10.1002/humu.20795.

Mullen, L.M. and Hoekstra, H.E. Natural selection along an environmental gradient: a classic cline in mouse pigmentation. *Evolution*, 62(7):1555–1570, 2008.

Mulligan, C.J., Robin, R.W., Osier, M.V., Sambuughin, N., Goldfarb, L.G., Kittles, R.A., Hesselbrock, D., Goldman, D. and Long, J.C. Allelic variation at alcohol metabolism

genes (adh1b, adh1c, aldh2) and alcohol dependence in an american indian population. *Human genetics*, 113(4):325–336, 2003.

Mulligan, C.J., Hunley, K., Cole, S. and Long, J.C. Population genetics, history, and health patterns in native americans. *Annu. Rev. Genomics Hum. Genet.*, 5:295–315, 2004.

Mundy, N.I., Kelly, J., Theron, E. and Hawkins, K. Evolutionary genetics of the melanocortin-1 receptor in vertebrates. *Ann N Y Acad Sci*, 994:307–12, Jun 2003.

Murray, F.G. Pigmentation, sunlight, and nutritional disease. *American Anthropologist*, 36(3):438–445, 1934.

Murray, N., Norton, H.L. and Parra, E.J. Distribution of two oca2 polymorphisms associated with pigmentation in east-asian populations. *Hum Genome Var*, 2:15058, 2015. doi: 10.1038/hgv.2015.58.

Mychaleckyj, J.C., Havt, A., Nayak, U., Pinkerton, R., Farber, E., Concannon, P., Lima, A.A. and Guerrant, R.L. Genome-wide analysis in brazilians reveals highly differentiated native american genome regions. *Mol Biol Evol*, 34(3):559–574, 03 2017. doi: 10.1093/molbev/msw249.

Nadkarni, N.A., Weale, M.E., Von Schantz, M. and Thomas, M.G. Evolution of a length polymorphism in the human per3 gene, a component of the circadian system. *Journal of biological rhythms*, 20(6):490–499, 2005.

Nagahara, N., Ito, T., Kitamura, H. and Nishino, T. Tissue and subcellular distribution of mercaptopyruvate sulfurtransferase in the rat: confocal laser fluorescence and immunoelectron microscopic studies combined with biochemical analysis. *Histochem Cell Biol*, 110(3):243–50, Sep 1998.

Nakayama, K., Soemantri, A., Jin, F., Dashnyam, B., Ohtsuka, R., Duanchang, P., Isa, M.N., Settheetham-Ishida, W., Harihara, S. and Ishida, T. Identification of novel functional variants of the melanocortin 1 receptor gene originated from asians. *Hum Genet*, 119(3):322–30, Apr 2006. doi: 10.1007/s00439-006-0141-1.

Nam, D., Kim, J., Kim, S.Y. and Kim, S. Gsa-snp: a general approach for gene set analysis of polymorphisms. *Nucleic Acids Res*, 38(Web Server issue):W749–54, Jul 2010. doi: 10.1093/nar/gkq428.

Nan, H., Kraft, P., Qureshi, A.A., Guo, Q., Chen, C., Hankinson, S.E., Hu, F.B., Thomas, G., Hoover, R.N., Chanock, S. et al. Genome-wide association study of tanning phenotype in a population of european ancestry. *Journal of Investigative Dermatology*, 129 (9):2250–2257, 2009.

Nariai, N., Kojima, K., Saito, S., Mimori, T., Sato, Y., Kawai, Y., Yamaguchi-Kabata, Y., Yasuda, J. and Nagasaki, M. Hla-vbseq: accurate hla typing at full resolution from whole-genome sequencing data. *BMC Genomics*, 16 Suppl 2:S7, 2015. doi: 10.1186/1471-2164-16-S2-S7.

Naysmith, L., Waterston, K., Ha, T., Flanagan, N., Bisset, Y., Ray, A., Wakamatsu, K., Ito, S. and Rees, J.L. Quantitative measures of the effect of the melanocortin 1 receptor on human pigmentary status. *J Invest Dermatol*, 122(2):423–8, Feb 2004. doi: 10.1046/j.0022-202X.2004.22221.x.

Need, A.C. and Goldstein, D.B. Next generation disparities in human genomics: concerns and remedies. *Trends Genet*, 25(11):489–94, Nov 2009. doi: 10.1016/j.tig.2009.09.012.

Neel, J.V. Diabetes mellitus: a "thrifty" genotype rendered detrimental by "progress"? *American journal of human genetics*, 14(4):353, 1962.

Nelson, W.C., Pyo, C.W., Vogan, D., Wang, R., Pyon, Y.S., Hennessey, C., Smith, A., Pereira, S., Ishitani, A. and Geraghty, D.E. An integrated genotyping approach for hla and other complex genetic systems. *Hum Immunol*, 76(12):928–38, Dec 2015. doi: 10.1016/j.humimm.2015.05.001.

Nettle, D. Height and reproductive success in a cohort of british men. *Hum Nat*, 13(4): 473–91, Dec 2002. doi: 10.1007/s12110-002-1004-7.

Newton, J.M., Cohen-Barak, O., Hagiwara, N., Gardner, J.M., Davisson, M.T., King, R.A. and Brilliant, M.H. Mutations in the human orthologue of the mouse underwhite gene (uw) underlie a new form of oculocutaneous albinism, oca4. *Am J Hum Genet*, 69 (5):981–8, Nov 2001. doi: 10.1086/324340.

Ng, M.C.Y., Graff, M., Lu, Y., Justice, A.E., Mudgal, P., Liu, C.T., Young, K., Yanek, L.R., Feitosa, M.F., Wojczynski, M.K. et al. Discovery and fine-mapping of adiposity loci using high density imputation of genome-wide association studies in individuals of african ancestry: African ancestry anthropometry genetics consortium. *PLoS Genet*, 13 (4):e1006719, Apr 2017. doi: 10.1371/journal.pgen.1006719.

Nielsen, R., Hellmann, I., Hubisz, M., Bustamante, C. and Clark, A.G. Recent and ongoing selection in the human genome. *Nat Rev Genet*, 8(11):857–68, Nov 2007. doi: 10.1038/nrg2187.

Nielsen, R., Akey, J.M., Jakobsson, M., Pritchard, J.K., Tishkoff, S. and Willerslev, E. Tracing the peopling of the world through genomics. *Nature*, 541(7637):302–310, 01 2017. doi: 10.1038/nature21347.

Nikolskiy, P. and Pitulko, V. Evidence from the yana palaeolithic site, arctic siberia, yields clues to the riddle of mammoth hunting. *Journal of Archaeological Science*, 40 (12):4189–4197, 2013.

Norton, H.L., Friedlaender, J.S., Merriwether, D.A., Koki, G., Mgone, C.S. and Shriver, M.D. Skin and hair pigmentation variation in island melanesia. *Am J Phys Anthropol*, 130(2):254–68, Jun 2006. doi: 10.1002/ajpa.20343.

Norton, H.L., Kittles, R.A., Parra, E., McKeigue, P., Mao, X., Cheng, K., Canfield, V.A., Bradley, D.G., McEvoy, B. and Shriver, M.D. Genetic evidence for the convergent

evolution of light skin in europeans and east asians. *Mol Biol Evol*, 24(3):710–22, Mar 2007. doi: 10.1093/molbev/msl203.

Norton, H.L., Edwards, M., Krithika, S., Johnson, M., Werren, E.A. and Parra, E.J. Quantitative assessment of skin, hair, and iris variation in a diverse sample of individuals and associated genetic variation. *American journal of physical anthropology*, 160(4):570–581, 2016.

Novembre, J. and Barton, N.H. Tread lightly interpreting polygenic tests of selection. *Genetics*, 208(4):1351–1355, Apr 2018. doi: 10.1534/genetics.118.300786.

Ohta, S., Ohsawa, I., Kamino, K., Ando, F. and Shimokata, H. Mitochondrial aldh2 deficiency as an oxidative stress. *Ann N Y Acad Sci*, 1011:36–44, Apr 2004.

Okada, Y., Momozawa, Y., Sakaue, S., Kanai, M., Ishigaki, K., Akiyama, M., Kishikawa, T., Arai, Y., Sasaki, T., Kosaki, K. et al. Deep whole-genome sequencing reveals recent selection signatures linked to evolution and disease risk of japanese. *Nat Commun*, 9 (1):1631, Apr 2018. doi: 10.1038/s41467-018-03274-0.

Olalde, I., Allentoft, M.E., Sánchez-Quinto, F., Santpere, G., Chiang, C.W.K., DeGiorgio, M., Prado-Martinez, J., Rodríguez, J.A., Rasmussen, S., Quilez, J. et al. Derived immune and ancestral pigmentation alleles in a 7,000-year-old mesolithic european. *Nature*, 507(7491):225–8, Mar 2014. doi: 10.1038/nature12960.

Oleksyk, T.K., Smith, M.W. and O'Brien, S.J. Genome-wide scans for footprints of natural selection. *Philos Trans R Soc Lond B Biol Sci*, 365(1537):185–205, Jan 2010. doi: 10.1098/rstb.2009.0219.

Orlando, L., Ginolhac, A., Zhang, G., Froese, D., Albrechtsen, A., Stiller, M., Schubert, M., Cappellini, E., Petersen, B., Moltke, I. et al. Recalibrating equus evolution using the genome sequence of an early middle pleistocene horse. *Nature*, 499(7456):74–8, Jul 2013. doi: 10.1038/nature12323.

Owers, K.A., Sjödin, P., Schlebusch, C.M., Skoglund, P., Soodyall, H. and Jakobsson, M. Adaptation to infectious disease exposure in indigenous southern african populations. *Proc Biol Sci*, 284(1852), Apr 2017. doi: 10.1098/rspb.2017.0226.

Parenti, F., Mercier, N. and Valladas, H. The oldest hearths of pedra furada, brasil: thermoluminescence analysis of heated stones. *Current Research in the Pleistocene*, 7: 36–38, 1990.

Park, J.H., Yamaguchi, T., Watanabe, C., Kawaguchi, A., Haneji, K., Takeda, M., Kim, Y.I., Tomoyasu, Y., Watanabe, M., Oota, H. et al. Effects of an asian-specific nonsynonymous edar variant on multiple dental traits. *J Hum Genet*, 57(8):508–14, Aug 2012. doi: 10.1038/jhg.2012.60.

Park, S., Morya, V.K., Nguyen, D.H., Singh, B.K., Lee, H.B. and Kim, E.K. Unrevealing the role of p-protein on melanosome biology and structure, using sirna-mediated down regulation of oca2. *Mol Cell Biochem*, 403(1-2):61–71, May 2015. doi: 10.1007/s11010-015-2337-y.

Parra, E.J. Human pigmentation variation: evolution, genetic basis, and implications for public health. *Am J Phys Anthropol*, Suppl 45:85–105, 2007. doi: 10.1002/ajpa.20727.

Partonen, T., Treutlein, J., Alpman, A., Frank, J., Johansson, C., Depner, M., Aron, L., Rietschel, M., Wellek, S., Soronen, P. et al. Three circadian clock genes per2, arntl, and npas2 contribute to winter depression. *Ann Med*, 39(3):229–38, 2007. doi: 10.1080/07853890701278795.

Pasaniuc, B., Sankararaman, S., Torgerson, D.G., Gignoux, C., Zaitlen, N., Eng, C., Rodriguez-Cintron, W., Chapela, R., Ford, J.G., Avila, P.C. et al. Analysis of latino populations from gala and mec studies reveals genomic loci with biased local ancestry estimation. *Bioinformatics*, 29(11):1407–15, Jun 2013. doi: 10.1093/bioinformatics/btt166.

Patterson, N., Hattangadi, N., Lane, B., Lohmueller, K.E., Hafler, D.A., Oksenberg, J.R., Hauser, S.L., Smith, M.W., O'Brien, S.J., Altshuler, D. et al. Methods for high-density admixture mapping of disease genes. *Am J Hum Genet*, 74(5):979–1000, May 2004. doi: 10.1086/420871.

Patterson, N., Price, A.L. and Reich, D. Population structure and eigenanalysis. *PLoS genetics*, 2(12):e190, 2006.

Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T. and Reich, D. Ancient admixture in human history. *Genetics*, 192(3):1065–93, Nov 2012. doi: 10.1534/genetics.112.145037.

Pavlidis, P., Jensen, J.D. and Stephan, W. Searching for footprints of positive selection in whole-genome snp data from nonequilibrium populations. *Genetics*, 185(3):907–22, Jul 2010. doi: 10.1534/genetics.110.116459.

Pavlidis, P., Jensen, J.D., Stephan, W. and Stamatakis, A. A critical assessment of storytelling: gene ontology categories and the importance of validating genomic scans. *Mol Biol Evol*, 29(10):3237–48, Oct 2012. doi: 10.1093/molbev/mss136.

Peng, Q., Li, J., Tan, J., Yang, Y., Zhang, M., Wu, S., Liu, Y., Zhang, J., Qin, P., Guan, Y. et al. Edarv370a associated facial characteristics in uyghur population revealing further pleiotropic effects. *Hum Genet*, 135(1):99–108, Jan 2016. doi: 10.1007/s00439-015-1618-6.

Pennings, P.S. and Hermisson, J. Soft sweeps ii–molecular population genetics of adaptation from recurrent mutation or migration. *Mol Biol Evol*, 23(5):1076–84, May 2006a. doi: 10.1093/molbev/msj117.

Pennings, P.S. and Hermisson, J. Soft sweeps iii: the signature of positive selection from recurrent mutation. *PLoS Genet*, 2(12):e186, Dec 2006b. doi: 10.1371/journal.pgen.0020186.

Perry, G.H., Dominy, N.J., Claw, K.G., Lee, A.S., Fiegler, H., Redon, R., Werner, J., Villanea, F.A., Mountain, J.L., Misra, R. et al. Diet and the evolution of human amylase gene copy number variation. *Nature genetics*, 39(10):1256, 2007.

Perry, J.R., Day, F., Elks, C.E., Sulem, P., Thompson, D.J., Ferreira, T., He, C., Chasman, D.I., Esko, T., Thorleifsson, G. et al. Parent-of-origin-specific allelic associations among 106 genomic loci for age at menarche. *Nature*, 514(7520):92–97, Oct 2014. doi: 10.1038/nature13545.

Peter, B.M., Huerta-Sanchez, E. and Nielsen, R. Distinguishing between selective sweeps from standing variation and from a de novo mutation. *PLoS Genet*, 8(10):e1003011, 2012. doi: 10.1371/journal.pgen.1003011.

Petrovski, S. and Goldstein, D.B. Unequal representation of genetic variation across ancestry groups creates healthcare inequality in the application of precision medicine. *Genome Biol*, 17(1):157, 07 2016. doi: 10.1186/s13059-016-1016-y.

Phan, J. and Reue, K. Lipin, a lipodystrophy and obesity gene. *Cell Metab*, 1(1):73–83, Jan 2005. doi: 10.1016/j.cmet.2004.12.002.

Pickrell, J.K. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am J Hum Genet*, 94(4):559–73, Apr 2014. doi: 10.1016/j.ajhg.2014.03.004.

Pickrell, J.K., Coop, G., Novembre, J., Kudaravalli, S., Li, J.Z., Absher, D., Srinivasan, B.S., Barsh, G.S., Myers, R.M., Feldman, M.W. et al. Signals of recent positive selection in a worldwide sample of human populations. *Genome research*, 19(5):826–837, 2009.

Pickrell, J.K., Berisa, T., Liu, J.Z., Ségurel, L., Tung, J.Y. and Hinds, D.A. Detection and interpretation of shared genetic influences on 42 human traits. *Nat Genet*, 48(7):709–17, 07 2016. doi: 10.1038/ng.3570.

Pierron, D., Razafindrazaka, H., Pagani, L., Ricaut, F.X., Antao, T., Capredon, M., Sambo, C., Radimilahy, C., Rakotoarisoa, J.A., Blench, R.M. et al. Genome-wide evidence of austronesian-bantu admixture and cultural reversion in a hunter-gatherer group of madagascar. *Proc Natl Acad Sci U S A*, 111(3):936–41, Jan 2014. doi: 10.1073/pnas.1321860111.

Pierron, D., Heiske, M., Razafindrazaka, H., Pereda-Loth, V., Sanchez, J., Alva, O., Arachiche, A., Boland, A., Olaso, R., Deleuze, J.F. et al. Strong selection during the last millennium for african ancestry in the admixed population of madagascar. *Nat Commun*, 9(1):932, 03 2018. doi: 10.1038/s41467-018-03342-5.

Polanowski, A.M., Robinson-Laverick, S.M., Paton, D. and Jarman, S.N. Variation in the tyrosinase gene associated with a white humpback whale (megaptera novaeangliae). *J Hered*, 103(1):130–3, 2012. doi: 10.1093/jhered/esr108.

Polimanti, R., Yang, B.Z., Zhao, H. and Gelernter, J. Evidence of polygenic adaptation in the systems genetics of anthropometric traits. *PLoS One*, 11(8):e0160654, 2016. doi: 10.1371/journal.pone.0160654.

Popejoy, A.B. and Fullerton, S.M. Genomics is failing on diversity. *Nature*, 538(7624):161–164, 10 2016. doi: 10.1038/538161a.

Porter, S., Clark, I.M., Kevorkian, L. and Edwards, D.R. The adamts metalloproteinases. *Biochem J*, 386(Pt 1):15–27, Feb 2005. doi: 10.1042/BJ20040424.

Pośpiech, E., Draus-Barini, J., Kupiec, T., Wojas-Pelc, A. and Branicki, W. Gene-gene interactions contribute to eye colour variation in humans. *J Hum Genet*, 56(6):447–55, Jun 2011. doi: 10.1038/jhg.2011.38.

Price, A.L., Weale, M.E., Patterson, N., Myers, S.R., Need, A.C., Shianna, K.V., Ge, D., Rotter, J.I., Torres, E., Taylor, K.D. et al. Long-range ld can confound genome scans in admixed populations. *Am J Hum Genet*, 83(1):132–5; author reply 135–9, Jul 2008. doi: 10.1016/j.ajhg.2008.06.005.

Price, A.L., Tandon, A., Patterson, N., Barnes, K.C., Rafaels, N., Ruczinski, I., Beaty, T.H., Mathias, R., Reich, D. and Myers, S. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet*, 5(6):e1000519, Jun 2009. doi: 10.1371/journal.pgen.1000519.

Pritchard, J.K., Stephens, M. and Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–59, Jun 2000.

Pritchard, J.K., Pickrell, J.K. and Coop, G. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Curr Biol*, 20(4):R208–15, Feb 2010. doi: 10.1016/j.cub.2009.11.055.

Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P.H., de Filippo, C. et al. The complete genome sequence of a neanderthal from the altai mountains. *Nature*, 505(7481):43–9, Jan 2014. doi: 10.1038/nature12886.

Pybus, M., Luisi, P., Dall'Olio, G.M., Uzkudun, M., Laayouni, H., Bertranpetit, J. and Engelken, J. Hierarchical boosting: a machine-learning framework to detect and classify hard selective sweeps in human populations. *Bioinformatics*, 31(24):3946–52, Dec 2015. doi: 10.1093/bioinformatics/btv493.

Quillen, E.E. The evolution of tanning needs its day in the sun. *Human biology*, 87(4): 352–360, 2015.

Quillen, E.E., Bauchet, M., Bigham, A.W., Delgado-Burbano, M.E., Faust, F.X., Klimentidis, Y.C., Mao, X., Stoneking, M. and Shriver, M.D. Oprm1 and egfr contribute to skin pigmentation differences between indigenous americans and europeans. *Hum Genet*, 131(7):1073–80, Jul 2012. doi: 10.1007/s00439-011-1135-1.

Quillen, E.E., Chen, X.D., Almasy, L., Yang, F., He, H., Li, X., Wang, X.Y., Liu, T.Q., Hao, W., Deng, H.W. et al. Aldh2 is associated to alcohol dependence and is the major genetic determinant of "daily maximum drinks" in a gwas study of an isolated rural chinese sample. *Am J Med Genet B Neuropsychiatr Genet*, 165B(2):103–10, Mar 2014. doi: 10.1002/ajmg.b.32213.

Racimo, F., Sankararaman, S., Nielsen, R. and Huerta-Sánchez, E. Evidence for archaic adaptive introgression in humans. *Nat Rev Genet*, 16(6):359–71, Jun 2015. doi: 10.1038/nrg3936.

Racimo, F., Gokhman, D., Fumagalli, M., Ko, A., Hansen, T., Moltke, I., Albrechtsen, A., Carmel, L., Huerta-Sánchez, E. and Nielsen, R. Archaic adaptive introgression in tbx15/wars2. *Mol Biol Evol*, 34(3):509–524, Mar 2017. doi: 10.1093/molbev/msw283.

Racimo, F., Berg, J.J. and Pickrell, J.K. Detecting polygenic adaptation in admixture graphs. *Genetics*, 208(4):1565–1584, 04 2018a. doi: 10.1534/genetics.117.300489.

Racimo, F., Berg, J.J. and Pickrell, J.K. Detecting polygenic adaptation in admixture graphs. *Genetics*, 2018b. ISSN 0016-6731. doi: 10.1534/genetics.117.300489. URL `http://www.genetics.org/content/early/2018/01/18/genetics.117.300489`.

Raghavan, M., Skoglund, P., Graf, K.E., Metspalu, M., Albrechtsen, A., Moltke, I., Rasmussen, S., Stafford, Jr, T.W., Orlando, L., Metspalu, E. et al. Upper palaeolithic siberian genome reveals dual ancestry of native americans. *Nature*, 505(7481):87–91, Jan 2014. doi: 10.1038/nature12736.

Raghavan, M., Steinrücken, M., Harris, K., Schiffels, S., Rasmussen, S., DeGiorgio, M., Albrechtsen, A., Valdiosera, C., Ávila-Arcos, M.C., Malaspinas, A.S. et al. Population genetics. genomic evidence for the pleistocene and recent population history of native americans. *Science*, 349(6250):aab3884, Aug 2015. doi: 10.1126/science.aab3884.

Rahman, M.M., Ng, J.C. and Naidu, R. Chronic exposure of arsenic via drinking water and its adverse health impacts on humans. *Environmental geochemistry and health*, 31 (1):189–200, 2009.

Rajsbaum, R., Stoye, J.P. and O'Garra, A. Type i interferon-dependent and -independent expression of tripartite motif proteins in immune cells. *Eur J Immunol*, 38(3):619–30, Mar 2008. doi: 10.1002/eji.200737916.

Rana, B.K., Hewett-Emmett, D., Jin, L., Chang, B.H., Sambuughin, N., Lin, M., Watkins, S., Bamshad, M., Jorde, L.B., Ramsay, M. et al. High polymorphism at the human melanocortin 1 receptor locus. *Genetics*, 151(4):1547–57, Apr 1999.

Ransohoff, K.J., Wu, W., Cho, H.G., Chahal, H.C., Lin, Y., Dai, H.J., Amos, C.I., Lee, J.E., Tang, J.Y., Hinds, D.A. et al. Two-stage genome-wide association study identifies a novel susceptibility locus associated with melanoma. *Oncotarget*, 8(11):17586–17592, Mar 2017. doi: 10.18632/oncotarget.15230.

Raqib, R., Ahmed, S., Sultana, R., Wagatsuma, Y., Mondal, D., Hoque, A.M.W., Nermell, B., Yunus, M., Roy, S., Persson, L.A. et al. Effects of in utero arsenic exposure on child immunity and morbidity in rural bangladesh. *Toxicol Lett*, 185(3):197–202, Mar 2009. doi: 10.1016/j.toxlet.2009.01.001.

Rasmussen, M., Li, Y., Lindgreen, S., Pedersen, J.S., Albrechtsen, A., Moltke, I., Metspalu, M., Metspalu, E., Kivisild, T., Gupta, R. et al. Ancient human genome

sequence of an extinct palaeo-eskimo. *Nature*, 463(7282):757–62, Feb 2010. doi: 10.1038/nature08835.

Rasmussen, M., Anzick, S.L., Waters, M.R., Skoglund, P., DeGiorgio, M., Stafford, Jr, T.W., Rasmussen, S., Moltke, I., Albrechtsen, A., Doyle, S.M. et al. The genome of a late pleistocene human from a clovis burial site in western montana. *Nature*, 506(7487): 225–9, Feb 2014. doi: 10.1038/nature13025.

Rasmussen, M., Sikora, M., Albrechtsen, A., Korneliussen, T.S., Moreno-Mayar, J.V., Poznik, G.D., Zollikofer, C.P.E., de León, M.P., Allentoft, M.E., Moltke, I. et al. The ancestry and affiliations of kennewick man. *Nature*, 523(7561):455–458, Jul 2015. doi: 10.1038/nature14625.

Rawofi, L., Edwards, M., Krithika, S., Le, P., Cha, D., Yang, Z., Ma, Y., Wang, J., Su, B., Jin, L. et al. Genome-wide association study of pigmentary traits (skin and iris color) in individuals of east asian ancestry. *PeerJ*, 5:e3951, 2017. doi: 10.7717/peerj.3951.

Reed, T.E. Caucasian genes in american negroes. *Science*, 165(3895):762–768, 1969.

Rees, J.L. The melanocortin 1 receptor (mc1r): more than just red hair. *Pigment Cell Res*, 13(3):135–40, Jun 2000.

Rees, J.L. Genetics of hair and skin color. *Annual review of genetics*, 37(1):67–90, 2003.

Rees, J.L. The genetics of sun sensitivity in humans. *Am J Hum Genet*, 75(5):739–51, Nov 2004. doi: 10.1086/425285.

Rees, J.L. and Harding, R.M. Understanding the evolution of human pigmentation: recent contributions from population genetics. *J Invest Dermatol*, 132(3 Pt 2):846–53, Mar 2012. doi: 10.1038/jid.2011.358.

Reggiani, C., Coppens, S., Sekhara, T., Dimov, I., Pichon, B., Lufin, N., Addor, M.C., Belligni, E.F., Digilio, M.C., Faletra, F. et al. Novel promoters and coding first exons in dlg2 linked to developmental disorders and intellectual disability. *Genome Med*, 9(1): 67, 07 2017. doi: 10.1186/s13073-017-0452-y.

Reich, D., Green, R.E., Kircher, M., Krause, J., Patterson, N., Durand, E.Y., Viola, B., Briggs, A.W., Stenzel, U., Johnson, P.L.F. et al. Genetic history of an archaic hominin group from denisova cave in siberia. *Nature*, 468(7327):1053–60, Dec 2010. doi: 10.1038/nature09710.

Reich, D., Patterson, N., Kircher, M., Delfin, F., Nandineni, M.R., Pugach, I., Ko, A.M.S., Ko, Y.C., Jinam, T.A., Phipps, M.E. et al. Denisova admixture and the first modern human dispersals into southeast asia and oceania. *Am J Hum Genet*, 89(4):516–28, Oct 2011. doi: 10.1016/j.ajhg.2011.09.005.

Reich, D., Patterson, N., Campbell, D., Tandon, A., Mazieres, S., Ray, N., Parra, M.V., Rojas, W., Duque, C., Mesa, N. et al. Reconstructing native american population history. *Nature*, 488(7411):370–4, Aug 2012. doi: 10.1038/nature11258.

Reinhardt, J.A., Kolaczkowski, B., Jones, C.D., Begun, D.J. and Kern, A.D. Parallel geographic variation in drosophila melanogaster. *Genetics*, 197(1):361–373, 2014.

Relethford, J.H. Apportionment of global human genetic diversity based on craniometrics and skin color. *Am J Phys Anthropol*, 118(4):393–8, Aug 2002. doi: 10.1002/ajpa.10079.

Relethford, J.H. *Human population genetics*, volume 7. John Wiley & Sons, 2012.

Reue, K., Xu, P., Wang, X.P. and Slavin, B.G. Adipose tissue deficiency, glucose intolerance, and increased atherosclerosis result from mutation in the mouse fatty liver dystrophy (fld) gene. *J Lipid Res*, 41(7):1067–76, Jul 2000.

Reue, K. The lipin family: mutations and metabolism. *Curr Opin Lipidol*, 20(3):165–70, Jun 2009. doi: 10.1097/MOL.0b013e32832adee5.

Reue, K. and Zhang, P. The lipin protein family: dual roles in lipid biosynthesis and gene expression. *FEBS Lett*, 582(1):90–6, Jan 2008. doi: 10.1016/j.febslet.2007.11.014.

Reymond, A., Meroni, G., Fantozzi, A., Merla, G., Cairo, S., Luzi, L., Riganelli, D., Zanaria, E., Messali, S., Cainarca, S. et al. The tripartite motif family identifies cell compartments. *EMBO J*, 20(9):2140–51, May 2001. doi: 10.1093/emboj/20.9.2140.

Reynolds, J., Weir, B.S. and Cockerham, C.C. Estimation of the coancestry coefficient: basis for a short-term genetic distance. *Genetics*, 105(3):767–79, Nov 1983.

Richards, M.P., Jacobi, R., Cook, J., Pettitt, P.B. and Stringer, C.B. Isotope evidence for the intensive use of marine foods by late upper palaeolithic humans. *J Hum Evol*, 49 (3):390–4, Sep 2005. doi: 10.1016/j.jhevol.2005.05.002.

Richman, A. Evolution of balanced genetic polymorphism. *Mol Ecol*, 9(12):1953–63, Dec 2000.

Richter, D., Grün, R., Joannes-Boyau, R., Steele, T.E., Amani, F., Rué, M., Fernandes, P., Raynal, J.P., Geraads, D., Ben-Ncer, A. et al. The age of the hominin fossils from jebel irhoud, morocco, and the origins of the middle stone age. *Nature*, 546(7657):293–296, 06 2017. doi: 10.1038/nature22335.

Rinchik, E.M., Bultman, S.J., Horsthemke, B., Lee, S.T., Strunk, K.M., Spritz, R.A., Avidano, K.M., Jong, M.T. and Nicholls, R.D. A gene for the mouse pink-eyed dilution locus and for human type ii oculocutaneous albinism. *Nature*, 361(6407):72–6, Jan 1993. doi: 10.1038/361072a0.

Ringholm, A., Klovins, J., Rudzish, R., Phillips, S., Rees, J.L. and Schiöth, H.B. Pharmacological characterization of loss of function mutations of the human melanocortin 1 receptor that are associated with red hair. *J Invest Dermatol*, 123(5):917–23, Nov 2004. doi: 10.1111/j.0022-202X.2004.23444.x.

Rishishwar, L., Conley, A.B., Wigington, C.H., Wang, L., Valderrama-Aguirre, A. and Jordan, I.K. Ancestry, admixture and fitness in colombian genomes. *Sci Rep*, 5:12376, Jul 2015. doi: 10.1038/srep12376.

Rivas, M.A., Avila, B.E., Koskela, J., Huang, H., Stevens, C., Pirinen, M., Haritunians, T., Neale, B.M., Kurki, M., Ganna, A. et al. Insights into the genetic epidemiology of crohn's and rare diseases in the ashkenazi jewish population. *PLoS Genet*, 14(5): e1007329, 05 2018. doi: 10.1371/journal.pgen.1007329.

Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J. et al. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–30, Feb 2015. doi: 10.1038/nature14248.

Robins, A.H. The evolution of light skin color: role of vitamin d disputed. *American journal of physical anthropology*, 139(4):447–450, 2009.

Robinson, M.R., Hemani, G., Medina-Gomez, C., Mezzavilla, M., Esko, T., Shakhbazov, K., Powell, J.E., Vinkhuyzen, A., Berndt, S.I., Gustafsson, S. et al. Population genetic differentiation of height and body mass index across europe. *Nat Genet*, 47(11):1357–62, Nov 2015. doi: 10.1038/ng.3401.

Rodríguez, J.A., Marigorta, U.M. and Navarro, A. Integrating genomics into evolutionary medicine. *Curr Opin Genet Dev*, 29:97–102, Dec 2014. doi: 10.1016/j.gde.2014.08.009.

Rogers, A., Iltis, D. and Wooding, S. Genetic variation at the mc1r locus and the time since loss of human body hair. *Current Anthropology*, 45(1):105–108, 2004.

Romano, R.A., Li, H., Tummala, R., Maul, R. and Sinha, S. Identification of basonuclin2, a dna-binding zinc-finger protein expressed in germ tissues and skin keratinocytes. *Genomics*, 83(5):821–33, May 2004. doi: 10.1016/j.ygeno.2003.11.009.

Ronen, R., Udpa, N., Halperin, E. and Bafna, V. Learning natural selection from the site frequency spectrum. *Genetics*, 195(1):181–93, Sep 2013. doi: 10.1534/genetics.113.152587.

Rooryck, C., Roudaut, C., Robine, E., Müsebeck, J. and Arveiler, B. Oculocutaneous albinism with tyrp1 gene mutations in a caucasian patient. *Pigment Cell & Melanoma Research*, 19(3):239–242, 2006.

Rooryck, C., Morice-Picard, F., Elçioglu, N.H., Lacombe, D., Taieb, A. and Arveiler, B. Molecular diagnosis of oculocutaneous albinism: new mutations in the oca1-4 genes and practical aspects. *Pigment Cell Melanoma Res*, 21(5):583–7, Oct 2008. doi: 10.1111/j.1755-148X.2008.00496.x.

Rosenberger, A., Friedrichs, S., Amos, C.I., Brennan, P., Fehringer, G., Heinrich, J., Hung, R.J., Muley, T., Müller-Nurasyid, M., Risch, A. et al. Meta-gsa: Combining findings from gene-set analyses across several genome-wide association studies. *PLoS One*, 10 (10):e0140179, 2015. doi: 10.1371/journal.pone.0140179.

Rudic, R.D., McNamara, P., Curtis, A.M., Boston, R.C., Panda, S., Hogenesch, J.B. and Fitzgerald, G.A. Bmal1 and clock, two essential components of the circadian clock, are

involved in glucose homeostasis. *PLoS Biol*, 2(11):e377, Nov 2004. doi: 10.1371/journal.pbio.0020377.

Ruiz-Linares, A., Ortíz-Barrientos, D., Figueroa, M., Mesa, N., Múnera, J.G., Bedoya, G., Vélez, I.D., García, L.F., Pérez-Lezaun, A., Bertranpetit, J. et al. Microsatellites provide evidence for y chromosome diversity among the founders of the new world. *Proc Natl Acad Sci U S A*, 96(11):6312–7, May 1999.

Ruiz-Linares, A. How genes have illuminated the history of early americans and latino americans. *Cold Spring Harb Perspect Biol*, 7(6), Sep 2014. doi: 10.1101/cshperspect.a008557.

Ruiz-Linares, A., Adhikari, K., Acuña-Alonzo, V., Quinto-Sanchez, M., Jaramillo, C., Arias, W., Fuentes, M., Pizarro, M., Everardo, P., de Avila, F. et al. Admixture in latin america: geographic structure, phenotypic diversity and self-perception of ancestry based on 7,342 individuals. *PLoS Genet*, 10(9):e1004572, Sep 2014. doi: 10.1371/journal.pgen.1004572.

Rundshagen, U., Zühlke, C., Opitz, S., Schwinger, E. and Käsmann-Kellner, B. Mutations in the matp gene in five german patients affected by oculocutaneous albinism type 4. *Hum Mutat*, 23(2):106–10, Feb 2004. doi: 10.1002/humu.10311.

Sabeti, P.C., Reich, D.E., Higgins, J.M., Levine, H.Z.P., Richter, D.J., Schaffner, S.F., Gabriel, S.B., Platko, J.V., Patterson, N.J., McDonald, G.J. et al. Detecting recent positive selection in the human genome from haplotype structure. *Nature*, 419(6909):832–7, Oct 2002a. doi: 10.1038/nature01140.

Sabeti, P.C., Reich, D.E., Higgins, J.M., Levine, H.Z., Richter, D.J., Schaffner, S.F., Gabriel, S.B., Platko, J.V., Patterson, N.J., McDonald, G.J. et al. Detecting recent positive selection in the human genome from haplotype structure. *Nature*, 419(6909):832, 2002b.

Sabeti, P.C., Schaffner, S.F., Fry, B., Lohmueller, J., Varilly, P., Shamovsky, O., Palma, A., Mikkelsen, T., Altshuler, D. and Lander, E. Positive natural selection in the human lineage. *science*, 312(5780):1614–1620, 2006.

Sabeti, P.C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E.H., McCarroll, S.A., Gaudet, R. et al. Genome-wide detection and characterization of positive selection in human populations. *Nature*, 449(7164):913–8, Oct 2007. doi: 10.1038/nature06250.

Sanchez-Albornoz, N. *The population of Latin America. A history.* Berkeley, CA (USA) Univ. of California Press, 1974.

Sánchez-Más, J., Sánchez-Laorden, B.L., Guillo, L.A., Jiménez-Cervantes, C. and García-Borrón, J.C. The melanocortin-1 receptor carboxyl terminal pentapeptide is essential for mc1r function and expression on the cell surface. *Peptides*, 26(10):1848–57, Oct 2005. doi: 10.1016/j.peptides.2004.11.030.

Sanjak, J.S., Sidorenko, J., Robinson, M.R., Thornton, K.R. and Visscher, P.M. Evidence of directional and stabilizing selection in contemporary humans. *Proc Natl Acad Sci U S A*, 115(1):151–156, Jan 2018. doi: 10.1073/pnas.1707227114.

Sankararaman, S., Sridhar, S., Kimmel, G. and Halperin, E. Estimating local ancestry in admixed populations. *Am J Hum Genet*, 82(2):290–303, Feb 2008. doi: 10.1016/j.ajhg.2007.09.022.

Sankararaman, S., Mallick, S., Dannemann, M., Prüfer, K., Kelso, J., Pääbo, S., Patterson, N. and Reich, D. The genomic landscape of neanderthal ancestry in present-day humans. *Nature*, 507(7492):354–7, Mar 2014. doi: 10.1038/nature12961.

Savage, S.A., Gerstenblith, M.R., Goldstein, A.M., Mirabello, L., Fargnoli, M.C., Peris, K. and Landi, M.T. Nucleotide diversity and population differentiation of the melanocortin 1 receptor gene, mc1r. *BMC Genet*, 9:31, Apr 2008. doi: 10.1186/1471-2156-9-31.

Scheib, C.L., Li, H., Desai, T., Link, V., Kendall, C., Dewar, G., Griffith, P.W., Mörseburg, A., Johnson, J.R., Potter, A. et al. Ancient human parallel lineages within north america contributed to a coastal expansion. *Science*, 360(6392):1024–1027, Jun 2018. doi: 10.1126/science.aar6851.

Schiöth, H.B., Kuusinen, A., Muceniece, R., Szardenings, M., Keinänen, K. and Wikberg, J.E. Expression of functional melanocortin 1 receptors in insect cells. *Biochem Biophys Res Commun*, 221(3):807–14, Apr 1996. doi: 10.1006/bbrc.1996.0678.

Schlebusch, C.M., Skoglund, P., Sjödin, P., Gattepaille, L.M., Hernandez, D., Jay, F., Li, S., De Jongh, M., Singleton, A., Blum, M.G.B. et al. Genomic variation in seven khoe-san groups reveals adaptation and complex african history. *Science*, 338(6105):374–9, Oct 2012. doi: 10.1126/science.1227721.

Schlebusch, C.M., Sjödin, P., Skoglund, P. and Jakobsson, M. Stronger signal of recent selection for lactase persistence in maasai than in europeans. *Eur J Hum Genet*, 21(5):550–3, May 2013. doi: 10.1038/ejhg.2012.199.

Schlebusch, C.M., Gattepaille, L.M., Engström, K., Vahter, M., Jakobsson, M. and Broberg, K. Human adaptation to arsenic-rich environments. *Molecular biology and evolution*, 32(6):1544–1555, 2015.

Schlenke, T.A. and Begun, D.J. Strong selective sweep associated with a transposon insertion in drosophila simulans. *Proceedings of the National Academy of Sciences of the United States of America*, 101(6):1626–1631, 2004.

Schmid, F., Schmid, M., Müssel, C., Sträng, J.E., Buske, C., Bullinger, L., Kraus, J.M. and Kestler, H.A. Giant: gene set uncertainty in enrichment analysis. *Bioinformatics*, 32(12):1891–4, 06 2016. doi: 10.1093/bioinformatics/btw030.

Schmidt-Küntzel, A., Eizirik, E., O'Brien, S.J. and Menotti-Raymond, M. Tyrosinase and tyrosinase related protein 1 alleles specify domestic cat coat color phenotypes of the albino and brown loci. *J Hered*, 96(4):289–301, 2005. doi: 10.1093/jhered/esi066.

Schollen, E., Pardon, E., Heykants, L., Renard, J., Doggett, N.A., Callen, D.F., Cassiman, J.J. and Matthijs, G. Comparative analysis of the phosphomannomutase genes pmm1, pmm2 and pmm2psi: the sequence variation in the processed pseudogene is a reflection of the mutations found in the functional gene. *Hum Mol Genet*, 7(2):157–64, Feb 1998.

Schrider, D.R. and Kern, A.D. S/hic: Robust identification of soft and hard sweeps using machine learning. *PLoS Genet*, 12(3):e1005928, Mar 2016. doi: 10.1371/journal.pgen.1005928.

Schrider, D.R. and Kern, A.D. Soft sweeps are the dominant mode of adaptation in the human genome. *Mol Biol Evol*, 34(8):1863–1877, Aug 2017a. doi: 10.1093/molbev/msx154.

Schrider, D.R. and Kern, A.D. Soft sweeps are the dominant mode of adaptation in the human genome. *Molecular biology and evolution*, 34(8):1863–1877, 2017b.

Schrider, D.R. and Kern, A.D. Supervised machine learning for population genetics: A new paradigm. *Trends Genet*, 34(4):301–312, Apr 2018. doi: 10.1016/j.tig.2017.12.005.

Schrider, D.R., Navarro, F.C., Galante, P.A., Parmigiani, R.B., Camargo, A.A., Hahn, M.W. and de Souza, S.J. Gene copy-number polymorphism caused by retrotransposition in humans. *PLoS genetics*, 9(1):e1003242, 2013.

Schroeder, H., Sikora, M., Gopalakrishnan, S., Cassidy, L.M., Maisano Delser, P., Sandoval Velasco, M., Schraiber, J.G., Rasmussen, S., Homburger, J.R., Ávila-Arcos, M.C. et al. Origins and genetic legacies of the caribbean taino. *Proc Natl Acad Sci U S A*, 115(10):2341–2346, Mar 2018. doi: 10.1073/pnas.1716839115.

Seaman, M.N., Marcusson, E.G., Cereghino, J.L. and Emr, S.D. Endosome to golgi retrieval of the vacuolar protein sorting receptor, vps10p, requires the function of the vps29, vps30, and vps35 gene products. *J Cell Biol*, 137(1):79–92, Apr 1997.

Seaman, M.N.J., Harbour, M.E., Tattersall, D., Read, E. and Bright, N. Membrane recruitment of the cargo-selective retromer subcomplex is catalysed by the small gtpase rab7 and inhibited by the rab-gap tbc1d5. *J Cell Sci*, 122(Pt 14):2371–82, Jul 2009. doi: 10.1242/jcs.048686.

Sheehan, S. and Song, Y.S. Deep learning for population genetic inference. *PLoS Comput Biol*, 12(3):e1004845, Mar 2016a. doi: 10.1371/journal.pcbi.1004845.

Sheehan, S. and Song, Y.S. Deep learning for population genetic inference. *PLoS computational biology*, 12(3):e1004845, 2016b.

Shi, H., Kichaev, G. and Pasaniuc, B. Contrasting the genetic architecture of 30 complex traits from summary association data. *Am J Hum Genet*, 99(1):139–53, Jul 2016. doi: 10.1016/j.ajhg.2016.05.013.

Shiina, T., Hosomichi, K., Inoko, H. and Kulski, J.K. The hla genomic loci map: expression, interaction, diversity and disease. *J Hum Genet*, 54(1):15–39, Jan 2009. doi: 10.1038/jhg.2008.5.

Shriver, M.D., Parra, E.J., Dios, S., Bonilla, C., Norton, H., Jovel, C., Pfaff, C., Jones, C., Massac, A., Cameron, N. et al. Skin pigmentation, biogeographical ancestry and admixture mapping. *Hum Genet*, 112(4):387–99, Apr 2003. doi: 10.1007/s00439-002-0896-y.

Shriver, M.D., Kennedy, G.C., Parra, E.J., Lawson, H.A., Sonpar, V., Huang, J., Akey, J.M. and Jones, K.W. The genomic distribution of population substructure in four populations using 8,525 autosomal snps. *Hum Genomics*, 1(4):274–86, May 2004.

Shungin, D., Winkler, T.W., Croteau-Chonka, D.C., Ferreira, T., Locke, A.E., Mägi, R., Strawbridge, R.J., Pers, T.H., Fischer, K., Justice, A.E. et al. New genetic loci link adipose and insulin biology to body fat distribution. *Nature*, 518(7538):187–196, Feb 2015. doi: 10.1038/nature14132.

Sibai, B.M. Thrombophilia and severe preeclampsia: time to screen and treat in future pregnancies? *Hypertension*, 46(6):1252–3, Dec 2005. doi: 10.1161/01.HYP.0000188904.47575.7e.

Simons, Y.B., Turchin, M.C., Pritchard, J.K. and Sella, G. The deleterious mutation load is insensitive to recent population history. *Nat Genet*, 46(3):220–4, Mar 2014. doi: 10.1038/ng.2896.

Skoglund, P. and Reich, D. A genomic view of the peopling of the americas. *Curr Opin Genet Dev*, 41:27–35, Dec 2016. doi: 10.1016/j.gde.2016.06.016.

Skoglund, P., Mallick, S., Bortolini, M.C., Chennagiri, N., Hünemeier, T., Petzl-Erler, M.L., Salzano, F.M., Patterson, N. and Reich, D. Genetic evidence for two founding populations of the americas. *Nature*, 525(7567):104–8, Sep 2015. doi: 10.1038/nature14895.

Skotte, L., Korneliussen, T.S. and Albrechtsen, A. Estimating individual admixture proportions from next generation sequencing data. *Genetics*, 195(3):693–702, Nov 2013. doi: 10.1534/genetics.113.154138.

Slatkin, M. Gene flow and selection in a cline. *Genetics*, 75(4):733–756, 1973.

Slominski, A., Wortsman, J., Plonka, P.M., Schallreuter, K.U., Paus, R. and Tobin, D.J. Hair follicle pigmentation. *Journal of Investigative Dermatology*, 124(1):13–21, 2005.

Smith, J.M. and Haigh, J. The hitch-hiking effect of a favourable gene. *Genet Res*, 23(1):23–35, Feb 1974.

Smith, J.M. "haldane's dilemma" and the rate of evolution. *Nature*, 219(5159):1114, 1968.

Smith, M.W. and O'Brien, S.J. Mapping by admixture linkage disequilibrium: advances, limitations and guidelines. *Nat Rev Genet*, 6(8):623–32, Aug 2005. doi: 10.1038/nrg1657.

Soejima, M. and Koda, Y. Population differences of two coding snps in pigmentation-related genes slc24a5 and slc45a2. *International journal of legal medicine*, 121(1):36–39, 2007.

Sohail, M., Maier, R.M., Ganna, A., Bloemendal, A., Martin, A.R., Turchin, M.C., Chiang, C.W.K., Hirschhorn, J.N., Daly, M.J., Patterson, N. et al. Signals of polygenic adaptation on height have been overestimated due to uncorrected population structure in genome-wide association studies. *bioRxiv*, 2018. doi: 10.1101/355057. URL `https://www.biorxiv.org/content/early/2018/07/09/355057`.

Solovieff, N., Cotsapas, C., Lee, P.H., Purcell, S.M. and Smoller, J.W. Pleiotropy in complex traits: challenges and strategies. *Nat Rev Genet*, 14(7):483–95, Jul 2013. doi: 10.1038/nrg3461.

Song, Q., Li, C., Feng, X., Yu, A., Tang, H., Peng, Z. and Wang, X. Decreased expression of scube2 is associated with progression and prognosis in colorectal cancer. *Oncol Rep*, 33(4):1956–64, Apr 2015. doi: 10.3892/or.2015.3790.

Speed, D., Hemani, G., Johnson, M.R. and Balding, D.J. Improved heritability estimation from genome-wide snps. *Am J Hum Genet*, 91(6):1011–21, Dec 2012. doi: 10.1016/j.ajhg.2012.10.010.

Spencer, C.C.A., Su, Z., Donnelly, P. and Marchini, J. Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS Genet*, 5(5):e1000477, May 2009. doi: 10.1371/journal.pgen.1000477.

Staels, B. When the clock stops ticking, metabolic syndrome explodes. *Nat Med*, 12(1):54–5; discussion 55, Jan 2006. doi: 10.1038/nm0106-54.

Stearns, S.C., Byars, S.G., Govindaraju, D.R. and Ewbank, D. Measuring selection in contemporary human populations. *Nat Rev Genet*, 11(9):611–22, Sep 2010. doi: 10.1038/nrg2831.

Stephan, W. Signatures of positive selection: from selective sweeps at individual loci to subtle allele frequency changes in polygenic adaptation. *Molecular ecology*, 25(1):79–88, 2016.

Stokowski, R.P., Pant, P.V.K., Dadd, T., Fereday, A., Hinds, D.A., Jarman, C., Filsell, W., Ginger, R.S., Green, M.R., van der Ouderaa, F.J. et al. A genomewide association study of skin pigmentation in a south asian population. *Am J Hum Genet*, 81(6):1119–32, Dec 2007. doi: 10.1086/522235.

Stoneking, M. *An introduction to molecular anthropology.* John Wiley & Sons, 2016.

Stram, D.O. *Design, analysis, and interpretation of genome-wide association scans.* Springer, 2016.

Stranger, B.E., Stahl, E.A. and Raj, T. Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics*, 187(2):367–83, Feb 2011. doi: 10.1534/genetics.110.120907.

Sturm, R.A. Molecular genetics of human pigmentation diversity. *Hum Mol Genet*, 18(R1):R9–17, Apr 2009. doi: 10.1093/hmg/ddp003.

Sturm, R.A. and Frudakis, T.N. Eye colour: portals into pigmentation genes and ancestry. *Trends Genet*, 20(8):327–32, Aug 2004. doi: 10.1016/j.tig.2004.06.010.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, 102(43):15545–50, Oct 2005. doi: 10.1073/pnas.0506580102.

Subramanian, A., Kuehn, H., Gould, J., Tamayo, P. and Mesirov, J.P. Gsea-p: a desktop application for gene set enrichment analysis. *Bioinformatics*, 23(23):3251–3, Dec 2007. doi: 10.1093/bioinformatics/btm369.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M. et al. Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med*, 12(3): e1001779, Mar 2015. doi: 10.1371/journal.pmed.1001779.

Sugden, L.A., Atkinson, E.G., Fischer, A.P., Rong, S., Henn, B.M. and Ramachandran, S. Localization of adaptive variants in human genomes using averaged one-dependence estimation. *Nat Commun*, 9(1):703, Feb 2018. doi: 10.1038/s41467-018-03100-7.

Sulem, P., Gudbjartsson, D.F., Stacey, S.N., Helgason, A., Rafnar, T., Magnusson, K.P., Manolescu, A., Karason, A., Palsson, A., Thorleifsson, G. et al. Genetic determinants of hair, eye and skin pigmentation in europeans. *Nat Genet*, 39(12):1443–52, Dec 2007. doi: 10.1038/ng.2007.13.

Sulem, P., Gudbjartsson, D.F., Stacey, S.N., Helgason, A., Rafnar, T., Jakobsdottir, M., Steinberg, S., Gudjonsson, S.A., Palsson, A., Thorleifsson, G. et al. Two newly identified genetic determinants of pigmentation in europeans. *Nat Genet*, 40(7):835–7, Jul 2008. doi: 10.1038/ng.160.

Sumi, D. and Himeno, S. Role of arsenic (+ 3 oxidation state) methyltransferase in arsenic metabolism and toxicity. *Biological and Pharmaceutical Bulletin*, 35(11):1870–1875, 2012.

Sun, H.J., Rathinasabapathi, B., Wu, B., Luo, J., Pu, L.P. and Ma, L.Q. Arsenic and selenium toxicity and their interactive effects in humans. *Environment International*, 69:148–158, 2014.

Suviolahti, E., Reue, K., Cantor, R.M., Phan, J., Gentile, M., Naukkarinen, J., Soro-Paavonen, A., Oksanen, L., Kaprio, J., Rissanen, A. et al. Cross-species analyses implicate lipin 1 involvement in human glucose metabolism. *Hum Mol Genet*, 15(3):377–86, Feb 2006. doi: 10.1093/hmg/ddi448.

Sverrisdóttir, O.Ó., Timpson, A., Toombs, J., Lecoeur, C., Froguel, P., Carretero, J.M., Arsuaga Ferreras, J.L., Götherström, A. and Thomas, M.G. Direct estimates of natural selection in iberia indicate calcium absorption was not the only driver of lactase persistence in europe. *Mol Biol Evol*, 31(4):975–83, Apr 2014. doi: 10.1093/molbev/msu049.

Szabó, G., Gerald, A.B., Pathak, M.A. and Fitzpatrick, T.B. Racial differences in the fate of melanosomes in human epidermis. *Nature*, 222(5198):1081, 1969.

Szolek, A., Schubert, B., Mohr, C., Sturm, M., Feldhahn, M. and Kohlbacher, O. Optitype: precision hla typing from next-generation sequencing data. *Bioinformatics*, 30(23):3310–6, Dec 2014. doi: 10.1093/bioinformatics/btu548.

Szpak, M., Mezzavilla, M., Ayub, Q., Chen, Y., Xue, Y. and Tyler-Smith, C. Finemav: prioritizing candidate genetic variants driving local adaptations in human populations. *Genome Biol*, 19(1):5, 01 2018. doi: 10.1186/s13059-017-1380-2.

Szpiech, Z.A. and Hernandez, R.D. selscan: an efficient multithreaded program to perform ehh-based scans for positive selection. *Mol Biol Evol*, 31(10):2824–7, Oct 2014. doi: 10.1093/molbev/msu211.

Tackney, J.C., Potter, B.A., Raff, J., Powers, M., Watkins, W.S., Warner, D., Reuther, J.D., Irish, J.D. and O'Rourke, D.H. Two contemporaneous mitogenomes from terminal pleistocene burials in eastern beringia. *Proc Natl Acad Sci U S A*, 112(45):13833–8, Nov 2015. doi: 10.1073/pnas.1511903112.

Tajima, F. Statistical method for testing the neutral mutation hypothesis by dna polymorphism. *Genetics*, 123(3):585–95, Nov 1989.

Takeuchi, F., Isono, M., Nabika, T., Katsuya, T., Sugiyama, T., Yamaguchi, S., Kobayashi, S., Ogihara, T., Yamori, Y., Fujioka, A. et al. Confirmation of aldh2 as a major locus of drinking behavior and of its variants regulating multiple metabolic phenotypes in a japanese population. *Circ J*, 75(4):911–8, 2011.

Takiwaki, H., Shirai, S., Kanno, Y., Watanabe, Y. and Arase, S. Quantification of erythema and pigmentation using a videomicroscope and a computer. *British Journal of Dermatology*, 131(1):85–92, 1994.

Tamm, E., Kivisild, T., Reidla, M., Metspalu, M., Smith, D.G., Mulligan, C.J., Bravi, C.M., Rickards, O., Martinez-Labarga, C., Khusnutdinova, E.K. et al. Beringian standstill and spread of native american founders. *PLoS One*, 2(9):e829, Sep 2007. doi: 10.1371/journal.pone.0000829.

Tamura, T. and Picciano, M.F. Folate and human reproduction–. *The American journal of clinical nutrition*, 83(5):993–1016, 2006.

Tan, J., Yang, Y., Tang, K., Sabeti, P.C., Jin, L. and Wang, S. The adaptive variant edarv370a is associated with straight hair in east asians. *Hum Genet*, 132(10):1187–91, Oct 2013. doi: 10.1007/s00439-013-1324-1.

Tan, J., Peng, Q., Li, J., Guan, Y., Zhang, L., Jiao, Y., Yang, Y., Wang, S. and Jin, L. Characteristics of dental morphology in the xinjiang uyghurs and correlation with the edarv370a variant. *Sci China Life Sci*, 57(5):510–8, May 2014. doi: 10.1007/s11427-014-4654-x.

Tang, H., Peng, J., Wang, P. and Risch, N.J. Estimation of individual admixture: analytical and study design considerations. *Genet Epidemiol*, 28(4):289–301, May 2005. doi: 10.1002/gepi.20064.

Tang, H., Coram, M., Wang, P., Zhu, X. and Risch, N. Reconstructing genetic ancestry blocks in admixed individuals. *Am J Hum Genet*, 79(1):1–12, Jul 2006. doi: 10.1086/504302.

Tang, H., Choudhry, S., Mei, R., Morgan, M., Rodriguez-Cintron, W., Burchard, E.G. and Risch, N.J. Recent genetic selection in the ancestral admixture of puerto ricans. *Am J Hum Genet*, 81(3):626–33, Sep 2007. doi: 10.1086/520769.

Taylor, J. and Tibshirani, R. A tail strength measure for assessing the overall univariate significance in a dataset. *Biostatistics*, 7(2):167–81, Apr 2006. doi: 10.1093/biostatistics/kxj009.

Thomas, H. *The slave trade: The story of the Atlantic slave trade: 1440-1870*. Simon and Schuster, 1997.

Thornton, R. *American Indian holocaust and survival: A population history since 1492*, volume 186. University of Oklahoma Press, 1987.

Tintle, N.L., Borchers, B., Brown, M. and Bekmetjev, A. Comparing gene set analysis methods on single-nucleotide polymorphism data from genetic analysis workshop 16. *BMC Proc*, 3 Suppl 7:S96, Dec 2009.

Tishkoff, S.A. and Kidd, K.K. Implications of biogeography of human populations for'race'and medicine. *Nature genetics*, 36(11s):S21, 2004.

Tishkoff, S.A., Reed, F.A., Ranciaro, A., Voight, B.F., Babbitt, C.C., Silverman, J.S., Powell, K., Mortensen, H.M., Hirbo, J.B., Osman, M. et al. Convergent adaptation of human lactase persistence in africa and europe. *Nat Genet*, 39(1):31–40, Jan 2007. doi: 10.1038/ng1946.

Torgerson, D.G., Ampleford, E.J., Chiu, G.Y., Gauderman, W.J., Gignoux, C.R., Graves, P.E., Himes, B.E., Levin, A.M., Mathias, R.A., Hancock, D.B. et al. Meta-analysis of genome-wide association studies of asthma in ethnically diverse north american populations. *Nature genetics*, 43(9):887, 2011.

Torroni, A., Schurr, T.G., Yang, C.C., Szathmary, E.J., Williams, R.C., Schanfield, M.S., Troup, G.A., Knowler, W.C., Lawrence, D.N. and Weiss, K.M. Native american mitochondrial dna analysis indicates that the amerind and the nadene populations were founded by two independent migrations. *Genetics*, 130(1):153–62, Jan 1992.

Torroni, A., Schurr, T.G., Cabell, M.F., Brown, M.D., Neel, J.V., Larsen, M., Smith, D.G., Vullo, C.M. and Wallace, D.C. Asian affinities and continental radiation of the four founding native american mtdnas. *Am J Hum Genet*, 53(3):563–90, Sep 1993.

Toulza, E., Mattiuzzo, N.R., Galliano, M.F., Jonca, N., Dossat, C., Jacob, D., de Daruvar, A., Wincker, P., Serre, G. and Guerrin, M. Large-scale identification of human genes implicated in epidermal barrier function. *Genome biology*, 8(6):R107, 2007.

Tsai, M.T., Cheng, C.J., Lin, Y.C., Chen, C.C., Wu, A.R., Wu, M.T., Hsu, C.C. and Yang, R.B. Isolation and characterization of a secreted, cell-surface glycoprotein scube2 from humans. *Biochem J*, 422(1):119–28, Jul 2009. doi: 10.1042/BJ20090341.

Turchin, M.C., Chiang, C.W.K., Palmer, C.D., Sankararaman, S., Reich, D., Genetic Investigation of ANthropometric Traits (GIANT) Consortium and Hirschhorn, J.N. Evidence of widespread selection on standing variation in europe at height-associated snps. *Nat Genet*, 44(9):1015–9, Sep 2012. doi: 10.1038/ng.2368.

Underhill, P.A., Jin, L., Zemans, R., Oefner, P.J. and Cavalli-Sforza, L.L. A pre-columbian y chromosome-specific transition and its implications for human evolutionary history. *Proc Natl Acad Sci U S A*, 93(1):196–200, Jan 1996.

Vågene, Å.J., Herbig, A., Campana, M.G., Robles García, N.M., Warinner, C., Sabin, S., Spyrou, M.A., Andrades Valtueña, A., Huson, D., Tuross, N. et al. Salmonella enterica genomes from victims of a major sixteenth-century epidemic in mexico. *Nat Ecol Evol*, 2(3):520–528, Mar 2018. doi: 10.1038/s41559-017-0446-6.

Valverde, G., Zhou, H., Lippold, S., de Filippo, C., Tang, K., López Herráez, D., Li, J. and Stoneking, M. A novel candidate region for genetic adaptation to high altitude in andean populations. *PLoS One*, 10(5):e0125444, 2015. doi: 10.1371/journal.pone.0125444.

Valverde, P., Healy, E., Jackson, I., Rees, J.L. and Thody, A.J. Variants of the melanocyte-stimulating hormone receptor gene are associated with red hair and fair skin in humans. *Nat Genet*, 11(3):328–30, Nov 1995. doi: 10.1038/ng1195-328.

van der Harst, P. and Verweij, N. Identification of 64 novel genetic loci provides an expanded view on the genetic architecture of coronary artery disease. *Circ Res*, 122(3): 433–443, Feb 2018. doi: 10.1161/CIRCRESAHA.117.312086.

van der Horst, G.T., Muijtjens, M., Kobayashi, K., Takano, R., Kanno, S., Takao, M., de Wit, J., Verkerk, A., Eker, A.P., van Leenen, D. et al. Mammalian cry1 and cry2 are essential for maintenance of circadian rhythms. *Nature*, 398(6728):627–30, Apr 1999. doi: 10.1038/19323.

van Dijk, M., Mulders, J., Poutsma, A., Könst, A.A.M., Lachmeijer, A.M.A., Dekker, G.A., Blankenstein, M.A. and Oudejans, C.B.M. Maternal segregation of the dutch preeclampsia locus at 10q22 with a new member of the winged helix gene family. *Nat Genet*, 37(5):514–9, May 2005. doi: 10.1038/ng1541.

van Dijk, M., van Bezu, J., van Abel, D., Dunk, C., Blankenstein, M.A., Oudejans, C.B.M. and Lye, S.J. The stox1 genotype associated with pre-eclampsia leads to a reduction of trophoblast invasion by alpha-t-catenin upregulation. *Hum Mol Genet*, 19(13):2658–67, Jul 2010. doi: 10.1093/hmg/ddq152.

Vernot, B. and Akey, J.M. Resurrecting surviving neandertal lineages from modern human genomes. *Science*, 343(6174):1017–21, Feb 2014. doi: 10.1126/science.1245938.

Vernot, B., Tucci, S., Kelso, J., Schraiber, J.G., Wolf, A.B., Gittelman, R.M., Dannemann, M., Grote, S., McCoy, R.C., Norton, H. et al. Excavating neandertal and denisovan dna from the genomes of melanesian individuals. *Science*, 352(6282):235–9, Apr 2016. doi: 10.1126/science.aad9416.

Vilhjálmsson, B.J., Yang, J., Finucane, H.K., Gusev, A., Lindström, S., Ripke, S., Genovese, G., Loh, P.R., Bhatia, G., Do, R. et al. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *Am J Hum Genet*, 97(4):576–92, Oct 2015. doi: 10.1016/j.ajhg.2015.09.001.

Villarreal-Molina, M.T., Aguilar-Salinas, C.A., Rodríguez-Cruz, M., Riaño, D., Villalobos-Comparan, M., Coral-Vazquez, R., Menjivar, M., Yescas-Gomez, P., Königsoerg-Fainstein, M., Romero-Hidalgo, S. et al. The atp-binding cassette transporter a1 r230c variant affects hdl cholesterol levels and bmi in the mexican population: association with obesity and obesity-related comorbidities. *Diabetes*, 56(7):1881–7, Jul 2007. doi: 10.2337/db06-0905.

Villarreal-Molina, M.T., Flores-Dorantes, M.T., Arellano-Campos, O., Villalobos-Comparan, M., Rodríguez-Cruz, M., Miliar-García, A., Huertas-Vazquez, A., Menjivar, M., Romero-Hidalgo, S., Wacher, N.H. et al. Association of the atp-binding cassette transporter a1 r230c variant with early-onset type 2 diabetes in a mexican population. *Diabetes*, 57(2):509–513, 2008.

Visconti, A., Duffy, D.L., Liu, F., Zhu, G., Wu, W., Chen, Y., Hysi, P.G., Zeng, C., Sanna, M., Iles, M.M. et al. Genome-wide association study in 176,678 europeans reveals genetic loci for tanning response to sun exposure. *Nat Commun*, 9(1):1684, May 2018. doi: 10.1038/s41467-018-04086-y.

Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A. and Yang, J. 10 years of gwas discovery: Biology, function, and translation. *Am J Hum Genet*, 101(1):5–22, Jul 2017. doi: 10.1016/j.ajhg.2017.06.005.

Visser, M., Kayser, M. and Palstra, R.J. Herc2 rs12913832 modulates human pigmentation by attenuating chromatin-loop formation between a long-range enhancer and the oca2 promoter. *Genome Res*, 22(3):446–55, Mar 2012. doi: 10.1101/gr.128652.111.

Visser, M., Palstra, R.J. and Kayser, M. Human skin color is influenced by an intergenic dna polymorphism regulating transcription of the nearby bnc2 pigmentation gene. *Hum Mol Genet*, 23(21):5750–62, Nov 2014. doi: 10.1093/hmg/ddu289.

Visser, M., Palstra, R.J. and Kayser, M. Allele-specific transcriptional regulation of irf4 in melanocytes is mediated by chromatin looping of the intronic rs12203592 enhancer to the irf4 promoter. *Hum Mol Genet*, 24(9):2649–61, May 2015. doi: 10.1093/hmg/ddv029.

Vitti, J.J., Grossman, S.R. and Sabeti, P.C. Detecting natural selection in genomic data. *Annu Rev Genet*, 47:97–120, 2013. doi: 10.1146/annurev-genet-111212-133526.

Voight, B.F., Kudaravalli, S., Wen, X. and Pritchard, J.K. A map of recent positive selection in the human genome. *PLoS Biol*, 4(3):e72, Mar 2006. doi: 10.1371/journal.pbio.0040072.

von Cramon-Taubadel, N., Strauss, A. and Hubbe, M. Evolutionary population history of early paleoamerican cranial morphology. *Sci Adv*, 3(2):e1602289, Feb 2017. doi: 10.1126/sciadv.1602289.

Wall, J.D., Andolfatto, P. and Przeworski, M. Testing models of selection and demography in drosophila simulans. *Genetics*, 162(1):203–216, 2002.

Wall, J.D., Yang, M.A., Jay, F., Kim, S.K., Durand, E.Y., Stevison, L.S., Gignoux, C., Woerner, A., Hammer, M.F. and Slatkin, M. Higher levels of neanderthal ancestry in east asians than in europeans. *Genetics*, 194(1):199–209, May 2013. doi: 10.1534/genetics.112.148213.

Wall, T.L., Luczak, S.E. and Hiller-Sturmhöfel, S. Biology, genetics, and environment: Underlying factors influencing alcohol metabolism. *Alcohol Res*, 38(1):59–68, 2016.

Wallace, D.C., Garrison, K. and Knowler, W.C. Dramatic founder effects in amerindian mitochondrial dnas. *Am J Phys Anthropol*, 68(2):149–55, Oct 1985. doi: 10.1002/ajpa.1330680202.

Walsh, S. and Kayser, M. A practical guide to the hirisplex system: Simultaneous prediction of eye and hair color from dna. *Methods Mol Biol*, 1420:213–31, 2016. doi: 10.1007/978-1-4939-3597-0_17.

Walsh, S., Lindenbergh, A., Zuniga, S.B., Sijen, T., de Knijff, P., Kayser, M. and Ballantyne, K.N. Developmental validation of the irisplex system: determination of blue and brown iris colour for forensic intelligence. *Forensic Science International: Genetics*, 5 (5):464–471, 2011.

Walsh, S., Liu, F., Wollstein, A., Kovatsi, L., Ralf, A., Kosiniak-Kamysz, A., Branicki, W. and Kayser, M. The hirisplex system for simultaneous prediction of hair and eye colour from dna. *Forensic Sci Int Genet*, 7(1):98–115, Jan 2013. doi: 10.1016/j.fsigen.2012.07.005.

Wang, K., Li, M. and Bucan, M. Pathway-based approaches for analysis of genomewide association studies. *Am J Hum Genet*, 81(6):1278–83, Dec 2007a. doi: 10.1086/522374.

Wang, S., Lewis, C.M., Jakobsson, M., Ramachandran, S., Ray, N., Bedoya, G., Rojas, W., Parra, M.V., Molina, J.A., Gallo, C. et al. Genetic variation and population structure in native americans. *PLoS Genet*, 3(11):e185, Nov 2007b. doi: 10.1371/journal.pgen.0030185.

Wang, S., Ray, N., Rojas, W., Parra, M.V., Bedoya, G., Gallo, C., Poletti, G., Mazzotti, G., Hill, K., Hurtado, A.M. et al. Geographic patterns of genome admixture in latin american mestizos. *PLoS Genet*, 4(3):e1000037, Mar 2008. doi: 10.1371/journal.pgen.1000037.

Wang, Z.Q., Si, L., Tang, Q., Lin, D., Fu, Z., Zhang, J., Cui, B., Zhu, Y., Kong, X., Deng, M. et al. Gain-of-function mutation of kit ligand on melanin synthesis causes familial progressive hyperpigmentation. *Am J Hum Genet*, 84(5):672–7, May 2009. doi: 10.1016/j.ajhg.2009.03.019.

Ware, E.B., Schmitz, L.L., Faul, J.D., Gard, A., Mitchell, C., Smith, J.A., Zhao, W., Weir, D. and Kardia, S.L. Heterogeneity in polygenic scores for common human traits. *bioRxiv*, 2017. doi: 10.1101/106062. URL `https://www.biorxiv.org/content/early/2017/02/05/106062`.

Waterson, R.H., Lander, E.S., Wilson, R.K. et al. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, 437(7055):69, 2005.

Wegmann, D., Leuenberger, C., Neuenschwander, S. and Excoffier, L. Abctoolbox: a versatile toolkit for approximate bayesian computations. *BMC bioinformatics*, 11(1): 116, 2010.

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L. et al. The nhgri gwas catalog, a curated resource of snp-trait associations. *Nucleic acids research*, 42(D1):D1001–D1006, 2013.

Westgate, G.E., Botchkareva, N.V. and Tobin, D.J. The biology of hair diversity. *International journal of cosmetic science*, 35(4):329–336, 2013.

Whiteman, D.C., Parsons, P.G. and Green, A.C. Determinants of melanocyte density in adult human skin. *Archives of dermatological research*, 291(9):511–516, 1999.

Wiedmann, S., Fischer, M., Koehler, M., Neureuther, K., Riegger, G., Doering, A., Schunkert, H., Hengstenberg, C. and Baessler, A. Genetic variants within the lpin1 gene, encoding lipin, are influencing phenotypes of the metabolic syndrome in humans. *Diabetes*, 57(1):209–17, Jan 2008. doi: 10.2337/db07-0083.

Wilde, S., Timpson, A., Kirsanow, K., Kaiser, E., Kayser, M., Unterländer, M., Hollfelder, N., Potekhina, I.D., Schier, W., Thomas, M.G. et al. Direct evidence for positive selection of skin, hair, and eye pigmentation in europeans during the last 5,000 y. *Proc Natl Acad Sci U S A*, 111(13):4832–7, Apr 2014. doi: 10.1073/pnas.1316513111.

Williams, G.C. Pleiotropy, natural selection, and the evolution of senescence. *evolution*, 11(4):398–411, 1957.

Winkler, C.A., Nelson, G.W. and Smith, M.W. Admixture mapping comes of age. *Annu Rev Genomics Hum Genet*, 11:65–89, 2010. doi: 10.1146/annurev-genom-082509-141523.

Wistow, G., Bernstein, S.L., Ray, S., Wyatt, M.K., Behal, A., Touchman, J.W., Bouffard, G., Smith, D. and Peterson, K. Expressed sequence tag analysis of adult human iris for the neibank project: steroid-response factors and similarities with retinal pigment epithelium. *Mol Vis*, 8:185–95, Jun 2002.

Wojcik, G., Graff, M., Nishimura, K.K., Tao, R., Haessler, J., Gignoux, C.R., Highland, H.M., Patel, Y.M., Sorokin, E.P., Avery, C.L. et al. Genetic diversity turns a new page in our understanding of complex traits. *bioRxiv*, 2017. doi: 10.1101/188094. URL `https://www.biorxiv.org/content/early/2017/09/15/188094`.

Wolf, A.B. and Akey, J.M. Outstanding questions in the study of archaic hominin admixture. *PLoS Genet*, 14(5):e1007349, 05 2018. doi: 10.1371/journal.pgen.1007349.

Wolf, M., Shah, A., Jimenez-Kimble, R., Sauk, J., Ecker, J.L. and Thadhani, R. Differential risk of hypertensive disorders of pregnancy among hispanic women. *J Am Soc Nephrol*, 15(5):1330–8, May 2004.

Wollstein, A., Walsh, S., Liu, F., Chakravarthy, U., Rahu, M., Seland, J.H., Soubrane, G., Tomazzoli, L., Topouzis, F., Vingerling, J.R. et al. Novel quantitative pigmentation phenotyping enhances genetic association, epistasis, and prediction of human eye colour. *Sci Rep*, 7:43359, Feb 2017. doi: 10.1038/srep43359.

Workman, P., Blumberg, B. and Cooper, A. Selection, gene migration and polymorphic stability in a us white and negro population. *American journal of human genetics*, 15 (4):429, 1963.

Wray, N.R. Allele frequencies and the r2 measure of linkage disequilibrium: impact on design and interpretation of association studies. *Twin Res Hum Genet*, 8(2):87–94, Apr 2005. doi: 10.1375/1832427053738827.

Wright, S. The genetical structure of populations. *Annals of Human Genetics*, 15(1): 323–354, 1949.

Xu, X., Dong, G.X., Hu, X.S., Miao, L., Zhang, X.L., Zhang, D.L., Yang, H.D., Zhang, T.Y., Zou, Z.T., Zhang, T.T. et al. The genetic basis of white tigers. *Curr Biol*, 23(11): 1031–5, Jun 2013. doi: 10.1016/j.cub.2013.04.054.

Yamaguchi, K., Watanabe, C., Kawaguchi, A., Sato, T., Naka, I., Shindo, M., Moromizato, K., Aoki, K., Ishida, H. and Kimura, R. Association of melanocortin 1 receptor gene (mc1r) polymorphisms with skin reflectance and freckles in japanese. *J Hum Genet*, 57 (11):700–8, Nov 2012. doi: 10.1038/jhg.2012.96.

Yanai, H., Chen, H.M., Inuzuka, T., Kondo, S., Mak, T.W., Takaoka, A., Honda, K. and Taniguchi, T. Role of ifn regulatory factor 5 transcription factor in antiviral immunity and tumor suppression. *Proc Natl Acad Sci U S A*, 104(9):3402–7, Feb 2007. doi: 10.1073/pnas.0611559104.

Yang, J., Lee, S.H., Goddard, M.E. and Visscher, P.M. Gcta: a tool for genome-wide complex trait analysis. *Am J Hum Genet*, 88(1):76–82, Jan 2011a. doi: 10.1016/j.ajhg.2010.11.011.

Yang, J., Weedon, M.N., Purcell, S., Lettre, G., Estrada, K., Willer, C.J., Smith, A.V., Ingelsson, E., O'Connell, J.R., Mangino, M. et al. Genomic inflation factors under

polygenic inheritance. *Eur J Hum Genet*, 19(7):807–12, Jul 2011b. doi: 10.1038/ejhg.2011.39.

Yang, J., Ferreira, T., Morris, A.P., Medland, S.E., Genetic Investigation of ANthropometric Traits (GIANT) Consortium, DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, Madden, P.A.F., Heath, A.C., Martin, N.G., Montgomery, G.W. et al. Conditional and joint multiple-snp analysis of gwas summary statistics identifies additional variants influencing complex traits. *Nat Genet*, 44(4):369–75, S1–3, Mar 2012. doi: 10.1038/ng.2213.

Yang, Z., Zhong, H., Chen, J., Zhang, X., Zhang, H., Luo, X., Xu, S., Chen, H., Lu, D., Han, Y. et al. A genetic mechanism for convergent skin lightening during recent human evolution. *Mol Biol Evol*, 33(5):1177–87, 05 2016. doi: 10.1093/molbev/msw003.

Yi, X., Liang, Y., Huerta-Sanchez, E., Jin, X., Cuo, Z.X.P., Pool, J.E., Xu, X., Jiang, H., Vinckenbosch, N., Korneliussen, T.S. et al. Sequencing of 50 human exomes reveals adaptation to high altitude. *Science*, 329(5987):75–8, Jul 2010. doi: 10.1126/science.1190371.

Yoon, S., Nguyen, H.C.T., Yoo, Y.J., Kim, J., Baik, B., Kim, S., Kim, J., Kim, S. and Nam, D. Efficient pathway enrichment and network analysis of gwas summary data using gsa-snp2. *Nucleic Acids Res*, 46(10):e60, Jun 2018. doi: 10.1093/nar/gky175.

Yu, R., Broady, R., Huang, Y., Wang, Y., Yu, J., Gao, M., Levings, M., Wei, S., Zhang, S., Xu, A. et al. Transcriptome analysis reveals markers of aberrantly activated innate immunity in vitiligo lesional and non-lesional skin. *PLoS One*, 7(12):e51040, 2012. doi: 10.1371/journal.pone.0051040.

Zaidi, A.A., Mattern, B.C., Claes, P., McEcoy, B., Hughes, C. and Shriver, M.D. Investigating the case of human nose shape and climate adaptation. *PLoS genetics*, 13(3):e1006616, 2017.

Zeggini, E., Scott, L.J., Saxena, R., Voight, B.F., Marchini, J.L., Hu, T., de Bakker, P.I.W., Abecasis, G.R., Almgren, P., Andersen, G. et al. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat Genet*, 40(5):638–45, May 2008. doi: 10.1038/ng.120.

Zegura, S.L., Karafet, T.M., Zhivotovsky, L.A. and Hammer, M.F. High-resolution snps and microsatellite haplotypes point to a single, recent entry of native american y chromosomes into the americas. *Mol Biol Evol*, 21(1):164–75, Jan 2004. doi: 10.1093/molbev/msh009.

Zhang, F., Gu, W., Hurles, M.E. and Lupski, J.R. Copy number variation in human health, disease, and evolution. *Annual review of genomics and human genetics*, 10:451–481, 2009.

Zhang, K., Cui, S., Chang, S., Zhang, L. and Wang, J. i-gsea4gwas: a web server for identification of pathways/gene sets associated with traits by applying an improved

gene set enrichment analysis to genome-wide association study. *Nucleic Acids Res*, 38 (Web Server issue):W90–5, Jul 2010a. doi: 10.1093/nar/gkq324.

Zhang, M., Song, F., Liang, L., Nan, H., Zhang, J., Liu, H., Wang, L.E., Wei, Q., Lee, J.E., Amos, C.I. et al. Genome-wide association studies identify several new loci associated with pigmentation traits and skin cancer risk in european americans. *Hum Mol Genet*, 22(14):2948–59, Jul 2013a. doi: 10.1093/hmg/ddt142.

Zhang, R., Jiang, F., Hu, C., Yu, W., Wang, J., Wang, C., Ma, X., Tang, S., Bao, Y., Xiang, K. et al. Genetic variants of lpin1 indicate an association with type 2 diabetes mellitus in a chinese population. *Diabet Med*, 30(1):118–22, Jan 2013b. doi: 10.1111/j.1464-5491.2012.03758.x.

Zhang, S., Zhu, X. and Zhao, H. On a semiparametric test to detect associations between quantitative traits and candidate genes using unrelated individuals. *Genetic epidemiology*, 24(1):44–56, 2003.

Zhang, Z., Ersoz, E., Lai, C.Q., Todhunter, R.J., Tiwari, H.K., Gore, M.A., Bradbury, P.J., Yu, J., Arnett, D.K., Ordovas, J.M. et al. Mixed linear model approach adapted for genome-wide association studies. *Nature genetics*, 42(4):355, 2010b.

Zhao, W., Rasheed, A., Tikkanen, E., Lee, J.J., Butterworth, A.S., Howson, J.M.M., Assimes, T.L., Chowdhury, R., Orho-Melander, M., Damrauer, S. et al. Identification of new susceptibility loci for type 2 diabetes and shared etiological pathways with coronary heart disease. *Nat Genet*, 49(10):1450–1457, Oct 2017. doi: 10.1038/ng.3943.

Zhou, D., Udpa, N., Ronen, R., Stobdan, T., Liang, J., Appenzeller, O., Zhao, H.W., Yin, Y., Du, Y., Guo, L. et al. Whole-genome sequencing uncovers the genetic basis of chronic mountain sickness in andean highlanders. *Am J Hum Genet*, 93(3):452–62, Sep 2013a. doi: 10.1016/j.ajhg.2013.07.011.

Zhou, Q., Zhao, L. and Guan, Y. Strong selection at mhc in mexicans since admixture. *PLoS Genet*, 12(2):e1005847, Feb 2016. doi: 10.1371/journal.pgen.1005847.

Zhou, X., Carbonetto, P. and Stephens, M. Polygenic modeling with bayesian sparse linear mixed models. *PLoS genetics*, 9(2):e1003264, 2013b.

Zhu, J., Shang, Y. and Zhang, M. Mechanistic basis of maguk-organized complexes in synaptic development and signalling. *Nat Rev Neurosci*, 17(4):209–23, Apr 2016. doi: 10.1038/nrn.2016.18.

Zoledziewska, M., Sidore, C., Chiang, C.W.K., Sanna, S., Mulas, A., Steri, M., Busonero, F., Marcus, J.H., Marongiu, M., Maschio, A. et al. Height-reducing variants and selection for short stature in sardinia. *Nat Genet*, 47(11):1352–1356, Nov 2015. doi: 10.1038/ng.3403.

Appendix A

Detecting signatures of selection in Native Americans

Table A.1: Worldwide populations used to assess the geographic distribution of the top selected SNPs.

Population	Country of origin	Major geographical region	Sample size	Source (reference)
ESN	Nigeria	Africa	95	1000 Genomes Project
GuiGhanaKgal	Botswana	Africa	14	Schlebusch et al. 2012
GWD	Gambia	Africa	111	1000 Genomes Project
Juhoansi	Namibia	Africa	15	Schlebusch et al. 2012
Khwe	Namibia	Africa	14	Schlebusch et al. 2012
LWK	Kenya	Africa	73	1000 Genomes Project
MSL	Sierra Leone	Africa	69	1000 Genomes Project
Bantu	South Africa	Africa	19	Schlebusch et al. 2012
Xun	Angola	Africa	19	Schlebusch et al. 2012
YRI	Nigeria	Africa	101	1000 Genomes Project
Aymara	Bolivia	Americas	13	Chacon-Duque et al. 2018
Colla	Argentina	Americas	11	Eichstaedt et al. 2014
Embera	Colombia	Americas	14	Chacon-Duque et al. 2018
Mixe	Mexico	Americas	16	Chacon-Duque et al. 2018
Nahua	Mexico	Americas	17	Chacon-Duque et al. 2018
Quechua	Peru/Bolivia	Americas	11	Chacon-Duque et al. 2018
Wichi	Argentina	Americas	15	Eichstaedt et al. 2014
CHB	China	Asia (Eastern Eurasia)	103	1000 Genomes Project
CHS-FuJian	China	Asia (Eastern Eurasia)	25	1000 Genomes Project
CHS-HuNan	China	Asia (Eastern Eurasia)	59	1000 Genomes Project
CHS	China	Asia (Eastern Eurasia)	13	1000 Genomes Project
JPT	Japan	Asia (Eastern Eurasia)	104	1000 Genomes Project
Papuan	Papua New Guinea	Asia (Eastern Eurasia)	15	Simons Genome Diversity Project
BEB	Bangladesh	Asia (Eastern Eurasia)	83	1000 Genomes Project
GIH	India	Asia (Eastern Eurasia)	96	1000 Genomes Project
ITU	India	Asia (Eastern Eurasia)	98	1000 Genomes Project
PJL	Pakistan	Asia (Eastern Eurasia)	86	1000 Genomes Project
STU	Sri Lanka	Asia (Eastern Eurasia)	95	1000 Genomes Project
Bajo	Indonesia	Asia (Eastern Eurasia)	31	Moerseburg et al. 2016
Burmese	Myanmar	Asia (Eastern Eurasia)	20	Moerseburg et al. 2016
CDX	China	Asia (Eastern Eurasia)	82	1000 Genomes Project

. . . continued on next page. . .

. . . continued from previous page . . .

Population	**Country of origin**	**Major geographical region**	**Sample size**	**Source (reference)**
Dusun	Brunei	Asia (Eastern Eurasia)	20	Moerseburg et al. 2016
Igorot	Phillippines	Asia (Eastern Eurasia)	21	Moerseburg et al. 2016
KHV	Vietnam	Asia (Eastern Eurasia)	98	1000 Genomes Project
Lebbo	Indonesia	Asia (Eastern Eurasia)	15	Moerseburg et al. 2016
Malay	Singapore	Asia (Eastern Eurasia)	25	Moerseburg et al. 2016
Murut	Brunei	Asia (Eastern Eurasia)	17	Moerseburg et al. 2016
Vietnamese	Vietnam	Asia (Eastern Eurasia)	18	Moerseburg et al. 2016
Evenki	Russia	Asia (Eastern Eurasia)	16	Carmona et al. 2014
Even	Russia	Asia (Eastern Eurasia)	14	Carmona et al. 2014
Koryak	Russia	Asia (Eastern Eurasia)	15	Carmona et al. 2014
Eskimo	Russia	Asia (Eastern Eurasia)	10	Carmona et al. 2014
CEU	USA	Europe (Western Eurasia)	91	1000 Genomes Project
FIN	Finland	Europe (Western Eurasia)	99	1000 Genomes Project
France	France	Europe (Western Eurasia)	10	Chacon-Duque et al. 2018
Cornwall	UK	Europe (Western Eurasia)	29	1000 Genomes Project
Kent	UK	Europe (Western Eurasia)	31	1000 Genomes Project
Orkney	UK	Europe (Western Eurasia)	21	1000 Genomes Project
Germany	Germany	Europe (Western Eurasia)	10	Chacon-Duque et al. 2018
Catilla y Leon	Spain	Europe (Western Eurasia)	12	1000 Genomes Project
Catalunya	Spain	Europe (Western Eurasia)	10	1000 Genomes Project
Valencia	Spain	Europe (Western Eurasia)	14	1000 Genomes Project
Portugal (Central)	Portugal	Europe (Western Eurasia)	11	Chacon-Duque et al. 2018
Portugal (North)	Portugal	Europe (Western Eurasia)	13	Chacon-Duque et al. 2018
Portugal (South)	Portugal	Europe (Western Eurasia)	12	Chacon-Duque et al. 2018
Andalucia	Spain	Europe (Western Eurasia)	15	Chacon-Duque et al. 2018
Basque	Spain	Europe (Western Eurasia)	14	Chacon-Duque et al. 2018
Canary Island	Spain	Europe (Western Eurasia)	14	Chacon-Duque et al. 2018
Spanish (Central)	Spain	Europe (Western Eurasia)	15	Chacon-Duque et al. 2018
TSI	Italy	Europe (Western Eurasia)	106	1000 Genomes Project
Jordan	Jordania	Europe (Western Eurasia)	15	Chacon-Duque et al. 2018
Libya	Libya	Europe (Western Eurasia)	15	Chacon-Duque et al. 2018
Morocco	Morocco	Europe (Western Eurasia)	14	Chacon-Duque et al. 2018
Tunisia	Tunisia	Europe (Western Eurasia)	14	Chacon-Duque et al. 2018

Table A.2: Significantly enriched Gene Ontology (GO) categories based on the ranking of the maximum PBS score of gene regions in 168 Native Americans (P-value < 0.01, FDR q-value < 0.1, enrichment score > 5.)

GO Term	Description	P-value	FDR q-value	Enrichment	Genes
GO:0007156	Homophilic cell adhesion	1.63E-14	2.24E-10	10.27	*PCDHGA7, PCDHGA6, PCDHGA10, PCDHGA9,*
	via plasma membrane adhesion				*PCDHGA3, PCDHGB4, PCDHGA2, PCDHGA5,*
	molecules				*PCDHGA4, PCDHGA1, PCDHGA12, PCDHGB3,*
					PCDHGA8, PCDHGB2, PCDHGB1, PCDHGA11,
					PCDHGC4, PCDHGB7, PCDHGB6, PCDHGB5,
					PCDHGC3, PCDHGC5
GO:0098742	Cell-cell adhesion	7.25E-13	4.98E-09	7.65	*PCDHGA7, PCDHGA6, PCDHGA10, PCDHGA9,*
	via plasma-membrane adhesion				*PCDHGA3, PCDHGB4, PCDHGA2, PCDHGA5,*
	molecules				*PCDHGA4, PCDHGA1, PCDHGA12, PCDHGB3,*
					PCDHGA8, PCDHGB2, PCDHGB1, PCDHGA11,
					PCDHGC4, PCDHGB7, PCDHGB6, PCDHGB5,
					PCDHGC3, PCDHGC5
GO:0007399	nervous system development	3.96E-10	1.82E-06	5.64	*PCDHGA7, PCDHGA6, PCDHGA10, PCDHGA9,*
					PCDHGA3, PCDHGB4, PCDHGA2, PCDHGA5,
					PCDHGA4, PCDHGA1, PCDHGA12, PCDHGB3,
					PCDHGA8, PCDHGB2, PCDHGB1, PCDHGA11,
					PCDHGC4, PCDHGB7, PCDHGB6, PCDHGB5,
					PCDHGC3, PCDHGC5

Table A.3: Significantly enriched Gene Ontology (GO) categories based on the ranking of the maximum PBS score of gene regions in the Andean Native American population (P-value < 0.01, FDR q-value < 0.1, enrichment score > 5.)

GO Term	**Description**	**P-value**	**FDR q-value**	**Enrichment**	**Genes**
GO:0052696	Flavonoid glucuronidation	2.93E-14	2.16E-10	37.48	*UGT1A3,UGT1A1,UGT1A4,UGT1A6,UGT1A5,*
					UGT1A8,UGT1A7,UGT1A10,UGT1A9
GO:0052697	Xenobiotic glucuronidation	2.93E-14	4.31E-10	37.48	*UGT1A3,UGT1A1,UGT1A4,UGT1A6,UGT1A5,*
					UGT1A8,UGT1A7,UGT1A10,UGT1A9
GO:0009812	Flavonoid metabolic process	3.63E-11	1.78E-07	25.95	*UGT1A3,UGT1A1,UGT1A4,UGT1A6,UGT1A5,*
					UGT1A8,UGT1A7,UGT1A10,UGT1A9
GO:0052695	Cellular glucuronidation	1.11E-10	4.08E-07	24.09	*UGT1A3,UGT1A1,UGT1A4,UGT1A6,UGT1A5,*
					UGT1A8,UGT1A7,UGT1A10,UGT1A9
GO:0019585	Glucuronate metabolic process	1.70E-10	4.16E-07	19.73	*UGT1A3,UGT1A1,SORD,UGT1A4,UGT1A6,*
					UGT1A5,UGT1A8,UGT1A7,UGT1A10,UGT1A9
GO:0006063	Uronic acid metabolic process	1.70E-10	4.99E-07	19.73	*UGT1A3,UGT1A1,SORD,UGT1A4,UGT1A6,*
					UGT1A5,UGT1A8,UGT1A7,UGT1A10,UGT1A9
GO:0006805	Xenobiotic metabolic process	3.14E-07	5.77E-04	5.88	*MARC1,AS3MT,CYB5B,AADAC,UGT1A6,*
					UGT1A5,UGT1A8,UGT1A7,UGT1A10,UGT1A3,
					UGT1A1,UGT1A4,EPHX1,GGT1,CYP1B1,UGT1A9
GO:0051552	Flavone metabolic process	2.93E-07	6.16E-04	31.23	*UGT1A1,UGT1A8,UGT1A7,UGT1A10,UGT1A9*

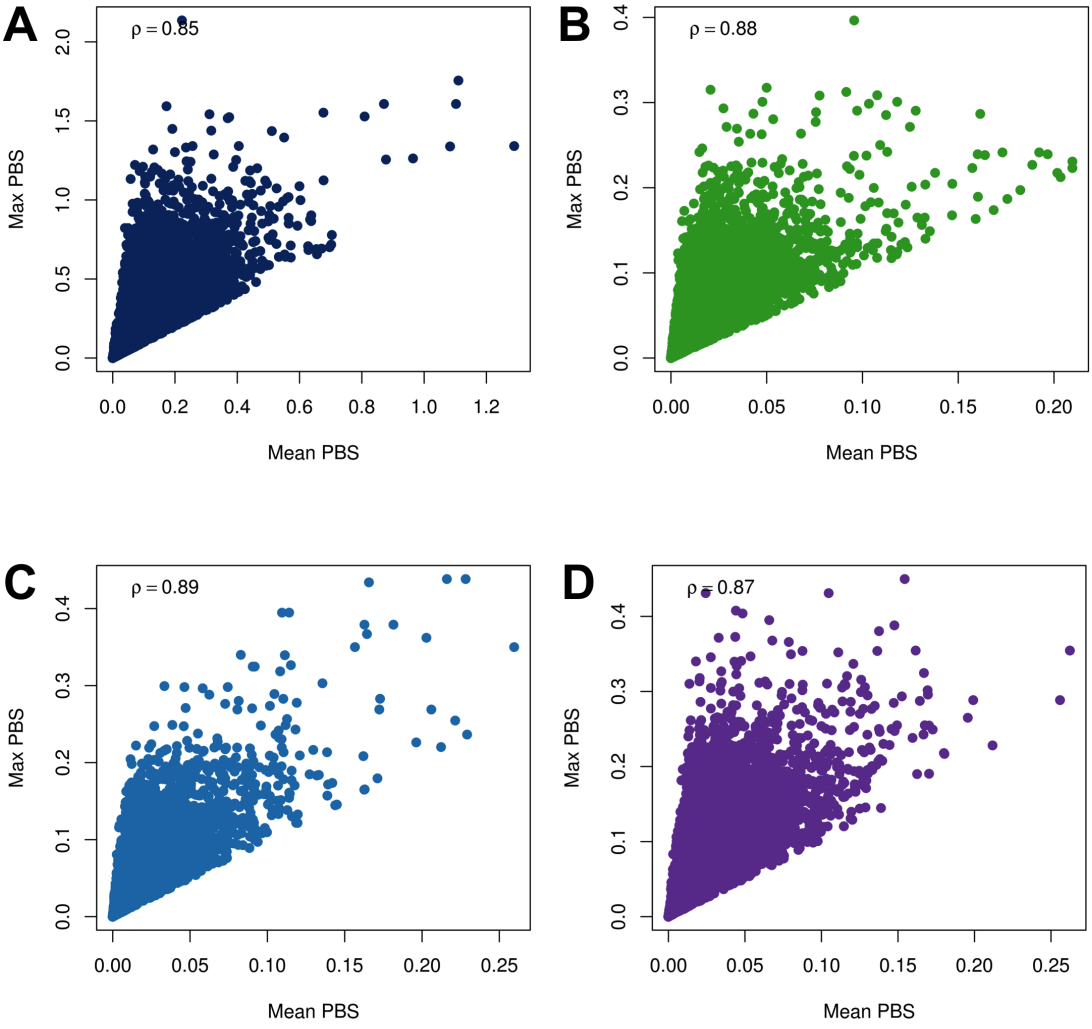

Figure A.1: Correlation between maximum and mean PBS score at each gene region. The spearman rank correlation (rho) was used to assess the correlation between the maximum and mean PBS score at each gene region in A) Native Americans, B) Meso-American Native Americans, C) Andean Native Americans and D) Mapuche Native Americans. The positive correlation indicates that the highest scoring SNP is a good representative of the selection pattern in a gene.

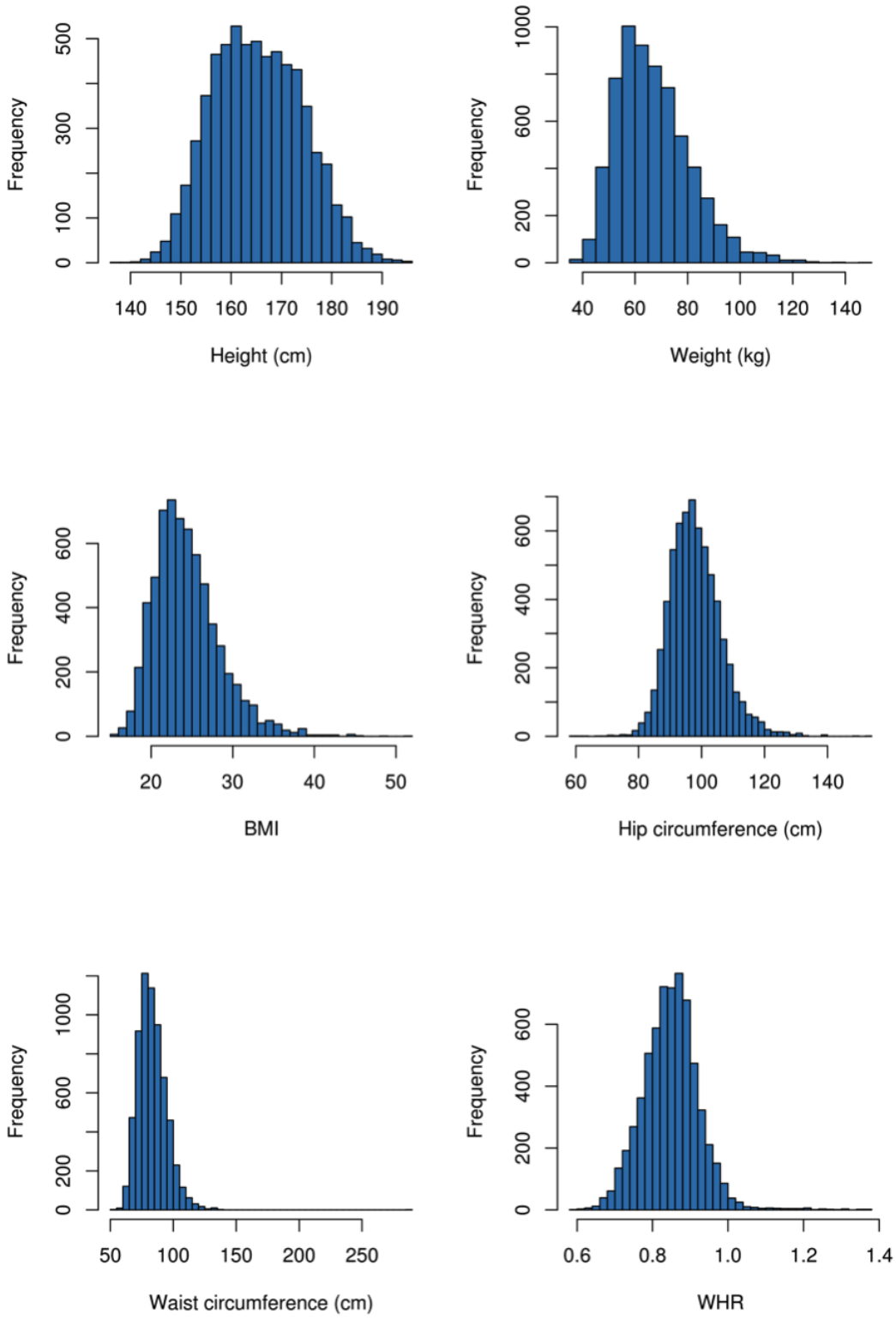

Figure A.2: Distribution of anthropometric phenotypes in the CANDELA sample.

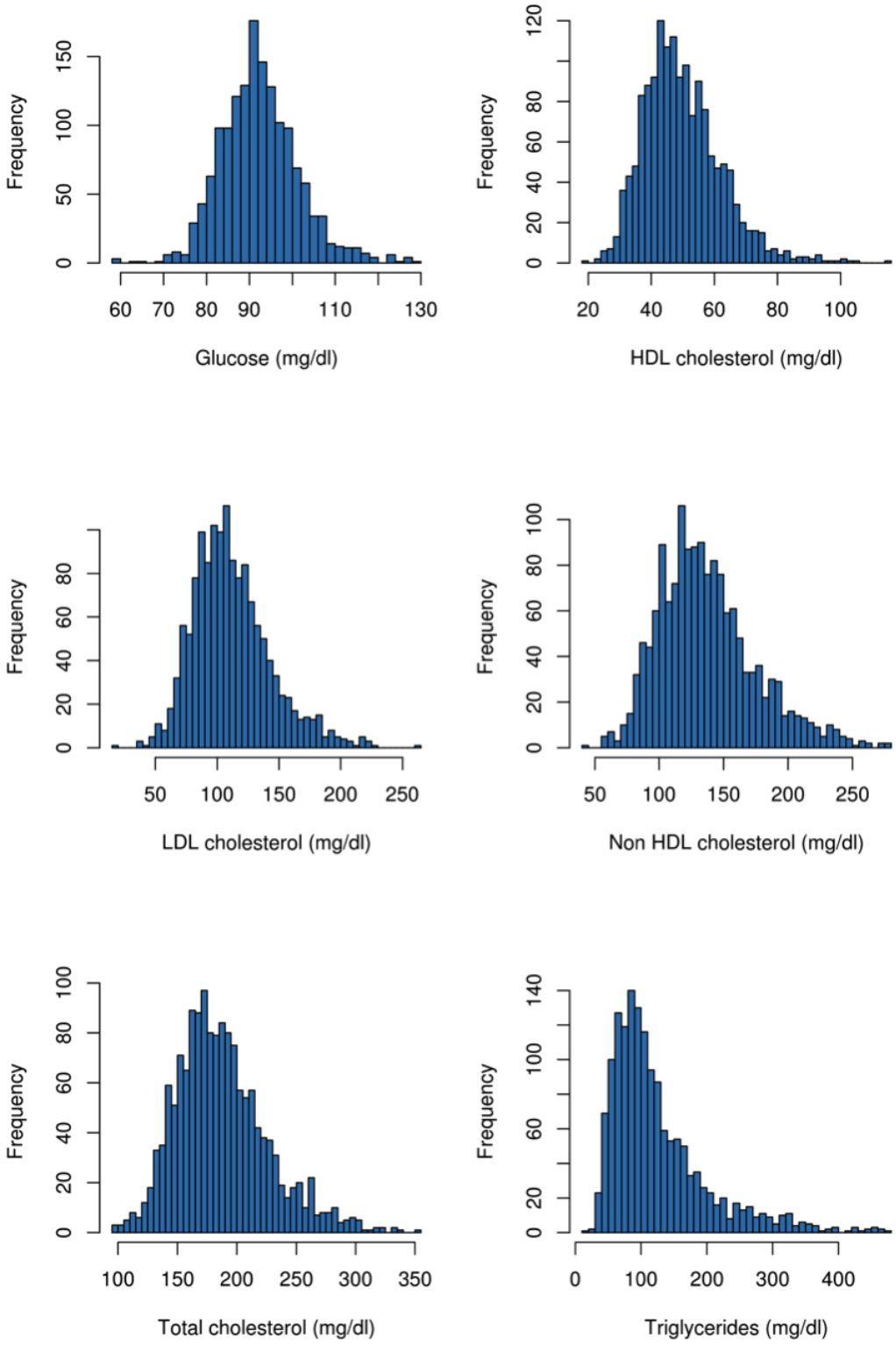

Figure A.3: Distribution of anthropometric phenotypes in Mexican volunteers from the CANDELA sample.

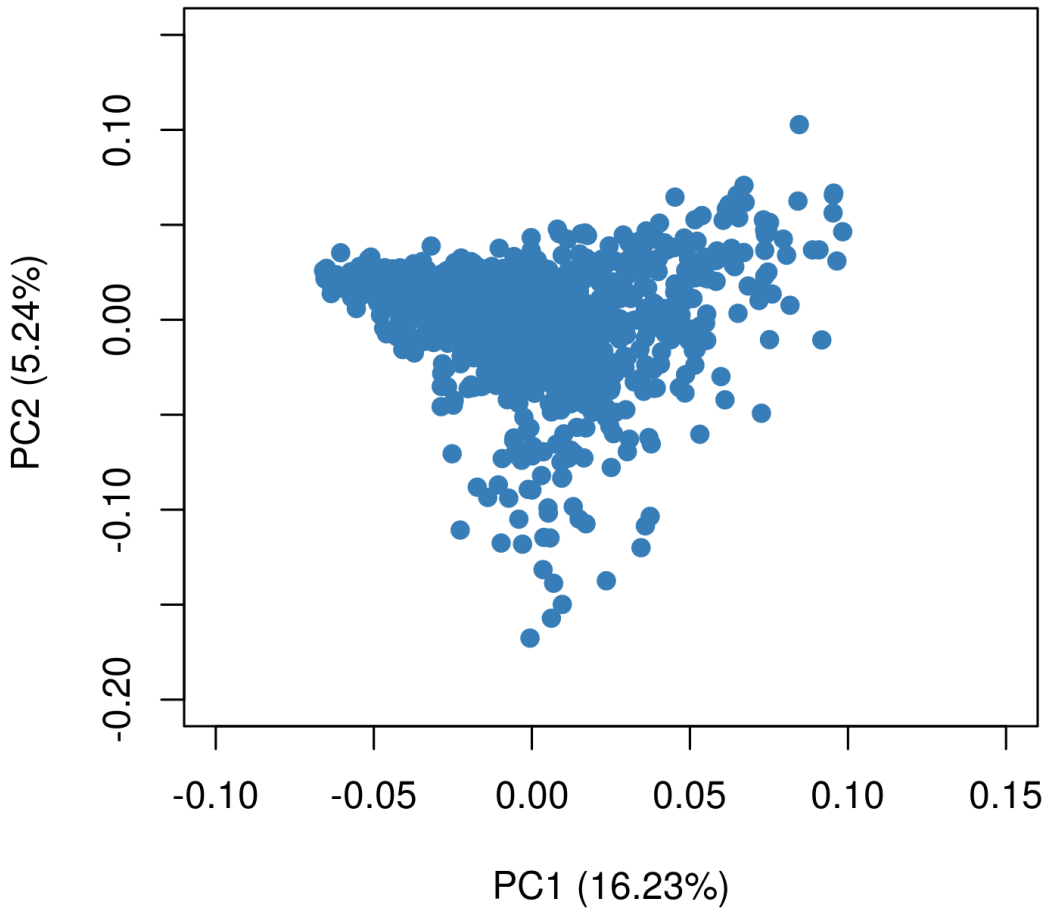

Figure A.4: Principal Component Analysis (PCA) of admixed Mexican individuals from the CANDELA sample.

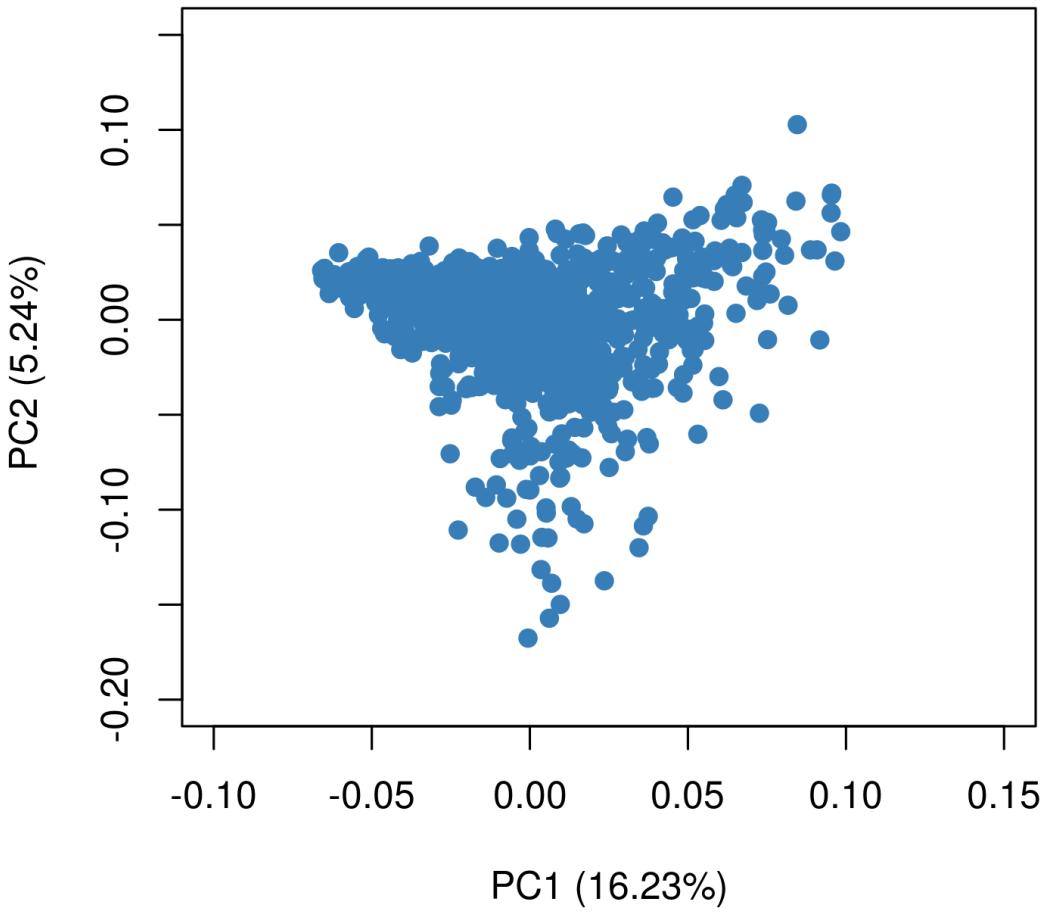

Figure A.5: Principal Component Analysis (PCA) of admixed Mexican individuals from the CANDELA sample.

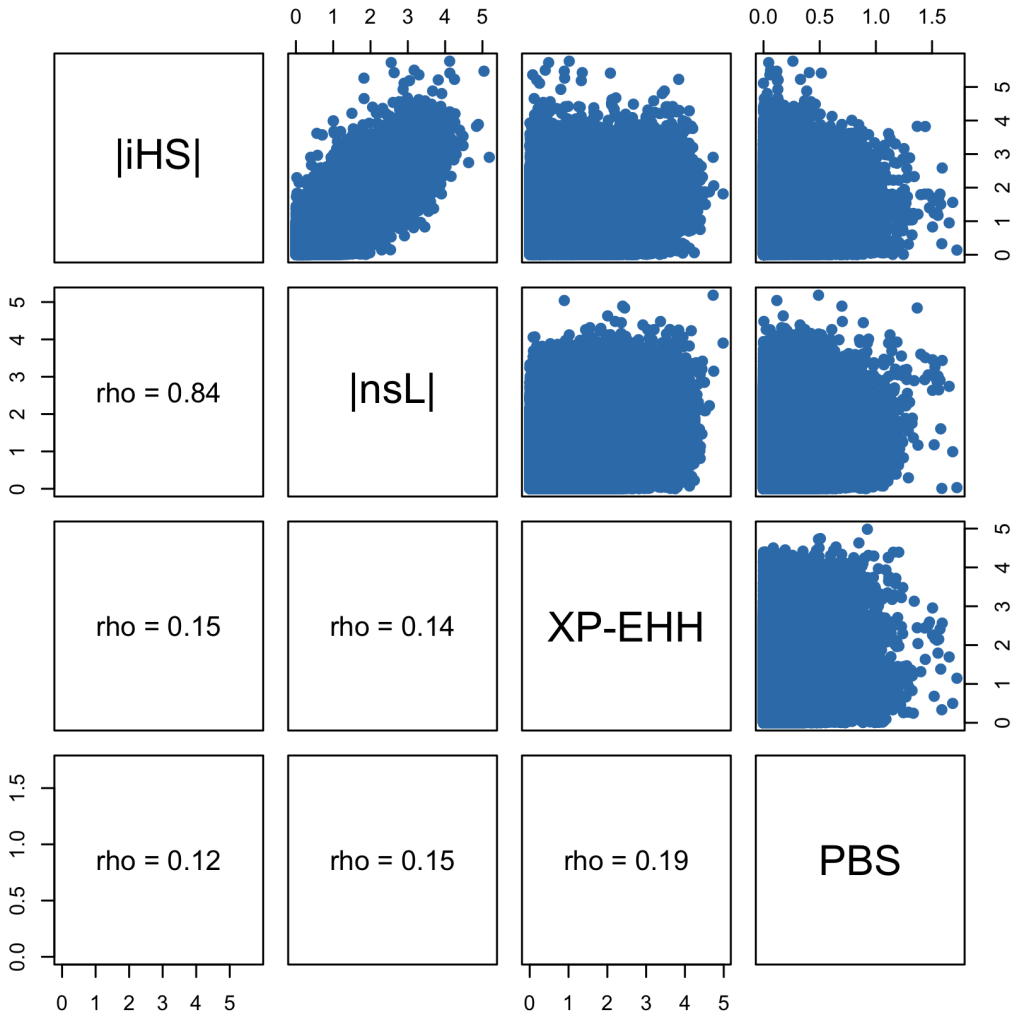

Figure A.6: Correlation between selection statistics. Spearman correlation coefficient (rho) was computed between the four selection statistics computed in 168 Native Americans.

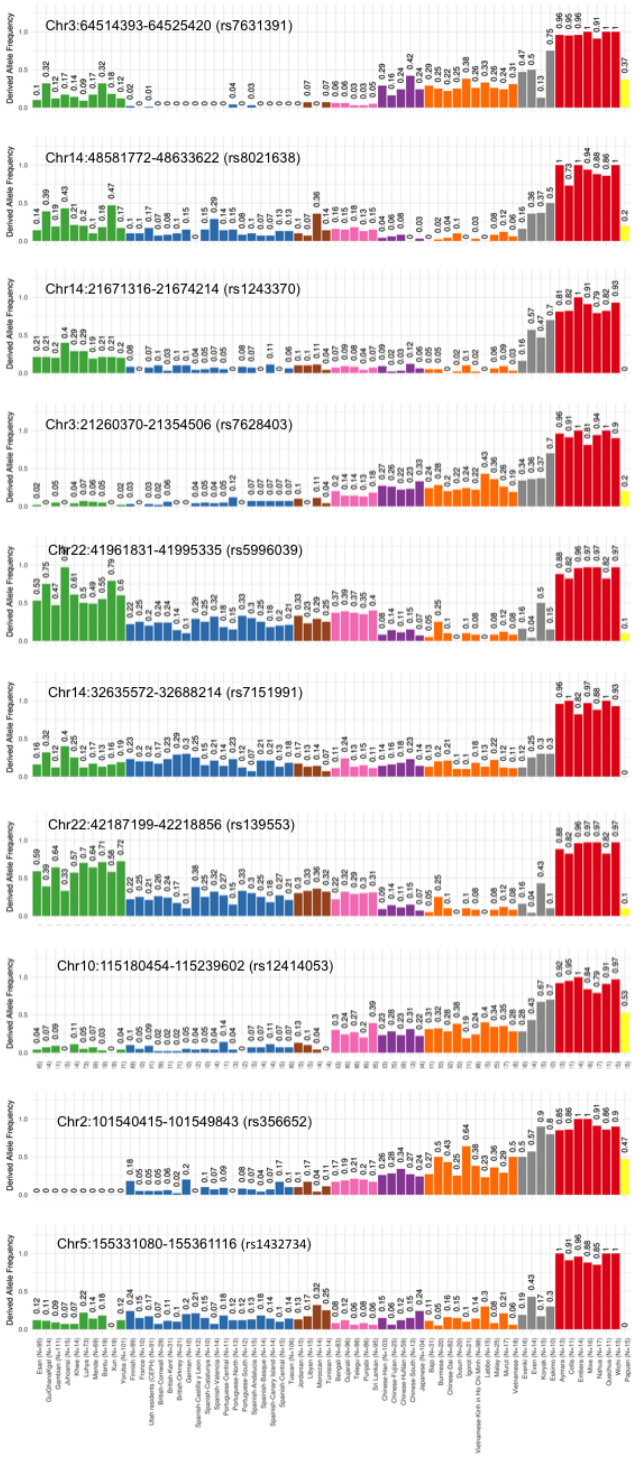

Figure A.7: Worldwide allele frequencies of the top PBS SNPs detected in Native Americans. The allele frequencies are estimated from 2,391 unrelated individuals collated from several public databases. The colors of the bars reflect the geographic origin of the populations for which the allele frequencies were estimated: Africa (green), Europe (blue), Middle East and North Africa (brown), South Asia (pink), East Asia (purple), South East Asia (orange), Siberian(grey), America (red) and Oceania (yellow). The number of individuals in each population (N) is given next to the population name and the derived allele frequency is shown at the top of each bar.

Appendix B

Detecting signals of selection post-admixture in Latin Americans

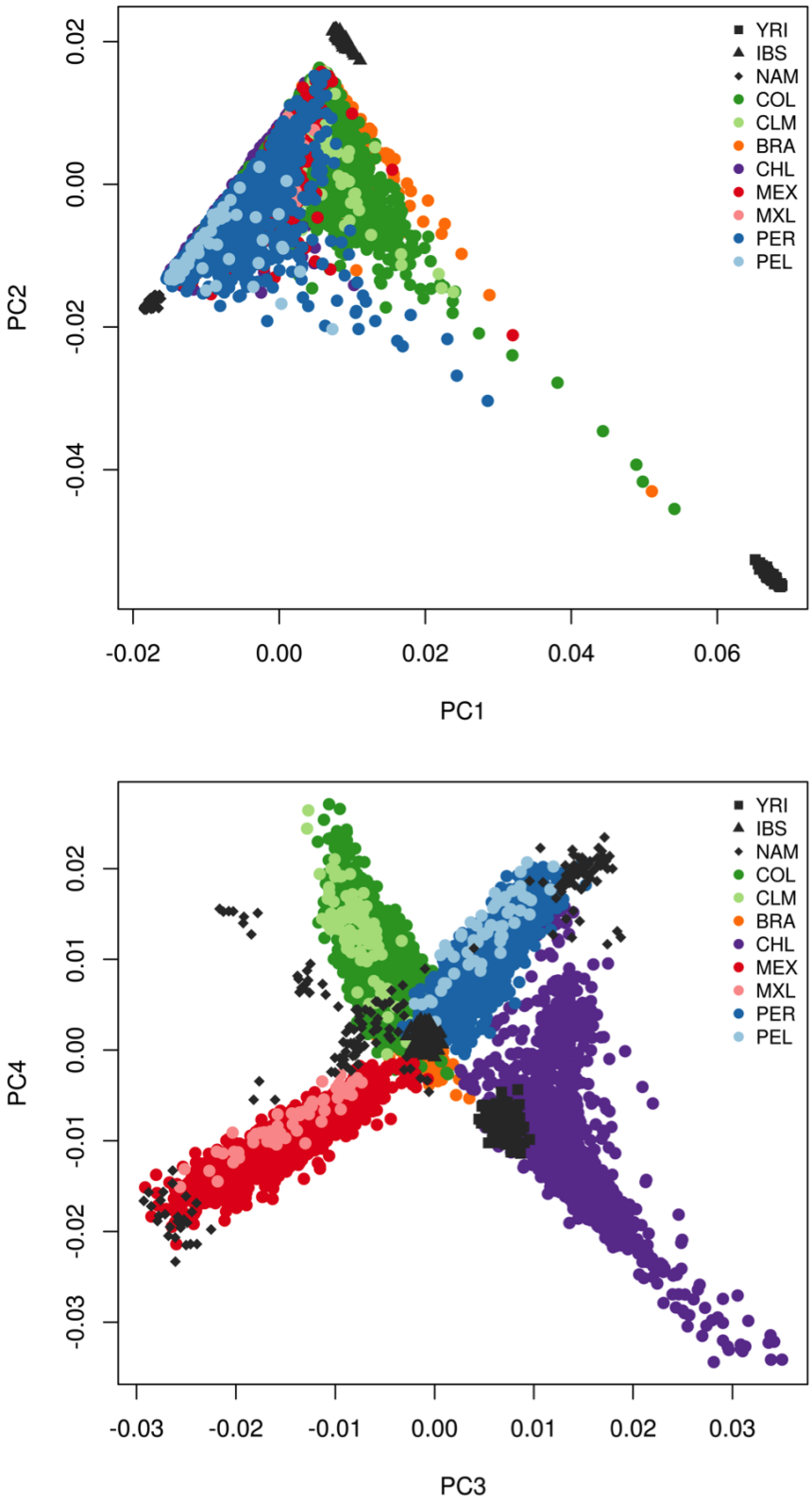

Figure B.1: Principal Component Analysis (PCA) of admixed Latin American individuals and continental reference panels. Each individual is represented as a point colored by country of origin. Abbreviations: NAM, Native Americans; IBS, Spanish Southern Europeans; YRI, West Africans; BRA, Brazilians from CANDELA; CHL, Chileans from CANDELA; COL, Colombians from CANDELA; CLM, Colombians from 1000 Genomes Project; MEX, Mexicans from CANDELA; MXL, Mexicans from 1000 Genomes Project; PER, Peruvians from CANDELA; PEL, Peruvians from 1000 Genomes Project.

Appendix C

Genetic determinants of pigmentation in Latin Americans

Table C.1: Features of the CANDELA sample.

Total	**Total**	**Colombia**	**Brazil**	**Chile**	**Mexico**	**Peru**
Sample size	6357	1507	651	1745	1207	1247
Percentage	100	23.7	10.2	27.5	19	19.6
% Female	54	55.9	68.5	39.6	60.3	58.4
Age (years)						
Min	18	18	18	18	18	18
Mean	24.2	24	25.8	25.2	24.4	22.2
Max	45	40	45	45	44	44
S.D.	5.7	5.3	6.3	5.8	5.6	5.2
Age, for Males (years)						
Min	18	18	18	18	18	18
Mean	24.9	24.7	25.8	25.3	25.1	23
Max	45	40	45	45	44	44
S.D.	5.7	5.5	6.4	5.5	5.6	5.7
Age, for Females (years)						
Min	18	18	18	18	18	18
Mean	23.8	23.5	25.4	25.2	24	21.6
Max	45	40	44	45	41	42
S.D.	5.7	5	4.2	6.2	4.7	4.7

Table C.2: Inflation factor and Tail statistic for pigmentation phenotypes.

Trait	**TS**	**Lambda**
Skin pigmentation (MI)	0.08	1.11
Hair color (categorical)	0.04	1.05
Eye color (categorical)	0.05	1.07
L (Brightness)	0.05	1.07
C (Saturation)	0.03	1.04
cosH (Hue)	0.05	1.05

Table C.3: Proportion of trait variation explained by each genome-wide associated SNP.

			Skin	Hair	Eye			
Region	**Candidate gene**	**SNP**	**MI**	**Categorical**	**Categorical**	**L (Brightness)**	**C (Saturation)**	**cos(H) (Hue)**
1q32	*DSTYK*	rs3795556	0.02	0	0	0.1	0.45	0.02
5p13	*SLC45A2*	rs16891982	6.07	3.78	0.87	0.89	0.4	0.24
6p25	*IRF4*	rs12203592	0.44	0.71	0.64	0.73	0.17	0.07
9p23	*TYRP1*	rs10809826	0.13	0.06	0.54	0.85	0.48	0.11
10q26	*EMX2*	rs11198112	0.48	0	0.01	0.01	0	0.01
11q14	*GRM5*	rs7118677	0.45	0.3	0	0	0.01	0.01
11q14	*TYR*	rs1042602	0.46	0.3	0	0.01	0.07	0
11q14	*TYR*	rs1126809	0.44	0.28	0.2	0.28	0.05	0.19
15q13	*OCA2*	rs4778219	0	0	0.05	0.04	0	0.03
15q13	*OCA2*	rs1800407	0.41	0.05	0.09	0.04	0.44	0.35
15q13	*OCA2*	rs1800404	0.52	0.1	0.61	1.01	0.38	0.07
15q13	*HERC2*	rs12913832	0.9	5.95	25.64	26.74	0.4	6.35
15q13	*HERC2*	rs4778249	0.27	0.14	0.56	1.15	0.98	0.01
15q21	*SLC24A5*	rs1426654	6.57	1.01	1.51	2.8	3.19	0.01
16q24	*MC1R*	rs885479	0.33	0.05	0	0	0	0
19p13	*MFSD12*	rs2240751	0.53	0	0.01	0	0.04	0
20q13	*WFDC5*	rs17422688	0.01	0	0	0.02	0	0.52
22q12	*MPST*	rs5756492	0.1	0	0.07	0.09	0.45	0.03

Table C.4: Unconditional GWAS P-values for each genome-wide associated SNPs.

			Skin	Hair	Eye			
Region	**Candidate gene**	**SNP**	**MI**	**Categorical**	**Categorical**	**L (Brightness)**	**C (Saturation)**	**cos(H) (Hue)**
1q32	DSTYK	rs3795556	0.7	0	0.2	2.2	7	0.6
5p13	SLC45A2	rs16891982	116.9	65.2	14.9	16.4	7.8	3.7
6p25	IRF4	rs12203592	9.5	12.7	11.9	13.5	2.9	1.3
9p23	TYRP1	rs10809826	3	1.5	10	15.3	7.7	1.9
10q26	EMX2	rs11198112	9.4	0.2	0.4	0.3	0.1	0.3
11q14	GRM5	rs7118677	8.9	5.5	0.2	0.1	0.3	0.3
11q14	TYR	rs1042602	9	5.6	0.1	0.4	1.4	0.1
11q14	*TYR*	rs1126809	8.6	5.2	3.9	5.3	1.1	3.1
15q13	*OCA2*	rs4778219	0.1	0.1	1.3	1	0.2	0.7
15q13	*OCA2*	rs1800407	8.2	1.3	2	1.1	6.9	5.3
15q13	*OCA2*	rs1800404	10.3	2.2	10.9	18.3	9.5	1.4
15q13	*HERC2*	rs12913832	17	104.1	200	200	6.2	91.9
15q13	*HERC2*	rs4778249	5.6	2.9	9.9	19.6	14.4	0.3
15q21	*SLC24A5*	rs1426654	129.8	18	26	49.1	44.2	0.4
16q24	*MC1R*	rs885479	6.7	1.3	0.3	0	0.1	0
19p13	*MFSD12*	rs2240751	10.3	0.1	0.5	0	0.9	0
20q13	*WFDC5*	rs17422688	0.3	0.2	0.1	0.7	0	7.5
22q12	*MPST*	rs5756492	2.3	0	1.6	2	6.9	0.8

Table C.5: Allele frequencies of the genome-wide associated SNPs in worldwide populations. CEU, IBS, CHB and YRI allele frequencies were retrieved from the 1000 Genomes Project. Allele frequencies for Native Americans are based on a subset of individuals from Ruiz-Linares et al. (2014).

Region	SNP	Anc/Der	CEU	IBS	CHB	YRI	NAM	CAN
1q32	rs3795556	T/C	0.23	0.26	0.41	0.36	0.42	0.33
5p13	rs16891982	C/G	0.98	0.82	0.01	0	0.03	0.49
6p25	rs12203592	C/T	0.16	0.13	0	0	0	0.07
9p23	rs10809826	C/G	0.62	0.56	0.01	0.06	0.03	0.29
10q26	rs11198112	C/T	0.17	0.14	0.14	0.17	0.19	0.16
11q14	rs7118677	G/T	0.67	0.74	0.33	0.21	0.05	0.4
11q14	rs1042602	C/A	0.4	0.39	0	0	0.01	0.25
11q14	rs1126809	G/A	0.25	0.29	0	0	0	0.11
15q13	rs4778219	T/C	0.88	0.8	0.11	0.84	0.64	0.75
15q13	rs1800407	C/T	0.08	0.1	0	0	0	0.05
15q13	rs1800404	C/T	0.82	0.73	0.39	0.07	0.31	0.58
15q13	rs12913832	A/G	0.77	0.32	0	0	0.03	0.23
15q13	rs4778249	T/A	0.98	0.98	0.98	0.28	1	0.97
15q21	rs1426654	G/A	1	1	0.03	0.01	0.02	0.56
16q24	rs885479	G/A	0.08	0.02	0.64	0	0.69	0.34
19p13	rs2240751	A/G	0.01	0	0.4	0	0.3	0.19
20q13	rs17422688	G/A	0.15	0.2	0.01	0	0	0.08
22q12	rs5756492	G/A	0.29	0.35	0.54	0.28	0.26	0.26

Abbreviations: Anc, Ancestral; Der, Derived. CAN, CANDELA sample.

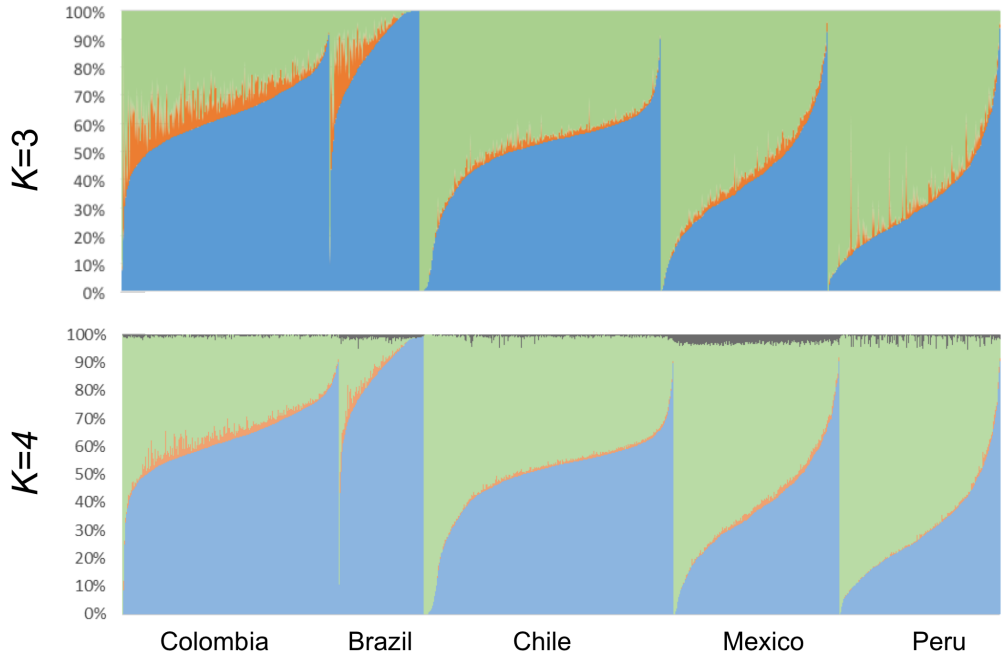

Figure C.1: Continental ancestry in the CANDELA sample. Ancestry values were estimated from a set of LD pruned SNPs via supervised ADMIXTURE analysis for $K = 3$ and $K = 4$. References populations from African, European, East Asian and Native American groups were chosen from the 1000 Genomes Project and selected Native Americans populations as described in Adhikari et al., 2016. Individual barplots for each country are shown. European, African, Native American, and East Asian ancestry is represented as light blue, orage, light green and grey colors. Individuals are sorted by European ancestry. From Adhikari & Mendoza-Revilla et al. (2018).

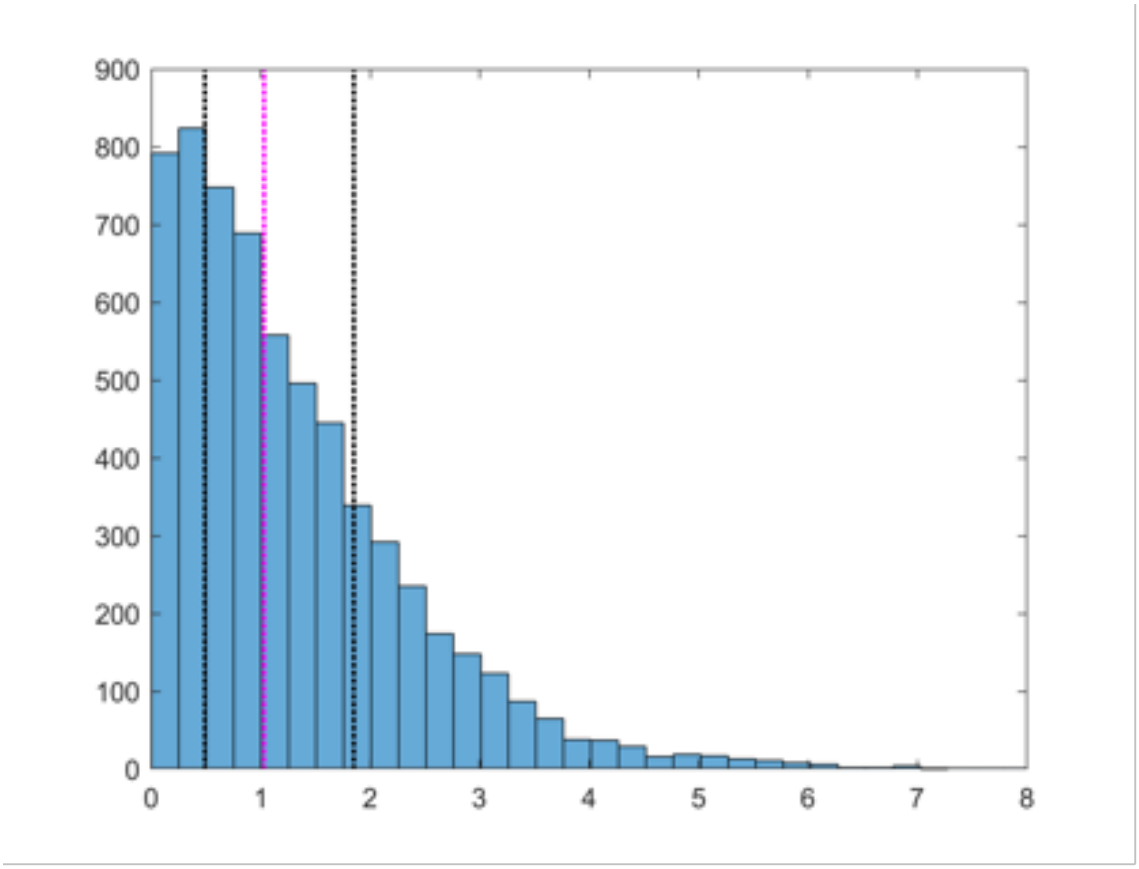

Figure C.2: Distribution of Melanin Index variability in the CANDELA sample. Measurements across the two arms were compared for each individual to assess the variability of the Melanin Index (MI). The absolute difference of the MI for an individual was taken as a measurement of variability. The distribution showed a maximum of 8 MI units. Median and quantiles are shown as a purple and black dotted lines. From Adhikari & Mendoza-Revilla et al. (2018).

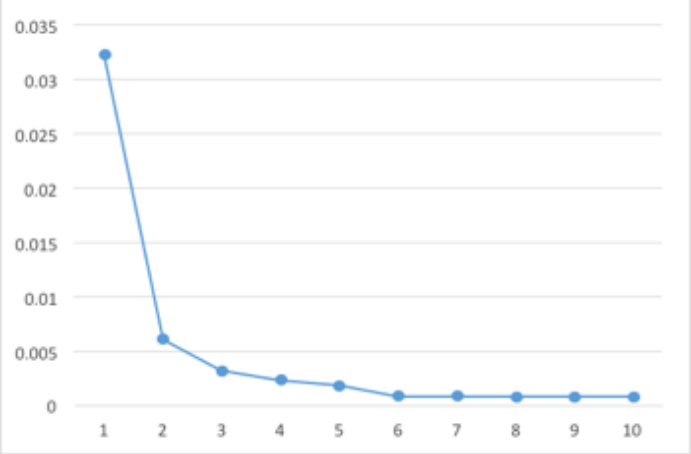

Figure C.3: Continental ancestry in the CANDELA sample Principal components were extracted from an LD-pruned SNP dataset. The proportion of variance explained by each PC is shown. From Adhikari & Mendoza-Revilla et al. (2018).

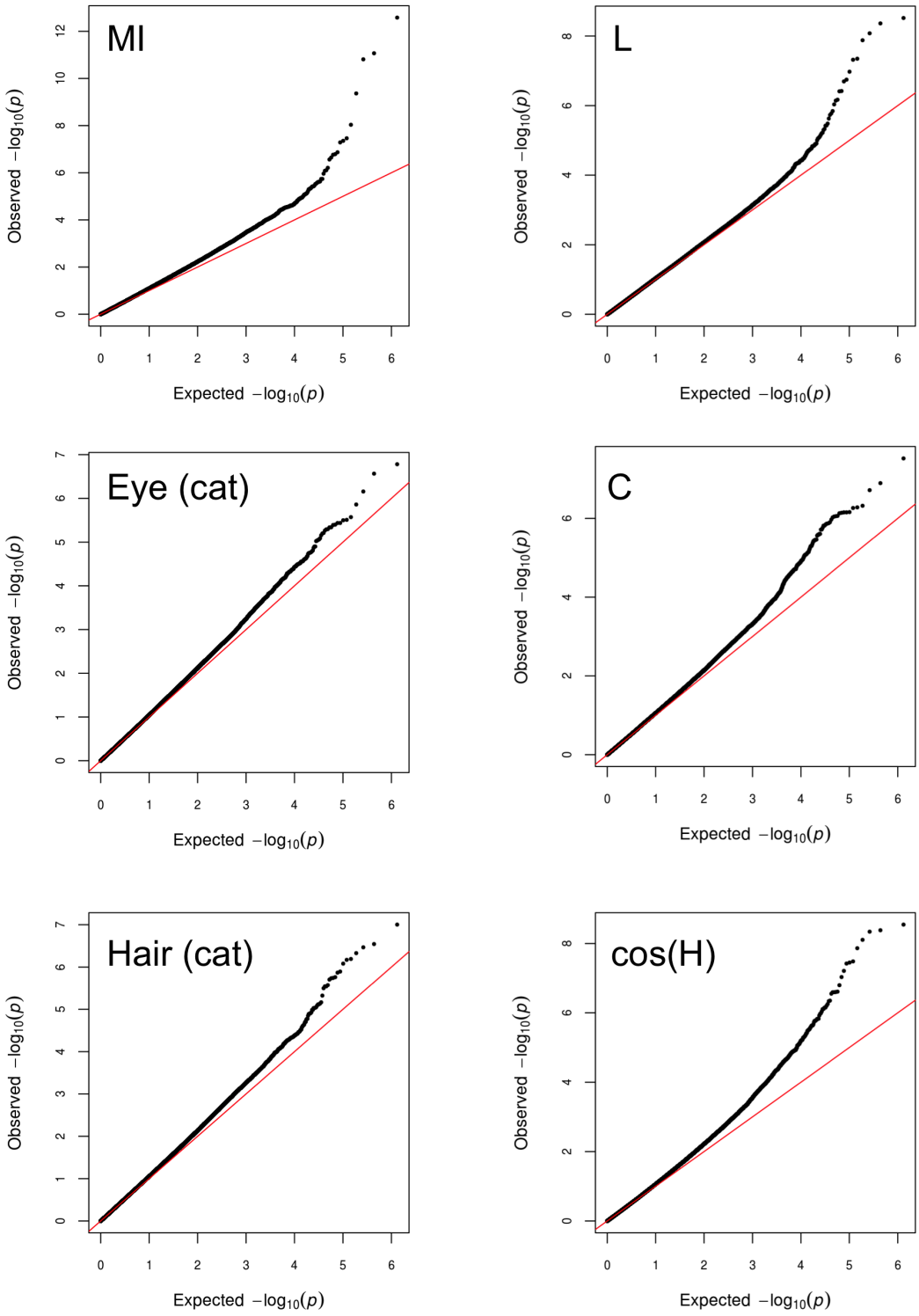

Figure C.4: GWAS quantile-quantile (QQ) plots of pigmentation phenotypes. From Adhikari & Mendoza-Revilla et al. (2018).

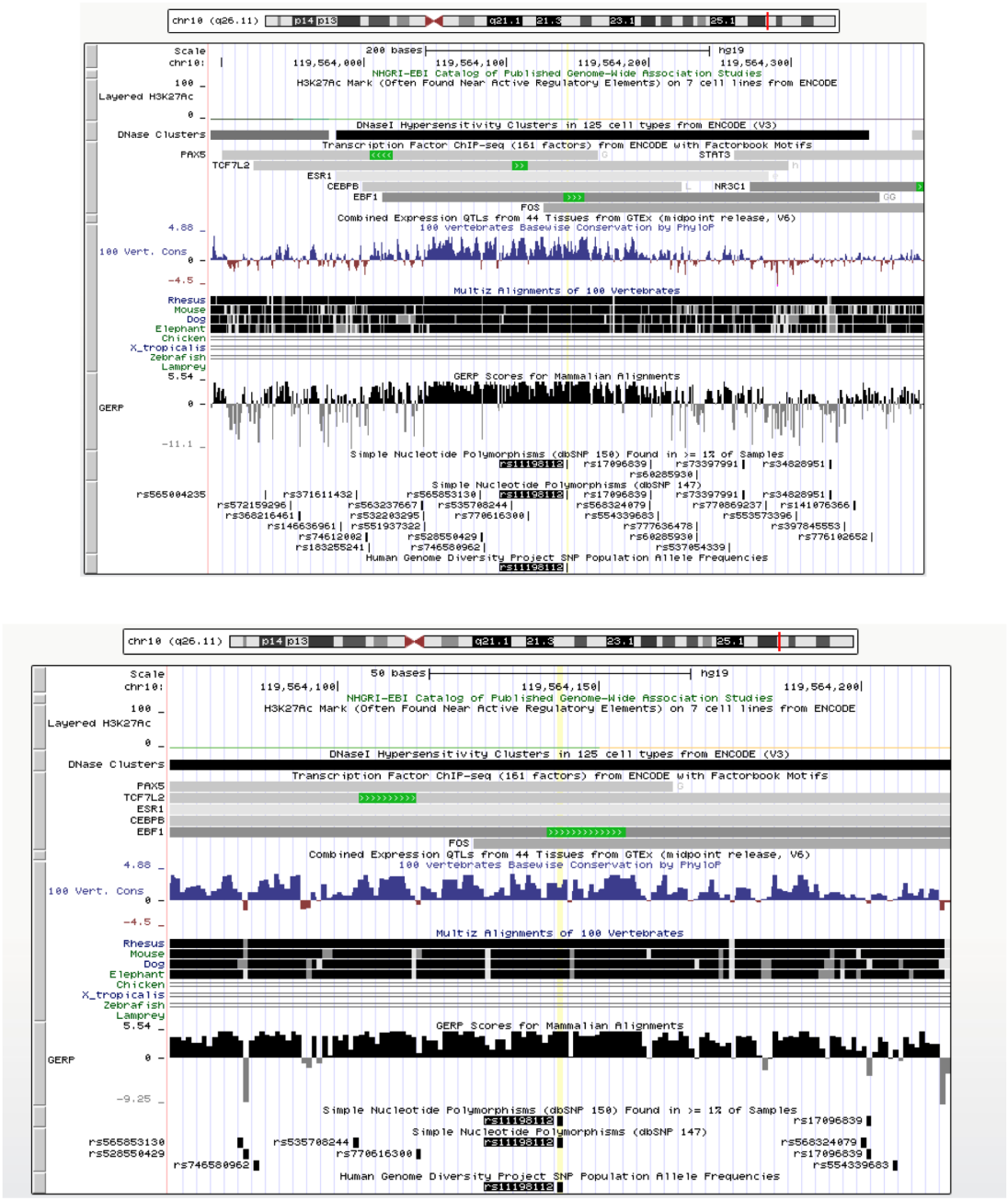

Figure C.5: Genomic annotation in the 10q26 intergenic region around SNP rs11198112. A UCSC genome-browser screenshot for the 10q26 genomic region around SNP rs11198112. SNP rs11198112 is included in the binding site for the Early B-cell factor (EBF1) transcription factor. Additionally, this region shows enriched PhyloP and GERP conversvation scores. Full view (upper panel) and zoomed in view (lower panel) are shown below. From Adhikari & Mendoza-Revilla et al. (2018).

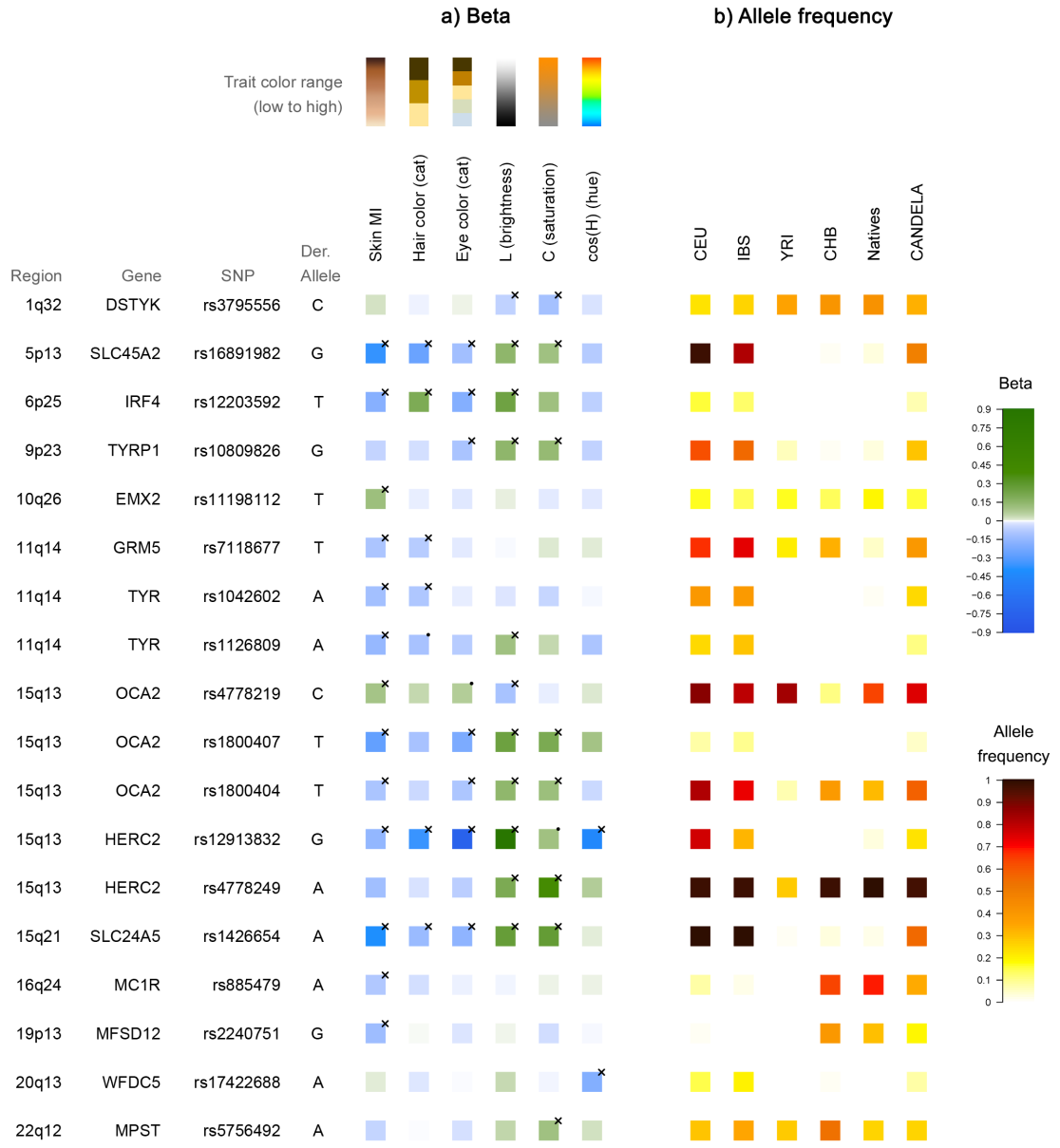

Figure C.6: Phenotypic effects (regression beta coefficients) and derived allele frequencies for the associated SNPs to pigmentation phenotypes in the CANDELA sample. a) Pigmentation phenotypes are shown at the top, with illustrative color ranges. Beta coefficients have been standardized to facilitate comparison across traits. Positive betas are shown in green and negative betas in blue (with color intensity reflecting beta values as indicated on the scale to the right). Significant betas are marked with a cross (×). In b) allele frequencies are shown for the CEU, IBS CHB and YRI samples from the 1KG, the CANDELA samples and selected Native Americans as described in the text. On the right is shown the color scale used to represent allele frequencies. From Adhikari & Mendoza-Revilla et al. (2018).

Appendix D

Exploring the convergent evolution of lighter skin pigmentation in Eurasia

Table D.1: Worldwide populations included in the correlation analysis between allele frequency at skin pigmentation associated loci and solar radiation.

Population	Country of origin	Major geographical region	Sample size	Longitude	Latitude	Solar radiation	Source (reference)
ESN	Nigeria	Africa	95	5.61	6.33	4.81	1000 Genomes Project
GuiGhanaKgal	Botswana	Africa	14	26.00	-22.00	5.85	Schlebusch et al. 2012
GWD	Gambia	Africa	111	-16.32	13.24	5.56	1000 Genomes Project
Juhoansi	Namibia	Africa	15	18.00	-18.00	6.03	Schlebusch et al. 2012
Khwe	Namibia	Africa	14	19.00	-13.00	5.58	Schlebusch et al. 2012
LWK	Kenya	Africa	73	34.77	0.62	5.91	1000 Genomes Project
MSL	Sierra Leone	Africa	69	-10.69	8.21	5.14	1000 Genomes Project
Bantu	South Africa	Africa	19	27.00	-22.00	5.8	Schlebusch et al. 2012
Xun	Angola	Africa	19	15.00	-13.00	5.65	Schlebusch et al. 2012
YRI	Nigeria	Africa	101	3.94	7.38	4.89	1000 Genomes Project
Aymara	Bolivia	Americas	13	-68.20	-16.50	5.3	Chacon-Duque et al. 2018
Colla	Argentina	Americas	11	-66.32	-24.23	6.2	Eichstaedt et al. 2014
Embera	Colombia	Americas	14	-76.00	7.00	4.31	Chacon-Duque et al. 2018
Mixe	Mexico	Americas	16	-96.58	16.95	5.26	Chacon-Duque et al. 2018
Nahua	Mexico	Americas	17	-99.08	17.63	6.05	Chacon-Duque et al. 2018
Quechua	Peru/Bolivia	Americas	11	-72.00	-13.50	5.34	Chacon-Duque et al. 2018
Wichi	Argentina	Americas	15	-64.10	-23.22	4.6	Eichstaedt et al. 2014
CHB	China	Asia (Eastern Eurasia)	103	116.40	39.91	4.4	1000 Genomes Project
CHS-FuJian	China	Asia (Eastern Eurasia)	25	118.12	25.49	3.63	1000 Genomes Project
CHS-HuNan	China	Asia (Eastern Eurasia)	59	111.65	27.54	3.29	1000 Genomes Project
CHS	China	Asia (Eastern Eurasia)	13	114.00	32.30	3.83	1000 Genomes Project
JPT	Japan	Asia (Eastern Eurasia)	104	138.09	35.81	3.63	1000 Genomes Project
Papuan	Papua New Guinea	Asia (Eastern Eurasia)	15	143.00	-4.00	4.91	Simons Genome Diversity Project
BEB	Bangladesh	Asia (Eastern Eurasia)	83	90.41	23.81	4.65	1000 Genomes Project
GIH	India	Asia (Eastern Eurasia)	96	71.00	23.16	5.16	1000 Genomes Project
ITU	India	Asia (Eastern Eurasia)	98	79.14	17.88	5.17	1000 Genomes Project
PJL	Pakistan	Asia (Eastern Eurasia)	86	74.35	31.56	5.33	1000 Genomes Project
STU	Sri Lanka	Asia (Eastern Eurasia)	95	80.77	7.59	5.34	1000 Genomes Project
Bajo	Indonesia	Asia (Eastern Eurasia)	31	122.52	-4.00	5.5	Moerseburg et al. 2016
Burmese	Myanmar	Asia (Eastern Eurasia)	20	96.60	21.63	4.78	Moerseburg et al. 2016
CDX	China	Asia (Eastern Eurasia)	82	100.82	22.02	4.61	1000 Genomes Project

. . . continued on next page. . .

. . . continued from previous page . . .

Population	Country of origin	Major geographical region	Sample size	Longitude	Latitude	Solar radiation	Source (reference)
Dusun	Brunei	Asia (Eastern Eurasia)	20	114.68	4.64	5.65	Moerseburg et al. 2016
Igorot	Phillippines	Asia (Eastern Eurasia)	21	121.27	16.57	4.67	Moerseburg et al. 2016
KHV	Vietnam	Asia (Eastern Eurasia)	98	106.65	10.81	5.14	1000 Genomes Project
Lebbo	Indonesia	Asia (Eastern Eurasia)	15	117.28	1.45	4.8	Moerseburg et al. 2016
Malay	Singapore	Asia (Eastern Eurasia)	25	103.86	1.35	4.49	Moerseburg et al. 2016
Murut	Brunei	Asia (Eastern Eurasia)	17	115.17	4.61	5.65	Moerseburg et al. 2016
Vietnamese	Vietnam	Asia (Eastern Eurasia)	18	108.41	14.48	4.56	Moerseburg et al. 2016
Evenki	Russia	Asia (Eastern Eurasia)	16	92.88	56.01	3.06	Carmona et al. 2014
Even	Russia	Asia (Eastern Eurasia)	14	151.29	59.58	2.82	Carmona et al. 2014
Koryak	Russia	Asia (Eastern Eurasia)	15	159.23	62.03	2.61	Carmona et al. 2014
Eskimo	Russia	Asia (Eastern Eurasia)	10	-173.05	64.68	2.34	Carmona et al. 2014
CEU	USA	Europe (Western Eurasia)	91	6.00	52.00	2.69	1000 Genomes Project
FIN	Finland	Europe (Western Eurasia)	99	24.94	60.17	2.73	1000 Genomes Project
France	France	Europe (Western Eurasia)	10	2.00	46.00	3.34	Chacon-Duque et al. 2018
Cornwall	UK	Europe (Western Eurasia)	29	-4.78	50.47	3.11	1000 Genomes Project
Kent	UK	Europe (Western Eurasia)	31	0.84	51.22	2.81	1000 Genomes Project
Orkney	UK	Europe (Western Eurasia)	21	-3.28	58.79	2.53	1000 Genomes Project
Germany	Germany	Europe (Western Eurasia)	10	10.64	51.11	2.71	Chacon-Duque et al. 2018
Catilla y Leon	Spain	Europe (Western Eurasia)	12	-4.73	41.65	4.06	1000 Genomes Project
Catalunya	Spain	Europe (Western Eurasia)	10	2.93	41.95	4.11	1000 Genomes Project
Valencia	Spain	Europe (Western Eurasia)	14	-0.50	39.47	4.98	1000 Genomes Project
Portugal (Central)	Portugal	Europe (Western Eurasia)	11	-8.19	39.57	4.32	Chacon-Duque et al. 2018
Portugal (North)	Portugal	Europe (Western Eurasia)	13	-8.59	41.15	4.35	Chacon-Duque et al. 2018
Portugal (South)	Portugal	Europe (Western Eurasia)	12	-7.92	37.02	4.88	Chacon-Duque et al. 2018
Andalucia	Spain	Europe (Western Eurasia)	15	-4.81	37.54	4.78	Chacon-Duque et al. 2018
Basque	Spain	Europe (Western Eurasia)	14	0.00	43.00	3.71	Chacon-Duque et al. 2018
Canary Island	Spain	Europe (Western Eurasia)	14	-16.31	28.49	5.4	Chacon-Duque et al. 2018
Spanish (Central)	Spain	Europe (Western Eurasia)	15	-3.69	40.40	4.4	Chacon-Duque et al. 2018
TSI	Italy	Europe (Western Eurasia)	106	11.06	43.50	3.91	1000 Genomes Project
Jordan	Jordania	Europe (Western Eurasia)	15	35.94	31.95	5.17	Chacon-Duque et al. 2018
Libya	Libya	Europe (Western Eurasia)	15	17.55	27.00	5.89	Chacon-Duque et al. 2018
Morocco	Morocco	Europe (Western Eurasia)	14	-8.01	31.63	5.24	Chacon-Duque et al. 2018
Tunisia	Tunisia	Europe (Western Eurasia)	14	10.18	36.81	4.95	Chacon-Duque et al. 2018

Table D.2: The estimated Predictive Error using the median point estimate based on a cross-validation sample of 100. The estimates were insensitive to difference tolerance rates.

Acceptance rate	Selection time (T)	Selection coefficient (s)
0.001	0.33	0.61
0.005	0.34	0.62
0.01	0.35	0.62
0.05	0.46	0.65

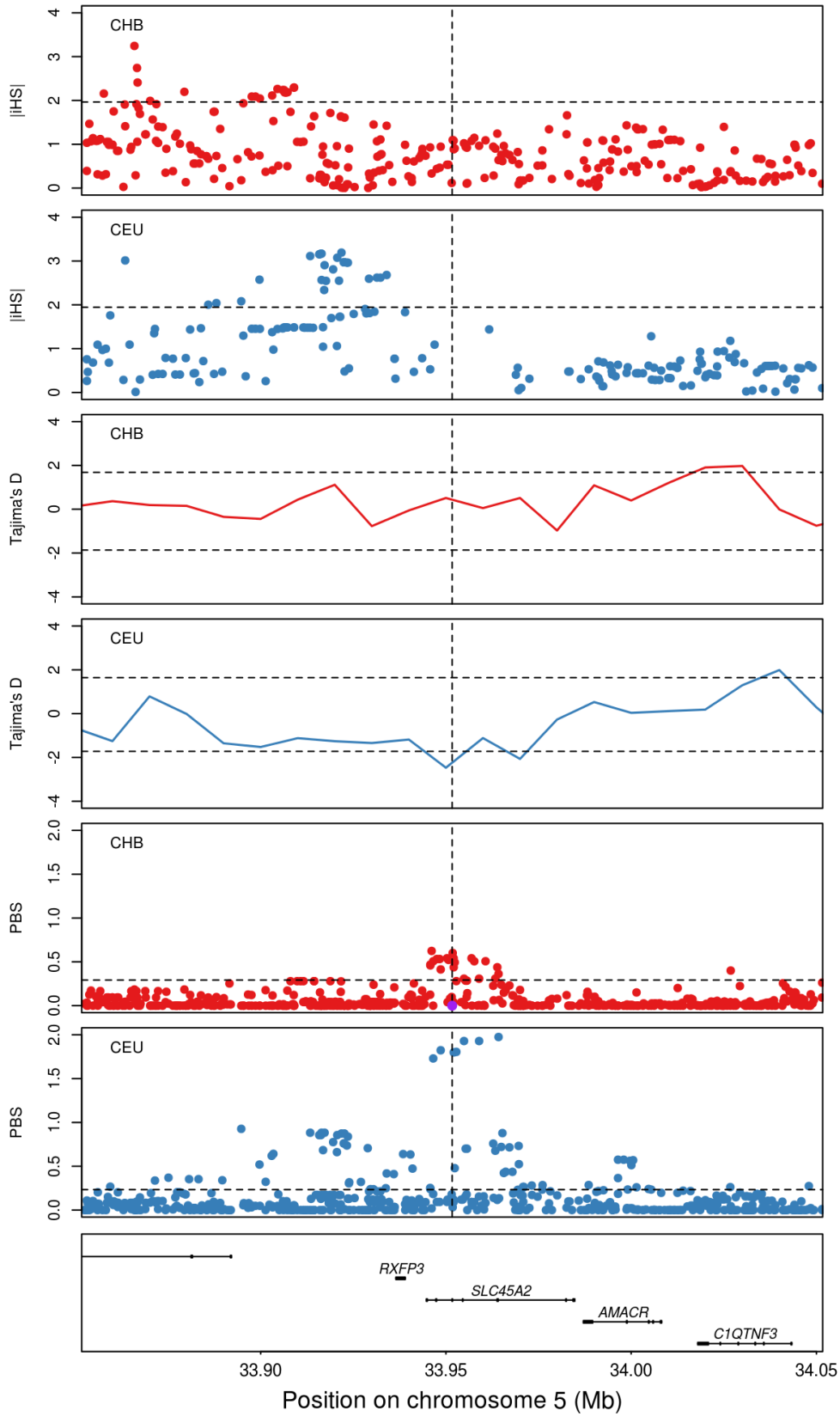

Figure D.1: Selection scans around candidate gene *SLC45A2* at SNP rs16891982 in Eurasian populations. iHS score distributions in CHB and CEU (first panel). Tajima's D score distribution in CHB and CEU (second panel). PBS score distribution in CHB and CEU (third panel). The 95th percentile threshold is shown with a horizontal dashed black line and additionally the 5th for Tajima's D. If present, the genome-wide associated SNP at the region is highlighted in purple and its position represented with a dashed vertical line. UCSC RefSeq genes and genomic coordinates (fourth panel). From Adhikari & Mendoza-Revilla et al. (2018).

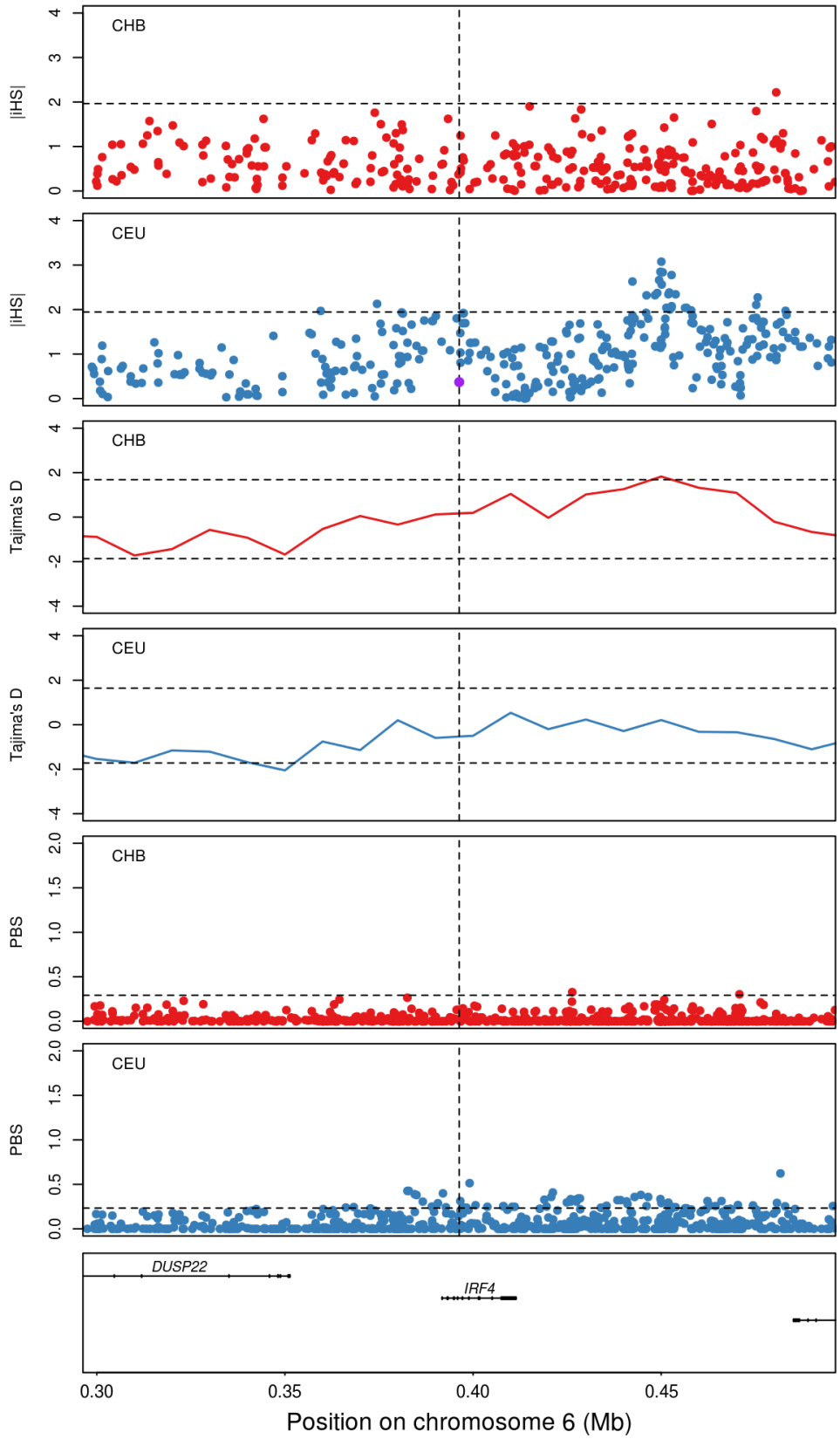

Figure D.2: Selection scans around candidate gene *IRF4* in Eurasian populations. iHS score distributions in CHB and CEU (first panel). Tajima's D score distribution in CHB and CEU (second panel). PBS score distribution in CHB and CEU (third panel). The 95th percentile threshold is shown with a horizontal dashed black line and additionally the 5th for Tajima's D. If present, the genome-wide associated SNP at the region is highlighted in purple and its position represented with a dashed vertical line. UCSC RefSeq genes and genomic coordinates (fourth panel). From Adhikari & Mendoza-Revilla et al. (2018).

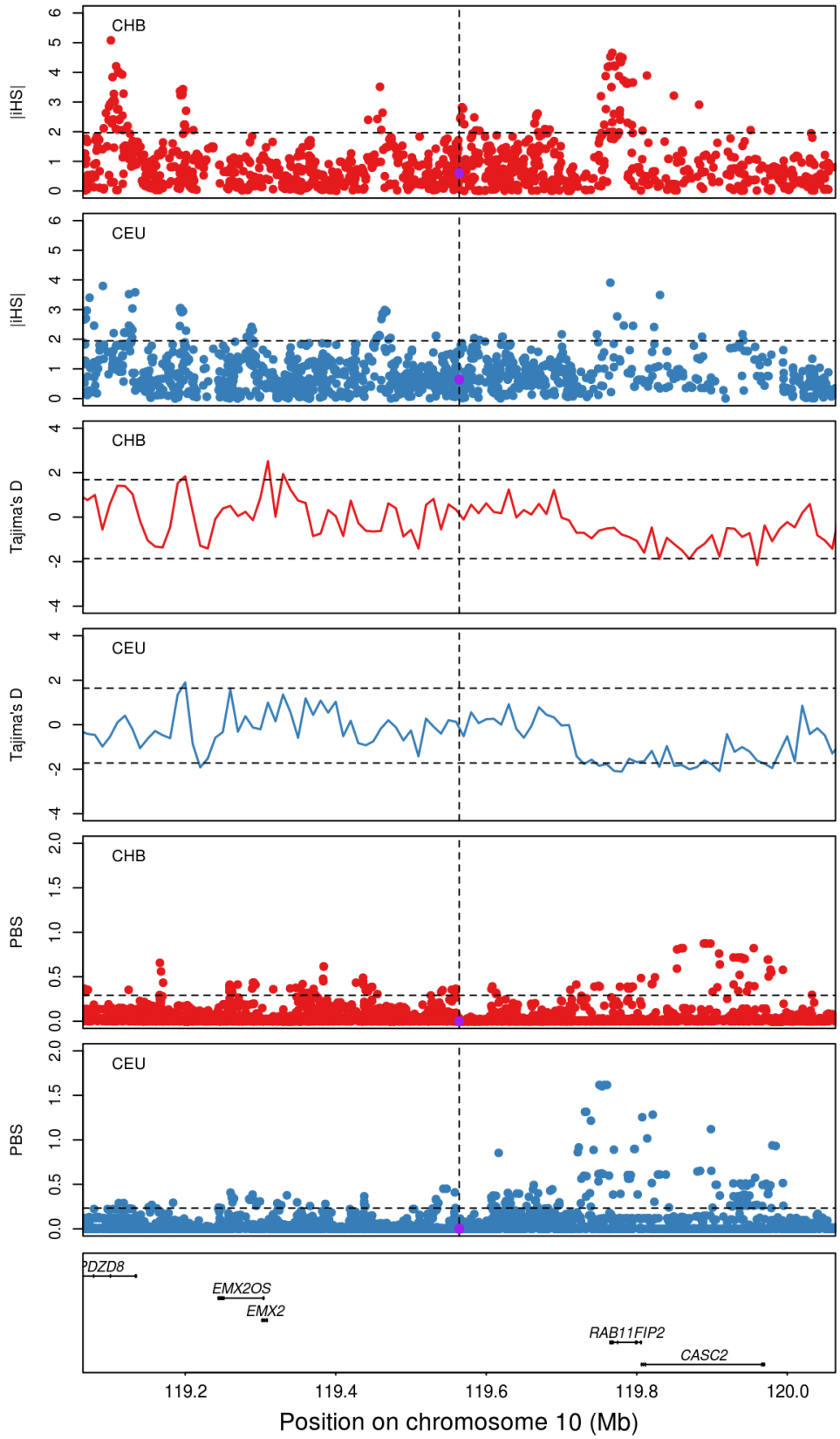

Figure D.3: Selection scans around candidate gene *EMX2* at SNP rs11198112 in Eurasian populations. iHS score distributions in CHB and CEU (first panel). Tajima's D score distribution in CHB and CEU (second panel). PBS score distribution in CHB and CEU (third panel). The 95th percentile threshold is shown with a horizontal dashed black line and additionally the 5th for Tajima's D. If present, the genome-wide associated SNP at the region is highlighted in purple and its position represented with a dashed vertical line. UCSC RefSeq genes and genomic coordinates (fourth panel). From Adhikari & Mendoza-Revilla et al. (2018).

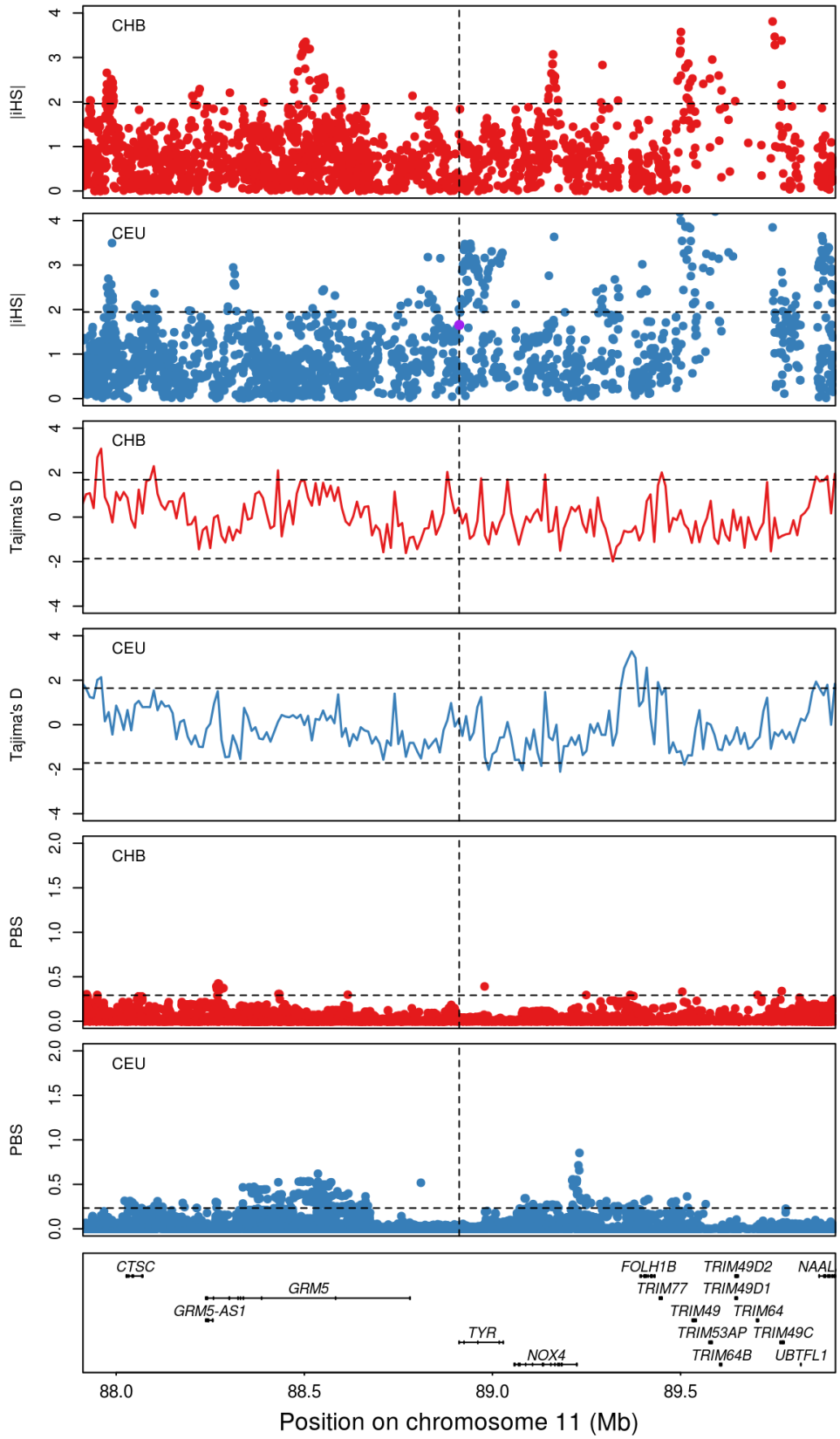

Figure D.4: Selection scans around candidate gene *TYR* at SNP rs1042602 in Eurasian populations. iHS score distributions in CHB and CEU (first panel). Tajima's D score distribution in CHB and CEU (second panel). PBS score distribution in CHB and CEU (third panel). The 95th percentile threshold is shown with a horizontal dashed black line and additionally the 5th for Tajima's D. If present, the genome-wide associated SNP at the region is highlighted in purple and its position represented with a dashed vertical line. UCSC RefSeq genes and genomic coordinates (fourth panel). From Adhikari & Mendoza-Revilla et al. (2018).

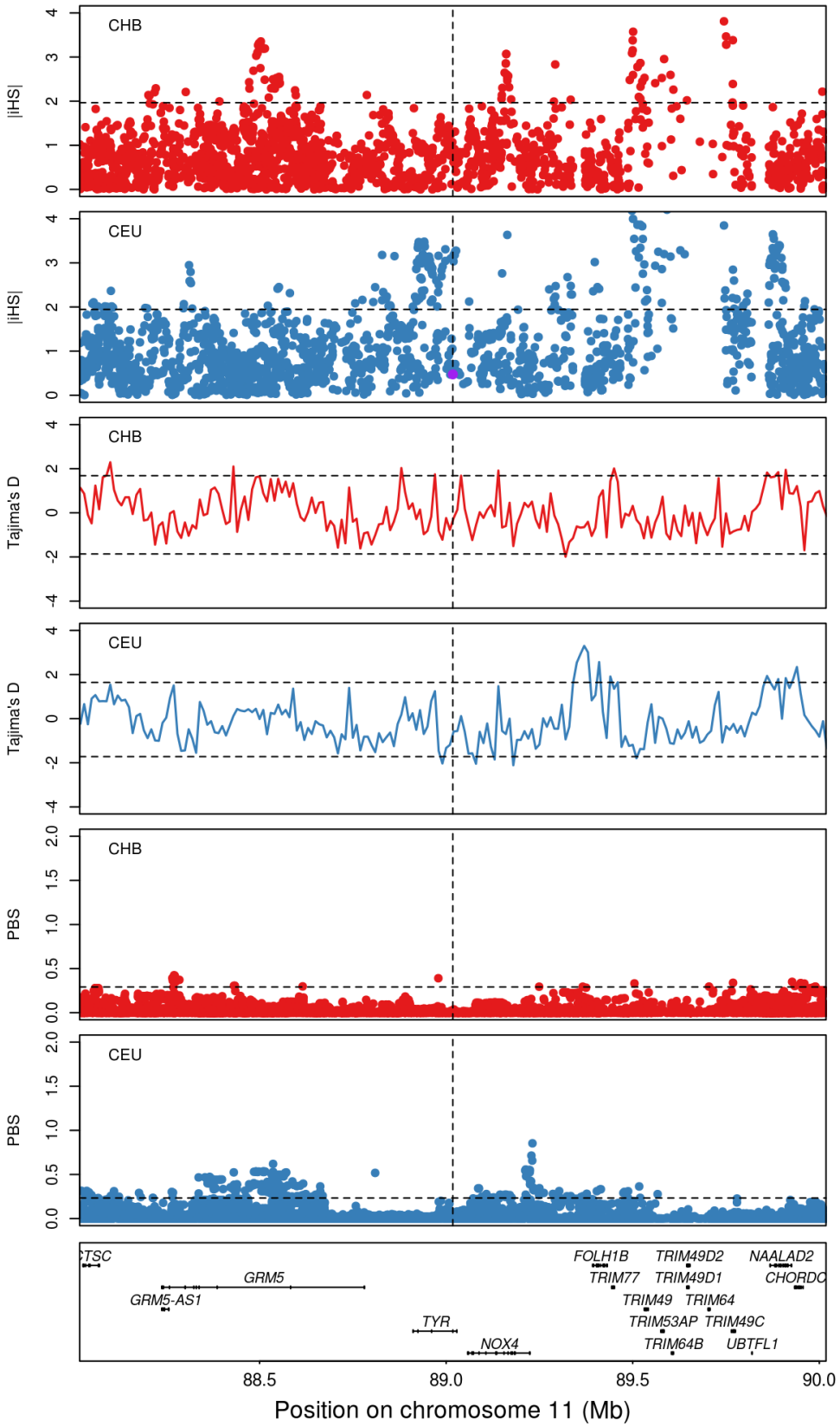

Figure D.5: Selection scans around candidate gene *TYR* at SNP rs1126809 in Eurasian populations. iHS score distributions in CHB and CEU (first panel). Tajima's D score distribution in CHB and CEU (second panel). PBS score distribution in CHB and CEU (third panel). The 95th percentile threshold is shown with a horizontal dashed black line and additionally the 5th for Tajima's D. If present, the genome-wide associated SNP at the region is highlighted in purple and its position represented with a dashed vertical line. UCSC RefSeq genes and genomic coordinates (fourth panel). From Adhikari & Mendoza-Revilla et al. (2018).

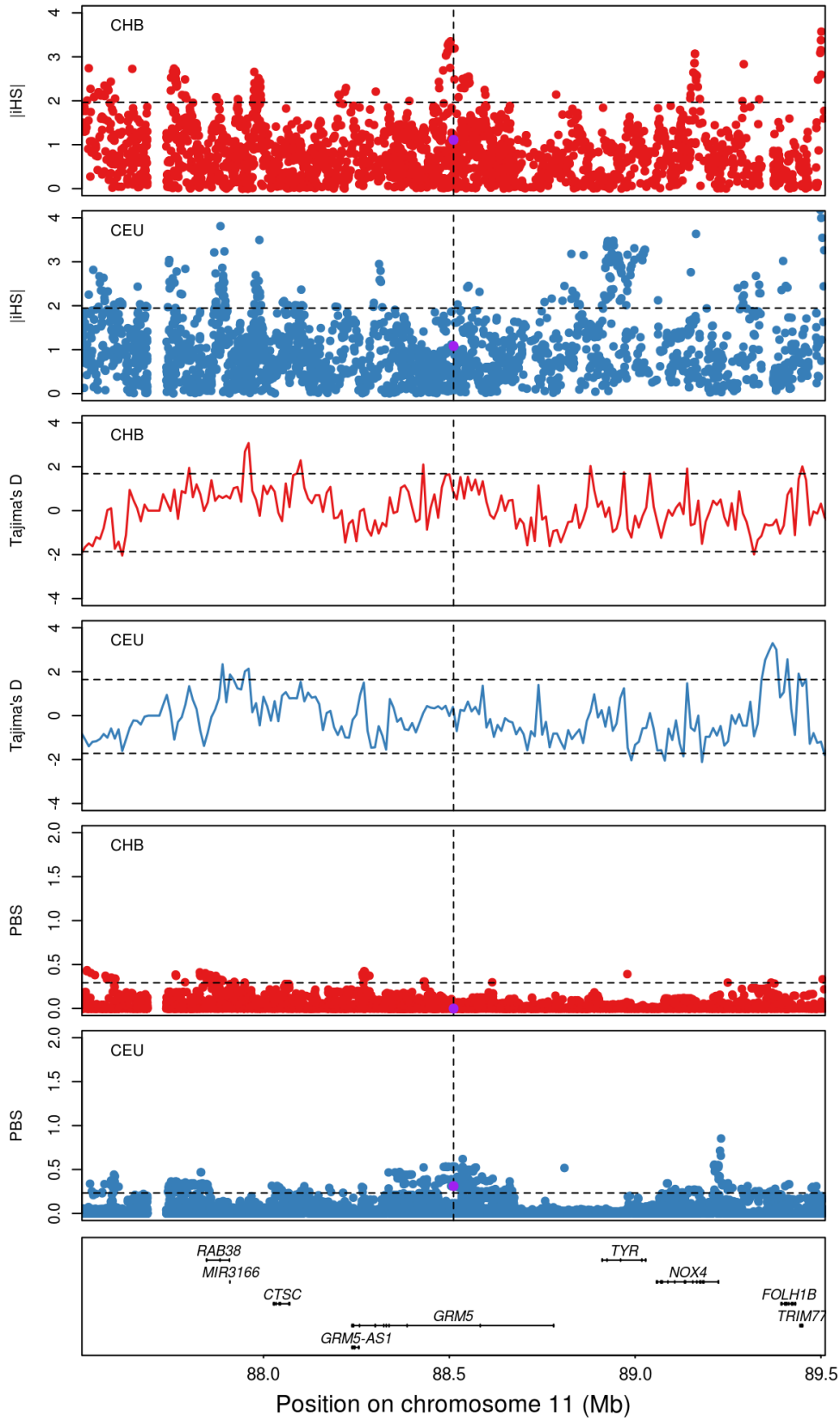

Figure D.6: Selection scans around candidate gene *GRM5* at SNP rs7118677 in Eurasian populations. iHS score distributions in CHB and CEU (first panel). Tajima's D score distribution in CHB and CEU (second panel). PBS score distribution in CHB and CEU (third panel). The 95th percentile threshold is shown with a horizontal dashed black line and additionally the 5th for Tajima's D. If present, the genome-wide associated SNP at the region is highlighted in purple and its position represented with a dashed vertical line. UCSC RefSeq genes and genomic coordinates (fourth panel). From Adhikari & Mendoza-Revilla et al. (2018).

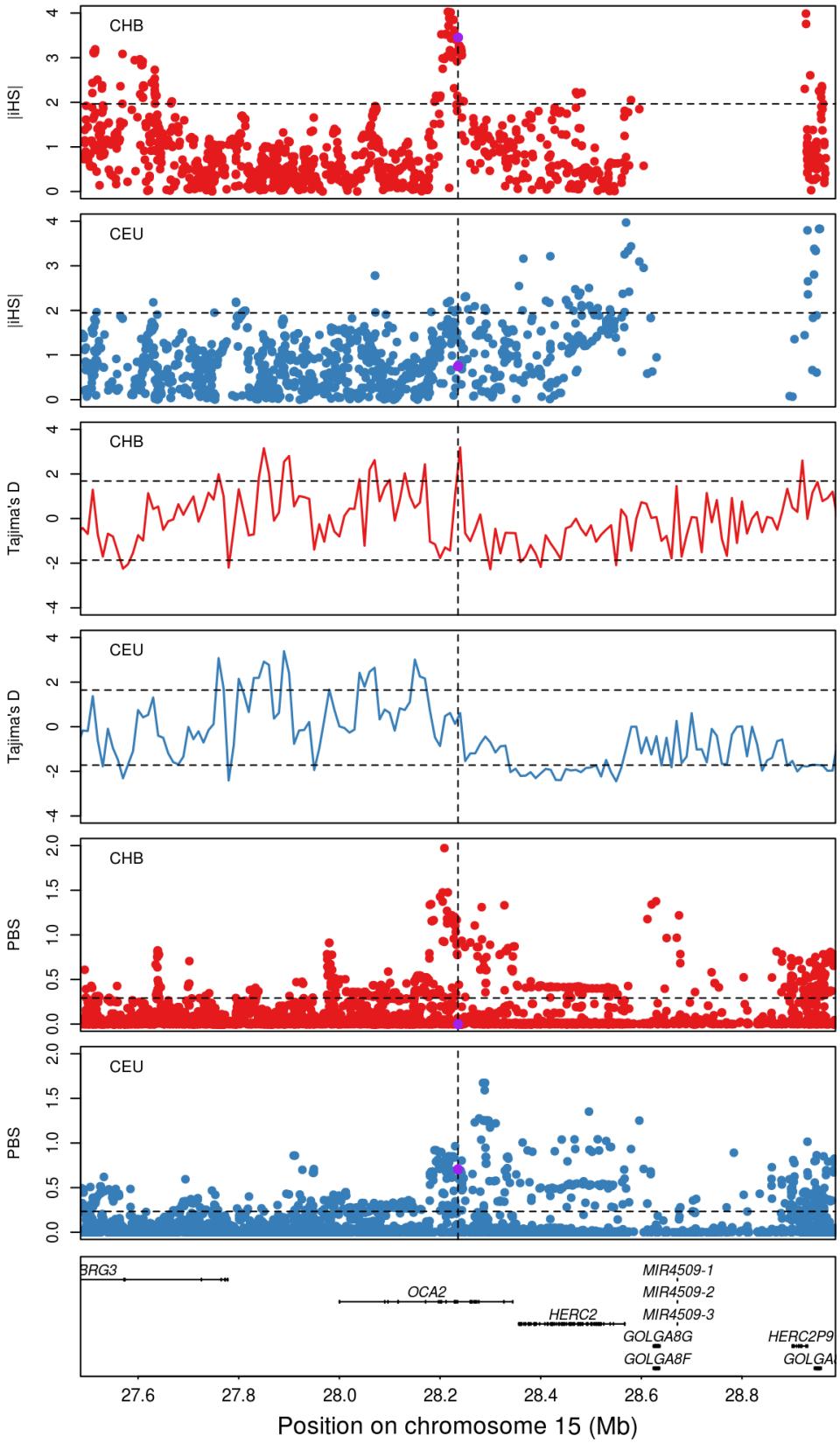

Figure D.7: Selection scans around candidate gene *OCA2* at SNP rs1800404 in Eurasian populations. iHS score distributions in CHB and CEU (first panel). Tajima's D score distribution in CHB and CEU (second panel). PBS score distribution in CHB and CEU (third panel). The 95th percentile threshold is shown with a horizontal dashed black line and additionally the 5th for Tajima's D. If present, the genome-wide associated SNP at the region is highlighted in purple and its position represented with a dashed vertical line. UCSC RefSeq genes and genomic coordinates (fourth panel). From Adhikari & Mendoza-Revilla et al. (2018).

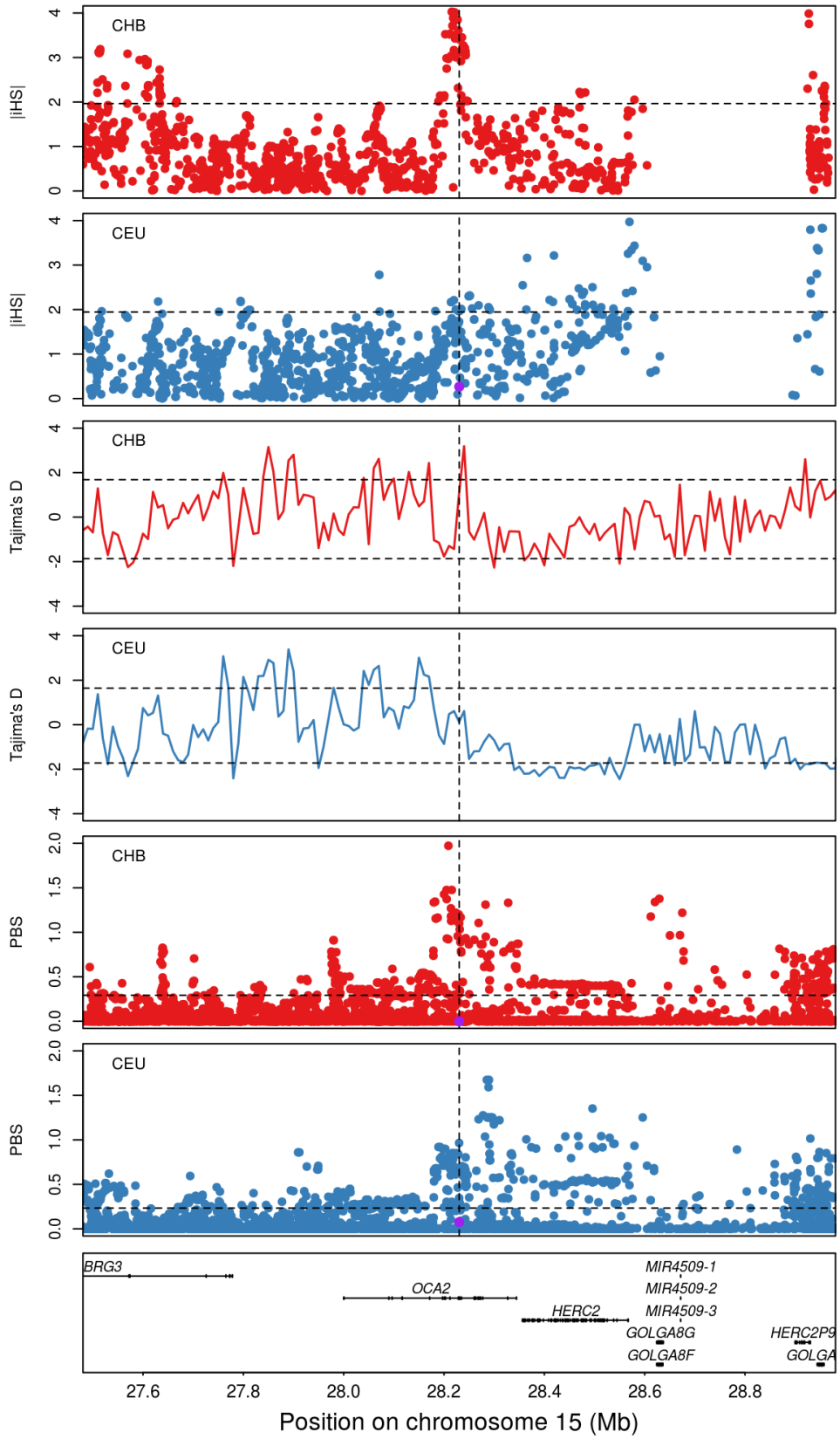

Figure D.8: Selection scans around candidate gene *OCA2* at SNP rs1800407 in Eurasian populations. iHS score distributions in CHB and CEU (first panel). Tajima's D score distribution in CHB and CEU (second panel). PBS score distribution in CHB and CEU (third panel). The 95th percentile threshold is shown with a horizontal dashed black line and additionally the 5th for Tajima's D. If present, the genome-wide associated SNP at the region is highlighted in purple and its position represented with a dashed vertical line. UCSC RefSeq genes and genomic coordinates (fourth panel). From Adhikari & Mendoza-Revilla et al. (2018).

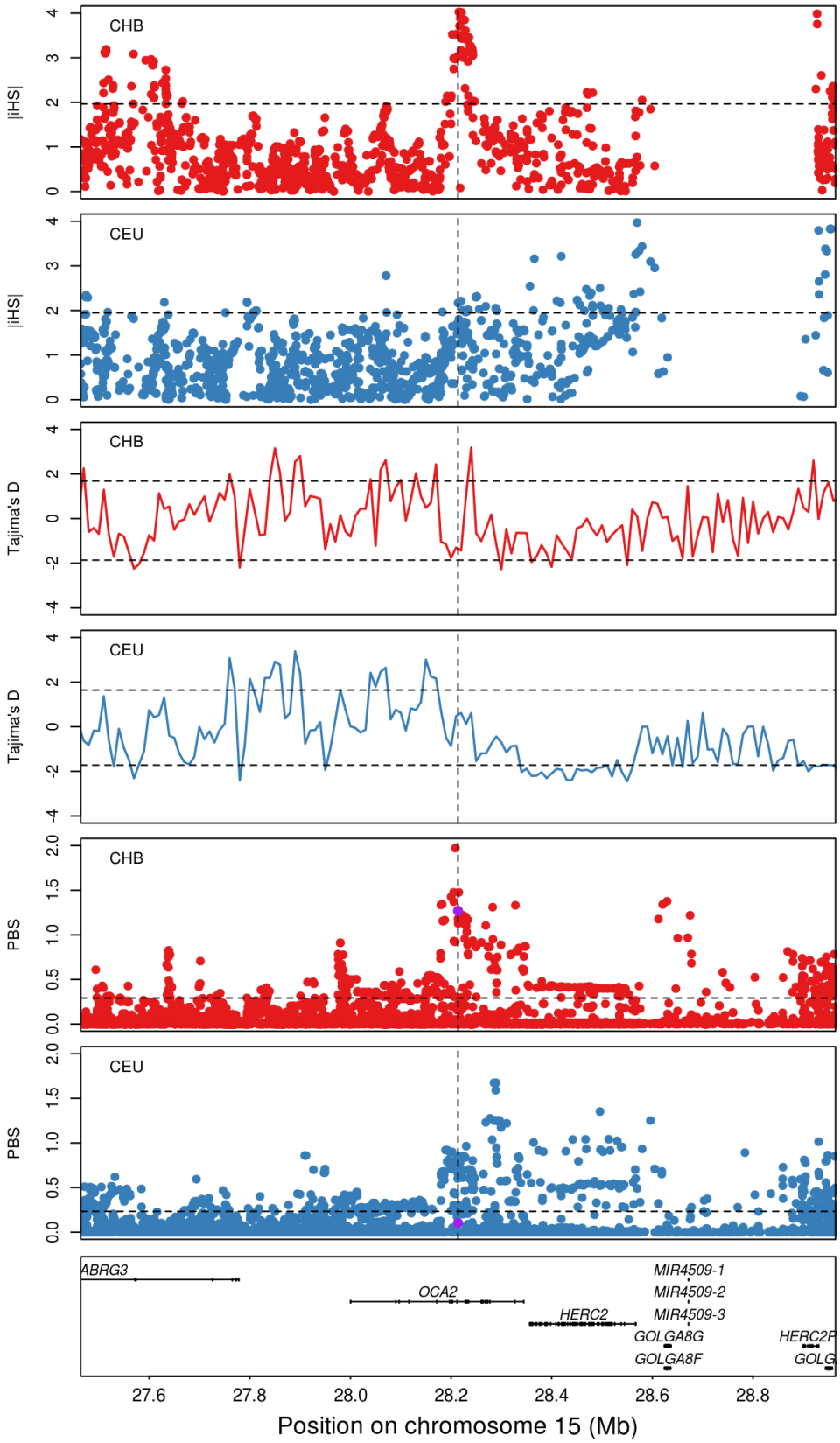

Figure D.9: Selection scans around candidate gene *OCA2* at SNP rs4778219 in Eurasian populations. iHS score distributions in CHB and CEU (first panel). Tajima's D score distribution in CHB and CEU (second panel). PBS score distribution in CHB and CEU (third panel). The 95th percentile threshold is shown with a horizontal dashed black line and additionally the 5th for Tajima's D. If present, the genome-wide associated SNP at the region is highlighted in purple and its position represented with a dashed vertical line. UCSC RefSeq genes and genomic coordinates (fourth panel). From Adhikari & Mendoza-Revilla et al. (2018).

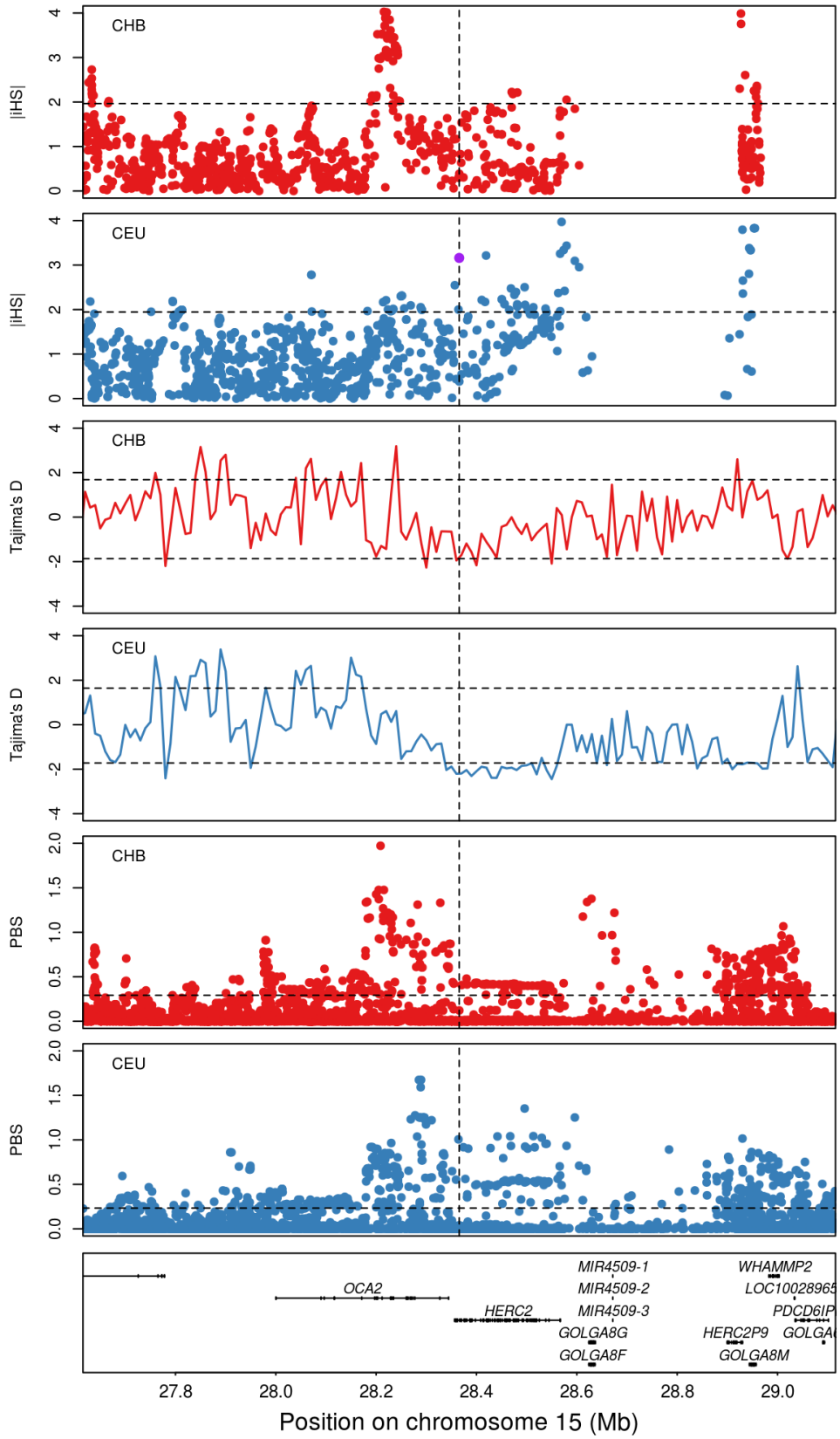

Figure D.10: Selection scans around candidate gene *HERC2* at SNP rs12913832 in Eurasian populations. iHS score distributions in CHB and CEU (first panel). Tajima's D score distribution in CHB and CEU (second panel). PBS score distribution in CHB and CEU (third panel). The 95th percentile threshold is shown with a horizontal dashed black line and additionally the 5th for Tajima's D. If present, the genome-wide associated SNP at the region is highlighted in purple and its position represented with a dashed vertical line. UCSC RefSeq genes and genomic coordinates (fourth panel). From Adhikari & Mendoza-Revilla et al. (2018).

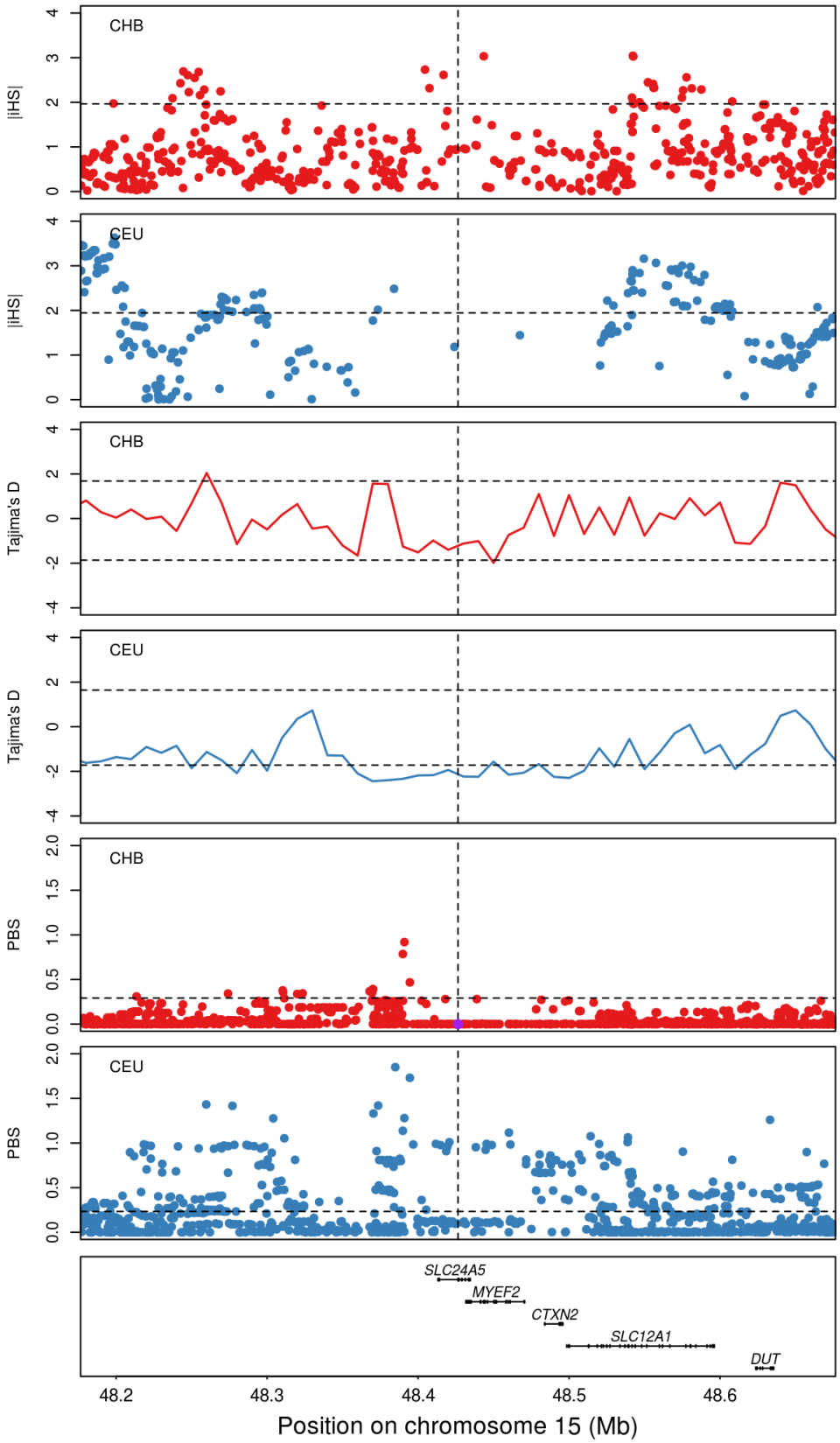

Figure D.11: Selection scans around candidate gene *SLC24A5* at SNP rs1426654 in Eurasian populations. iHS score distributions in CHB and CEU (first panel). Tajima's D score distribution in CHB and CEU (second panel). PBS score distribution in CHB and CEU (third panel). The 95th percentile threshold is shown with a horizontal dashed black line and additionally the 5th for Tajima's D. If present, the genome-wide associated SNP at the region is highlighted in purple and its position represented with a dashed vertical line. UCSC RefSeq genes and genomic coordinates (fourth panel). From Adhikari & Mendoza-Revilla et al. (2018).

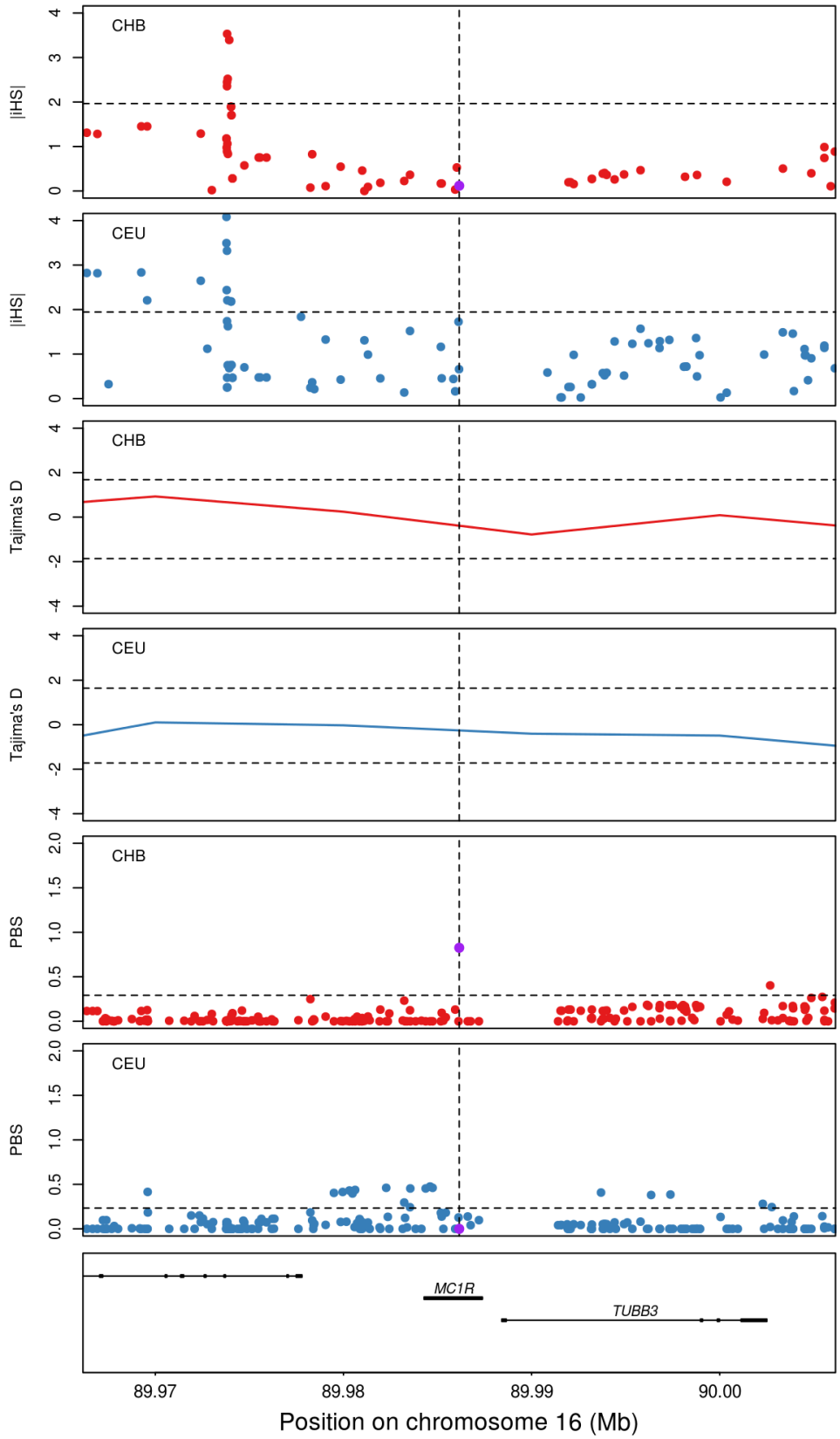

Figure D.12: Selection scans around candidate gene *MC1R* at SNP rs885479 in Eurasian populations. iHS score distributions in CHB and CEU (first panel). Tajima's D score distribution in CHB and CEU (second panel). PBS score distribution in CHB and CEU (third panel). The 95th percentile threshold is shown with a horizontal dashed black line and additionally the 5th for Tajima's D. If present, the genome-wide associated SNP at the region is highlighted in purple and its position represented with a dashed vertical line. UCSC RefSeq genes and genomic coordinates (fourth panel). From Adhikari & Mendoza-Revilla et al. (2018).

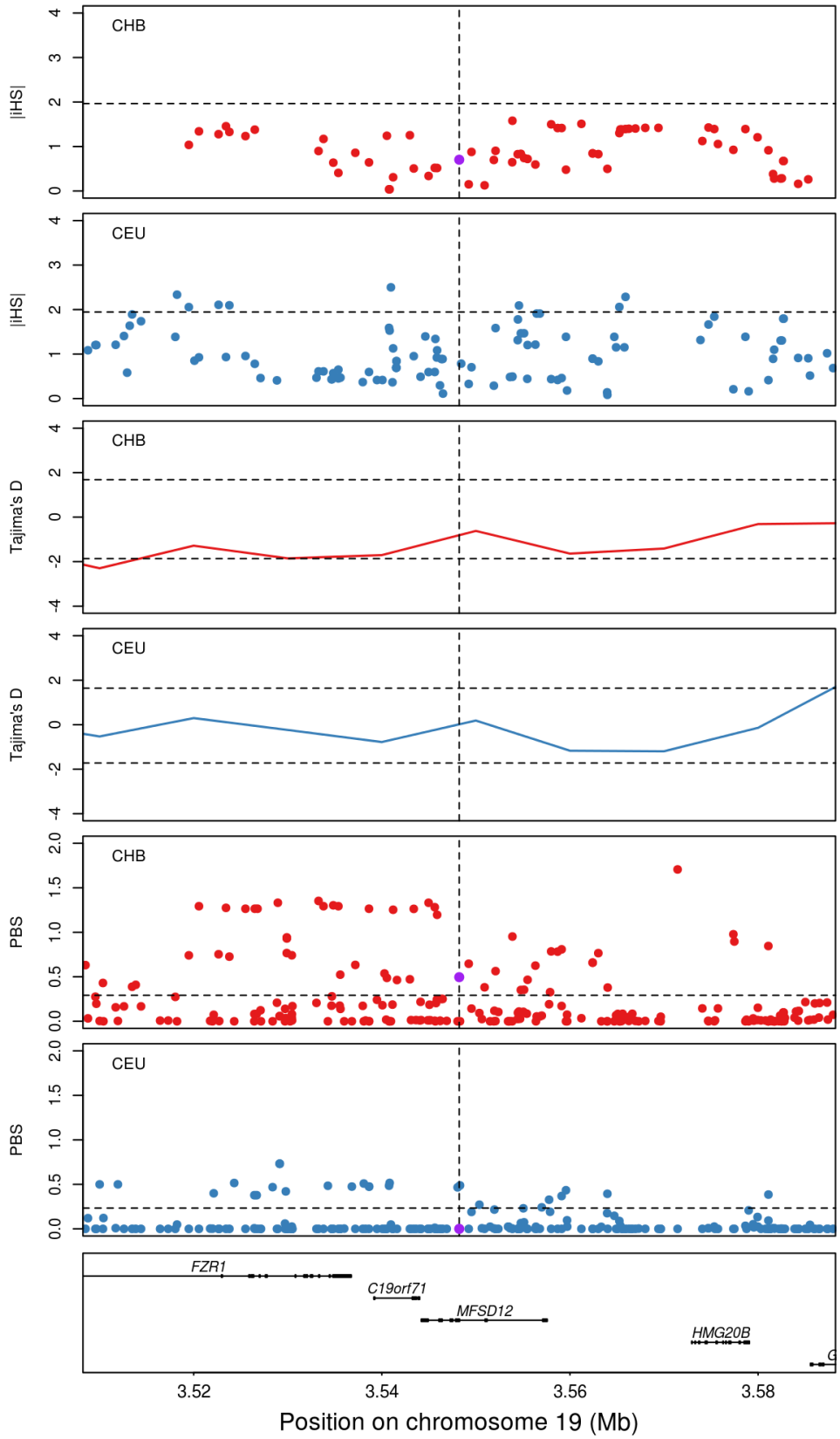

Figure D.13: Selection scans around candidate gene *MFSD12* at SNP rs2240751 in Eurasian populations. iHS score distributions in CHB and CEU (first panel). Tajima's D score distribution in CHB and CEU (second panel). PBS score distribution in CHB and CEU (third panel). The 95th percentile threshold is shown with a horizontal dashed black line and additionally the 5th for Tajima's D. If present, the genome-wide associated SNP at the region is highlighted in purple and its position represented with a dashed vertical line. UCSC RefSeq genes and genomic coordinates (fourth panel). From Adhikari & Mendoza-Revilla et al. (2018).

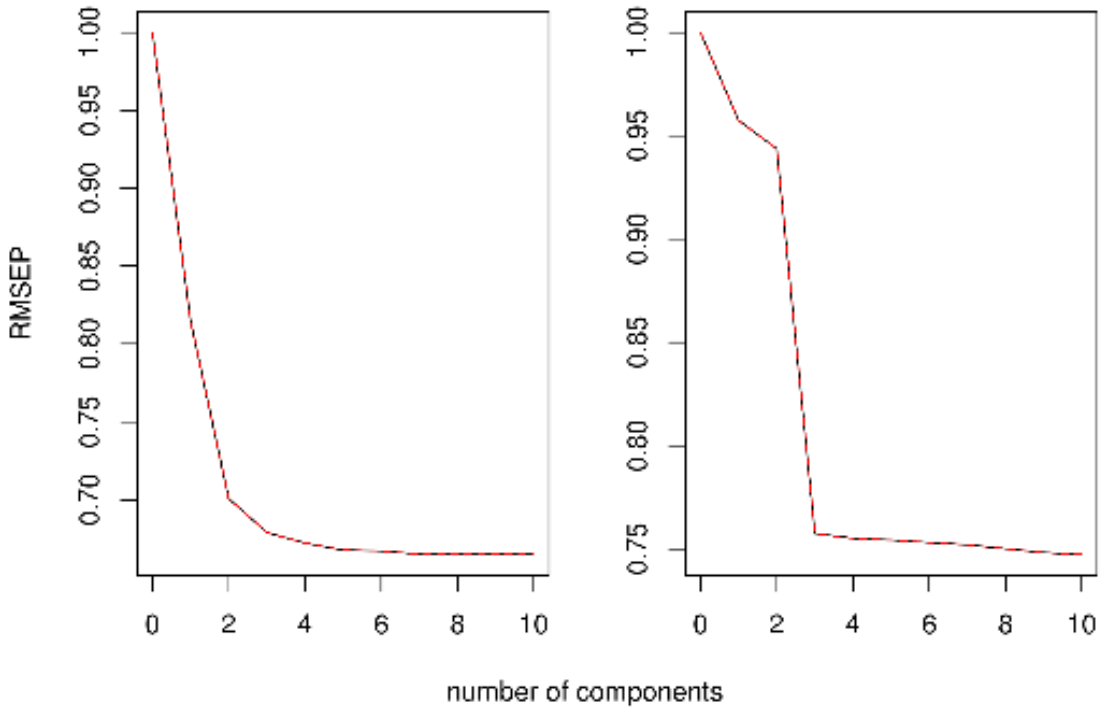

Figure D.14: RMSE plots. Information contained within each PLS component for the starting time of selection (left) and selection coefficient (right). From Adhikari & Mendoza-Revilla et al. (2018).

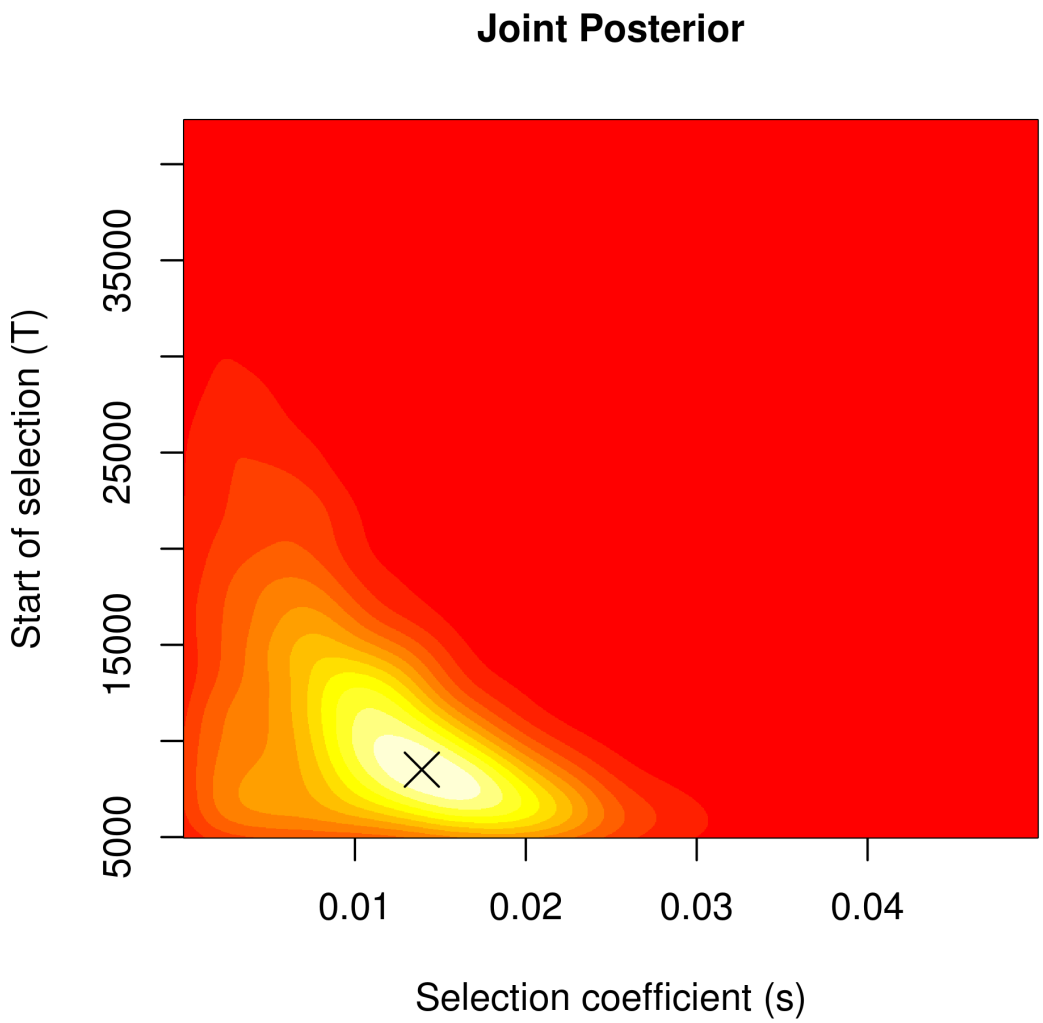

Figure D.15: Joint estimation of the starting time of selection (T) and selection coefficient (s) at the *MFSD12* gene region. Joint inference of the starting time of selection (T) and selection coefficient (s) was done using an ABC approach. The white, yellow, and red colors mark areas of high, moderate, and low joint density, respectively. The black cross indicates the joint maximum a posteriori (MAP at s=0.139 and T=8,508 years ago. From Adhikari & Mendoza-Revilla et al. (2018).