NNT/NL : 0000AIXM0000/000ED000

THÈSE DE DOCTORAT

Soutenue à Aix-Marseille Université
le 28 février 2023 par

Sagnik Palmal

Phenotypic prediction from genotype data: applications to pigmentation in forensics, anthropology and genetic mapping

Discipline

Sciences de l'environnement

Spécialité

Anthropologie biologique

École doctorale

ED 251 – SCIENCES DE L'ENVIRONNEMENT

Laboratoire/Partenaires de recherche

UMR 7268 ADES : Anthropologie Droit Ethique et Santé

Composition du jury

Sijia WANG Chinese Academy of Sciences	Rapporteur
Pedro MORAL CASTRILLO Universitat de Barcelona	Rapporteur
Kaustubh ADHIKARI The Open University	Examinateur
Anaïs Baudot Aix-Marseille Université	Présidente du jury
Andrés RUIZ-LINARES Aix-Marseille Université	Directeur de thèse
Pierre FAUX Aix-Marseille Université	Co-directeur de thèse

Affidavit

I, undersigned, Sagnik PALMAL, hereby declare that the work presented in this manuscript is my own work, carried out under the scientific direction of Andrés Ruiz-Linares and Pierre Faux, in accordance with the principles of honesty, integrity and responsibility inherent to the research mission. The research work and the writing of this manuscript have been carried out in compliance with both the French national charter for Research Integrity and the Aix-Marseille University charter on the fight against plagiarism.

This work has not been submitted previously either in this country or in another country in the same or in a similar version to any other examination body.

Place Marseille, date 16 décembre 2022

Cette œuvre est mise à disposition selon les termes de la Licence Creative Commons Attribution - Pas d'Utilisation Commerciale - Pas de Modification 4.0 International.

List of Publications and Conferences

List of publications relevant to the thesis:

[Published] S. Palmal *et al.,* Prediction of eye, hair and skin colour in Latin Americans, Forensic Science International: Genetics, 2021.

[In preparation] S. Palmal *et al.*, The evolution of the skin and eye pigmentation in humans: a survey based on genetic predictions across time and space.

[In preparation] S. Palmal *et al.*, Transcriptome-wide association study reveals novel calcium channel gene influencing skin pigmentation in Latin Americans.

Other publications:

[Published] B. Bonfante *et al.*, A GWAS in Latin Americans identifies novel face shape loci, implicating VPS13B and a Denisovan introgressed region in facial variation, Science Advances, 2021.

[Published] Y. Chen *et al.,* A genome-wide association study identifies novel gene associations with facial skin wrinkling and mole count in Latin Americans, British Journal of Dermatology, 2021.

Conference participation:

K. Adhikari *et al.*, Prediction of pigmentation in ancient DNA samples – a survey through time and space, Poster, EMGM, 2022.

S. Palmal *et al.*, Prediction of eye, hair and skin color in admixed populations of Latin America, Poster, ESHG, 2021.

S. Palmal *et al.*, Predictions of pigmentation in admixed populations, Oral presentation, EDSE, 2021.

K. Adhikari *et al.*, Pigmentation prediction in admixed Latin Americans, Poster, EMGM, 2021.

Résumé

La pigmentation humaine est une composante majeure de l'apparence physique, très importante pour les études médico-légales, anthropologiques et médicales. Il existe plusieurs ensembles de SNP pour la prédiction médico-légale de la pigmentation. Cependant, les études se sont principalement basées sur des cohortes d'origine européenne, limitant ainsi leur applicabilité à d'autres populations. Outre les phénotypes, les méthodes de prédiction basées sur les génotypes sont également appliquées à l'expression de gènes en utilisant des bases de données transcriptomiques. Ces approches ont gagné en popularité et se sont révélées prometteuses pour améliorer la cartographie de phénotypes complexes. Dans cette thèse, j'ai étudié l'utilisation de la prédiction basée sur le génotype pour deux cas, les phénotypes de pigmentation et les niveaux d'expression des gènes, afin d'améliorer les connaissances actuelles sur la biologie de la pigmentation humaine. Mes principaux objectifs étaient (1) d'établir des modèles de prédiction de la pigmentation applicables à diverses populations, (2) de les étendre à des échantillons d'ADN anciens, et (3) de prédire les niveaux d'expression des gènes et de tester leur association aux phénotypes de pigmentation.

J'ai utilisé les données du projet CANDELA, lequel inclut les génotypes de >7,000 individus d'Amérique latine avec un degré variable de métissage et des phénotypes quantitatifs de pigmentation. J'ai abordé l'objectif (1) en développant une méthode pour construire des ensembles de SNP prédictifs associés à la pigmentation avec une grande précision de prédiction. Les nouveaux ensembles de SNP ainsi développés ont donné de meilleurs résultats que ceux existants pour la couleur de la peau sur la population métissée (caractère quantitatif) et à égalité pour la couleur des yeux et des cheveux (caractère catégorique). Cela souligne l'importance de la concordance entre l'origine des échantillons d'entraînement et des échantillons cibles. J'ai exploré plusieurs modèles statistiques et d'apprentissage automatique et trouvé que la régression linéaire était plus performante pour les traits polygéniques comme la couleur de la peau.

Pour répondre à l'objectif (2), j'ai adapté les ensembles de SNP pour la couleur de la peau et des yeux afin de couvrir des origines peu représentées dans la cohorte CANDELA. J'ai également amélioré les modèles de prédiction pour tenir compte de deux spécificités de l'ADN ancien : la rareté et la pseudo-haploïdie. Ces modèles ont d'abord été validés sur 2,803 échantillons contemporains (1000 Genomes Phase 3 et HGDP), puis appliqués à ~3,600 individus anciens disponibles publiquement (base de données AADR) datant de 50,000 av. JC à 1,500 apr. JC. De par la nature quantitative des phénotypes prédits, j'ai obtenu des tendances graduelles commençant il y a ~15,000 ans pour l'indice de mélanine et la couleur des yeux. J'ai remarqué une augmentation récente de l'indice de mélanine dans le nord-est de l'Eurasie qui est en phase avec les prédictions contemporaines. J'ai prédit un indice de mélanine plus élevé que prévu pour les résidents de la région polaire, ce qui pourrait être lié à un régime enrichi en vitamine D.

Pour répondre à l'objectif (3), j'ai exploré les modèles statistiques PrediXcan. Ceux-ci fournissent une prédiction de l'expression des gènes basée sur les SNP en exploitant les données GTEx et les associations phénotype-transcriptome. La prédiction inter-tissus basée sur les résumés statistiques de GWAS a identifié une nouvelle association entre le gène CACNA2D2 et l'indice de mélanine. L'étude de ce gène chez la souris a également donné des résultats prometteurs.

En conclusion, les modèles de prédiction basés sur le génotype nous ont aidés à mieux comprendre l'architecture des facteurs génomiques ayant un impact sur les traits de pigmentation, ainsi que les tendances évolutives et les forces de sélection qui agissent sur eux.

Mots clés : Prédiction du phénotype, Génétiques des populations, Génétique des caractères complexes

Abstract

Human pigmentation is a major component of physical appearance, of great importance to forensic, anthropological and disease studies. There exists several SNP sets for forensic, genotype-based prediction of pigmentation phenotypes. However, studies have mostly focused on European ancestry cohorts, limiting therefore their applicability to genetically distant groups. Beside phenotypes, genotype-based prediction methods are also applied to gene expression using transcriptomic databases. These approaches have been gaining popularity and has shown promises for enhancing the mapping of complex phenotypes; In this thesis, I investigated the use of genotype-based prediction for two cases, pigmentation phenotypes and gene expression levels, to further enhance the current knowledge of the biology of human pigmentation. My main objectives were (1) to establish pigmentation prediction models applicable to genetically diverse populations, (2) to extent them to ancient DNA samples, and (3) to predict gene expression levels and test their association to pigmentation phenotypes.

I used the data from the CANDELA project, with genotypes of >7,000 Latin American individuals with varying degree of continental admixture, and quantitative pigmentation phenotypes. I addressed the first objective by developing a method to construct predictive SNP sets of pigmentation-associated genetic variants with high prediction accuracy. Novel SNP sets performed better than existing ones for quantitative skin colour on the admixed population and at par for categorical eye and hair colour. The importance of concordance between the ancestries of model training and target samples was highlighted. I explored several statistical and machine learning models and found linear regression to perform better than machine learning models for polygenic traits like skin colour.

To address the second objective, I adapted the SNP sets for skin and eye colour to account for two specificities of ancient DNA: sparsity and pseudo-haploidy, and improved the prediction models to encounter ancestries scarcely represented in the CANDELA set. These models were first validated on 2,803 contemporary samples (100 Genomes Phase 3 & HGDP), then applied to ~3,600 publicly-available ancient individuals (AADR database) dating from 50,000 BCE to 1,500 CE. Thanks to the quantitative nature of the predicted phenotypes, I obtained gradual trends starting ~15,000 years ago for melanin index and eye colour. Interestingly, I also predicted a recent increase of melanin index in North-Eastern Eurasia that aligns with the prediction on contemporary populations. I predicted higher than expected Melanin Index for residents in the polar region that may relate to a vitamin D-enriched diet.

To address the third objective, I explored the class of PrediXcan statistical models, which provide SNP-based prediction of gene expression, leveraging the GTEx dataset and phenotype-transcriptome association testing. The cross-tissue prediction based on GWAS summary statistics identified a novel association between CACNA2D2 in 3p21 and melanin index. Silencing this gene in mice has further shown promising results.

In conclusion, genotype-based prediction models helped us to better understand the architecture of the genomic factors impacting the pigmentation traits, as well as the evolutionary trends and selection forces working on them.

Keywords: Phenotype prediction, Population genetics, Complex traits genetics

Acknowledgement

First, I would like to thank my thesis supervisors Andrés Ruiz-Linares and Pierre Faux. Andrés thank you for allowing me to take part in this project, for guiding me and for sharing your knowledge and ideas with me. Pierre thank you for teaching me so much, for being so available, ingenious, and passionate about science.

I would like to thank the rapporteurs and examiners of this work, Sijia Wang, Pedro Moral Castrillo, Anaïs Baudot and Kaustubh Adhikari for accepting to be part of my thesis jury.

I would like to thank all the people who participated in the Consortium for the Analysis and Diversity and Evolution (CANDELA). Thank you to the volunteers, students and researchers who made it possible to gather all this data without them I could not have done this work. Special thanks to Kaustubh Adhikari for providing answers to my questions since I started working on this project.

I must acknowledge Centre de Calcul Intensif d'Aix-Marseille for granting me access to its high-performance computing resources.

I would like to thank the members of the ADES laboratory, in particular the blood group biology team for their welcome and all the good times shared during these last years. Thank you to my doctoral colleagues from the office, Pauline and Betty for their help and our discussions, whether scientific or personal.

Finally, a big thank you to my family and friends for their support and encouragement. Thank you Saswati for your patience. Thanks to my parents for everything.

Table of contents

Affidavit 3
List of Publications and Conferences 5
Résumé 7
Abstract 9
Acknowledgement 11
Table of contents 13
Figures 17
Tables 19
Abbreviations 21
1. General Introduction 23
1.1. Cellular biology and biochemistry of human pigmentation 23
1.1.1. Cellular aspects 23
1.1.2. Synthesis of melanin 24
1.1.3. Melanin molecules and their relation to pigmentation phenotypes 25
1.2. Evolution of pigmentation 27
1.2.1. Loss of body hair in humans and UV radiation 28
1.2.2. Genetic drift, selection and admixture 28
1.3. Prediction methods 30
1.3.1. Linear statistical models 30
1.3.2. Machine learning models 31
1.3.3. Polygenic risk score models 31
1.4. Forensic phenotype prediction 31
1.4.1. Prediction of human phenotypes 32
1.4.2. Forensic DNA phenotyping 32
1.5. Ancient DNA and anthropological genetics 33
1.5.1. Ancient DNA and human population history 34
1.5.2. Ancient DNA and the evolution of human phenotypes 35
1.6. Identification of genes for pigmentation variation 36
1.6.1. Mendelian genetics 37
1.6.2. GWAS methods and results for pigmentation 37
1.6.3. TWAS methods and results for phenotypes 38
1.7. Aim of the thesis 39
2. Materials and methods 41
2.1. Study cohort 41
2.1.1. Sample collection 41
2.1.2. Sample phenotyping 42

2.1.3. Genotype data 46
2.2. Prediction algorithms 48
2.2.1. Linear Regression (LR) 49
2.2.2. Stepwise Regression (SR) 50
2.2.3. Multinomial Logistic Regression (MLR) 51
2.2.4. Ordinal regression (OR) 52
2.2.5. Ensemble Learning models 53
2.2.6. Artificial Neural Network (ANN) 54
2.2.7. Least Absolute Shrinkage and Selection Operator (LASSO) & Elastic Net 56
3. Results 59
3.1. Article 1: Prediction of eye, hair and skin colour in Latin Americans 59
3.1.1. Introduction 60
3.1.2. Methods 62
3.1.3. Results 66
3.1.4. Discussion 74
3.1.5. References 78
3.2. Article 2: The evolution of skin and eye pigmentation in humans: a survey based on genetic predictions across time and space 85
3.2.1. Introduction 86
3.2.2. Methods 87
3.2.3. Results 91
3.2.4. Discussion 101
3.2.5. References 104
3.3. Article 3: Transcriptome-wide association study reveals novel calcium channel gene influencing skin pigmentation in Latin Americans 107
3.3.1. Introduction 108
3.3.2. Methods 110
3.3.3. Results and discussion 113
3.3.4. Conclusion 117
3.3.5. References 119
4. General discussion 123
4.1. Summary of the thesis 123
4.2. Variation of prediction models and inference 124
4.3. TWAS finding 125
4.4. Pigmentation evolution 126
4.5. Future perspectives 127
4.6. Importance of ancestry inclusion 129
4.7. Concluding remarks 130
References 131
Appendices 151
A. Article 1: Supplementary information 151

Supplementary note: Details on the construction of CAN-E/S/H SNP sets 151
Supplementary note: Statistical and Machine learning models used for prediction 153
Supplementary note: Benchmark Prediction Strategies 155
Supplementary note: Diversity index 157
Supplementary figures 158
Supplementary tables 164
B. Article 2: Supplementary information 165
Supplementary Figures 165
Supplementary Notes 168
Supplementary Tables 172
C. Article 3: Supplementary information 173
Supplementary Figures 173
Supplementary Tables 180

Figures

Fig. 1 Melanosome structure variation and transfer pathway 24
Fig. 2 Biochemistry of the melanin synthesis 26
Fig. 3 Effect of latitude on skin reflectance 28
Fig. 4 Global variation of human pigmentation 29
Fig. 5 Putative migration waves out of Africa 34
Fig. 6 Birthplace maps of the individuals recruited in CANDELA study 42
Fig. 7 Pigmentation categories of the CANDELA individuals 43
Fig. 8 Quantitative eye colours. 44
Fig. 9 Bicone HCL colour space of all the retained iris colours 45
Fig. 10 Skin colour distribution of the CANDELA individuals 46
Fig. 11 Continental ancestry in the CANDELA individuals 47
Fig. 12 Continental admixture in the CANDELA individuals with additional East Asian ancestry 48
Fig. 13 Schematic of a double hidden layer artificial neural network 55

Article 1, Fig. 1 Study overview 65
Article 1, Fig. 2 Overlap between SNP sets used for prediction of pigmentation traits. 67
Article 1, Fig. 3 Prediction accuracy in relation to models, methods and pigmentation SNP sets 68
Article 1, Fig. 4 Prediction accuracy for individuals with varying Native American/European admixture 71
Article 1, Fig. 5 Portability of prediction models trained in highly European/Native American cohorts 73

Article 1, Fig. 6 Solar radiation levels and skin pigmentation (MI) predicted with CAN-S and HIrisplex-S SNP sets 74
Article 2, Fig. 1 Prediction on the diploid contemporary datasets 92
Article 2, Fig. 2 Prediction comparison on the diploid and haploid genotypes using contemporary samples 94
Article 2, Fig. 3 Correlation between absolute latitude and predicted pigmentation phenotypes 96
Article 2, Fig. 4 Spatio-temporal prediction on the haploid ancient human samples 98
Article 2, Fig. 5 Allelic age compared to effect on pigmentation 100

Article 3, Fig. 1 Methods summary 113
Article 3, Fig. 2 Manhattan plots for Skin Melanin Index using S-MultiXcan 115
Article 3, Fig. 3 Overlap of TWAS and GWAS signals 116

Tables

Table 1: Sampling of CANDELA cohort.....47

Article 1, Table 1 Overall accuracy (%) and trait AUC (%) for categorical (A) eye, (B) hair and (C) skin colour obtained here and in other studies.....77

Article 2, Table 1 Number of ancient human samples examined for five different time periods and 15 geographic subregions.....103

Abbreviations

AM: admixture mapping
ANN: artificial neural network
DNA: deoxyribonucleic acid
CANDELA: consortium for the analysis of the diversity and evolution of Latin America
FDP: forensic DNA phenotyping
FDR: false discovery rate
GRM: genomic relationship matric
GWA: genome-wide association
GWAS: genome-wide association study
HapMap: haplotype map of the human genome
HCL: hue chroma luminance
IBD: identity by descendent
KYA: thousand years ago
LD: linkage disequilibrium
LR: linear regression
MAF: minor allele frequency
MLR: multinomial logistic regression
OR: ordinal regression
PC: principal component
PCA: principal component analysis
QC: quality control
QTL: quantitative trait loci
RF: random forest
SNP: single nucleotide polymorphism
SR: stepwise regression
TWAS: transcriptome-wide association study
UVR: ultraviolet radiation
XGB/XGBoost: extreme gradient boosting
YA: years ago

1. General Introduction

I have always been fascinated with human. The thought process, human interaction, behaviour within societal norms has always intrigued me. I have always seen the species as one big unit, albeit with variation – both within and between subgroups. As I grew old, the variations became more prominent, and learned about human pigmentation being one of the main reasons of this segregation. As Professor Nina Jablonski puts it "*We are united, and divided, by our skin color*"[1].

I started my thesis with the aim to establish the oneness within the human species, how we started from Eastern Africa and spread across the earth. There were certain adaptations because of the movement, and it explains the variation we have today as an outcome of the journey. Since the 19th century, researchers are developing theories on the emergence of the human from the ape family and explaining the variation through evolution. Charles Darwin (1809-1882) and Alfred Russel Wallace (1823-1913) independently came up with their evolution by natural selection theory[2]. Evolution enacted on the human population to change physical appearances and pigmentation is a large component of human appearance; hence variation of pigmentation plays a large part in diversity of human appearance. Therefore, it is important to understand the mechanism of human pigmentation and its evolution from the ancient times.

1.1. Cellular biology and biochemistry of human pigmentation

Human pigmentation is the outcome of melanin biosynthesis within melanosomes and the eye, hair and skin colour variation is determined by the type and amount of the melanin produced[3]. Pigmentation dictates social cues, affects visual acuity and shields against the ultra-violet radiation (UVR), while deficiencies cause various forms of albinism, elevates the risk of skin cancer, patterned disorder such as vitiligo or piebaldism[4,5]. Despite the importance of human pigmentation, we are yet to understand the complete biological basis underlying its full complexity drawing attention towards the formation and distribution of the melanin pigments[3].

1.1.1. Cellular aspects

The melanin pigments in the human eye, hair and skin are synthesised, deposited and transported in a membrane-bound organelle called melanosomes[6,7]. Melanosomes are the pigment granules responsible for the photoprotection against ionizing radiation and colour of the tissues[8]. Melanosomes are synthesised in choroidal melanocytes and retinal

pigment epithelial cells in the eye and in melanocyte cells in the case of skin[7,8]. These heterogeneous group of cells reside in the bottom layer of epidermis and hair follicles of the human skin[9], the middle layer of the eye (uvea)[10], inner ear[10] among other places in the body (see Figure 1A).

The visible phenotype we observe in the feathers, skin or fur is largely shaped by melanocytes. Melanosomes are exported from melanocytes to the adjacent keratinocytes, where most melanin pigments are found (see Figure 1B). The variation in the phenotype is not related to the melanocyte numbers in the tissues, as they largely remain constant, rather due to the number, size and distribution of melanosomes[10].

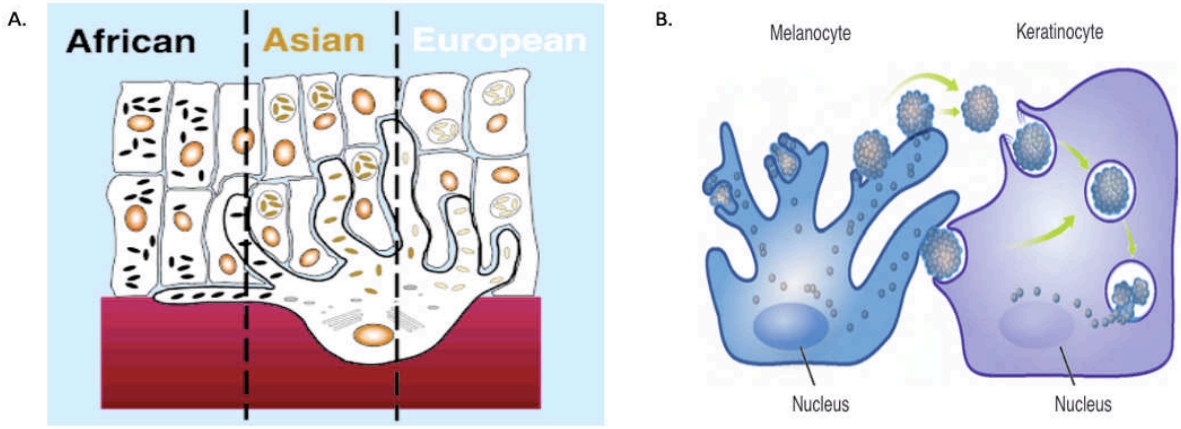


Fig. 1 Melanosome structure variation and transfer pathway

*(**A**) Schematic representation of melanosome structure variation and distribution in various groups. This figure shows the gradient in melanosome size and number in three continental ancestry skin – African, Asian and European. The dispersion of melanosomes in the darkest skin is also notable (taken from Sturm et al 1998[11]). **(B)** Melanocyte plasma membranes encapsulates the globules of melanosomes, which are then released into the extracellular space and phagocytosed by keratinocytes and dispersed into the perinuclear area (taken from Ando et al 2012[12]).*

1.1.2. Synthesis of melanin

Melanocytes produce the melanin pigments within melanosomes through a complex synthesis process called melanogenesis[13]. The ultraviolet radiation is understood to be the main physiological stimulus for human melanogenesis[14]. Although multiple signalling pathways have been found to be involved in the regulation of melanogenesis, the cyclic AMP (cAMP)/protein kinase A (PKA) signalling pathway remains one of the most significant pathways[15]. The melanocortin-1 receptor (MC1R gene in region 16q24) works as one of the important receptors on melanocytes and modulates their function. On the membrane of melanocytes when α-melanocyte-stimulating hormone (α-MSH) binds to MC1R, adenylate cyclase is activated, which in turn affects levels of intracellular cAMP positively and tyrosinase activity. Oxidation of tyrosine by tyrosinase (TYR) is an essential step for pigment types. The membrane-associated transport protein (MATP) and the pink-eyed dilution protein (P) are melanosomal membrane compounds that controls the quantity of pigment synthesis within melanosomes[16]. Keratinocytes and melanocytes are

the main components of the epidermal melanin unit which affects skin melanogenesis. Keratinocytes are responsible for the several paracrine factors secretion most of which acts on melanocytes to promote or inhibit melanogenesis[15]. Additionally, studies have found that DNA repair mechanisms can be controlled by paracrine factors by activating different receptors and signalling pathways which helps to maintain melanocyte homeostasis and prevent UV mutagenesis[17].

1.1.3. Melanin molecules and their relation to pigmentation phenotypes

Human eye, hair and skin colour phenotypes are mainly controlled by two components of melanin pigments – eumelanin and pheomelanin[13]. Eumelanin dictates on the black to brown colour gradient while pheomelanin controls red to yellow axis[18,19]. Skin colour is a highly polygenic trait (i.e. that is controlled by multiple genetic variants) but hair and eye colours are monitored by a handful of strong variants[20]. Nevertheless, MC1R has been strongly associated with pale human skin, freckling and red hair[21,22], also characterised as a melanoma susceptibility gene[17,23]. Primary effect of the gene is to promote eumelanin synthesis, at the expense of pheomelanin synthesis and the loss-of-function mutation causes it to produce more pheomelanin than eumelanin which causes the phenotypic change of carrot red hair[11,16]. However, variation of all three pigmentation phenotypes are resultant of different proportions of eumelanin and pheomelanin[19,24,25]. The biochemistry of eumelanin and pheomelanin synthesis has been elaborated in the Figure 2.

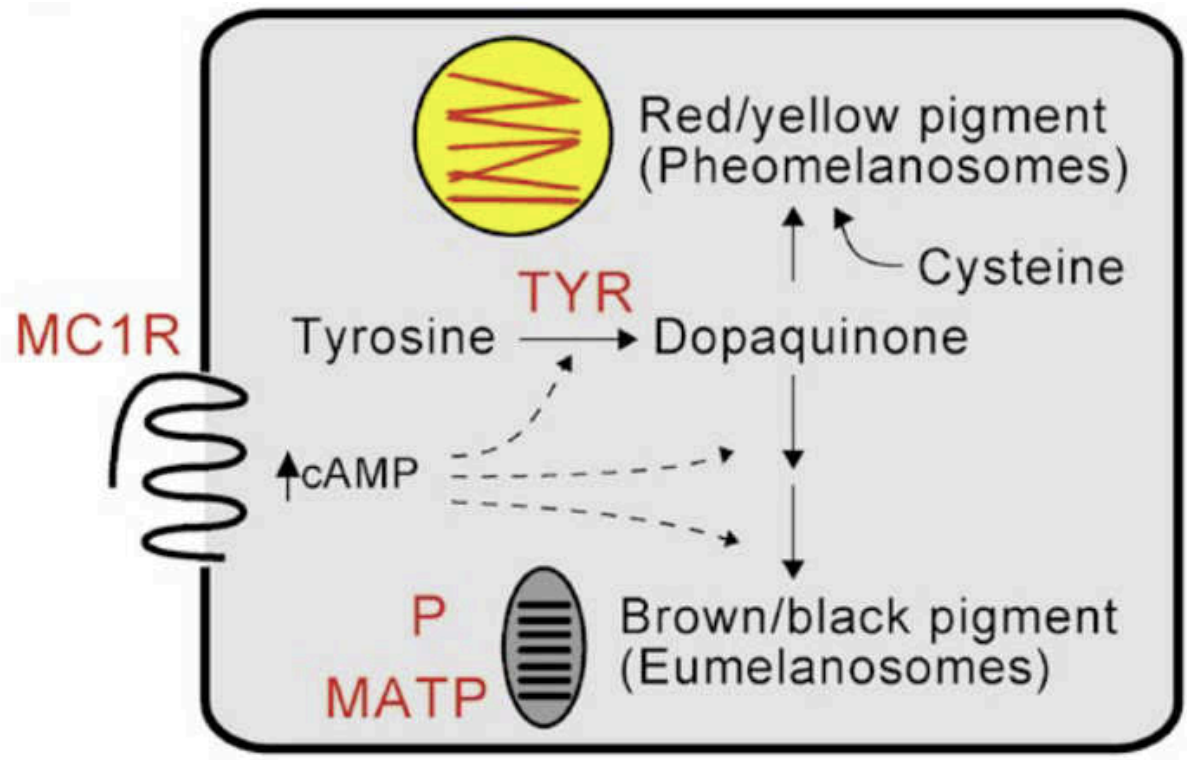


Fig. 2 Biochemistry of the melanin synthesis

Melanocortin 1 receptor (MC1R) controls the synthesis of eumelanin and pheomelanin. Oxidation of tyrosine by tyrosinase (TYR) is essential for both melanin pigments. Pink-eyed dilution protein (P) and membrane-associated transport protein (MATP) controls melanosomes with the extent of the pigment synthesis (figure taken from Barsh GS (2003)[16]).

Eye colour is the outcome of the pigmentation of the iris which is the circular band around the small black circle at the centre of the eyes (the pupil). Eye colours are often recorded as a categorical variable with categories as blue/gray, honey, green, light brown and dark brown/black. This is a direct consequence of the melanin content in the iris[26]. Human iris samples have revealed the existence of both eumelanin and pheomelanin. Pheomelanin majority eyes would be more towards the green end of the spectrum whereas brown eyes are dominated by presence of eumelanin. Other eye colours can be categorised as a proportional variation of both pigments[27]. In the human society, brown is the most common eye colour[20]. Alternatively, eye colour can also be measured in a quantitative way using the HCL (Hue, Chroma and Luminance) colour space[28] which represents the human perception of colour. We have discussed in detail about the collection of categorical and quantitative eye colours from the CANDELA cohort in Chapter 2 along with the details of the cohort.

Hair colour can vary from black, brown, blond to red. This spectrum of hair colour is a consequence of quantity and ratio of eumelanin and pheomelanin. If we order the observed hair colours in terms of decreasing eumelanin content (with a trace of pheomelanin), we will get black, dark brown, brown, light brown and blond, whereas, red hair is an outcome of equal or higher proportion of pheomelanin than eumelanin [19,29,30].

The final individuals included in our analysis from the CANDELA cohort exhibit three kinds of hair colour – blond, dark blond/light brown and brown/black.

Human skin is made up of three layers – outer thin layer called epidermis, thicker middle layer called dermis and innermost fatty layer called hypodermis. Epidermis contains melanin, which is produced by melanocytes. Total melanin content of a human skin dictates the skin tone – higher melanin results in darker skin colours[19,31]. Additionally, the skin melanin pigments also vary in size and can be accumulated in different numbers causing variation in human skin colours. Other chemicals such as keratin also impacts skin colour variation[32].

The components of melanin work differently - eumelanin works as a photoprotective agent, whereas, pheomelanin is a phototoxic agent. Some articles have observed that eumelanin and pheomelanin exists in a constant ratio in all human epidermis (except dark skin individuals where eumelanin proportion is higher than rest of the skin tones)[18,31] contradictory to the notion of higher concentration of pheomelanin in the lighter skinned people and higher proportion of eumelanin in darker skinned individuals[19,32,33]. Nonetheless it is well established that people with pale skin colour are more prone to suffer from skin related problems as lack of melanin (especially eumelanin) in the epidermis cannot screen the harmful effects of the UVR[4,16,18,25,34]. Thus we find more people with lighter skin pigmentation in the higher latitudes than the darker pigmented population inhabiting in the tropical regions. The variation of human pigmentation can be attributed to evolution through natural selection to adapt for the UVR[4].

1.2. Evolution of pigmentation

Pigmentation exhibits wide range of colouration both within and between species and plays a very important role in the functioning of animal kingdom. It holds a crucial part in physiological processes, mate attraction, thermoregulation, concealment through camouflage which helps both in predation avoidance and food hunting and communication[35–38].

As a result research in pigmentation genetics have garnered significant interest for over a century[39,40]. Knowledge of cellular interaction and signalling pathways is essential to comprehend the complex pigmentation patterns[41]. Recent studies have benefited from the information on pigmentation pattern related genes and the genetic changes those animals have been through to arrive in the current situation[42]. Several mechanisms affect the variation in melanin type, distribution and density which effectively gives rise to pigmentation patterns within vertebrates. During embryogenesis, neural crest cells are differentiated into melanoblasts, which migrates along the body axis and enters the epidermis. In most of the species melanoblasts moves to the hair follicles and differentiated into melanocytes. These melanocytes produce pigment upon maturation and the pigments are packaged into melanosomes, which are transferred to the developing hair keratinocytes[43,44]. Human skin has significantly less hair density

compared to the nearest primates[45] and yet out of the all the mammalian species, human skin shows the greatest variation of colours[46], which makes it interesting to learn about impact of strong UV radiation and other evolutionary forces on human pigmentation.

1.2.1. Loss of body hair in humans and UV radiation

The earliest humans started to lose body hair (compared to archaic hominins) to maintain the body temperature through sweating in the hot environments (low latitudes of Africa) [47]. Added to that the lack of body coverings and scarcity of protective sheltering forced the human skin to function as a barrier to protect human body from harmful ultra-violet radiation (UVR) from the sun and the external pathogens. Additionally, when the early humans started to migrate from the tropical Africa, which was under strong exposure to sun and UVR, to southern Africa or Eurasia (i.e. places far from the equator with lower sun exposure), skin colour came under strong natural selection[32]. Depigmentation of the skin also facilitated the synthesis of Vitamin D, which is a UV-dependent process[48,49] (see Figure 3). Genes associated with human pigmentation have been affected by population bottlenecks during human migration[50,51].

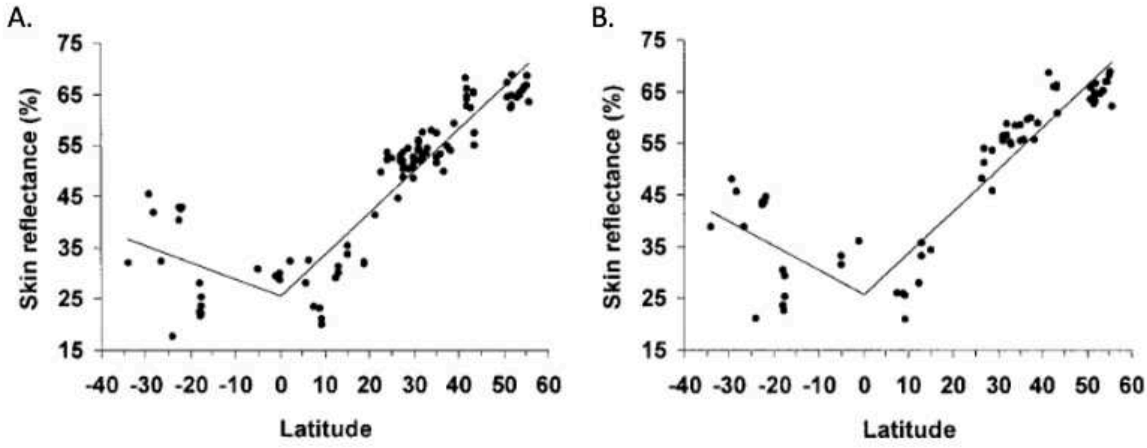


Fig. 3 Effect of latitude on skin reflectance

*Non-linear piecewise regression showing the relation of latitude and skin reflectance(%) using **(A)** 102 males (R^2=0.85) **(B)** 65 females (R^2=0.85) (image taken from Relethford JH (1997)[52]).*

1.2.2. Genetic drift, selection and admixture

Studies have concluded that strong and multiple evolutionary forces along with selection processes have shaped the modern human[50]. Today we can observe variation in the human skin, hair and eye colour traits within and between subpopulations thanks to those forces acting on the species throughout prehistory[53]. Where eye and hair colour has been found to be more variable within continental Europe, skin colour tends to show more variation outside of Europe (see Figure 4). Nonetheless it was not only strong selection forces acting on human pigmentation – but also ancestral admixture, and

population bottlenecks have shaped the trait over time and space[54]. Skin pigmentation phenotypes have long been identified as polygenic traits[55]. As different combination of the genetic markers can produce similar skin colour palettes, genetic drift induced genotypic loss did not impact the phenotypic variability[50]. A big testament to that is the convergent evolution of the European and East Asian population towards lighter skin pigmentation[28]. But the similar arguments do not hold true for eye and hair colour. Contrary to skin colour which has been shaped by natural selection in a large part, eye and hair colour have been affected by sexual selection[50,56,57]. Although it has been argued that sexual selection plays a part for skin colour as well. There are multiple cultures across the world where finding a fairer mating partner is the norm[58,59]. In an extension to that, several authors have claimed that eye and hair pigmentation evolution is owned by the pleiotropic effects of the genes involved in the skin colour evolution[53,55].

Understanding the human history depends on the DNA-based pigmentation prediction from the human remains[60] and genetic pigmentation prediction depicts how human population may have looked in the past and unfolds the evolutionary history[55].

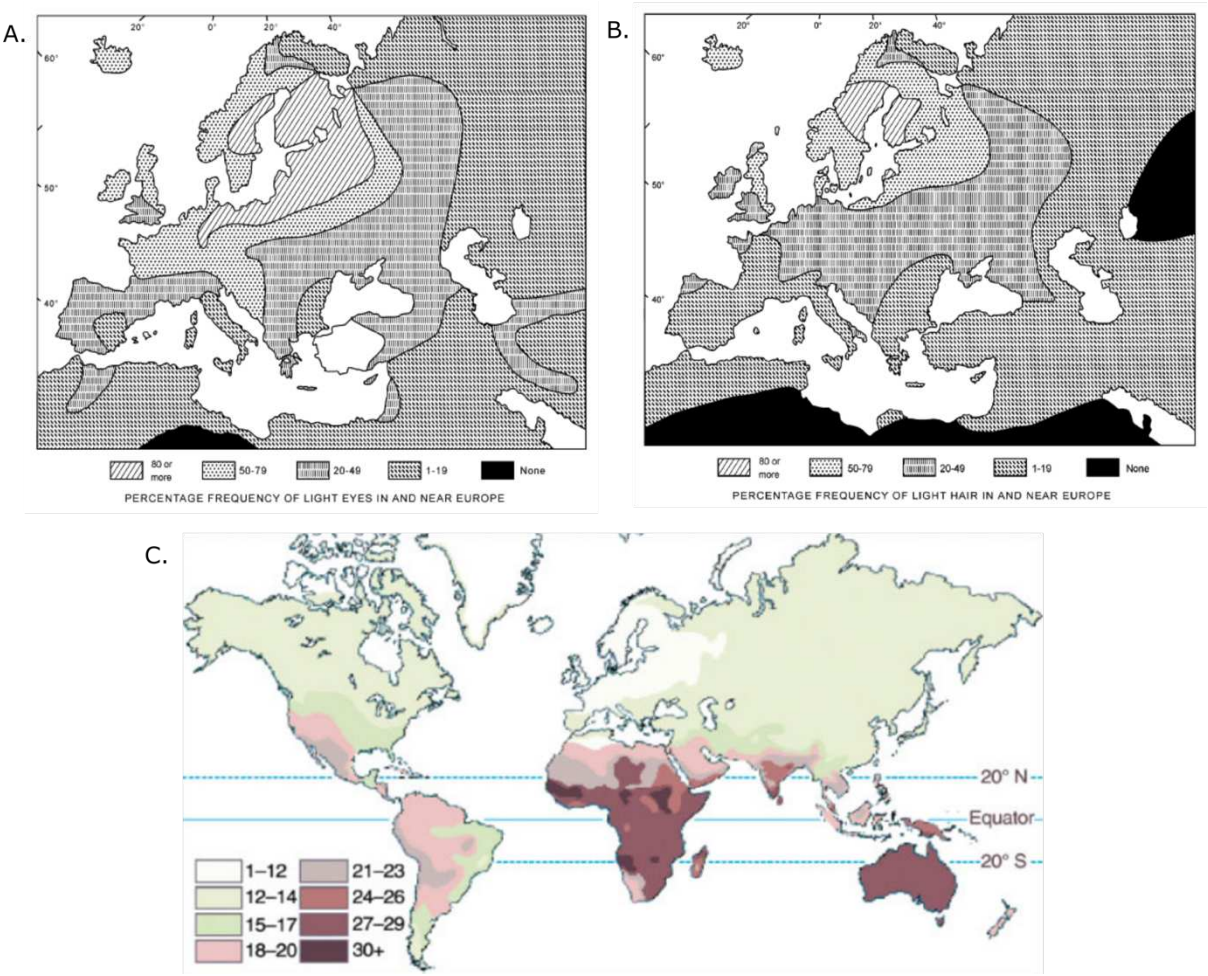


Fig. 4 Global variation of human pigmentation

(A) *The proportion of light eyes around Europe* ***(B)*** *The proportion of light eyes around Europe* ***(C)*** *Distribution of skin colour around the world, higher numbers represent darker skin. (image A,B taken from Beals et al. (1965)[61], image C is a traditional skin color map by Biasutti, based on http://anthro.palomar.edu/vary/)*

1.3. Prediction methods

In the last decade and a half, Genome-wide Association Studies (GWAS) have been instrumental in finding the association of genetic variants with complex human traits[62]. Although it has not fully uncovered the complete genomic structures of the complex human phenotypes, the outcomes from the GWAS results have been found to be instrumental in predicting phenotypes, calculating the disease risk for an individual or prescribing highly individualised therapeutics[63]. Human phenotype is shaped by the environment, individual genes, the mutual interaction of those genes as well as their interaction with the environment[64,65]. Therefore, given the genetic variation, we try to capture the phenotypic variation through phenotype prediction[66]. In personalised medicine and healthcare, we have experienced the impact of precise prediction of phenotypes[67–69]. In agricultural science, prediction of grain yield from the genomic and environmental variables has improved the plant breeding which has resulted in better production[70–72]. In the case of bacterial science, predicting phenotypes (e.g. antibiotic resistance) is an essential step to better understand the bacteria[73,74].

A broad distinction of the existing prediction methods can be made based on the source of genomic data – one approach is to utilise the individual genotype and phenotype data whereas the other approach incorporates GWAS summary statistics to bypass difficulties to obtain the individual-level datasets[75].

1.3.1. Linear statistical models

Traditionally, the parametric regression model is the most common framework for testing statistical association and prediction using individual data. The use of linear[76,77] or multinomial logistic regression[78–80] is quite prevalent in this area as it can easily handle quantitative phenotype (such as height, blood pressure or body mass index) and categorical phenotype (such as presence or absence of disease, categorical eye and hair colour). To address the stratification and the confounding effects of the population, age, sex and ancestries are included as covariates in the model[63], although this reduces power of the binary classification models (e.g. case-control studies) to detect genetic variants[81]. In this regard, linear mixed models can improve statistical power for genomic discovery by robustly accounting for population stratification and relatedness through inclusion of random effects for a genetic relationship matrix. Classical methods such as BLUP, which was originally used in the livestock genetics thanks to the availability of the family information (pedigree), has been successfully migrated into the human genetics since genomic data provides robust estimates of relatedness between individuals [82]. Subsequent adaptations of the BLUP method relaxed the assumption of assigning same variance to each SNP effect and methods such as, adaptive MultiBLUP[83], compressed BLUP[84], weighted-GBLUP[85] obtained even better performances in phenotype prediction.

1.3.2. Machine learning models

Recently, machine learning models are also becoming more popular in the phenotype prediction scenarios thanks to their simpler framework than parametric models[86,87] and relaxed adherence to the genomic structure of the phenotypes[75]. Linear models with penalised regression, e.g. LASSO[88,89], tree-ensemble models, e.g. random forest[90–92] or more complex models e.g. neural networks[93–96] are able to capture the non-linear relationship of the genetic markers or the non-explicit interactive effects, added to that these models are adept to overcome the significant problem of the classical statistical models of "large p, small n" (sample size smaller than the parameters to be estimated)[74,88].

1.3.3. Polygenic risk score models

Nonetheless, the access to individual-level data is not always feasible. The logistic, ethical and privacy concerns restrict the sharing of the datasets[97]. To navigate those shortcomings, summary level information from the GWAS studies become helpful. GWAS Catalog[98] contains variant association from almost 5,700 studies on more than 3,300 phenotypes which help us to understand the respective phenotypes' underlying biology, estimate genetic heritability of the phenotypes, compute genetic correlations and predict the genetic risk of the individuals for a certain disease[63]. The latter part corresponds to polygenic risk score (PRS), which aggregates the weighted sum of the number of risk alleles carried by an individual, where the weights are the effect sizes obtained from GWAS[99]. PRS can be computed in several ways, the simplest approach is to prune and threshold, which select a subset of the SNPs based on the trait-association pvalue[100]. The more advanced approaches take the LD structure, functional information into account to produce improved PRS prediction with more accurate marginal effect sizes[101–103]. A recent study has shown that genomic risk predicted from the genome-wide PRSs can identify disease risk for coronary artery disease, atrial fibrillation, type 2 diabetes, inflammatory bowel disease and breast cancer as well as monogenic risk prediction methods based on rare, highly penetrant mutations[104]. The disease risk calculated from the gene disposition of an individual enables us to make preventive intervention and prepares the path for personalised treatment in the future[104–107]. Nonetheless, one should be attentive to the population structure in the PRS models, as not accounting for the population stratification can lead to biased PRSs and models might not be portable to other ancestries[108,109].

1.4. Forensic phenotype prediction

The discovery of the novel and efficient methods has made prediction from human remains easier, more commonplace and accurate. Modern genotyping methods has become more affordable and hence genotype-to-phenotype prediction has picked up pace for forensic research. Identity determination from human remains is an investigative

procedure and is a crucial block in the medicolegal processes, especially in high mortality cases, such as armed conflicts, migrant crises, natural mass disasters, epidemic etc.

1.4.1. Prediction of human phenotypes

Identification process includes the assertion of the physical attributes – height, age, gender, ancestry and depends heavily on the state of the remains found[110]. Height and age determination methods largely depend on the gender recognition, making it the crucial part of the whole identification process. Gender can be reliably determined from the hip bone[111]. Although in the extreme cases where fraction of the body is recovered, sex can be determined from the teeth and skull remains[112]. Computed tomography (CT) of the maxillary sinus[113] and pelvis skeleton[114], multidetector CT of patellar measurements[115], panoramic dental radiography[116] and fingerprints[117] have also shown good discriminative power for binary genders. To determine age of the human remains, bone length and morphological features of pelvis and skull is examined, more accurate measurements come from aspartic acid racemisation method[118]. Recent application of machine learning models on the teeth x-ray images have produced satisfactory results[119]. Lengths of human long bones can be instrumental to derive estimates of stature. Femur, humerus, tibia lengths have produced significant prediction accuracy even when measured from radiographs [120–122]. While most of the methods make assumption on the population or ancestry of those human samples to produce better accuracy, recent studies have attempted to relax those constraints[123]. Eye, hair and skin colours are also considered as integral parts of identity assertion, although most of the methods discussed above do not facilitate pigmentation prediction.

1.4.2. Forensic DNA phenotyping

With the help of statistical approaches and information on finite set of genetic variants, eye, hair and skin colour can be predicted with reasonable accuracy. In recent times, genotyping has become more affordable which has facilitated decoding the DNA samples retrieved from human remains. Altogether, the prediction of physical appearances (pigmentation is a big part of it) from the biological evidence found at the scene of crime or from unknown missing person is gaining popularity is forensic genetics. This is named as forensic DNA phenotyping (FDP) [124]. This enables the forensic investigators to trim down list of potential suspects with the help of biological substances recovered, also known as DNA 'intelligence'[125].

In the last decade several authors have tried to propose SNP assays to predict pigmentation categories from DNA samples without assessing the ancestry. Ruiz et al 2013[126] proposed an assay of 37 SNPs named Snipper (http://mathgene.usc.es/snipper/), which can predict all three pigmentation phenotypes. It was built adapting an online Bayesian classifier, the models were trained on six populations of north and south Europe. Spichenok et al. 2011[127] proposed a 7 SNP assay for eye and skin colour prediction based on a European majority sample. They used

decision tree method, starting with rs12913832 at the parent node. In a later update the authors added another SNP - rs12896399 (SLC24A4)to the existing 7-plex system and called the latest version 8-plex[128]. The new version was tested on 212 additional individuals and claimed to have substantially improved performance compared to the earlier version. Allwood et al. 2013[129] proposed a decision tree based approach based on 4 SNPs. This eye colour predicting assay was developed on the 101 unrelated individuals from New Zealand. The most well-known and widely used SNP assay is 6 SNP based IrisPlex[130] used for prediction of categorical eye colours. Later Walsh et al 2013[80] proposed a tool with additional 18 SNPs which could simultaneously predict eye and hair colour. The multinomial logistic regression-based approach used 22 SNPs to predict hair colour categories and 6 SNPs to predict eye colour categories. Their latest version of the tool named HIrisPlex-S[20] is also capable of skin colour prediction with total 41 markers (40 SNPs and one INDEL, 36 SNPs for skin colour prediction). This molecular tool has been widely used for the simultaneous prediction of the pigmentation phenotypes on contemporary samples[131–136]. Ancient samples have also been evaluated based on this tool, available online (https://hirisplex.erasmusmc.nl/)[60,137–139]. Except the latest version (HIrisPlex-S), the modelling population used for the IrisPlex and HIrisPlex assays has been collected from different parts of Europe. To predict skin colour more reliably the authors have introduced a small population from non-European ancestry in the study[140].

Most of the pigmentation prediction SNP assays are modelled on majority European samples and therefore it is suspected to have low portability to a genetically distant cohort which has been identified by multiple studies[126,131,141]. Additionally, it has been shown that all the assays perform poorly for the intermediate pigmentation categories[142]. As a resolution, new SNP sets have been proposed to cater the specific ancestries outside the European modelling cohort[126,127,141,143]. Nonetheless, admixed populations are largely unexplored for prediction, which also contains substantial intermediate pigmentation categories. Although FDP has shown promises to be implemented more globally and more commonly in real forensic situations, due to ethical, social and legal issues there has been restricted usage till date[144].

1.5. Ancient DNA and anthropological genetics

Ancient DNA (aDNA) is considered as a vital source of information which can be leveraged to understand the prehistory of human evolution. Human phenotypes are shaped by both genetics and environment; it remains a question about the predictability of the phenotypes given the environmental factors from the past are hard to estimate. Nonetheless, studies have shown that the estimates of the phenotypes in the past can be helpful to interpret evolutionary trends[145].

1.5.1. Ancient DNA and human population history

There are 2 competing theories about the origin of the modern human (*Homo Sapiens*) – multiregional evolution and recent replacement from Africa. According to multiregional evolution (MRE) theory, human species came into being around two million years ago and went through multiple stages of evolution. The same species encompass all the archaic hominin forms – Homo Erectus, Neanderthals till the latest form *Homo Sapiens* [146,147]. On the contrary, recent origin of modern humans have been heavily supported by the more recent genetic studies. It is said that modern human originated in Africa nearly 200,000 years ago (ya). Nearly 50,000-60,000 ya, they started moving out of Africa and went on to populate the globe[148–150]. Recent analyses with genetic data from ancient human samples, contemporary humans and human microbiome has suggested the time-frame for expansion out of Africa, also estimated the distinct migration waves out of Africa and the respective routes taken by those human clusters[149,151] (migration waves have been illustrated in Figure 5).

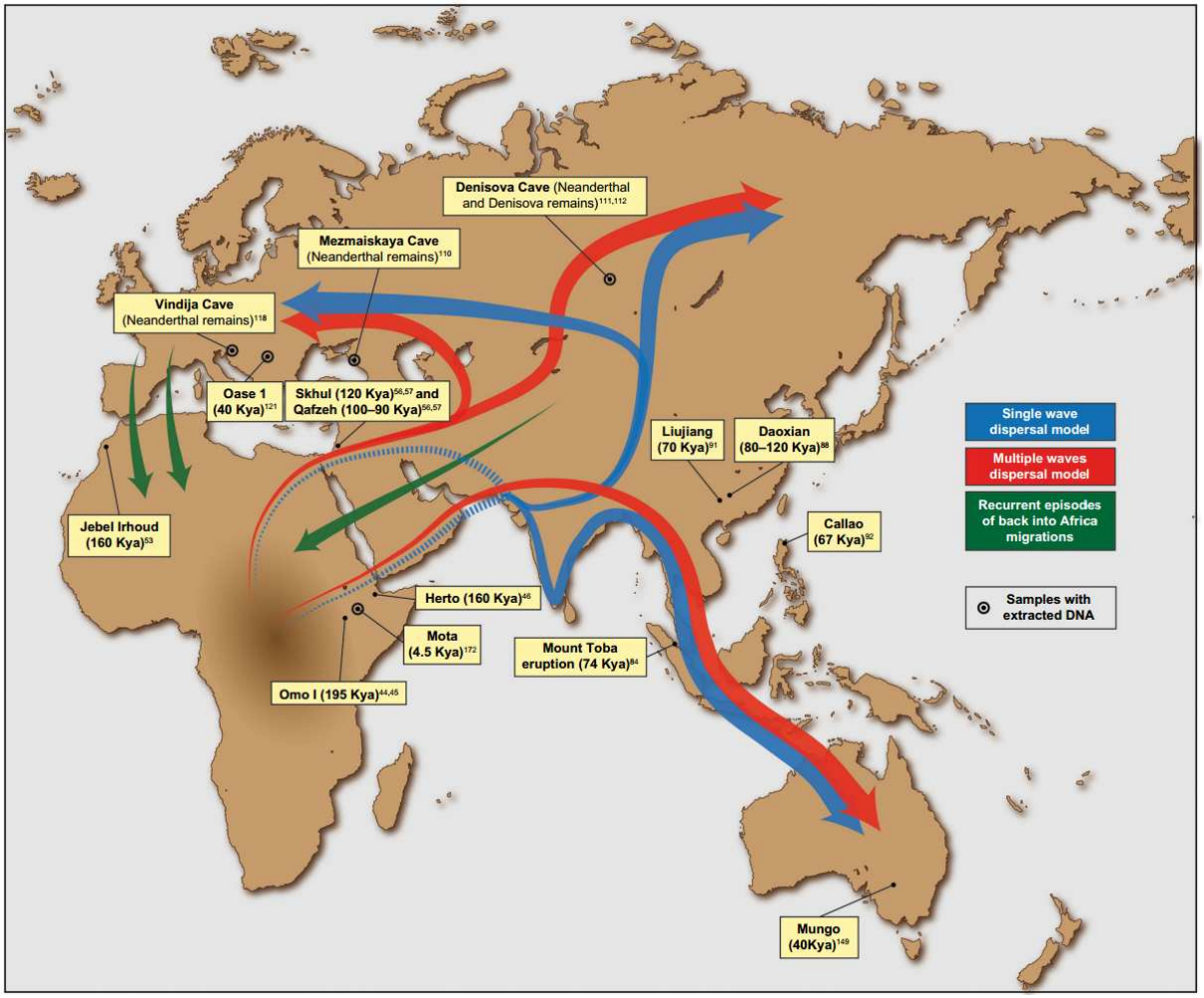


Fig. 5 Putative migration waves out of Africa

The map shows the putative migration out-of-Africa and back migrations into the continent. This also points the locations of major human and ancient hominin remains and archaeological sites (taken from López et al. 2015[152]).

Admixture with archaic hominins were suggested by early studies[153,154], although contradictory approaches were also present[155]. Then with DNA retrieval from three Neanderthal bones from the Vindija Cave in Croatia[156] and small finger bone from the Denisova Cave in southern Siberia[157] (see Figure 5), the replacement model came under revision. It has been shown that present-day individuals from Eurasia share 1-4% of genomes with Neanderthals[156]. Similarly in the Polynesia and south-east Asian population show some admixture (~5%) with the Denisovan population[157,158]. Studies initially claimed it is not because of the introgression from archaic hominins and suggested it could be replicated from the ancient population substructure[159]. Nonetheless the claim was refuted quickly saying the ancient structure alone could not explain the introgression signal[160,161].

Prof David Reich from Harvard University and his team there have curated a large database on archaic hominin and ancient human genomic database (https://reich.hms.harvard.edu/datasets), which is publicly available for analysis. The database contains both the genotype and archaeological information on more than 5,000 samples (in V44) from the history. Due to low-coverage of those samples and degradation of the retrieved genetic material, most of the ancient samples are coded as pseudo-haploids, i.e. only one allele of each of the genetic variants are recorded. We have utilized the database for our analysis of pigmentation evolution in Chapter 3.2.

1.5.2. Ancient DNA and the evolution of human phenotypes

Natural selection acts on a species by changing the allele frequencies in a specific direction such that the favourable phenotypes become more prevalent and helps the individuals get better suited for a certain environment. Thus the variation of the environment forces natural selection to act differently on the population and causes genetic diversity within modern human population which are detectable from present-day genomes[162–164]. Although to ascertain the precise time and situation of those changes, one need to look into the ancient DNA (aDNA) data for most phenotypes[165].

Human pigmentation are polygenic traits and several SNP associations show past evidence of selective sweeps in the contemporary genomes[166,167]. Recently studies have focused on predicting pigmentation phenotypes from the aDNA, although these reconstructions have been concentrated on the samples collected from western Eurasia, due to higher concentration of pigmentation related studies on European cohorts[168] and subsequent studies pointing out the difficulties of portability of trait-associated variants to non-European population cohorts[169,170]. Olalde et al 2014 concluded that lighter skin colour in the continental Europe was not present before the Neolithic age from the Mesolithic hunter-gatherer remains found in central and western Europe[171]. González-Fortes et al 2017 argued that Mesolithic individuals had dark hair, brown eyes and dark skin and Eneolithic individuals had dark hair, light eyes and lighter skin pigmentation from the Mesolithic and Eneolithic remains found in central Europe[172]. Wilde et al 2014

suggested that eye, hair and skin colour associated alleles have been under strong selection forces in Europe for the last 5,000 years[55]. Recently Ju and Mathieson 2021 claimed that lighter skin pigmentation evolution in Europe was due to the strong selection on the large effect sized variants associated with skin melanin, which constitute a rather small proportion of all the associated loci[173].

Studies have explored aDNA data to better understand the evolution through genetics for other human phenotypes as well. Human stature is one such example. Compared to pigmentation, height is even more polygenic with large number of alleles and small effect sizes[174,175]. Based on the genetic components of human height, studies have claimed that ancient west Eurasian population were more differentiated for this trait compared to the present-day habitants of the same region and this is more than what is predicted from genetic drift alone[176–178].

Contrary to human pigmentation or stature, lactase persistence or LP (the continuous expression of the milk sugar lactose-digesting enzyme, lactase into human adulthood) is considered as a single gene trait. It has been under strong selection for the last 10,000 years in multiple human population[179]. Within the Eurasian population, rs4988235-A is the primary allele responsible for LP[180]. Studies have shown that during Bronze and Iron ages this particular allele rose to significant frequency in the population, even though humans had started consuming milk from sheep, goats and cattle long back[176,181,182]. This rapid rise in frequency has been attributed to the migration of Pontic-Caspian steppe population starting 5,000 ya[183,184]. To better understand the reasons and impact of human adaptations, we need to identify the causal genes of the human phenotypes, as through the allele frequency changes of the respective genetic variants we can reliably estimate timelines of evolutionary forces and reconstruct ancient scenarios.

1.6. Identification of genes for pigmentation variation

Human pigmentation, being shaped by evolution, shows remarkable variation both within and between populations. Skin pigmentation has shown strong correlation with ultra violet radiation or latitude[47], whereas, eye and hair colour is particularly extensive in Europe[185]. Pigmentation phenotypes were assumed to be Mendelian traits for a long time, and few instances could be explained using Mendelian assumptions, although the variation and complexity of those traits did not adhere to those rules. Towards the end of last century with the arrival of genetic studies, handful pigmentation associated loci were identified. Once individual genome-wide data became accessible, GWAS started capturing more loci[79,186,187] and explaining the variation of the phenotypes. Nonetheless we are yet to achieve complete understanding of human pigmentation, transcriptomic data might be useful in that regard.

1.6.1. Mendelian genetics

Mendelian inheritance refers to the biological inheritance patterns where the phenotype of an offspring is fully determined from the parental alleles at a single locus. The alleles may have interaction (known as dominance/recessiveness) or may not (known as co-dominance) between them. These patterns follow the rules of segregation and independent assortment, where the gene inherited from either parent segregates into gametes at an equal frequency. Mendel assumed the traits to be influenced by a single gene and each gene to have two alleles[188]. In the initial periods of genetics studies, eye colour was conceived as a mendelian trait: blue was considered as the recessive trait and brown was the dominant trait[189]. This simple model works well in some cases assuming OCA2 working as a blue eye gene. However eye colour has more variation than only blue and brown[80] and Mendelian assumptions does not apply in the broader cases[24,190]. In the case of hair colour, red hair can essentially be considered as a Mendelian trait with 73% of the SNP heritability explained by a single gene – MC1R. Although in the true sense hair colour is not a Mendelian trait either as accounting for more trait-associated genes helps to explain red hair SNP heritability upto 90%. Moreover blonde hair has been associated with more than 200 genetic variants[191]. Human skin pigmentation has been long been ruled out as mendelian trait. Davenport and Davenport (1910)[192] noticed that skin colour does not segregates to the initial colours in the second generation and does not follow mendelism even in the first generation. Nonetheless, extreme phenotypes caused by rare mutations follow Mendelian inheritance patterns[16]. The notable example is oculocutaneous albinism (OCA) which is an autosomal recessive disorder. Till date 7 types of OCA (OCA1-7) have been discovered, out of which Type II (tyrosinase-positive) OCA is the most common albinism in the world[193]. In this case melanin production is affected in the cutaneous tissues and it is the most prevalent type of oculocutaneous albinism in African ancestry population[194]. A common mutation (a 2.7 kb deletion) in chromosomal region 15q11-12 has been attributed to ~78% of OCA2-causing mutations. As a precaution, parental testing has been provided in South Africa since 1970s and couples with high-risk for the recessive deletion have been accordingly advised[195]. More recently, studies have also identified autosomal dominant inheritance pattern of OCA in a Korean family study, where the patients attained normal pigmentation in their late twenties[196].

1.6.2. GWAS methods and results for pigmentation

Eye, hair and skin pigmentation are few of the highly visible phenotypes and hence comparatively to other traits, it is very easy to gather information on those phenotypes. Nonetheless most of our knowledge on human pigmentation has come from studies based on European population. Studies have identified genetic loci and polymorphisms responsible for the variation within European populations or between European and non-European populations. Recent studies have been diversified to include other ancestries under consideration and gene variation in African[197,198], east Asian[199–201], south Asian[202,203] and Latin American[28] has enriched our understanding of the genetic

landscape of pigmentary traits. Biobank level datasets and large-scale meta-analyses using GWAS summary results are still discovering novel associations for pigmentation traits[191,204–207].

GWAS studies have found that HERC2-OCA2, a 357Kb region on chromosomal region 15q13, explains most of the blue and brown eye colour inheritance[78]. The SNP rs12913832 in the intron of HERC2 acts as an enhancer of OCA2 transcription[208]. MC1R gene on the region 16q24, identified as the red hair gene[21,209], also plays a crucial role to protect the melanocytes from the UV damage. UV causes direct damage to pyrimidine bases within DNA mutagenesis unless cleared from the genome[23]. MC1R signalling works as a major discriminant for the amount and type of pigments produced by melanocytes. Under the impact of UV, MC1R signalling increases eumelanin production to protect from further damage[210].

1.6.3. TWAS methods and results for phenotypes

Transcriptome-wide association study (TWAS) is a gene-based association method that inspects associations between genetically regulated gene expression and complex phenotypes[211]. As measuring the gene expression for each of the study individuals is often not feasible, TWAS methods integrate the genotype data with publicly available gene expression datasets (e.g. GTEx database[212]) to impute gene expression levels to investigate causal genetic associations.

GWAS has been a successful method to find the trait associated SNPs[98]. However, majority(~90%) of the statistically significant associations come from the non-coding parts of the human genome[213]. Moreover, the associated genetic variants from the GWAS analyses are often not causal for the trait, but appear more likely due to linkage disequilibrium (LD) structure[214]. It becomes problematic to attach the non-coding variants to downstream affected genes. TWAS helps us to identify the genes associated with respective phenotypes through genetically regulated transcriptional activity. Additionally, the multiple testing burden is lower in TWAS compared to GWAS – while TWAS finds association for ~20,000 genes, GWAS typically works with millions of SNPs[211]. As a result, TWAS has been applied successfully for multiple human traits till date, most of them being diseases, such as breast cancer[215], colorectal cancer[216], pancreatic cancer[217], haematological measurements[218], Alzheimer's disease[219], type 1 diabetes and Crohn's disease[220]. Studies have identified novel and secondary disease associated genetic loci, and in some cases discovered underpowered GWAS loci[215,216]. However, TWAS method has mostly overlooked non-disease phenotypes, except height[221]. Human pigmentation and related traits are yet to get benefitted from this approach. The more advanced TWAS methods utilizing GWAS summary data and leveraging the cross-tissue expression[222] can further enrich our understanding of the pigmentation traits.

1.7. Aim of the thesis

The main aim of the thesis is to assess the potential of prediction methods in predicting phenotypes and gene expression levels from the individual genotypes. Eye, hair and skin colour from the CANDELA cohort[223] (details of this cohort has been provided in Chapter 2) serves as the phenotype examples for this thesis. The added complexity of the 3-way admixture in the population makes the study more intriguing. At the end of this thesis, better understanding of the pigmentation traits should be attained which would also help in better prediction applications.

A. First part of the thesis deals with the issues around prediction of the human pigmentation traits and explore the possibility of extend the model predictions beyond CANDELA individuals. Thus, aim is to
   I. Find the impact of genetic variables on the prediction of human pigmentation and figure out the best set of predictors.
   II. Compare the performance of the statistical and machine learning models and explore the genetic structure of the phenotypes to explain model performances.
   III. Explore the possibility of model portability to the ancestries not represented in the CANDELA cohort.
   IV. Provision for the ancestries not contributing to CANDELA cohort. Additionally, assess the adaptability of the original models (built on diploid data) to low-coverage pseudohaploid data.
   V. Validate the predictions from CANDELA data on out-of-training contemporary datasets (e.g. 1000 Genomes Phase 3, CEPH-HGDP) and extend to the aDNA data to robustly predict the trends of human pigmentation evolution.
B. Second part of the thesis aims to predict gene expression levels from the genotype data of CANDELA individuals. Here, the aim is to
   I. Implement TWAS methods to investigate gene association with pigmentation through predicted expression levels.
   II. Explore the possibility to assemble multiple tissues affecting human pigmentation under TWAS and obtain better understanding of the underlying genomic structure.

2. Materials and methods

In this chapter, I am first going to elaborate the population cohort I have used for this thesis and then will explain statistical and machine learning models used in this thesis. The models were mostly used for phenotype prediction using genotypes. I have explained the basic algorithms, relevance to our analyses, the usage of each of the methods in the literature and discussed the pros and cons of those approaches. I have highlighted the particular R[224] packages I have explored to implement those methods in this thesis.

2.1. Study cohort

This thesis is conceptualised and developed leveraging the data from an international and multidisciplinary collaboration, the Consortium for the ANalysis of the Diversity and Evolution of Latin America (CANDELA) consortium[223] which recruited more than 9,000 volunteers.

2.1.1. Sample collection

Individuals were recruited from five Latin American countries – Brazil, Chile, Colombia, Mexico and Peru, birthplaces of the volunteers have been indicated in Figure 6.

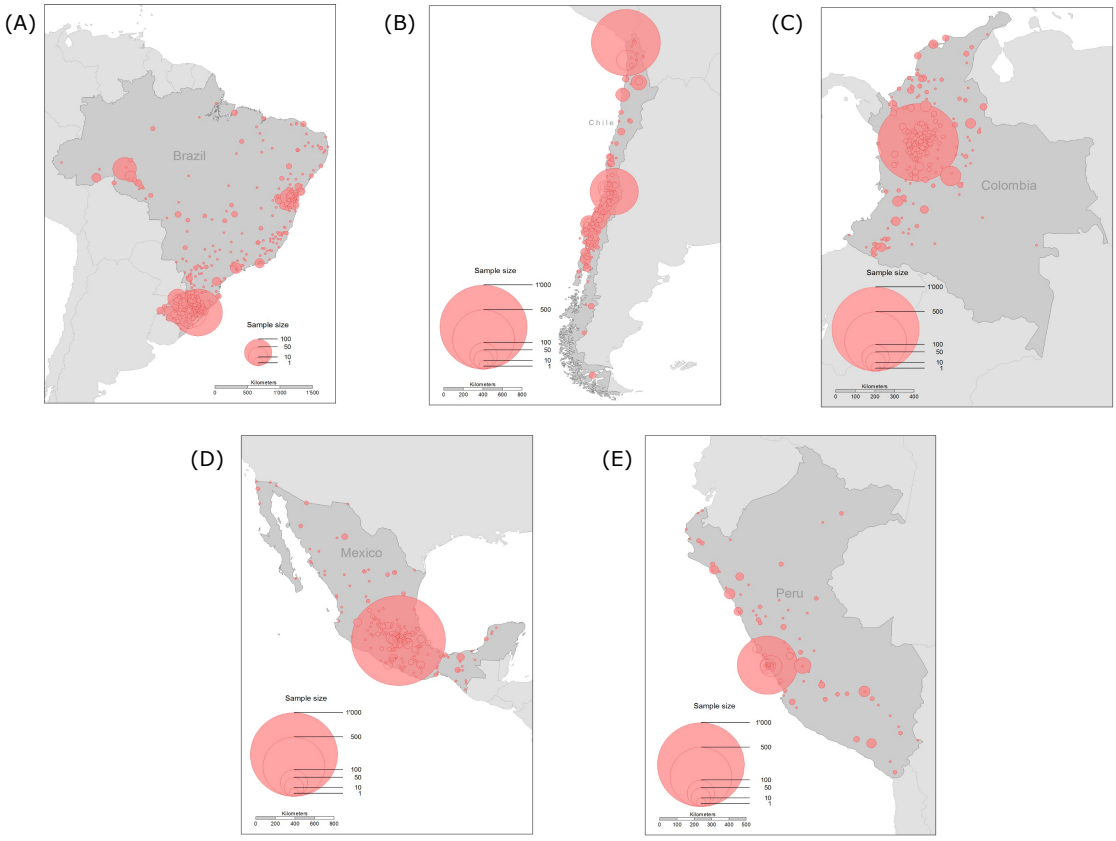


Fig. 6 Birthplace maps of the individuals recruited in CANDELA study
(A)*Brazil,* ***(B)*** *Chile,* ***(C)*** *Colombia,* ***(D)*** *Mexico and* ***(E)*** *Peru (taken from Ruiz-Linares et al. 2014* [223]*).*

Although being a convenience sample, main collection centres for each countries contribute majority of the individuals, Latin America has seen large-scale migration to the cities in recent times. Moreover, the countries covered in the study accounts for ~70% of the total population of Latin America, including three of its most populous countries (Brazil, Mexico and Colombia)[223]. Hence, this sampling can be considered as a representation of the region. For Chile, almost 2/3 of the volunteers were professional soldiers; for rest of the countries, most of them were students and staffs from the participating universities. Ethics approvals were obtained from Escuela Nacional de Antropología e Historia (México), Universidad de Antioquia (Colombia), Universidad Perúana Cayetano Heredia (Perú), Universidad de Tarapacá (Chile), Universidad Federal do Rio Grande do Sul (Brazil) and University College London (UK). All participants provided written informed consent.

2.1.2. Sample phenotyping

Data from the study volunteers were obtained

1. using photographs and measurements recorded on the individuals – this also includes the anthropological measurements. This process was followed using the same protocol and instruments across all recruitment sites.
2. from blood samples collected by a certified phlebotomist. These samples were later used for DNA extraction and subsequent genotyping.
3. from the questionnaires filled by the individuals – this includes queries on age, sex, socio-economic variables, and self-perception of ancestry.

For this thesis, our attention will be limited to three common pigmentation traits – eye, hair and skin colour. Categorical eye and hair colour were obtained from the physical examination of the volunteers. Eye and hair colours were recorded in five (1-blue/gray, 2-honey, 3-green, 4-light brown, 5-dark brown/ black) and four (1-red/reddish, 2-blond, 3-dark blond/light brown or 4-brown/black) categories (sample categories are illustrated in Figure 7), respectively[225]. Although red hair individuals were removed from the prediction analyses, as it is a rare in our sample (frequency less than 0.6%).

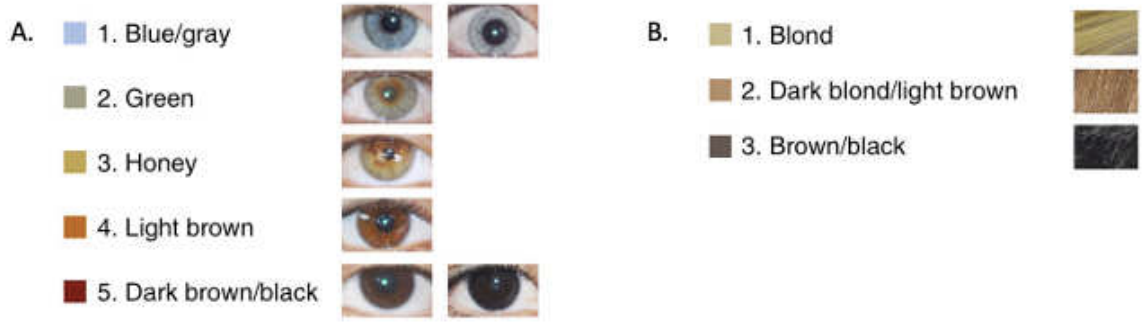


Fig. 7 Pigmentation categories of the CANDELA individuals

The figure represents the approximation of each of the (A) eye and (B) hair colour categories (reference Adhikari et al 2019[28]).

Along with the categorical eye colour assessment, quantitative variables related to the eye colour were obtained from the digital photographs of the individuals. Based on the image quality, one of the two eyes of every individual were selected to derive the quantitative variables. Colour of the iris (it is the ring-shaped membrane around the pupil, the central dark circle in the eye) is presumed to be the colour of the eyes. Therefore, using a graphical interface designed in MATLAB ten landmarks were plotted to extract the visible part of the iris (illustrated in Figure 8). Additional landmarks were plotted to derive the whitest part of the eyes and darkest part of the pupil. Using the extreme white and black references images were normalised and adjusted for the various colour casts or differences in illumination levels. Highlights (e.g. reflections on the iris) were removed using an adaptive threshold. Images were removed from further analysis if errors were found in the digitization steps following individual checks. 5,513 images were finally retained to extract RGB pixel colour values. Out of those images, 195 were selected to assess inter-rater variability of the obtained iris colour and median absolute difference between the raters was 3.3 unites (on a scale of 0-255) on the RGB scale.

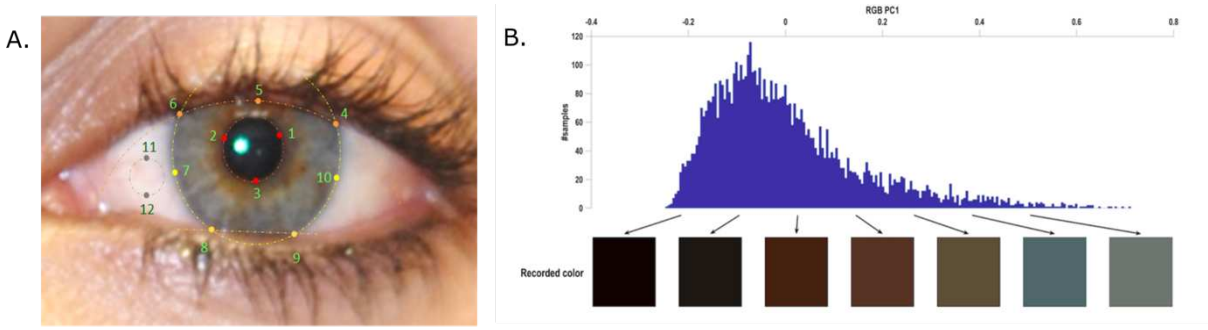


Fig. 8 Quantitative eye colours.

(A)*Landmarks and boundaries drawn on the image of an eye (figure taken from Adhikari et al 2019[28]).* ***(B)*** *The eye colour of the samples, recorded in the HCL space, were converted to the RGB space. The first principal component from the RGB space is used to plot them in a histogram and mapped to the visible colour blocks obtained from the RGB space.*

Although RGB values are convenient for digital imaging and previously been used for genetic association studies[226], they not necessarily agree with human perceptions and biological relevance. Out of the other options for colour spaces, for this thesis HCL (Hue, Chroma, Luminance) scheme was explored. This bicone colour space is directly derived from the RGB and therefore represents the three primary colours – red, green and blue in opposing corners. Moreover, in our study cohort, HCL values exhibit red-orange and cyan-blue colour hues (spread of the colour space is shown in Figure 9) which imparts more information than other colour spaces[28]. Being a circular variable, Hue was linearized with cosine and shifted by 15° so that most of the individuals fall within the range [0, 180°] and thus cos(Hue + 15) will be considered as the one of the quantitative eye colour traits.

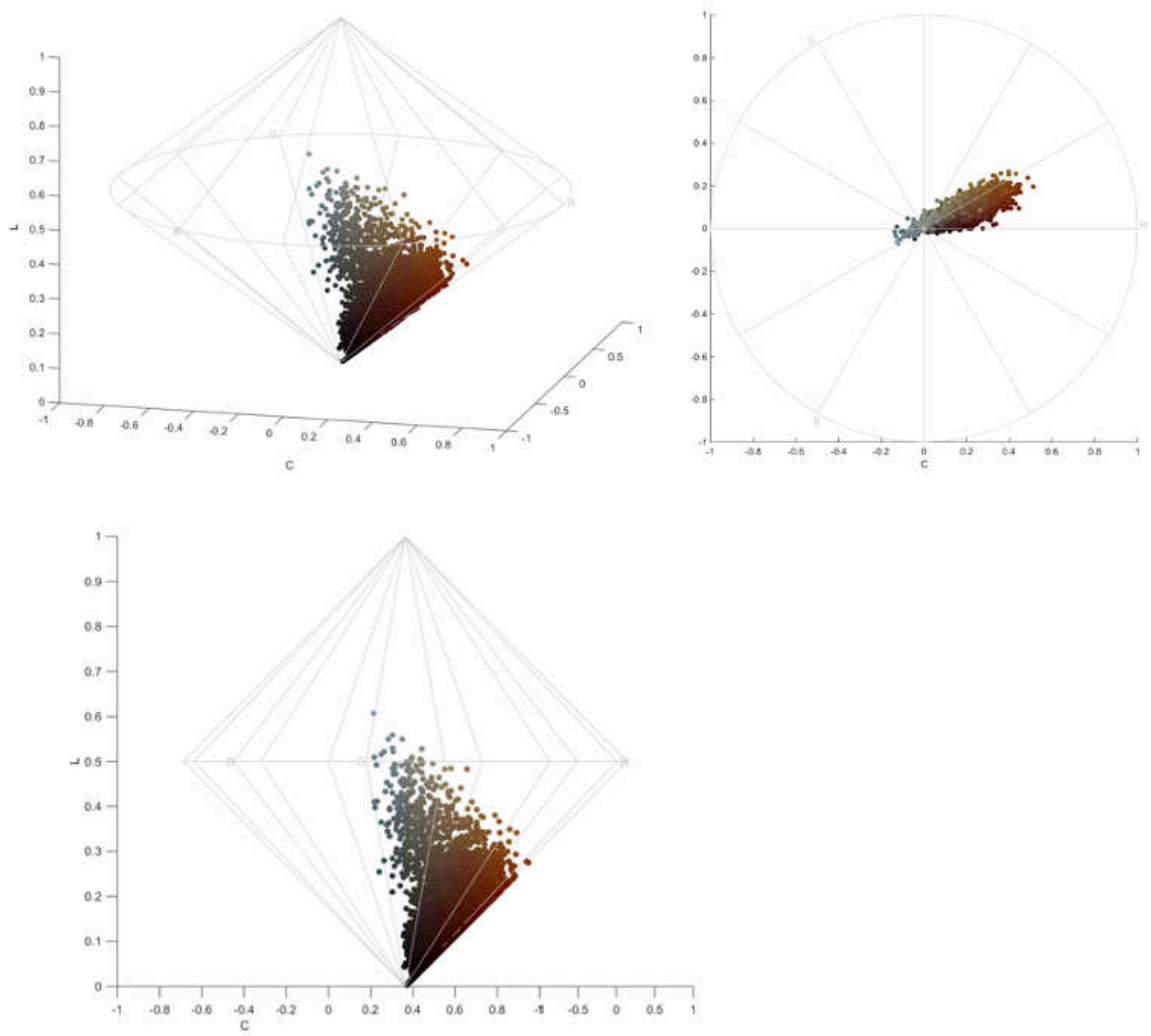


Fig. 9 Bicone HCL colour space of all the retained iris colours
(figure taken from Adhikari et al 2019[28]).

Constitutive skin pigmentation (the MI) was obtained using the DermaSpectrometer DSMEII reflectometer (Cortex Technology, Hadsund, Denmark) during the physical examination of the volunteers. Measurements were obtained from the both inner arms and average of the readings were considered as the final MI. To assess the variability of the measurements, the absolute difference of the readings were analysed and the median variability was 1.03 (in comparison, the final MI range is 20 to 65 for all the CANDELA individuals retained). In Figure 10, we illustrate the variability of MI in the data, converting the MI values to approximate RGB values.

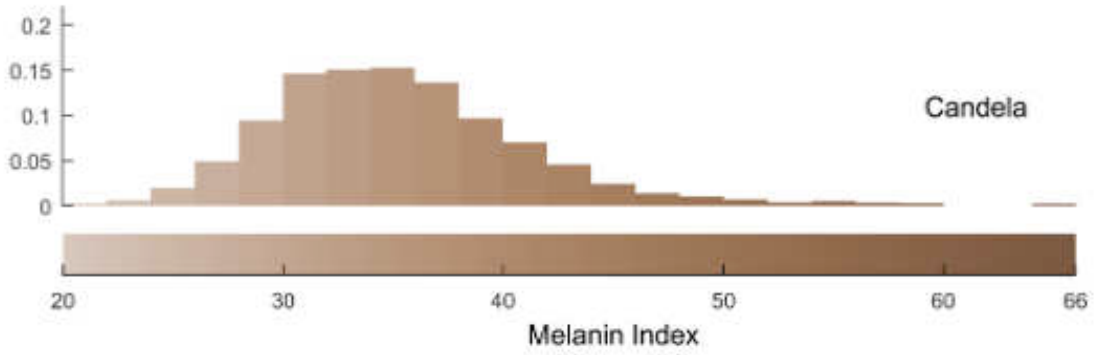


Fig. 10 Skin colour distribution of the CANDELA individuals

Histogram of the skin colour of the full CANDELA cohort. The mapping of the original skin tone and Melanin Index(MI) has been provided in the horizontal bar below. The MI values has been transformed to the RGB values approximately. On the x-axis the proportion of the cohort has been represented (taken from Adhikari et al 2019[28]).

2.1.3. Genotype data

DNA samples from more than 7,000 study volunteers were genotyped on the Illumina HumanOmniExpress chip that includes 730,525 SNPs. The preliminary data quality checks were performed using PLINK v1.9[227] which includes i) removing individuals with more than 5% missing SNPs, ii) removing SNPs with minor allele frequency less than 1%, iii) discarding related individuals estimated through Identity-by Descent (when IBD > 0.1), iv) excluding individuals with failed sex concordance check (recorded sex and X-chromosome heterozygosity estimated sex). At this stage, 669,462 SNPs and 6,35 individuals were remained for further analysis.

The chip genotype data was phased using SHAPEIT2[228] and then imputed using IMPUTE2[229] while using the 1000 Genomes Phase 1 data as the reference. Positions which were monomorphic for the Latin American samples in 1000 Genomes (CLM, MXL and PUR) were excluded, resulting 11,025,002 SNPs being imputed. 48,695 SNPs with poor imputation score (less than 0.4) were removed. IMPUTE2 provided median 'info' score 0.986 for the remaining SNPs as assessment of imputation certainty. With the help of PLINK v1.9, the genotype probabilities from IMPUTE2 were transformed into the most probable genotypes, with uncertainty less than 0.1, as in the default setting. Similar to genotypes, imputed SNPs with more than 5% uncalled genotypes or minor allele frequency less than 1% were also excluded. IMPUTE2 provides a concordance score for the chip genotyped SNPs, by masking the genotypes first and then imputing it from the nearby chip genotypes. SNPs with concordance score less than 0.7 were also removed, similar to the SNPs with large absolute difference in info and concordance values (>0.1). Finally 9,143,600 SNPs were obtained with median concordance value 0.994 for the genotyped SNPs.

An LD-pruned set of autosomal SNPs was used to estimate the individual African, European and Native American ancestry proportions (see Figure 11) using the supervised runs of ADMIXTURE[230], unsupervised runs produced similar results[231]. Along with the genetic variants, we have used estimated continental ancestries to predict phenotypes in Chapter 3.1. The estimated ancestry proportion of the genotyped CANDELA individuals is 51% European, 44% Native American and 5% African (see Table 1 for details split across countries).

Table 1: Sampling of CANDELA cohort

	Sampling count	Genotyped count	Mean % European ancestry	Mean % Native ancestry	Mean % African ancestry
Brazil	1,600	697	0.85	0.09	0.06
Chile	2,045	1,891	0.49	0.48	0.02
Colombia	2,273	1,857	0.62	0.29	0.09
Mexico	1,682	1,290	0.39	0.58	0.03
Peru	1,467	1,291	0.31	0.66	0.03
CANDELA	9,067	7,026	0.51	0.44	0.05

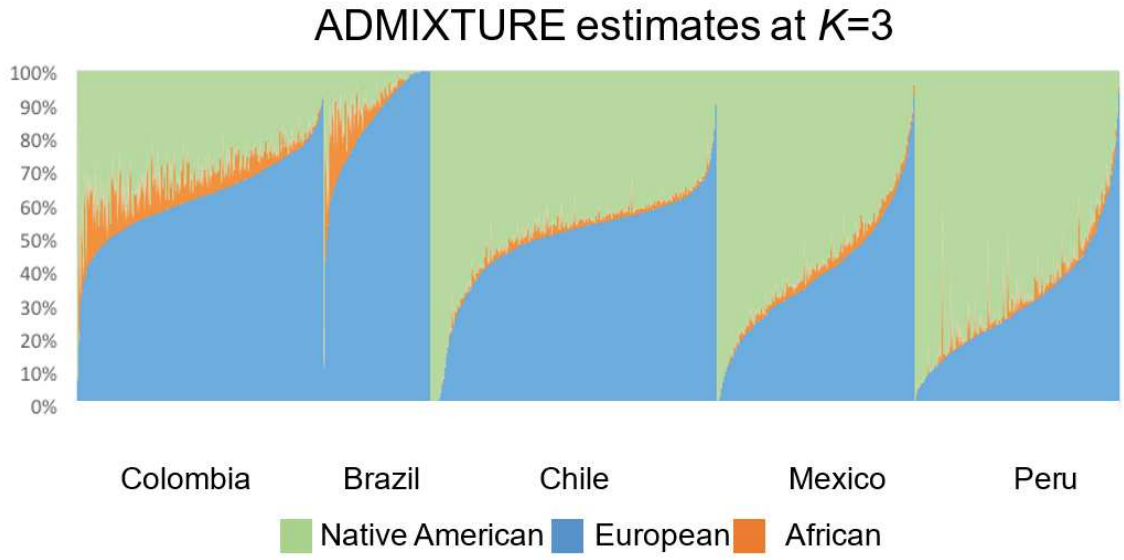


Fig. 11 Continental ancestry in the CANDELA individuals

Continental ancestry estimated from 160,858 autosomal SNPs (LD pruned) using supervised runs of the ADMIXTURE software – reference population of African , Native American, European, East Asian groups were taken from 1000 Genomes Phase 3 project (figure taken from Adhikari et al 2019[28]).

If a fourth component is allowed for the CANDELA individuals (K=4), we can observe an additional contribution from the East Asian ancestry (Figure 12). Further analysis

revealed there were 185 individuals in the cohort with more than 5% East Asian ancestry. Although cross-validation on ADMIXTURE points out K=3 is a better modelling choice, we shall use K=4 outputs and discuss why it is advantageous for us in that situation in Chapter 3.2.

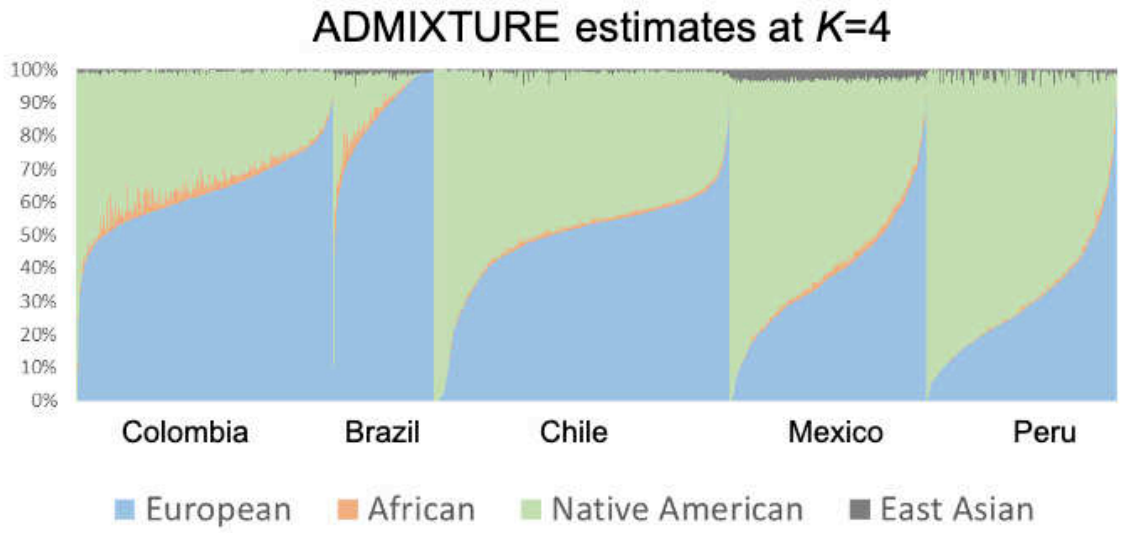


Fig. 12 Continental admixture in the CANDELA individuals with additional East Asian ancestry

(figure taken from Adhikari et al 2019[28]).

2.2. Prediction algorithms

Here I explain the statistical and machine learning models which we explored as the predictive algorithms using the genotype and phenotype data from the CANDELA study. While linear regression and stepwise regression are the statistical models used for prediction on the continuous phenotypes (e.g. skin melanin index, quantitative eye colours – HCL), multinomial logistic regression and ordinal regression have been explored for categorical phenotypes (e.g. categorical eye and hair colour). Machine learning models have been used in the prediction literature although there demand for larger datasets to capture the complex phenotypes have seen limited usage[232,233]. The large CANDELA cohort provides a perfect opportunity to explore few of the advanced models, including neural net based deep learning models. We assessed the performance of random forest, extreme gradient boosting and artificial neural network on the all the pigmentation phenotypes with distinct instructions for quantitative and categorical traits.

Once we obtain the genotype-regulated gene expression for respective tissues, penalised regression methods will be helpful to find the expression-associated genes. Least absolute shrinkage and selection operator (LASSO) and elastic net methods come

handy to select a handful of significant genes from a large set of candidates. We explain the model algorithms in detail here and use them extensively in Chapter 3.1, 3.2 and 3.3.

2.2.1. Linear Regression (LR)

Linear Regression is an approach to model the quantitative target variable and set of independent variables through a linear relationship, i.e. the model is linear in the parameters [234]. The independent variables can be either categorical or quantitative. In some cases, the quantitative variables may first be transformed by applying a mathematical function (log, power functions etc).

Let us assume there are *p* input variables, $X^T = (X_1, X_2, \ldots, X_p)$ and the dependent outcome variable is Y. If the dependent variable is linearly related to the input variables, then we can express Y as a function of X, i.e.

$$Y = f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j \tag{2.1}$$

As the relationship of X and Y is said to be linear, so the coefficients (β_j) are all linear.

Now, if the observed data comes in the form (x_1,y_1), (x_2,y_2), …, (x_n,y_n) for n samples where each x_i is a vector containing *p* elements $(x_{i1},x_{i2},\ldots,x_{ip})^T$, for i = 1,2,…,n. The coefficients $\beta = (\beta_0,\beta_1,\beta_2,\ldots,\beta_p)^T$ is obtained through minimising the residual sum of squares (RSS), where

$$RSS(\beta) = \sum_{i=1}^{n} (y_i - f(x_i))^2 \tag{2.2}$$

And the minimization yields the estimates of the parameters,

$$\hat{\beta} = (X^T X)^{-1} X^T y \tag{2.3}$$

The solution of β exists and is unique if $(X^TX)^{-1}$ exists. In actual genomic studies, we often find one independent variable is associated with one or more variables within the dataset. Therefore in such cases $(X^TX)^{-1}$ does not exist arising the problem of multicollinearity. To get rid of the problem, we need to find independent components that explains the variability of the matrix X or use alternative regression methods (e.g. LASSO) which can tackle the rank deficiency of the matrix (which arises due to presence of dependent columns within the matrix).

LR provides a strong starting point for the classical genome-wide association study (GWAS) setups where the phenotype in consideration is the output vector (y in equation

(2.1)) for n individuals and covariates (e.g. age, sex or vectors explaining population structure) and genetic variant or one single nucleotide polymorphism (SNP) as the input variables (X_j's in equation (2.1)). Typical equations look like below:

$$phenotype(y) \sim covariates(age, sex) + genetic\ components\ explaining\ population\ structure + SNP_i \quad (2.4)$$

where *i* denotes the ith SNP from a set of N SNPs genotyped/imputed for the population [235]. So for each of the SNPs the model gives an estimates of the corresponding coefficient of the SNP and also the significance of the association between the SNP and the phenotype.

On the other hand, LR is also quite useful as this is a supervised algorithm. Once the coefficients have been estimated, we can use the $\hat{\beta}$ vector to predict on the unobserved data. If we have information on the input variables Xi and the coefficients are estimated, then using equation (2.1) we can estimate the output vector $\hat{y}$.

For our analysis, we used the lm function from the stats package in R [224]. It offers the basic LR but if specified separately, it can also take care of the weighted samples. For our Chapter 3.1, we used the SNPs as predictors and phenotypes as output and checked for accuracy from the lm predicted results. For our Chapter 3.2, we specified sample weightage based on the genetic ancestry of the individuals. The simplistic algorithm, easy interpretability and fast runtime makes this method an obvious choice.

2.2.2. Stepwise Regression (SR)

Contrary to LR, Stepwise Regression does not necessarily keep all the predictors in the model. In a sequential search it either adds the best available variable or removes the least informative one. SR can work in two directions, namely forward or backward.

Forward stepwise regression starts with the intercept and then searches for the best available variable in terms of improvement in the fit. Once the first variable is selected, the model again searches from the remaining variables that can improve the fit further, and only the best among the rest is added to the existing model and it goes on searching until the point an addition does not minimise the AIC criteria. This is a greedy approach where at each step we look for the best option available which can be worse than a combination from the global set of predictors. But with increasing number of predictors this random search can be highly complex and time consuming, where forward and backward selection provides a comparable solution.

The backward stepwise selection starts with all the variables in the model. It sequentially searches for the one variable that has least effect on the fit of the model. This method also stops when further removing a predictor does not reduce the AIC. Compared to the Forward stepwise approach has one advantage over this method. Backward selection is restricted by the number of individuals (n). In case the number of parameters (p) is more than n, then the method cannot estimate the contribution of the predictors and hence the selection method stops. Forward approach can still work in this situation.

We used the R package in 'caret' which in turn uses the MASS package for the function "stepAIC" in Chapter 3.1 to perform stepwise regression. It is in fact a combination of both forward and backward methods. Stepwise regression or specific cases of forward or backward stepwise methods have been used in genetics for over two decades now [236,237] but it is also quite frequently been used in phenotypic prediction for the forensic science [76,78] for its usage in feature reduction and effective variable selection.

Like the quantitative phenotypes, categorical phenotypes were also explored for their predictability. Multinomial logistic regression is the adaptation of the logistic regression for the multi-level dependent variable, which models the log odds of outcomes.

2.2.3. Multinomial Logistic Regression (MLR)

Logistic regression is a linear model useful in the case of a binary target variable. Multinomial logistic regression is the generalisation when the target variable can assume multiple levels. We model the log odds as the combination of the independent variables X_i, e.g.

$$\ln \frac{\Pr(Y_i = p)}{\Pr(Y_i = q)} = \beta_{pq}.X_i \tag{2.5}$$

where the target variable Y_i assumes values p and q.

For ease of calculation, we pivot one level of the dependent variable and formulate the rest log odds of rest of the levels against the pivot level. If the last outcome (say, M) is the pivot, then the equations would look like:

$$\begin{aligned} \ln \frac{\Pr(Y_i = 1)}{\Pr(Y_i = M)} &= \beta_1.X_i \\ \ln \frac{\Pr(Y_i = 2)}{\Pr(Y_i = M)} &= \beta_2.X_i \\ &\dots \\ \ln \frac{\Pr(Y_i = M-1)}{\Pr(Y_i = M)} &= \beta_{K-1}.X_i \end{aligned} \tag{2.6}$$

Now as the sum of probabilities for all the M levels is 1, we can use this information to equate the above probability equations.

To deduce the coefficient parameters for the equations we need to maximize the conditional likelihood function for the for multinomial distribution. For simplicity we consider a binomial setup here. So the log-likelihood function would look like below:

$$l(\beta) = \sum_{i=1}^{n} (y_i \beta^T x_i - \log(1 + e^{\beta^T x_i}))$$

(2.7)

where β is the parameter vector for the independent variables. Here β is of size 2×1 and the a column within X is a vector of 1 corresponding to the intercept. To estimate the coefficient vector β, we need to differentiate the likelihood function it does not produce any closed form equation. So we implement Newton-Raphson iterative algorithm to obtain the final estimates based on some convergence criteria.

We have used the 'multinom' function nnet v7.3.12 package on R. This method has already been used for categorical phenotypes [238]. The Eye, Hair and Skin colour prediction models implemented in the HIrisPlex-S models uses the same algorithm [20].

2.2.4. Ordinal regression (OR)

Ordinal regression is an extension of MLR when the target variable is also ordered, i.e. the target variable can assume any value but the levels have a relative ordering. In Chapter 3.1, where we consider a categorical skin colour variable, the individuals can have any value on the melanin index scale within the levels but the relative ordering of fair, intermediate and dark skin is maintained.

If the ordinal target variable Y assumes M levels, then the odd of being less than a particular ordered level m can be defined as

$$\frac{P(Y \leq m)}{P(Y > m))}$$

(2.8)

for p = 1,2,3,…,M-1 since P(Y < 1) = P(Y ≥ M) = 0. The ordinal logistic regression model can be represented as

$$logit\big(P(Y \leq m)\big) = \beta_{m0} + \beta_{m1} x_1 + \cdots + \beta_{mp} x_p$$

(2.9)

where m assumes values 1,2,…,M-1 and assuming p independent predictors available. As the ordered levels are parallel to each other the slopes are same across levels but the intercept term differs. Hence the simplified equation (2.12) will look like

$$logit\big(P(Y \leq m)\big) = \beta_{m0} + \beta_1 x_1 + \cdots + \beta_p x_p$$

(2.10)

We used R function 'polr' from the package MASS [239] version 7.3.51.4 for the ordinal regression analysis. This model has been used in previously for the ordered variables [78] but we have also seen at times ordinal traits being treated as a continuous variable [28].

Along with the classic statistical models, more advanced machine learning models were also explored thanks to their prediction accuracy in various scenarios. For our prediction purposes we used broadly two kinds of methods – tree-based ensemble models where multiple trees are built either parallelly or sequentially and deep learning models where multiple nodes and edge weights are estimated for a neural network with hidden layers.

2.2.5. Ensemble Learning models

Ensemble learning builds on the approach of aggregating weaker models to form better prediction. Simpler models are trained on the full or part of the data and combining each of models to predict on the full data. There are two well-known approaches to this method – boosting and bagging. Boosting works to decrease bias and bagging works towards decreasing variance. Boosting works sequentially, where each subsequent model depends on the previous models. New model tries to amend the errors of the previous models. Bagging (bootstrap aggregation) models work parallelly where each basic model is built independently of the other such models. At the end when all the models are finalised, the aggregated result is taken as the outcome. We have used one models from each method for our analysis. Extreme gradient boosting was explored as part of the boosting approach where random forest represented bagging algorithms.

2.2.5.1. Extreme Gradient Boosting (XGBoost)

XGBoost [240] is a fast and scalable implementation of the gradient boosting decision tree algorithm. The basic aim of the model is to minimise the overall loss function. We first fit one weak model to the full data and calculate the first loss function. To address the loss function, we add a new predictor which would try to minimise it. This is a cyclic method where at each step we build a new tree based model to the capture the residual and add to the existing ensemble model and predict on all the samples. Then calculate the loss function again and try to capture the remaining error with a subsequent tree.

As we are trying to fit more and more weak models to capture the remaining traces of errors, we become prone to overfit to the training data. There are some parameters in the package ‘xgboost’ (in R) that helps to circumnavigate this issue. 'n_estimators’ dictates the number of sequential trees to be added in the full ensemble model. If the number of trees in the model is too high, then it performs well on the training data but fails to replicate the same on the hold-out test data. On the other hand, if the number the trees remain too low, then the model does not learn on the training data and consequently fails on the test data. ‘Early_stopping_rounds’ is another parameter which keeps a check on the overfitting problem. It mandates the consecutive number of steps the model does not improve the prediction on the validation data. ‘Learning_rate’ as a parameter does not control the number of trees but it helps to overcome overfitting. This parameter adds noise to the

prediction at each step, so the subsequent trees added to the ensemble performs less. Thus the whole process is slowed but the model performs better on the new data.

This method has already been used in genetics for gene expression prediction [241], RNA regulatory axis prediction [242], ancestry inference using multi-InDel [243] etc but we could not find literature of usage on phenotype prediction.

2.2.5.2. Random Forest (RF)

Random forest is a bagging ensemble model where multiple decision trees are trained simultaneously of bootstrapped data [244]. The model can be used for regression where the average of the prediction from each tree is taken as the final prediction. For classification problems each tree predicts a level for the categorical variable and the modal voting is chosen as the predicted class. RF has a significant advantage to XGBoost in terms of reconstruction of the models and interpreting the variable importance. Since the publication this method has been used to predict phenotypes [245], to identify variants [246,247] or genes and even to understand the complex genetic interactions [90,248].

The model creates multiple folds of the data, the subsets are not exclusive but exhaustive. A decision tree is built on a particular subset with the aim:

I. Select a random subset of variables from the full set.
II. Find the split for each of the variables and the note the resulting nodes.
III. Choose the variable with most significant split.

The same method is applied for the consecutive nodes of the tree until the node size reaches the specified minimum samples requirement by the parameter 'nodesize'. We have used the package 'randomForest' version 4.6-14 for the modelling in R. The important parameters we used for the training and to avoid overfitting are 'ntree' and 'mtry'. 'ntree' determines how many trees would be there in the forest and 'mtry' dictates the maximum number variables sampled before selecting the best split at each level. We searched over all the combinations of ntree and mtry options to find the optimal parameter choice for our models.

Allowing larger trees tend to overfit; to avoid that we need to prune the trees mandating certain sample size in the terminal nodes. Also the model itself does a validation on the hold-out samples to tackle overfitting. Though averaging over the output from each of the trees makes the prediction robust and reduces bias. RF has an advantage from its independent implementation as the trees can grow independently on their own., hence training the trees can be parallelized. Apart from prediction, we have also used this model to select significant variables for our studies.

2.2.6. Artificial Neural Network (ANN)

Artificial neural network is a connected graph of nodes across multiple layers where a data input is translated into a desired output. ANNs are made of three components – input layer, hidden layer(s) and output layer. This method can be implemented for multi-level classification problems where the output layer should have the exact number of nodes as

the class levels of the target variable. For a regression setup, the output layer will generally have single node for quantitative prediction, though these networks function well with multiple response nodes.

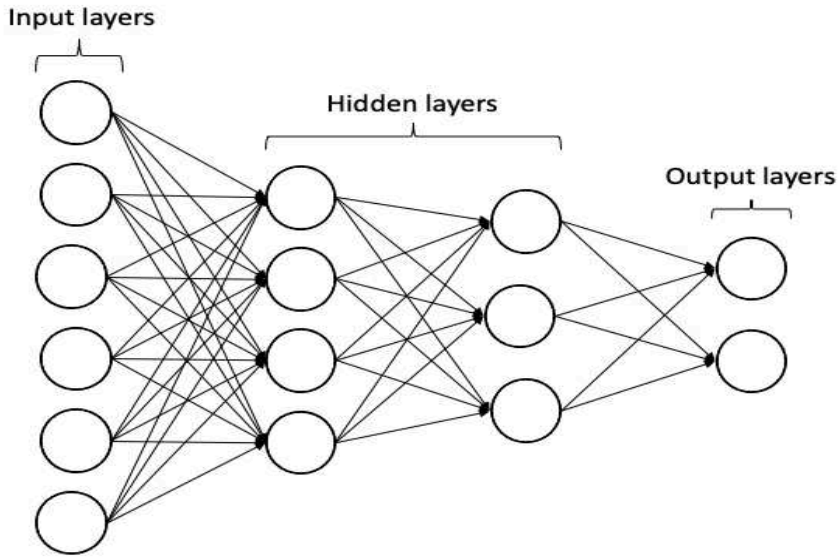


Fig. 13 Schematic of a double hidden layer artificial neural network

In figure 13, there is 6 input nodes, 2 hidden layers with 4 and 3 nodes respectively and a double node output layer. Each node is connected to both previous and next layers except for the extreme layers who are connected in only one direction.

Backpropagation: Each directional edge connecting two nodes carries a weightage which is calibrated to minimise the error calculated from actual and predicted values. This method is known as backpropagation where the weights are recursively updated and error terms are recalculated. There are several options in the R package 'Keras' version 2.2.5.0 for optimisation function. It involves the learning rate which controls the speed of convergence in backpropagation.

Activation Functions: For each of the neurons or edges connecting two nodes there is a activation function or a mathematical equation determining whether an information will be passed on to the next node. The feed forward nature of the neural network is leveraged send information from the initial input layers to the end layers and some information often does not pass through certain nodes as they are not 'activated'. The activation function introduces the non-linearity in the model. Among the several available options we used 'relu' and 'tanh' for continuous traits and additionally used 'softmax' for categorical target variables.

Neural networks are widely used in genetics today, from medical and healthcare to phenotype prediction in forensics , this is generally considered one of the most promising areas for future research. Though interpretability often hinders the progress of the promising models [253], work has been going on to increase intelligibility of the same [254].

2.2.7. Least Absolute Shrinkage and Selection Operator (LASSO) & Elastic Net

Lasso is a regression method used for both regularization and feature selection. In case of the predictor variable matrix not being full column rank, LASSO shrinks the coefficients of the correlated predictor variables by imposing a L_1 (coefficients are of order 1) penalty constraint for the shrinkage as below[234]

$$\hat{\beta}^{lasso} = \underset{\beta}{argmin} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \tag{2.11}$$

$$subject\ to \sum_{j=1}^{p} |\beta_j| \leq t \tag{2.12}$$

This subject condition can also be changed to L_2 (coefficients are of order 2) penalty norm and that method is called Ridge regression. In the case of data having more parameters to estimate than the number of samples ("p>n problem" where the coefficients cannot be uniquely determined) LASSO selects at most n covariates, even if more predictor variables contribute to the target variable. In a group of highly correlated predictor variables LASSO selects only one at random. Also in n>p situation, if the predictor variables are highly correlated ridge egression performs better than LASSO [255]. To tackle the issues, elastic net was proposed, which can be conceptualised a convex combination of LASSO and ridge regression. The constraint function (same as equation 2.15) becomes

$$\alpha \sum_{j=1}^{p} \beta_j^2 + (1-\alpha) \sum_{j=1}^{p} |\beta_j| \leq t \tag{2.13}$$

LASSO and elastic net has been used in the GWAS studies to prioritise genetic variants [256,257] or in the set of related individuals [258]. We have also encountered both methods in the transcriptome-wide genetic study approach named PrediXcan [259].

Within the methods discussed above, I have used linear regression, stepwise regression, multinomial logistic regression, ordinal regression, extreme gradient boosting, random forest and artificial neural network in Chapter 3.1, linear regression (with additional weightage scheme) in Chapter 3.2 and least absolute shrinkage and

selection operator (LASSO) & elastic net methods in Chapter 3.3. LR, SR and all three machine learning models were used for quantitative trait prediction – skin colour and eye colour hue, chroma and luminance, whereas, MLR, OR, along with the machine learning algorithms were explored for categorical eye and hair colours. LASSO and elastic net were utilised within the TWAS approach to find the gene expression associations in three cutaneous tissues.

3. Results

3.1. Article 1: Prediction of eye, hair and skin colour in Latin Americans

This article was published in Forensic Science International Genetics in July 2021 (article and supplementary materials in their published forms are available here (https://doi.org/10.1016/j.fsigen.2021.102517). Supplementary information is available in Appendix A. In this article my contributions are:

- Data preparation:
    - Phenotype quality control,
    - Samples quality control,
    - Genotype data curation
- SNP set preparation
- Prediction model preparation:
    - Statistical and machine learning model training,
    - Prediction on CANDELA individuals,
    - Comparison of model performance
- Impact of model portability
- Prediction on the native individuals
- Draft preparation and writing

Authors and affiliations

Sagnik Palmal[1*], Kaustubh Adhikari[2,3*], Javier Mendoza-Revilla[4,5], Macarena Fuentes-Guajardo[6], Caio C. Silva de Cerqueira[7], Betty Bonfante[1], Juan Camilo Chacón-Duque[8], Anood Sohail[9], Malena Hurtado[4], Valeria Villegas[4], Vanessa Granja[4], Claudia Jaramillo[9], William Arias[10], Rodrigo Barquera Lozano[11,12], Paola Everardo-Martínez[11], Jorge Gómez-Valdés[11], Hugo Villamil-Ramírez[13], Tábita Hünemeier[14], Virginia Ramallo[15,16], Rolando Gonzalez-José[16], Lavinia Schüler-Faccini[15], Maria-Cátira Bortolini[15], Victor Acuña-Alonzo[11], Samuel Canizales-Quinteros[13], Carla Gallo[4], Giovanni Poletti[4], Gabriel Bedoya[10], Francisco Rothhammer[17,18], David Balding[3,19], Pierre Faux[1#], Andrés Ruiz-Linares[1,3,20#]

† Contributed equally to this work.

* These authors jointly supervised this work.

[1] UMR 7268 ADES, CNRS, Aix-Marseille Université, EFS, Faculté de Médecine Timone, Marseille, 13005, France.

[2] School of Mathematics and Statistics, Faculty of Science, Technology, Engineering and Mathematics, The Open University, Milton Keynes, MK7 6AA, United Kingdom.

[3] Department of Genetics, Evolution and Environment, and UCL Genetics Institute, University College London, London, WC1E 6BT, UK

[4] Laboratorios de Investigación y Desarrollo, Facultad de Ciencias y Filosofía, Universidad Peruana Cayetano Heredia, Lima, 31, Perú.
[5] Unit of Human Evolutionary Genetics, Institut Pasteur, Paris, 75015, France.
[6] Departamento de Tecnología Médica, Facultad de Ciencias de la Salud, Universidad de Tarapacá, Arica, 1000000, Chile.
[7] Scientific Police of São Paulo State, Ourinhos-SP, 19900-109, Brazil.
[8] Division of Vertebrates and Anthropology, Department of Earth Sciences, Natural History Museum, London, SW7 5BD, UK.
[9] Department of Biotechnology, Kinnaird College for Women, 93 - Jail Road, Lahore, 54000, Pakistan.
[10] GENMOL (Genética Molecular), Universidad de Antioquia, Medellín, 5001000, Colombia.
[11] National Institute of Anthropology and History, Mexico City, 6600, Mexico.
[12] Department of Archaeogenetics, Max Planck Institute for the Science of Human History (MPI-SHH), Jena, 07745, Germany.
[13] Unidad de Genomica de Poblaciones Aplicada a la Salud, Facultad de Química, UNAM-Instituto Nacional de Medicina Genómica, Mexico City, 4510, Mexico.
[14] Departamento de Genética e Biologia Evolutiva, Instituto de Biociências, Universidade de São Paulo, São Paulo, SP, 05508-090, Brazil.
[15] Departamento de Genética, Universidade Federal do Rio Grande do Sul, Porto Alegre, 90040-060, Brasil.
[16] Instituto Patagónico de Ciencias Sociales y Humanas, Centro Nacional Patagónico, CONICET, Puerto Madryn, U9129ACD, Argentina.
[17] Instituto de Alta Investigación, Universidad de Tarapacá, Arica, Arica 1000000, Chile.
[18] Programa de Genetica Humana, ICBM, Facultad de Medicina, Universidad de Chile, Santiago, Arica 1000000, Chile.
[19] Melbourne Integrative Genomics, Schools of BioSciences and Mathematics & Statistics, University of Melbourne, Melbourne, VIC 3010, Australia.
[20] Ministry of Education Key Laboratory of Contemporary Anthropology and Collaborative Innovation Center of Genetics and Development, School of Life Sciences and Human Phenome Institute, Fudan University, Yangpu District, Shanghai, China.

Abstract

We report an evaluation of prediction accuracy for eye, hair and skin pigmentation based on genomic and phenotypic data for over 6,500 admixed Latin Americans (the CANDELA dataset). We examined the impact on prediction accuracy of three main factors: (i) The methods of prediction, including classical statistical methods and machine learning approaches, (ii) The inclusion of non-genetic predictors, continental genetic ancestry and pigmentation SNPs in the prediction models, and (iii) Compared two sets of pigmentation SNPs: the commonly-used HIrisPlex-S set (developed in Europeans) and novel SNP sets we defined here based on genome-wide association results in the CANDELA sample. We find that Random Forest or regression are globally the best performing methods. Although continental genetic ancestry has substantial power for prediction of pigmentation in Latin Americans, the inclusion of pigmentation SNPs increases prediction accuracy considerably, particularly for skin color. For hair and eye color, HIrisPlex-S has a similar performance to the CANDELA-specific prediction SNP sets. However, for skin pigmentation the performance of HIrisPlex-S is markedly lower than the SNP set defined here, including predictions in an independent dataset of Native American data. These results reflect the relatively high variation in hair and eye color among Europeans for whom HIrisPlex-S was developed, whereas their variation in skin pigmentation is comparatively lower. Furthermore, we show that the dataset used in the training of prediction models strongly impacts on the portability of these models across Europeans and Native Americans.

3.1.1. Introduction

There is growing interest in the use of genetic data for the prediction of physical appearance, particularly in forensic, historical and paleo-anthropological studies[1–3].

Strong impetus for these studies has been provided by Genome Wide Association Studies (GWAS) of traits such as pigmentation, which have enabled the identification of single nucleotide polymorphisms (SNPs) associated with eye, hair and skin color[4–6]. Early prediction studies evaluated the accuracy of predicting eye color from SNP data[7–10] and these analyses have been more recently extended to hair [11,12], and skin pigmentation[13]. As a result, sets of SNPs have now been proposed for the simultaneous prediction of eye, hair and skin color[14,15]. However, so far, these studies have had a strong European bias[7,10,11,14,16]. The pigmentation SNPs for prediction have been selected from GWAS performed mainly in Northern Europeans[4,5,17] and prediction models also optimized mostly in European datasets[17–21]. Furthermore, evaluation of the accuracy of these prediction tools has also been performed essentially in European-derived population samples[18–20,22].

Among major continental populations, Europeans show the largest variation in eye and hair color, whereas the variation in skin color in Europeans is relatively lower compared to other continental regions such as Africa or South Asia[23–28]. The other continental regions have narrower variation in eye and hair color, mostly within the brown-black range; although fine-scale quantitative measurement still allows us to discern this variability and identify genetic variants unique to those populations[27,29–31]. Interestingly, even though the underlying biological mechanisms are broadly similar for these different types of pigmentations, their genetic architecture is substantially different, as eye and hair pigmentation are chiefly controlled by a handful of large-effect variants, whereas skin pigmentation is observed to be more polygenic and the impact of any single variant is smaller[6,16,24–27]. Consequently, studies assessing the portability of pigmentation prediction systems derived primarily on European cohorts have reported good portability for eye and hair phenotypes in other non-European continental populations, where most samples were correctly predicted to be in the brown-black category, whereas portability for skin color in other continental populations was poorer[23,32,33]; for example, all light-skinned Japanese individuals were incorrectly predicted to be of dark skin color[23]. The convergent evolution of lighter skin color in different regions of Eurasia, as reported previously in the CANDELA cohort[31] and elsewhere[26,34,35], is the likely cause for this discrepancy, and highlights the importance of using diverse cohorts for building skin pigmentation prediction models.

Latin Americans represent one of the largest recently admixed populations world-wide. The history of Latin America has involved extensive admixture, particularly between Native Americans, Europeans and sub-Saharan Africans (with respective estimates of ancestry share: 38%, 52% and 6%[36]). Consistent with its partly European ancestry, a recent GWAS for pigmentation traits in Latin Americans in the CANDELA cohort detected phenotypic effects for a number of loci previously identified in Europeans[31]. In addition to these, novel pigmentation SNPs with genome-wide significant association were also identified in that study. These included SNPs polymorphic only in East Asians and Native Americans, consistent with the independent evolution of skin pigmentation in West and East Eurasia[26,34,35]. The admixed ancestry of Latin America and the finding in the region of pigmentation variants not present in

Europeans emphasizes the need to evaluate the accuracy of tools currently available for prediction of pigmentation traits in this population.

Here we aimed to evaluate the accuracy of prediction of pigmentation traits in Latin Americans in the large Latin American dataset (the CANDELA cohort) recently used in GWASs of physical appearance, including for hair, eye and skin pigmentation[31,36–41]. We focused on comparing the accuracy of prediction methods, the predictors included in models, and the SNP sets used for prediction. We find that Random Forest and regression are generally the method with best performance, depending on the predicted trait. We also find that the inclusion of pigmentation SNPs increases the accuracy of prediction models substantially over that obtained by the inclusion of other predictors. Finally, the HIrisPlex-S has a similar performance to a CANDELA-specific set of prediction SNPs except for skin pigmentation.

3.1.2. Methods

Study sample: phenotypes, genetic data and covariates

We analyzed data previously studied by the CANDELA consortium for GWAS of pigmentation traits[31,36–39]. The consortium gathered genetic and phenotypic data from over 6,500 individuals recruited in five Latin American countries: Mexico (N=~1,200), Colombia (N=~1,700), Peru(N=~1,230), Chile (N=~1,730) and Brazil (N=~630).

Pigmentation traits evaluated directly on the research subjects consists of (A) hair pigmentation (recorded in four categories: 1-red/reddish, 2-blond, 3-dark blond/light brown or 4-brown/black. However, due to their very low frequency (<0.6%), individuals in the 'red/reddish' category were not included here), (B) eye color, recorded as five ordered categories: 1-blue/gray, 2-honey, 3-green, 4-light brown, 5-dark brown/black. For increasing consistency with previous publications[10,13,42–44], here we recoded these data into just three categories: 1-Blue/Gray, 2-Intermediate (honey or green) and 3-Brown/Black (light brown or dark brown/black) and (C) a quantitative measure of skin pigmentation from an area unexposed to sunlight (the Melanin Index MI, obtained by reflectometry). We also had available additional measures of iris pigmentation, extracted from digital photographs, using the HCL color space (Hue, Chroma and Luminance). Hue being an angle (recorded in arc degree), we linearized this trait with cosine and shifted the angle by 15° in order to maximize the number of samples in the range [0,180°]; hence the trait considered is cos(Hue+15). The frequency distribution for these traits in the CANDELA dataset is shown in Supplementary Figure S1.

For comparison with other studies[13,42], we converted the skin MI into a three-level categorical trait. For this, we selected CANDELA individuals with 100% European ancestry (N=70) and examined their facial photographs in order to label them as very pale or pale. Similarly, we selected CANDELA individuals with highest African ancestry (>39%; N=23) and examined their photographs in order to label them as dark or very dark. We plotted the distribution of the MI for individuals in the very pale, pale, dark and very dark categories and identified MI values of 33 and 47 as thresholds for categorical skin color:

fair (MI <33; N= 2,506), intermediate (MI 33-47; N=3,840) and dark (MI>47; N=180) classified as dark (see Supplementary figure S2). These thresholds are in line with values obtained in another study of admixed Brazilians[45].

The genetic data consisted of ~9 million genotypes, ~700k of which were obtained experimentally by genotyping Illumina's Omni Express chip, the remainder obtained by imputation as described in Adhikari et al[31]. We applied several filters to the CANDELA dataset prior to the trait prediction analyses. Firstly, we retained only individuals aged 18 to 45. Secondly, we removed 8 pairs of individuals whose pairwise probability of IBD was estimated close to 1, to discard potential sample mix-ups (hence, 16 individuals removed), and individuals whose estimated African ancestry was more than European and native ancestry estimations (23 individuals), as those were considered as genetic outliers and thus excluded. Finally, we excluded all individuals with missing data on any of the covariates (age, sex, BMI). Note that BMI was considered as a covariate since we found it significantly correlated to some pigmentation phenotypes; that correlation is most likely a confounding effect of continental genetic ancestry. The final sample size used in the analyses was: 6,495 for hair color, 6,526 for MI, 6,529 for categorical eye color and 5,738 for quantitative eye color traits – Hue, Chroma and Luminance. These three eye color phenotypes constitute the bicone color space model – HCL scale for human perception of eye colors (previously explained by Adhikari et al[31]).

Pigmentation SNP sets used for prediction

We used two sets of SNPs for the prediction analyses of pigmentation traits.

Firstly, as a benchmark we used HIrisplex-S, a SNP set that has been developing over the years for the prediction of eye, hair and skin pigmentation mainly in Europeans[7,11,14] HIrisplex-S currently includes 41 SNPs, 22 of which were directly genotyped in the CANDELA samples (genotypes for the remaining 19 SNPs being imputed). In our analyses, we only retained SNPs with MAF >= 1% in the CANDELA data, reducing the HIrisplex-S set used here to 34 SNPs (see Supplementary table S1).

Secondly, we devised "CANDELA" (CAN) SNP sets for prediction of each pigmentation trait (E-eye; H-hair; S-skin) based on results from a GWAS conducted in the CANDELA sample[31]. To pre-select SNPs for each trait, we used the following protocol: (1) selection of all SNPs with GWAS association p-values <1E-5, (2) grouping (clumping) of SNPs in high LD and (3) for each SNP group, selection of the SNP with highest predictive power (it is well-known that the most significant marker is not necessarily the best predictive marker[46]). In this way we pre-selected 1,471 SNPs relevant for skin pigmentation prediction, 207 for hair pigmentation prediction and 701 for eye pigmentation prediction. These SNPs were then ranked based on conditional predictive power for each trait, in order to search for the set of predictive SNPs with the least SNPs included. Details of the approach used are provided in Supplementary method S1, and the resulting CAN-E, CAN-S and CAN-H sets described in the Results section.

Prediction methods and models evaluated

A broad array of statistical methods have been employed in the literature to predict pigmentation traits, such as (multiple) linear[17,47] or (multinomial) logistic regression[10,11], decision trees[17,48], neural networks[17,49], and naïve Bayes classifiers[33,42,50]. Each method has its advantages and disadvantages, and are better suited for certain types of traits, e.g. linear regression for quantitative traits[17,47] and logistic regression for categorical traits[10,11].

The overall strategy we used for performing pigmentation prediction in the CANDELA dataset is shown in Article 1, Figure 1. Linear Regression (LR) or Multinomial Logistic Regression (MLR) were used as the reference methods for quantitative or categorical traits, respectively. These two methods were used to evaluate three prediction models, incorporating an increasing number of predictors (Article 1, Figure 1A):

Using only non-genetic covariates as predictors:

$$y \sim Age + Sex + BMI$$

(equation 1)

Incorporating genetic ancestry to model 1:

$$y \sim Age + Sex + BMI + EUR_{ancestry} + AFR_{ancestry}$$

(equation 2)

Here we included as predictors the estimates of European and African ancestry (obtained by unsupervised admixture estimation on genome-wide data). Native American ancestry was omitted so as to avoid collinearity (since the three continental ancestries sum to 1).

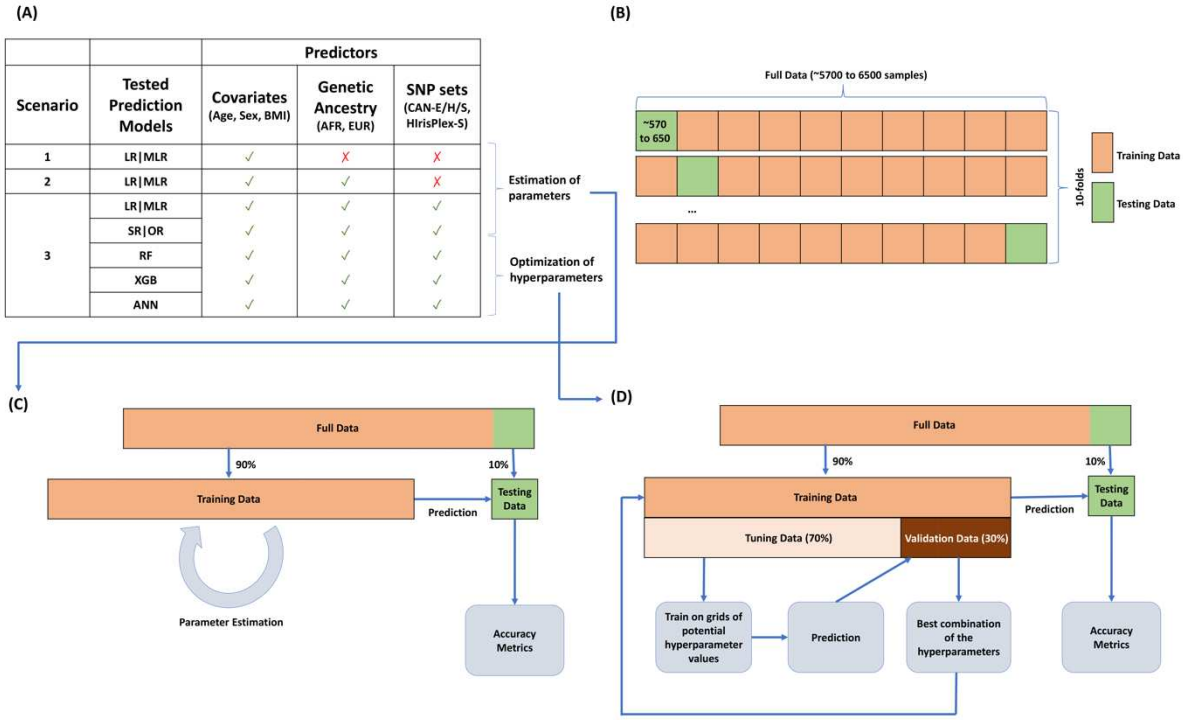
Incorporating pigmentation SNPs to model 2:

$$y \sim Age + Sex + BMI + EUR_{ancestry} + AFR_{ancestry} + \sum_{j \in SNPsets} SNP_j,$$

(equation 3)

where SNPsets refers to SNPs included either in the HIrisPlex-S or CAN sets defined above.

For the third (full) model, in addition to regression, the following statistical and machine-learning methods were used, in order to evaluate their relative performance for prediction: Random Forest (RF), Extreme Gradient Boosting (XGB), Artificial Neural Network (ANN), Ordinal regression (OR) and Stepwise regression (SR). We provide more details on their implementation in Supplementary method S2.



Article 1, Fig. 1 Study overview

(A) *Models tested, predictors used and prediction methods: multinomial regression (MLR), linear regression (LR), ordinal regression (OR), stepwise regression (SR), random forests (RF), extreme gradient boosting (XGB) and artificial neural network (ANN).* ***(B)*** *10-fold cross-validation: the full data is randomly split into 10 equally-sized data sub-groups. For each of the 10 sub-groups, the estimation of model parameters* ***(C)*** *or optimization of model hyperparameters* ***(D)*** *was performed on a pool of the nine remaining sub-groups.*

Evaluation of prediction accuracy

To measure prediction accuracy in quantitative traits, we used the coefficient of determination (R^2, the proportion of phenotype variance that is explained by the model, measured as $1 - SS_{res}/SS_{tot}$, where SS_{res} and SS_{tot} respectively stand for the residual and total sum of squares). For categorical traits, we used a metric denoted "accuracy" that is the proportion of correctly classified individuals in confusion matrices. For these traits, we also computed the Area Under the ROC Curve (AUC – ranging from 0.5 to 1) as well as the expected accuracies from two benchmark strategies (either using the categories' frequency – PropStrat, or always guessing the most frequent category – maxP; see Supplementary Method S3) for comparison.

Accuracy of prediction was evaluated using 10-fold cross validation (10-fold CV) (Fig. 1B). For this, the full dataset is split into 10 approximately equal subsets based on a stratified sampling on the trait (so that trait distribution is similar across subsets). Each of the 1/10th subsets is used as test data for evaluating prediction accuracy of methods trained using the other 9/10th of the data. For regression methods (MLR, LR, SR, OR), coefficient parameters are estimated in the training data (Fig. 1C). For Machine Learning models, the training data is further split into a tuning data (70% of the training data) and validation data (30% of the training data). The parameter space for these methods are

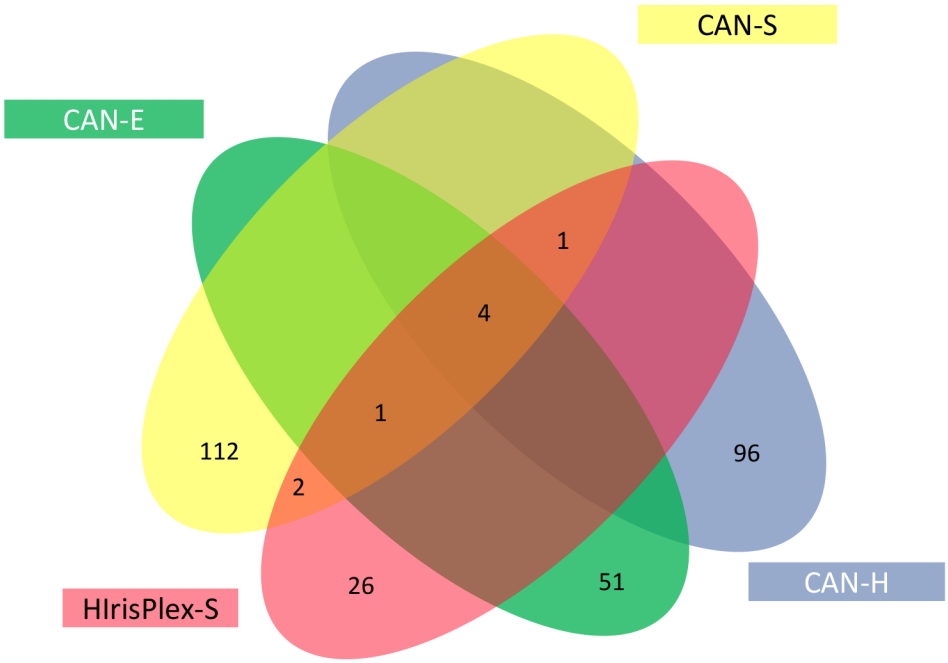
tuned creating a grid of all possible combinations of the hyperparameters and the particular combination producing best result on the validation data is selected as the set of optimal combination for the whole training data (see Supplementary Method S2). All possible permutations of binding nine folds for training produces 10 different train-test data combinations and the hyperparameters are tuned based on that subsequently gives rise to 10 prediction accuracy results (Fig. 1D). These measures of goodness of fit are used in a boxplot, or the average of them is used as a single prediction accuracy metric of the method. This helps us in avoiding inflation in the results and the predictions are more robust to small changes in the data.

Prediction of MI in Native American individuals of unknown phenotype

Genotype and geo-localization data for 104 Native American individuals (with an estimated >99% Native American ancestry) from 16 Native American populations were available from a previous study[40]. We analysed genotypes for these 'pure' Native American individuals, predicted their MI, and regressed these predicted values on the amount of solar radiation at the site of population sampling. We trained two RF models (one with CAN-S and one with HIrisPlex-S) using 550 CANDELA individuals with >= 80% native ancestry and using sex as the only covariate. Solar radiation levels were defined as insolation incident on a horizontal surface (in kWh/m2/day) as reported in the NASA Surface meteorology and Solar Energy (SSE) Web site (https://eosweb.larc.nasa.gov/sse/) (data previously used in[31]).

3.1.3. Results

Below we compare the performance of HIrisplex-S with SNP sets devised here for the prediction of eye (CAN-E), hair (CAN-H) and skin (CAN-S) from summary statistics of a pigmentation GWAS performed in the CANDELA sample (Materials and Methods and Supplementary method S1). The sets devised here consist of: 56 (CAN-E), 101 (CAN-H) and 120 (CAN-S) pigmentation-associated SNPs and are detailed in Supplementary table S2. Consistent with the genetic correlation of eye, hair and skin pigmentation, some SNPs are shared across the CAN-E/H/S sets, as well as with the HIrisPlex-S set. The overlap between these four SNP sets is shown in Article 1, Figure 2.

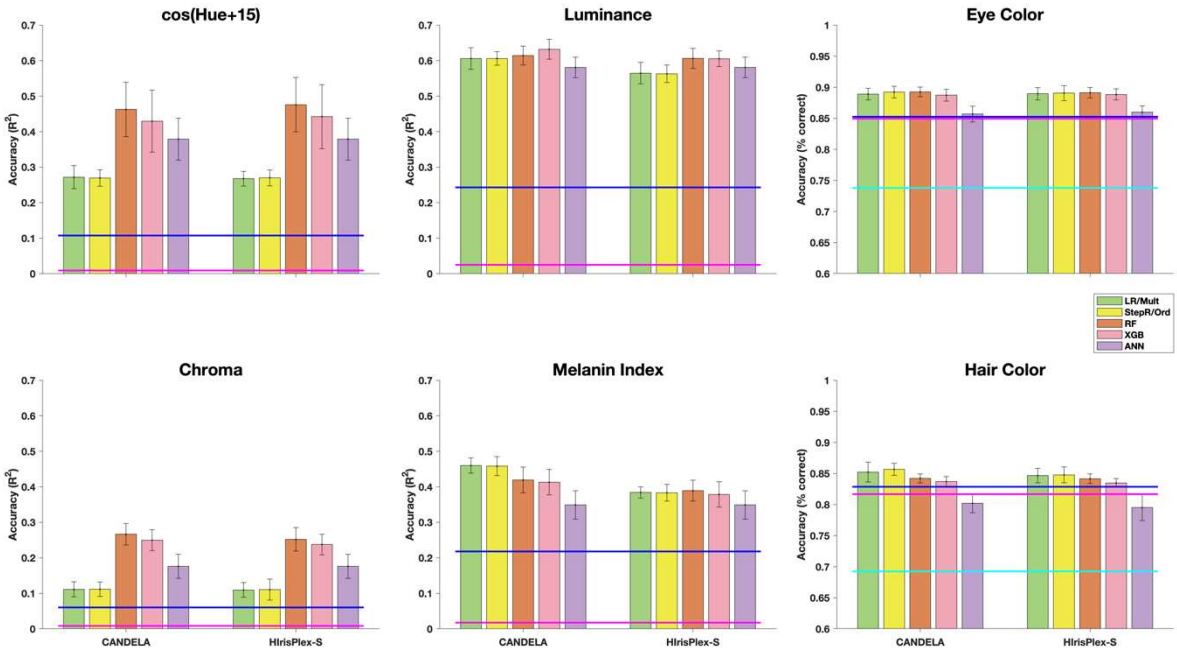


Article 1, Fig. 2 Overlap between SNP sets used for prediction of pigmentation traits.

CAN-E, CAN- S and CAN-H refer, respectively, to SNP sets designed here for the prediction of eye, skin and hair pigmentation, based on a GWAS performed in the CANDELA sample[31]. HIrisPlex-S is a SNP assay developed for simultaneous Eye, Hair and Skin color prediction[14]. Numbers refer to SNPs shared between SNP sets.

Prediction Accuracy in relation to models, methods and pigmentation SNP sets

Article 1, Figure 3 presents the accuracy of prediction for various phenotypes of eye, hair and skin color. For categorical traits, the baseline model (i.e. including only non-genetic predictors: age, sex and BMI), reaches 84.9% and 81.7% accuracy (proportion of correctly classified individuals) for eye and hair color, respectively. That level of accuracy is actually also reached by always guessing the phenotype to be the most frequent category (maxP strategy, see Supplementary table S3 and Supplementary Method S2). This high accuracy obtained by a deterministic strategy probably relates to the highly skewed trait distribution in the CANDELA individuals: ~ 82% having black/dark brown hair and 85% having Brown/Black eyes. Alternately, randomly guessing the phenotypes based solely on the frequency of the traits (PropStrat; cyan line in Article 1, Figure 3) also yields good levels of accuracy (~74% and ~69% for categorical eye and hair color, respectively).



Article 1, Fig. 3 Prediction accuracy in relation to models, methods and pigmentation SNP sets

For continuous traits (Hue (transformed), Luminance, Chroma and Melanin Index; top and middle panels) we used R^2 as measure of prediction accuracy. For categorical traits (Eye and Hair color; bottom panels) accuracy is the proportion of correctly classified individuals. Magenta and blue lines indicate the accuracy obtained when only non-genetic predictors or non-genetic + genomic ancestry are included in regression models, respectively. For categorical traits, the performance of a random guessing strategy (PropStrat) was also evaluated (cyan line). For these traits, the average accuracy of the deterministic maxP strategy is numerically the same as the accuracy obtained when only non-genetic predictors are used (magenta line), hence is not shown separately in this figure. For the full prediction model (non-genetic predictors + genetic ancestry + pigmentation SNPs) the performance of regression and four additional prediction methods was evaluated (bars are colored: green = LR/MLR; yellow = SR/OR; brown = RF; pink = XGB; purple = ANN). Detailed numerical values are given in Supplementary Table S3. The pigmentation SNP set incorporated in the prediction models is indicated at the bottom of the plots.

It is important to keep the maxP strategy in context when assessing prediction performance for categorical traits, since it represents how skewed the trait distribution is – a binary trait with a frequency distribution of 90% and 10% of the two categories will have 90% accuracy under the simplest maxP strategy (even though its sensitivity will be 0 for the rare category; see Supplementary Method S2). Thus, a skew in the trait distribution causes an upward shift in accuracy of prediction methods, especially those methods which are biased to the most frequent category, making them appear better-performing than they actually are. The PropStrat strategy is comparatively less biased as it gives proportional weight to the rare category (hence non-zero sensitivity for this category), and thus has lower accuracy than the maxP strategy. It is therefore a better

benchmark to compare the performance of other strategies for assessing their gain in accuracy. Conversely, a comparison of those strategies to the maxP benchmark better represents their relative change in incorrect classification rather than the relative gain in correct classification (i.e. gain in accuracy).

Although the accuracies of these basic strategies are already high due to our skewed trait distributions, adding genetic ancestry to the model has a further impact, especially for hair color: it decreases the proportion of error by ~6% (from 18.3% of error to 17.1% - detailed numbers in Supplementary table S3). Then the further addition of pigmentation SNPs has an even larger effect: the remaining proportion of errors decreases by another ~16% and ~27% respectively for hair and eye color. It is also noticeable that the gain in prediction brought by SNPs relatively to that brought by genetic ancestries is much larger for eye color (~14*x*) than for hair color (~3*x*).

For continuous traits, we observe a large increase in prediction accuracy (R^2) when genetic ancestry is incorporated in regression models, relative to the baseline (including only non-genetic predictors). Furthermore, when pigmentation SNP sets are incorporated in the regression model, accuracy usually more than doubles over that obtained with genetic ancestry plus non-genetic covariates (green bar versus blue line in Article 1, Figure 3). When using this full prediction model, lowest LR prediction accuracy was observed for Chroma (R^2 ~0.12) and highest for Luminance (R^2 ~0.58), two quantitative estimates of eye color variation.

Comparing different prediction methods for the full model (i.e. incorporating all predictors) we do not observe large differences in performance, with the exception of a relatively lower accuracy of regression for Hue and Chroma. For those two traits, RF markedly outperforms regression methods, more than doubling the accuracy of LR in the case of Chroma. For the two other continuous traits (Luminance and Melanin Index), regression models are as effective as machine learning models (Article 1, Figure 3). We also note that those rank similarly throughout categorical and continuous traits: RF is almost always better than the other tree-based model (extreme gradient boosting; XGB) and artificial neural networks (ANN) always underperform compared to tree-based models.

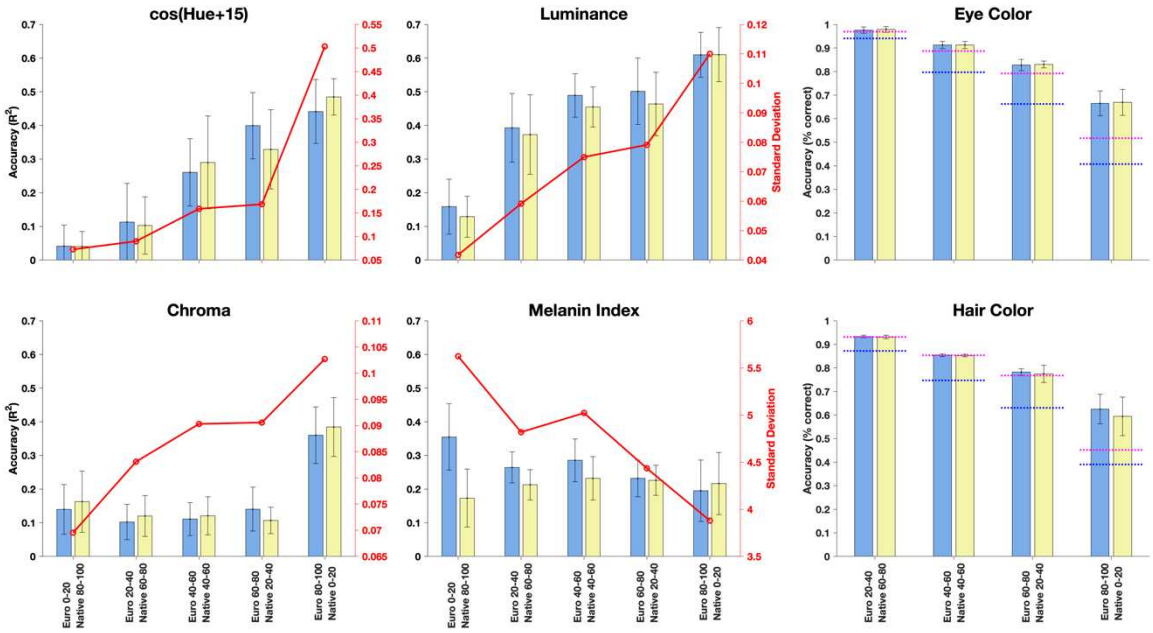
Regarding the SNP sets tested, we observe little difference in prediction accuracy across traits and methods although the number of SNPs is larger in the sets that we designed (56 to 120 SNPs) than in HIrisPlex-S (34 SNPs). Continuous skin pigmentation (Melanin Index) stands nonetheless as an exception: for that specific trait, CAN-S consistently outperforms HIrisPlex-S (particularly with regression methods). We have further assessed the impact of limiting the number of SNPs from the CAN sets to the top 10, 20 or 30 SNPs. A reduction in the number of SNPs had little effect for most traits, to the noticeable exception of Melanin Index for which considering the top 30 SNPs (i.e. 25% of the CAN-S SNP set) yields less than 90% of the prediction R^2 achieved with the whole set (see Supplementary table S7). Also, we ranked the SNPs by their variable importance (returned by the RF models) for each trait (see Supplementary table S4 and Supplementary figure S3). Three SNPs, from three known pigmentation genes (SLC45A2,

HERC2 and SLC24A5), are consistently among the top ones throughout traits and are common to HIrisPlex-S and the 3 CAN SNP sets.

Prediction accuracy at varying levels of European/Native American ancestry

Since a substantial fraction of individuals in the CANDELA sample have minimal African ancestry, we sought to evaluate prediction accuracy specifically for varying levels of European/Native American ancestry in the CANDELA sample. To this aim, we pooled individuals with negligible African ancestry in ~20% ancestry bins (so that the smallest pool size included at least 570 individuals) and examined prediction accuracy in the pools (see Supplementary table S5). Furthermore, for categorical traits, we ensured that at least two trait categories, each with >20 individuals, were observed in the pools, which led to withdraw the most Native-American pool. We assessed prediction accuracy using RF, a full model (i.e. equation 3 but without genetic ancestry as predictor) including only pigmentation SNPs having >1% MAF in the pool of individuals being tested.

For the categorical traits, there is a drop in prediction accuracy (from ~95% to ~70%) at increasing European ancestry (Article 1, Figure 4). However, as European ancestry increases there is greater accuracy relative to random guessing based on trait frequency, probably reflecting the trait being less variable at higher Native American ancestry levels. For the quantitative eye pigmentation variables (particularly H and L, Article 1, Figure 4), as the percentage of European ancestry increases there is a trend for an increase in trait variation (red line in Article 1, Figure 4) and also in prediction R^2. For skin pigmentation (MI) we observe an opposite trend in trait variability in relation to ancestry, relative to hair/eye color: variation in MI decreases at increasing European ancestry. There is also a trend towards an increase in the performance of the CAN-S SNP set at decreasing European ancestry: in individuals with <20% European ancestry CAN-S has an accuracy that is nearly twice that observed for HIrisPlex-S. Although CAN-S tends to outperform HIrisPlex-S in most comparisons, it is only for MI that such a large difference in performance was observed. In summary, across all pigmentation traits we observe a gain in prediction accuracy for the Native American/European ancestry bins showing greater phenotypic diversity, with CAN SNPs markedly outperforming HIrisPlex-S only for MI in individuals with low (<20%) European ancestry.



Article 1, Fig. 4 Prediction accuracy for individuals with varying Native American/European admixture

Prediction was assessed in individual bins varying ~20% in admixture (bottom axis; for eye and hair pigmentation <20 individuals with >80% Native Ancestry were available in each trait category, thus preventing estimation of prediction accuracy). Colored bars indicate accuracy obtained with Random Forest models using non-genetic + pigmentation SNP sets as predictors (R^2 being used as prediction measure for quantitative traits: Hue, Luminance, Chroma and Melanin index). Blue bars indicate the CAN-E/H/S SNP sets. Yellow bars indicate the HIrisPlex-S set. The standard deviation of the quantitative traits in each ancestry bin is indicated as a red line. For the categorical traits (Eye and Hair Color), accuracy (proportion of correctly classified individuals) is used as the metric for prediction measures. Accuracy obtained without genetic predictors using a guessing strategy is indicated with a horizontal blue line for Proportional Strategy (random guessing) and magenta line for maxP (deterministic guessing). Detailed numerical values are given in Supplementary Table S8 and S9.

Prediction accuracy in CANDELA relative to other population samples

Several studies have examined accuracy of prediction of the HIrisPlexS SNP set for categorical pigmentation traits using an online tool (https://hirisplex.erasmusmc.nl, referred to here as HIrisPlexS-Online), which implements MLR prediction models trained in a reference dataset (consisting mostly of Europeans)[7,11,14]. Article 1, Table 1 compares the published estimates with those we obtained here with the CAN and HIrisPlex-S SNP sets using our implementation of MLR models trained in the CANDELA data (for skin color, we transformed MI into a 3-level categorical variable: fair, intermediate and dark; as described in Materials and Methods). For a more accurate comparison to HIrisPlexS-Online, these models were trained using SNPs only, without basic covariates or ancestry.

Prediction accuracy estimates for eye color have been reported for HIrisPlex-S-Online in a Latin American sample (99 individuals from Venezuela and Brazil) and in a European sample[10,14,15]. The light (blue/gray) and dark (Brown/Black) color categories have similar prediction accuracies across studies (~90-93%), except for the prediction of light eye color reported for HIrisPlex-S-Online in the Venezuelan/Brazilian sample, where accuracy is much lower (85%). The main difference in eye-color prediction across studies lies in the intermediate category. No intermediate eye colors were predicted by HIrisPlex-S-Online in the Venezuelan/Brazilian sample. Our predictions in the CANDELA sample have markedly higher accuracy than reported for HIrisPlex-S-Online in Europeans, both when the prediction model was trained in the CANDELA data or in the reference HIrisPlex-S data (respectively 89% and 85%, versus 73% in Europeans). Prediction accuracy, in the CANDELA sample, of the CAN-S and HIrisPlex-S SNP sets was identical.

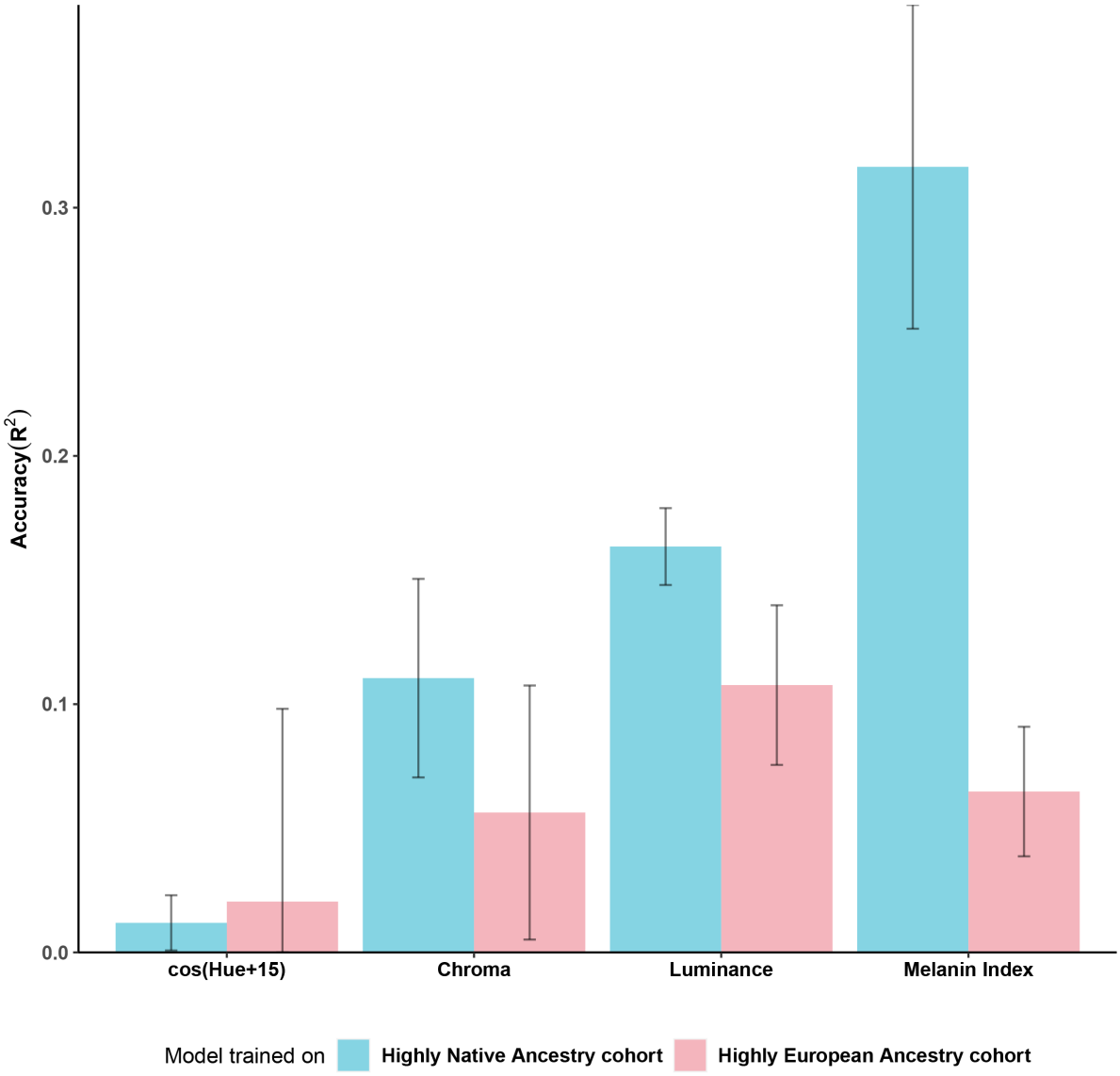
Estimates for hair color prediction accuracy using HIrisPlex-S-Online have been reported for a European sample[44]. Prediction accuracy estimates obtained here for the CANDELA sample are higher than reported in Europeans for all hair colors, except in the case of intermediate hair-color (i.e. brown) predicted with HIrisPlex-S-Online. The highest hair-color prediction accuracy was consistently obtained with the CAN-H SNP set, although the difference relative to the HIrisPlex-S, with models trained in the CANDELA data, is small.

Concerning categorical skin color, we observe that predictions from HIrisPlex-S-Online have markedly lower accuracy in the CANDELA sample than reported for a world-wide sample. Model training in the CANDELA dataset increases prediction accuracy substantially both for HIrisPlex-S and CAN-S, although the accuracy values obtained are still below the published estimates for HIrisPlex-S. As for hair color, the CAN-S SNP set consistently outperforms HIrisPlex-S, but only marginally.

Portability of models for pigmentation prediction in individuals with high Native Ancestry

Considering the impact of training datasets in the performance of HIrisPlex-S (Article 1, Table 1), we specifically examined the portability of RF models developed in two training datasets with extreme differences in ancestry (extracted from the CANDELA sample – see Supplementary figure S4): (i) a highly European training dataset (European ancestry >= 80% and Native American ancestry < 20%) and (ii) a highly Native training dataset (European ancestry < 20% and Native American ancestry >= 80%). We examined the performance of the resulting prediction models in a subset of the highly Native test dataset (Article 1, Figure 5) in a cross-validation scheme. We observe that models developed in the highly Native training dataset have a better performance than those developed in the highly European training dataset for Chrome, Luminance and MI. The most striking difference in performance is seen for MI, where the model trained with highly Native data has a prediction accuracy ~6 times that of the model trained in highly European data (Article 1, Figure 5). Hue is the only trait for which the model trained in

the highly European dataset outperforms the model trained in the highly Native dataset, but prediction accuracy in this case is extremely low (<2%).



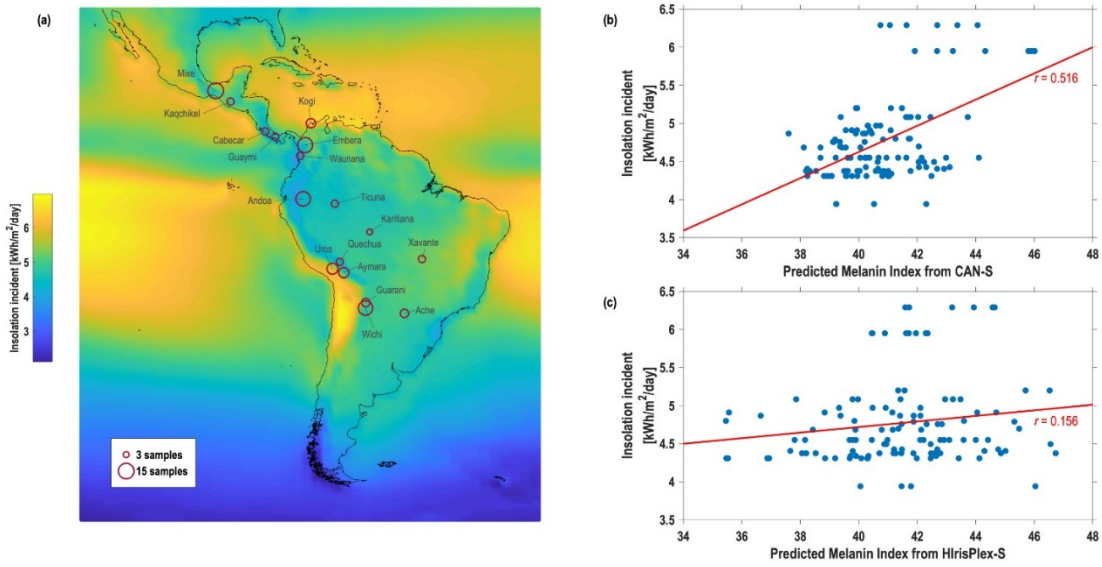
Article 1, Fig. 5 Portability of prediction models trained in highly European/Native American cohorts

For each continuous trait (cos(H+15), C, L and Melanin Index) we compare prediction accuracy on the same test data (a highly Native ancestry cohort). The prediction models were trained either on a highly Native (Blue) or highly European (Pink) cohort established from the CANDELA sample. For testing, we created equally-sized 4-folds for each pool of individuals. We built the RF models using three of the folds and evaluated prediction accuracy in the left-out fold from the Native ancestry cohort (see Supplementary figure S4).

Prediction of skin pigmentation in Native Americans

We examined prediction performance of the CAN-S and HIrisPlex-S sets in a highly Native dataset independent of CANDELA by predicting MI in 117 individuals from 17 Native American populations[40]. As above, we trained RF prediction models using

CANDELA individuals with >= 80% native ancestry. Since performance could not be measured directly in this dataset (due to the lack of phenotypic data), we examined the correlation of predicted skin pigmentation (MI) with solar radiation levels at the site of population sampling (Article 1, Figure 6). Previous surveys of skin pigmentation in native populations from across the world have found a correlation between skin pigmentation and solar radiation[51], an observation that has been interpreted as the result of selection throughout human evolution. In the Native Americans examined here we obtained correlations of 0.543 (p-value 2 x 10^{-9}) and 0.163 (p-value 0.1) with the CAN-S and HIrisPlex-S SNP sets, respectively (Article 1, Figure 6).



Article 1, Fig. 1 Solar radiation levels and skin pigmentation (MI) predicted with CAN-S and HIrisplex-S SNP sets

(a) *Annual average of insolation incident on a horizontal surface (kWh/m^2/day - data from NASA Surface meteorology and Solar Energy, 2008) and location of the Native American population sampled. The predicted MI and solar radiation levels at the sampling site for 104 individuals from 16 Native American populations is shown in* ***(b)*** *CAN-S and* ***(c)*** *HIrisPlex-S.*

3.1.4. Discussion

Our results agree with previous studies in finding that regression and RF generally outperform other prediction methods (MLR being the approach implemented in HIrisPlex-S-Online). Owing to their tree-based structures, RF and XGB implicitly model the underlying genetic interaction across SNPs, and studies have noted the presence of epistasis (SNP-SNP interactions) between major pigmentation SNPs to some degree for various pigmentation traits, suggesting that prediction accuracy can be further increased by specifically allowing for interaction between SNPs[23,24,52]. Since tree-based models

implicitly model epistasis including complex higher-order interactions, the difference of their prediction accuracy to additive linear/logistic models may also shed light on the genetic architecture of the traits. It is tempting to hypothesize that SNP interactions may have a more substantial impact on the genetic architecture of Hue and Chroma phenotypes[31,53], especially considering that these traits are relatively more non-linear in nature (e.g. Hue is an angle, i.e. a circular trait)[52]. Furthermore, the differences in prediction accuracy are much larger from linear to tree-based models (up to +0.19 gain in R^2 – see Supplementary table S3) than they are from HIrisPlex-S to CAN-E (up to +0.014 gain in R^2) whereas these two SNP sets only have 5 SNPs in common. We might therefore expect that such relevant interactions would be limited to a handful of SNPs, consistent with the literature which primarily observed interactions between major pigmentation SNPs[23,24,31]. Conversely, we may expect a greater additive polygenicity[23,25] in the case of skin pigmentation. Linear models overperform tree-based models for that trait (Article 1, Figure 3) and the number of SNPs in the model has an impact on the prediction accuracy, as suggests the +0.07 gain in R^2 from HIrisPlex-S (34 SNPs) to CAN-S (120 SNPs). In line with these observations, we also noted that prediction accuracies are relatively lower for Melanin Index than for other continuous traits when constraining the number of SNPs in the model (see Supplementary table S6 and S7).

Previous studies[31,36,54,55] have shown that genetic ancestry correlates with pigmentation, probably resulting from the variable allele frequency across continents of certain pigmentation-related alleles. Consistent with this, we observe that continental genetic ancestry has significant predictive power (Article 1, Figure 3, Supplementary table S3) and that the SNPs retained for prediction rank among the most correlated to continental ancestry components (see Supplementary table S2). Inclusion in the prediction models of pigmentation SNP sets selected from the GWAS further increases predictive power, markedly for quantitative pigmentation traits. Candidate SNPs on an average add twice the prediction accuracy of that brought in by the genetic ancestry for the quantitative traits. For the categorical traits, the difference is smaller but comparatively candidate SNPs decreases the proportion of incorrect classification by a larger amount than does continental genetic ancestry, especially for eye color.

The difference in predictive power for categorical traits is partly the result of their intrinsic lower informativeness relative to quantitative traits. This is particularly so for hair and eye color, as the CANDELA sample has a highly skewed phenotypic distribution for these traits: 82% of individuals in the CANDELA sample are assigned to the darkest category for eye and 85% for hair color. This skewed phenotypic distribution is consistent with the fact that world-wide lightly pigmented eyes and hair are essentially Western Eurasian traits, while there is considerable variation in skin pigmentation outside Europe[23–28]. That is, the occurrence of lightly pigmented hair and eyes in the CANDELA sample is essentially a reflection of its partly European ancestry. This may explain why the HIrisPlex-S SNP set built primarily in Europeans performs equally well to the CAN SNP set built on our more diverse admixed samples, and matches what has been observed in the literature for predicting eye and hair color in non-European populations[23,32,33].

By contrast, prediction of skin pigmentation is influenced by the genetic architecture of this trait in all three populations contributing to admixture in Latin America (Native Americans, Europeans and Africans). Consistent with this, our analysis along a Native American-European ancestry gradient show the highest gain in prediction accuracy for the ancestry bins with the greatest phenotypic diversity: that is, the highest European bin for eye/hair color and the highest Native American bin for skin color. In the bin with lowest European ancestry, there is hardly any gain in prediction accuracy for eye/hair color over the deterministic maxP strategy, as almost all individuals in that bin are in the highly pigmented category. On the other hand, this bin has the highest variation in skin pigmentation, also shows the greatest accuracy for the CAN-S, and this SNP set strongly outperforms HIrisPlex-S prediction for this ancestry bin (Article 1, Figure 4) – all these facts together emphasize that skin color prediction in admixed Latin Americans is substantially influenced by the genetic architecture of this trait in non-European populations. Our observation of a stronger correlation of predicted skin with solar radiation levels in pigmentation in Native Americans for the CAN-S set, relative to HIrisPlex-S is also consistent with this, and with literature reporting comparatively poor portability of European-based pigmentation prediction models on non-European populations[23,32,33].

Although quantitative traits are intrinsically more informative than categorical ones, anthropological and forensic applications of pigmentation prediction are usually interested in a few discrete categories, often just two (e.g. blue v. non-blue eyes) or three (light, intermediate or dark color). In that setting we find that for the CANDELA dataset there is little difference in the prediction performance of HIrisPlex-S (which, in addition, was shrunken from 41 SNPs to 34 polymorphic SNPs in our dataset), relative to the CAN SNP sets developed here. However, we observe a strong impact of the training dataset used for optimizing the prediction methods, both in the extreme scenario explored in Article 1, Figure 5, and when comparing the prediction accuracies we obtained here with those published previously (Article 1, Table 1). These observations question the portability across human populations of models developed using a training dataset with a different genetic makeup. This is most sharply illustrated by the difference in performance between the online implementation of HIrisPlex-S and the results we obtained when training models with this SNP set in the CANDELA data, particularly for skin pigmentation categories (Article 1, Table 1). These observations thus caution against the use of the online implementation of HIrisPlex-S for the prediction of pigmentation traits in non-Europeans.

Article 1, Table 1 Overall accuracy (%) and trait AUC (%) for categorical (A) eye, (B) hair and (C) skin colour obtained here and in other studies

(% frequency of trait in sample is shown in parentheses).

(A)

Sample	CANDELA N=6,529				Venezuela/Brazil[b] N = 99		Europeans[d] N=2,364	
SNP set		CAN-E[a]	HIrisPlex-S	HIrisPlexS-Online				
Overall Accuracy		89	89	87				83
1. Blue/gray	(3)	93	93	93	(12)	85	(68)	91
2. Intermediate	(13)	89	89	85	(23)	NA	(10)	73
3. Brown/black	(85)	92	92	90	(64)	91	(23)	93

(B)

Sample	CANDELA N=6,495				Europeans[c] N = 385	
SNP set		CAN-H	HIrisPlex-S	HIrisPlexS-Online		
Overall Accuracy		85	84	56		71
1.Red	(0)	NA	NA	NA	(25)	90
2. Blond	(3)	94	92	90	(54)	75
3. Brown	(16)	82	81	65	(9)	72
4. Black	(82)	87	85	80	(12)	78

(C)

Sample	CANDELA N=6,526				World-wide N=1,159[e]	
SNP set		CAN-S	HIrisPlex-S	HIrisPlex-S-Online		
Overall Accuracy		73	72	26		87
1. Fair	(38)	83	81	78	(92)	97
2. Intermediate	(59)	77	77	32	(3)	83
3. Dark	(3)	86	84	81	(5)	96

HIrisPlexS-Online: https://hirisplex.erasmusmc.nl

[a] Obtained using a model with only SNPs as predictors (no covariates) and with MLR as prediction method.

[b] Obtained using HIrisPlex as described in Freire-Aradas et al. 2014[43].

[c] Obtained using HIrisPlex-S as described in Branicki et al 2011[44], here accuracy is calculated from a 4-way classification setup.

[d] Obtained from the specificity reported in Liu et al 2009[10]

[e] This prediction values reported in Walsh et al 2017[13] were obtained from the prediction of the 194 samples from Maronas et al 2014[42]

3.1.5. References

[1] T.E. King, G.G. Fortes, P. Balaresque, M.G. Thomas, D. Balding, P.M. Delser, R. Neumann, W. Parson, M. Knapp, S. Walsh, L. Tonasso, J. Holt, M. Kayser, J. Appleby, P. Forster, D. Ekserdjian, M. Hofreiter, K. Schürer, Identification of the remains of King Richard III, Nat Commun. 5 (2014) 5631. https://doi.org/10.1038/ncomms6631.

[2] I. Olalde, M.E. Allentoft, F. Sánchez-Quinto, G. Santpere, C.W.K. Chiang, M. DeGiorgio, J. Prado-Martinez, J.A. Rodríguez, S. Rasmussen, J. Quilez, O. Ramírez, U.M. Marigorta, M. Fernández-Callejo, M.E. Prada, J.M.V. Encinas, R. Nielsen, M.G. Netea, J. Novembre, R.A. Sturm, P. Sabeti, T. Marquès-Bonet, A. Navarro, E. Willerslev, C. Lalueza-Fox, Derived immune and ancestral pigmentation alleles in a 7,000-year-old Mesolithic European, Nature. 507 (2014) 225–228. https://doi.org/10.1038/nature12960.

[3] C. Lalueza-Fox, H. Rompler, D. Caramelli, C. Staubert, G. Catalano, D. Hughes, N. Rohland, E. Pilli, L. Longo, S. Condemi, M. de la Rasilla, J. Fortea, A. Rosas, M. Stoneking, T. Schoneberg, J. Bertranpetit, M. Hofreiter, A Melanocortin 1 Receptor Allele Suggests Varying Pigmentation Among Neanderthals, Science. 318 (2007) 1453–1455. https://doi.org/10.1126/science.1147417.

[4] M. Kayser, F. Liu, A.C.J.W. Janssens, F. Rivadeneira, O. Lao, K. van Duijn, M. Vermeulen, P. Arp, M.M. Jhamai, W.F.J. van IJcken, J.T. den Dunnen, S. Heath, D. Zelenika, D.D.G. Despriet, C.C.W. Klaver, J.R. Vingerling, P.T.V.M. de Jong, A. Hofman, Y.S. Aulchenko, A.G. Uitterlinden, B.A. Oostra, C.M. van Duijn, Three Genome-wide Association Studies and a Linkage Analysis Identify HERC2 as a Human Iris Color Gene, The American Journal of Human Genetics. 82 (2008) 411–423. https://doi.org/10.1016/j.ajhg.2007.10.003.

[5] P. Sulem, D.F. Gudbjartsson, S.N. Stacey, A. Helgason, T. Rafnar, M. Jakobsdottir, S. Steinberg, S.A. Gudjonsson, A. Palsson, G. Thorleifsson, S. Pálsson, B. Sigurgeirsson, K. Thorisdottir, R. Ragnarsson, K.R. Benediktsdottir, K.K. Aben, S.H. Vermeulen, A.M. Goldstein, M.A. Tucker, L.A. Kiemeney, J.H. Olafsson, J. Gulcher, A. Kong, U. Thorsteinsdottir, K. Stefansson, Two newly identified genetic determinants of pigmentation in Europeans, Nat Genet. 40 (2008) 835–837. https://doi.org/10.1038/ng.160.

[6] P. Sulem, D.F. Gudbjartsson, S.N. Stacey, A. Helgason, T. Rafnar, K.P. Magnusson, A. Manolescu, A. Karason, A. Palsson, G. Thorleifsson, M. Jakobsdottir, S. Steinberg, S. Pálsson, F. Jonasson, B. Sigurgeirsson, K. Thorisdottir, R. Ragnarsson, K.R. Benediktsdottir, K.K. Aben, L.A. Kiemeney, J.H. Olafsson, J. Gulcher, A. Kong, U. Thorsteinsdottir, K. Stefansson, Genetic determinants of hair, eye and skin pigmentation in Europeans, Nat Genet. 39 (2007) 1443–1452. https://doi.org/10.1038/ng.2007.13.

[7] S. Walsh, F. Liu, K.N. Ballantyne, M. van Oven, O. Lao, M. Kayser, IrisPlex: A sensitive DNA tool for accurate prediction of blue and brown eye colour in the absence of ancestry information, Forensic Science International: Genetics. 5 (2011) 170–180. https://doi.org/10.1016/j.fsigen.2010.02.004.

[8] S. Walsh, A. Wollstein, F. Liu, U. Chakravarthy, M. Rahu, J.H. Seland, G. Soubrane, L. Tomazzoli, F. Topouzis, J.R. Vingerling, J. Vioque, A.E. Fletcher, K.N. Ballantyne, M. Kayser, DNA-based eye colour prediction across Europe with the IrisPlex system, Forensic Science International: Genetics. 6 (2012) 330–340. https://doi.org/10.1016/j.fsigen.2011.07.009.

[9] J. Purps, M. Geppert, M. Nagy, L. Roewer, Evaluation of the IrisPlex eye colour prediction tool in a German population sample, Forensic Science International: Genetics Supplement Series. 3 (2011) e202–e203. https://doi.org/10.1016/j.fsigss.2011.08.101.

[10] F. Liu, K. van Duijn, J.R. Vingerling, A. Hofman, A.G. Uitterlinden, A.C.J.W. Janssens, M. Kayser, Eye color and the prediction of complex phenotypes from genotypes, Current Biology. 19 (2009) R192–R193. https://doi.org/10.1016/j.cub.2009.01.027.

[11] S. Walsh, F. Liu, A. Wollstein, L. Kovatsi, A. Ralf, A. Kosiniak-Kamysz, W. Branicki, M. Kayser, The HIrisPlex system for simultaneous prediction of hair and eye colour from DNA, Forensic Science International: Genetics. 7 (2013) 98–115. https://doi.org/10.1016/j.fsigen.2012.07.005.

[12] S. Walsh, L. Chaitanya, L. Clarisse, L. Wirken, J. Draus-Barini, L. Kovatsi, H. Maeda, T. Ishikawa, T. Sijen, P. de Knijff, W. Branicki, F. Liu, M. Kayser, Developmental validation of the HIrisPlex system: DNA-based eye and hair colour prediction for forensic and anthropological usage, Forensic Science International: Genetics. 9 (2014) 150–161. https://doi.org/10.1016/j.fsigen.2013.12.006.

[13] S. Walsh, L. Chaitanya, K. Breslin, C. Muralidharan, A. Bronikowska, E. Pospiech, J. Koller, L. Kovatsi, A. Wollstein, W. Branicki, F. Liu, M. Kayser, Global skin colour prediction from DNA, Hum Genet. 136 (2017) 847–863. https://doi.org/10.1007/s00439-017-1808-5.

[14] L. Chaitanya, K. Breslin, S. Zuñiga, L. Wirken, E. Pośpiech, M. Kukla-Bartoszek, T. Sijen, P. de Knijff, F. Liu, W. Branicki, M. Kayser, S. Walsh, The HIrisPlex-S system for eye, hair and skin colour prediction from DNA: Introduction and forensic developmental validation, Forensic Science International: Genetics. 35 (2018) 123–135. https://doi.org/10.1016/j.fsigen.2018.04.004.

[15] K. Breslin, B. Wills, A. Ralf, M. Ventayol Garcia, M. Kukla-Bartoszek, E. Pospiech, A. Freire-Aradas, C. Xavier, S. Ingold, M. de La Puente, K.J. van der Gaag, N. Herrick, C. Haas, W. Parson, C. Phillips, T. Sijen, W. Branicki, S. Walsh, M. Kayser, HIrisPlex-S system for eye, hair, and skin color prediction from DNA: Massively parallel sequencing solutions for two common forensically used platforms, Forensic Science International: Genetics. 43 (2019) 102152. https://doi.org/10.1016/j.fsigen.2019.102152.

[16] S.I. Candille, D.M. Absher, S. Beleza, M. Bauchet, B. McEvoy, N.A. Garrison, J.Z. Li, R.M. Myers, G.S. Barsh, H. Tang, M.D. Shriver, Genome-Wide Association Studies of Quantitatively Measured Skin, Hair, and Eye Pigmentation in Four European Populations, PLoS ONE. 7 (2012) e48294. https://doi.org/10.1371/journal.pone.0048294.

[17] K. Zaorska, P. Zawierucha, M. Nowicki, Prediction of skin color, tanning and freckling from DNA in Polish population: linear regression, random forest and neural

network approaches, Hum Genet. 138 (2019) 635–647. https://doi.org/10.1007/s00439-019-02012-w.

[18] E. Pośpiech, Y. Chen, M. Kukla-Bartoszek, K. Breslin, A. Aliferi, J.D. Andersen, D. Ballard, L. Chaitanya, A. Freire-Aradas, K.J. van der Gaag, L. Girón-Santamaría, T.E. Gross, M. Gysi, G. Huber, A. Mosquera-Miguel, C. Muralidharan, M. Skowron, Á. Carracedo, C. Haas, N. Morling, W. Parson, C. Phillips, P.M. Schneider, T. Sijen, D. Syndercombe-Court, M. Vennemann, S. Wu, S. Xu, L. Jin, S. Wang, G. Zhu, N.G. Martin, S.E. Medland, W. Branicki, S. Walsh, F. Liu, M. Kayser, Towards broadening Forensic DNA Phenotyping beyond pigmentation: Improving the prediction of head hair shape from DNA, Forensic Science International: Genetics. 37 (2018) 241–251. https://doi.org/10.1016/j.fsigen.2018.08.017.

[19] V. Kastelic, K. Drobnič, A single-nucleotide polymorphism (SNP) multiplex system: the association of five SNPs with human eye and hair color in the Slovenian population and comparison using a Bayesian network and logistic regression model, Croat Med J. 53 (2012) 401–408. https://doi.org/10.3325/cmj.2012.53.401.

[20] K.L. Hart, S.L. Kimura, V. Mushailov, Z.M. Budimlija, M. Prinz, E. Wurmbach, Improved eye- and skin-color prediction based on 8 SNPs, Croat Med J. 54 (2013) 248–256. https://doi.org/10.3325/cmj.2013.54.248.

[21] V. Kastelic, E. Pośpiech, J. Draus-Barini, W. Branicki, K. Drobnič, Prediction of eye color in the Slovenian population using the IrisPlex SNPs, Croat Med J. 54 (2013) 381–386. https://doi.org/10.3325/cmj.2013.54.381.

[22] P. Dario, H. Mouriño, A.R. Oliveira, I. Lucas, T. Ribeiro, M.J. Porto, J. Costa Santos, D. Dias, F. Corte Real, Assessment of IrisPlex-based multiplex for eye and skin color prediction with application to a Portuguese population, Int J Legal Med. 129 (2015) 1191–1200. https://doi.org/10.1007/s00414-015-1248-5.

[23] E. Pośpiech, A. Wojas-Pelc, S. Walsh, F. Liu, H. Maeda, T. Ishikawa, M. Skowron, M. Kayser, W. Branicki, The common occurrence of epistasis in the determination of human pigmentation and its impact on DNA-based pigmentation phenotype prediction, Forensic Science International: Genetics. 11 (2014) 64–72. https://doi.org/10.1016/j.fsigen.2014.01.012.

[24] E.E. Quillen, H.L. Norton, E.J. Parra, F. Lona-Durazo, K.C. Ang, F.M. Illiescu, L.N. Pearson, M.D. Shriver, T. Lasisi, O. Gokcumen, I. Starr, Y. Lin, A.R. Martin, N.G. Jablonski, Shades of complexity: New perspectives on the evolution and genetic architecture of human skin, Am J Phys Anthropol. 168 (2019) 4–26. https://doi.org/10.1002/ajpa.23737.

[25] A.R. Martin, M. Lin, J.M. Granka, J.W. Myrick, X. Liu, A. Sockell, E.G. Atkinson, C.J. Werely, M. Möller, M.S. Sandhu, D.M. Kingsley, E.G. Hoal, X. Liu, M.J. Daly, M.W. Feldman, C.R. Gignoux, C.D. Bustamante, B.M. Henn, An Unexpectedly Complex Architecture for Skin Pigmentation in Africans, Cell. 171 (2017) 1340-1353.e14. https://doi.org/10.1016/j.cell.2017.11.015.

[26] L. Deng, S. Xu, Adaptation of human skin color in various populations, Hereditas. 155 (2018) 1. https://doi.org/10.1186/s41065-017-0036-2.

[27] H.L. Norton, M. Edwards, S. Krithika, M. Johnson, E.A. Werren, E.J. Parra, Quantitative assessment of skin, hair, and iris variation in a diverse sample of individuals and associated genetic variation: QUANTITATIVE ASSESSMENT OF PIGMENTARY PHENOTYPE, Am. J. Phys. Anthropol. 160 (2016) 570–581. https://doi.org/10.1002/ajpa.22861.

[28] M. Jonnalagadda, M.A. Faizan, S. Ozarkar, R. Ashma, S. Kulkarni, H.L. Norton, E. Parra, A Genome-Wide Association Study of Skin and Iris Pigmentation among Individuals of South Asian Ancestry, Genome Biology and Evolution. 11 (2019) 1066–1076. https://doi.org/10.1093/gbe/evz057.

[29] M. Edwards, D. Cha, S. Krithika, M. Johnson, G. Cook, E.J. Parra, 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 (2016) 141–162. https://doi.org/10.1111/pcmr.12435.

[30] L. Rawofi, M. Edwards, S. Krithika, P. Le, D. Cha, Z. Yang, Y. Ma, J. Wang, B. Su, L. Jin, H.L. Norton, E.J. Parra, Genome-wide association study of pigmentary traits (skin and iris color) in individuals of East Asian ancestry, PeerJ. 5 (2017) e3951. https://doi.org/10.7717/peerj.3951.

[31] K. Adhikari, J. Mendoza-Revilla, A. Sohail, M. Fuentes-Guajardo, J. Lampert, J.C. Chacón-Duque, M. Hurtado, V. Villegas, V. Granja, V. Acuña-Alonzo, C. Jaramillo, W. Arias, R.B. Lozano, P. Everardo, J. Gómez-Valdés, H. Villamil-Ramírez, C.C. Silva de Cerqueira, T. Hunemeier, V. Ramallo, L. Schuler-Faccini, F.M. Salzano, R. Gonzalez-José, M.-C. Bortolini, S. Canizales-Quinteros, C. Gallo, G. Poletti, G. Bedoya, F. Rothhammer, D.J. Tobin, M. Fumagalli, D. Balding, A. Ruiz-Linares, A GWAS in Latin Americans highlights the convergent evolution of lighter skin pigmentation in Eurasia, Nat Commun. 10 (2019) 358. https://doi.org/10.1038/s41467-018-08147-0.

[32] A. Pneuman, Z.M. Budimlija, T. Caragine, M. Prinz, E. Wurmbach, Verification of eye and skin color predictors in various populations, Legal Medicine. 14 (2012) 78–83. https://doi.org/10.1016/j.legalmed.2011.12.005.

[33] L. Yun, Y. Gu, H. Rajeevan, K.K. Kidd, Application of six IrisPlex SNPs and comparison of two eye color prediction systems in diverse Eurasia populations, Int J Legal Med. 128 (2014) 447–453. https://doi.org/10.1007/s00414-013-0953-1.

[34] H.L. Norton, R.A. Kittles, E. Parra, P. McKeigue, X. Mao, K. Cheng, V.A. Canfield, D.G. Bradley, B. McEvoy, M.D. Shriver, Genetic Evidence for the Convergent Evolution of Light Skin in Europeans and East Asians, Molecular Biology and Evolution. 24 (2006) 710–722. https://doi.org/10.1093/molbev/msl203.

[35] M. Edwards, A. Bigham, J. Tan, S. Li, A. Gozdzik, K. Ross, L. Jin, E.J. Parra, Association of the OCA2 Polymorphism His615Arg with Melanin Content in East Asian Populations: Further Evidence of Convergent Evolution of Skin Pigmentation, PLoS Genet. 6 (2010) e1000867. https://doi.org/10.1371/journal.pgen.1000867.

[36] A. Ruiz-Linares, K. Adhikari, V. Acuña-Alonzo, M. Quinto-Sanchez, C. Jaramillo, W. Arias, M. Fuentes, M. Pizarro, P. Everardo, F. de Avila, J. Gómez-Valdés, P. León-Mimila, T. Hunemeier, V. Ramallo, C.C. Silva de Cerqueira, M.-W. Burley, E. Konca, M.Z. de Oliveira,

M.R. Veronez, M. Rubio-Codina, O. Attanasio, S. Gibbon, N. Ray, C. Gallo, G. Poletti, J. Rosique, L. Schuler-Faccini, F.M. Salzano, M.-C. Bortolini, S. Canizales-Quinteros, F. Rothhammer, G. Bedoya, D. Balding, R. Gonzalez-José, Admixture in Latin America: Geographic Structure, Phenotypic Diversity and Self-Perception of Ancestry Based on 7,342 Individuals, PLoS Genet. 10 (2014) e1004572. https://doi.org/10.1371/journal.pgen.1004572.

[37] K. Adhikari, J. Mendoza-Revilla, J.C. Chacón-Duque, M. Fuentes-Guajardo, A. Ruiz-Linares, Admixture in Latin America, Current Opinion in Genetics & Development. 41 (2016) 106–114. https://doi.org/10.1016/j.gde.2016.09.003.

[38] K. Adhikari, M. Fuentes-Guajardo, M. Quinto-Sánchez, J. Mendoza-Revilla, J. Camilo Chacón-Duque, V. Acuña-Alonzo, C. Jaramillo, W. Arias, R.B. Lozano, G.M. Pérez, J. Gómez-Valdés, H. Villamil-Ramírez, T. Hunemeier, V. Ramallo, C.C. Silva de Cerqueira, M. Hurtado, V. Villegas, V. Granja, C. Gallo, G. Poletti, L. Schuler-Faccini, F.M. Salzano, M.-C. Bortolini, S. Canizales-Quinteros, M. Cheeseman, J. Rosique, G. Bedoya, F. Rothhammer, D. Headon, R. González-José, D. Balding, A. Ruiz-Linares, A genome-wide association scan implicates DCHS2, RUNX2, GLI3, PAX1 and EDAR in human facial variation, Nat Commun. 7 (2016) 11616. https://doi.org/10.1038/ncomms11616.

[39] K. Adhikari, T. Fontanil, S. Cal, J. Mendoza-Revilla, M. Fuentes-Guajardo, J.-C. Chacón-Duque, F. Al-Saadi, J.A. Johansson, M. Quinto-Sanchez, V. Acuña-Alonzo, C. Jaramillo, W. Arias, R. Barquera Lozano, G. Macín Pérez, J. Gómez-Valdés, H. Villamil-Ramírez, T. Hunemeier, V. Ramallo, C.C. Silva de Cerqueira, M. Hurtado, V. Villegas, V. Granja, C. Gallo, G. Poletti, L. Schuler-Faccini, F.M. Salzano, M.-C. Bortolini, S. Canizales-Quinteros, F. Rothhammer, G. Bedoya, R. Gonzalez-José, D. Headon, C. López-Otín, D.J. Tobin, D. Balding, A. Ruiz-Linares, A genome-wide association scan in admixed Latin Americans identifies loci influencing facial and scalp hair features, Nat Commun. 7 (2016) 10815. https://doi.org/10.1038/ncomms10815.

[40] J.-C. Chacón-Duque, K. Adhikari, M. Fuentes-Guajardo, J. Mendoza-Revilla, V. Acuña-Alonzo, R. Barquera, M. Quinto-Sánchez, J. Gómez-Valdés, P. Everardo Martínez, H. Villamil-Ramírez, T. Hünemeier, V. Ramallo, C.C. Silva de Cerqueira, M. Hurtado, V. Villegas, V. Granja, M. Villena, R. Vásquez, E. Llop, J.R. Sandoval, A.A. Salazar-Granara, M.-L. Parolin, K. Sandoval, R.I. Peñaloza-Espinosa, H. Rangel-Villalobos, C.A. Winkler, W. Klitz, C. Bravi, J. Molina, D. Corach, R. Barrantes, V. Gomes, C. Resende, L. Gusmão, A. Amorim, Y. Xue, J.-M. Dugoujon, P. Moral, R. González-José, L. Schuler-Faccini, F.M. Salzano, M.-C. Bortolini, S. Canizales-Quinteros, G. Poletti, C. Gallo, G. Bedoya, F. Rothhammer, D. Balding, G. Hellenthal, A. Ruiz-Linares, Latin Americans show wide-spread Converso ancestry and imprint of local Native ancestry on physical appearance, Nat Commun. 9 (2018) 5388. https://doi.org/10.1038/s41467-018-07748-z.

[41] K. Adhikari, G. Reales, A.J.P. Smith, E. Konka, J. Palmen, M. Quinto-Sanchez, V. Acuña-Alonzo, C. Jaramillo, W. Arias, M. Fuentes, M. Pizarro, R. Barquera Lozano, G. Macín Pérez, J. Gómez-Valdés, H. Villamil-Ramírez, T. Hunemeier, V. Ramallo, C.C. Silva de Cerqueira, M. Hurtado, V. Villegas, V. Granja, C. Gallo, G. Poletti, L. Schuler-Faccini, F.M. Salzano, M.-C. Bortolini, S. Canizales-Quinteros, F. Rothhammer, G. Bedoya, R. Calderón, J. Rosique, M. Cheeseman, M.F. Bhutta, S.E. Humphries, R. Gonzalez-José, D. Headon, D. Balding, A. Ruiz-Linares, A genome-wide association study identifies multiple loci for

variation in human ear morphology, Nat Commun. 6 (2015) 7500. https://doi.org/10.1038/ncomms8500.

[42] O. Maroñas, C. Phillips, J. Söchtig, A. Gomez-Tato, R. Cruz, J. Alvarez-Dios, M.C. de Cal, Y. Ruiz, M. Fondevila, Á. Carracedo, M.V. Lareu, Development of a forensic skin colour predictive test, Forensic Science International: Genetics. 13 (2014) 34–44. https://doi.org/10.1016/j.fsigen.2014.06.017.

[43] A. Freire-Aradas, Y. Ruiz, C. Phillips, O. Maroñas, J. Söchtig, A.G. Tato, J.Á. Dios, M.C. de Cal, V.N. Silbiger, A.D. Luchessi, A.D. Luchessi, M.A. Chiurillo, Á. Carracedo, M.V. Lareu, Exploring iris colour prediction and ancestry inference in admixed populations of South America, Forensic Science International: Genetics. 13 (2014) 3–9. https://doi.org/10.1016/j.fsigen.2014.06.007.

[44] W. Branicki, F. Liu, K. van Duijn, J. Draus-Barini, E. Pośpiech, S. Walsh, T. Kupiec, A. Wojas-Pelc, M. Kayser, Model-based prediction of human hair color using DNA variants, Hum Genet. 129 (2011) 443–454. https://doi.org/10.1007/s00439-010-0939-8.

[45] T.K.M. Leite, R.M.C. Fonseca, N.M. de França, E.J. Parra, R.W. Pereira, Genomic Ancestry, Self-Reported "Color" and Quantitative Measures of Skin Pigmentation in Brazilian Admixed Siblings, PLoS ONE. 6 (2011) e27162. https://doi.org/10.1371/journal.pone.0027162.

[46] A. Lo, H. Chernoff, T. Zheng, S.-H. Lo, Why significant variables aren't automatically good predictors, Proc Natl Acad Sci USA. 112 (2015) 13892–13897. https://doi.org/10.1073/pnas.1518285112.

[47] R.K. Valenzuela, M.S. Henderson, M.H. Walsh, N.A. Garrison, J.T. Kelch, O. Cohen-Barak, D.T. Erickson, F. John Meaney, J. Bruce Walsh, K.C. Cheng, S. Ito, K. Wakamatsu, T. Frudakis, M. Thomas, M.H. Brilliant, Predicting Phenotype from Genotype: Normal Pigmentation, Journal of Forensic Sciences. 55 (2010) 315–322. https://doi.org/10.1111/j.1556-4029.2009.01317.x.

[48] O. Spichenok, Z.M. Budimlija, A.A. Mitchell, A. Jenny, L. Kovacevic, D. Marjanovic, T. Caragine, M. Prinz, E. Wurmbach, Prediction of eye and skin color in diverse populations using seven SNPs, Forensic Science International: Genetics. 5 (2011) 472–478. https://doi.org/10.1016/j.fsigen.2010.10.005.

[49] B. Alipanahi, P. Fontanillas, 23andMe Resarch Team, S. Pitts, R. Gentleman, Pigmentor—Accurate prediction of multiple pigmentation phenotypes, (2017).

[50] Y. Ruiz, C. Phillips, A. Gomez-Tato, J. Alvarez-Dios, M. Casares de Cal, R. Cruz, O. Maroñas, J. Söchtig, M. Fondevila, M.J. Rodriguez-Cid, Á. Carracedo, M.V. Lareu, Further development of forensic eye color predictive tests, Forensic Science International: Genetics. 7 (2013) 28–40. https://doi.org/10.1016/j.fsigen.2012.05.009.

[51] N.G. Jablonski, G. Chaplin, The evolution of human skin coloration, Journal of Human Evolution. 39 (2000) 57–106. https://doi.org/10.1006/jhev.2000.0403.

[52] E. Pośpiech, J. Draus-Barini, T. Kupiec, A. Wojas-Pelc, W. Branicki, Gene–gene interactions contribute to eye colour variation in humans, J Hum Genet. 56 (2011) 447–455. https://doi.org/10.1038/jhg.2011.38.

[53] A. Wollstein, S. Walsh, F. Liu, U. Chakravarthy, M. Rahu, J.H. Seland, G. Soubrane, L. Tomazzoli, F. Topouzis, J.R. Vingerling, J. Vioque, S. Böhringer, A.E. Fletcher, M. Kayser, Novel quantitative pigmentation phenotyping enhances genetic association, epistasis, and prediction of human eye colour, Sci Rep. 7 (2017) 43359. https://doi.org/10.1038/srep43359.

[54] T.N. Frudakis, Molecular photofitting: predicting ancestry and phenotype using DNA, Elsevier/Academic Press, Amsterdam ; Boston, 2008.

[55] C.C.S. de Cerqueira, T. Hünemeier, J. Gomez-Valdés, V. Ramallo, C.D. Volasko-Krause, A.A.L. Barbosa, P. Vargas-Pinilla, R.C. Dornelles, D. Longo, F. Rothhammer, G. Bedoya, S. Canizales-Quinteros, V. Acuña-Alonzo, C. Gallo, G. Poletti, R. González-José, F.M. Salzano, S.M. Callegari-Jacques, L. Schuler-Faccini, A. Ruiz-Linares, M. Cátira Bortolini, for CANDELA (Consortium for the Analysis of the Diversity and Evolution of Latin America), Implications of the Admixture Process in Skin Color Molecular Assessment, PLoS ONE. 9 (2014) e96886. https://doi.org/10.1371/journal.pone.0096886.

3.2. Article 2: The evolution of skin and eye pigmentation in humans: a survey based on genetic predictions across time and space

This article is in preparation. Supplementary information is available in Appendix C. In this article my contributions are:

- Data preparation:
    - Phenotype quality control,
    - CANDELA data preparation,
    - Ancient DNA data preparation,
    - Cotemporary data preparation,
- SNP set preparation
- Prediction model preparation:
    - Prediction model pipelines for pseudohaploids,
    - Prediction on contemporary samples,
- Validation of the prediction on the contemporary samples
- Post prediction trend analysis
- Draft preparation and writing

Authors and affiliations

Sagnik Palmal[1*], Kaustubh Adhikari[2,3*], Javier Mendoza-Revilla[4,5], Macarena Fuentes-Guajardo[6], Caio C. Silva de Cerqueira[7], Pierre Justeau[1], Xiyang Cai[8], Betty Bonfante[1], Soumya Subhra Paria[2], Sijia Wang[8,9], Juan Camilo Chacón-Duque[10], Malena Hurtado[4], Valeria Villegas[4], Vanessa Granja[4], Claudia Jaramillo[11,12], William Arias[12], Rodrigo Barquera Lozano[13,14], Paola Everardo-Martínez[13], Jorge Gómez-Valdés[13], Hugo Villamil-Ramírez[15], Tábita Hünemeier[16], Virginia Ramallo[17,18], Maria-Laura Parolin[19], Rolando Gonzalez-José[18], Lavinia Schüler-Faccini[17], Maria-Cátira Bortolini[17], Victor Acuña-Alonzo[13], Samuel Canizales-Quinteros[15], Carla Gallo[4], Giovanni Poletti[4], Gabriel Bedoya[12], Francisco Rothhammer[20,21], Pierre Faux[1#], Andrés Ruiz-Linares[23,1,3#]

# These authors jointly supervised this work.

* Contributed equally to this work.

[1] UMR 7268 ADES, CNRS, Aix-Marseille Université, EFS, Faculté de Médecine Timone, Marseille, 13005, France.

[2] School of Mathematics and Statistics, Faculty of Science, Technology, Engineering and Mathematics, The Open University, Milton Keynes, MK7 6AA, United Kingdom.

[3] Department of Genetics, Evolution and Environment, and UCL Genetics Institute, University College London, London, WC1E 6BT, UK

[4] Laboratorios de Investigación y Desarrollo, Facultad de Ciencias y Filosofía, Universidad Peruana Cayetano Heredia, Lima, 31, Perú.

[5] Unit of Human Evolutionary Genetics, Institut Pasteur, Paris, 75015, France.

[6] Departamento de Tecnología Médica, Facultad de Ciencias de la Salud, Universidad de Tarapacá, Arica, 1000000, Chile.

[7] Scientific Police of São Paulo State, Ourinhos-SP, 19900-109, Brazil.

[8] CAS Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai, China
[9] Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming, China
[10] Division of Vertebrates and Anthropology, Department of Earth Sciences, Natural History Museum, London, SW7 5BD, UK.
[11] Department of Biotechnology, Kinnaird College for Women, 93 - Jail Road, Lahore, 54000, Pakistan.
[12] GENMOL (Genética Molecular), Universidad de Antioquia, Medellín, 5001000, Colombia.
[13] National Institute of Anthropology and History, Mexico City, 6600, Mexico.
[14] Department of Archaeogenetics, Max Planck Institute for the Science of Human History (MPI-SHH), Jena, 07745, Germany.
[15] Unidad de Genomica de Poblaciones Aplicada a la Salud, Facultad de Química, UNAM-Instituto Nacional de Medicina Genómica, Mexico City, 4510, Mexico.
[16] Departamento de Genética e Biologia Evolutiva, Instituto de Biociências, Universidade de São Paulo, São Paulo, SP, 05508-090, Brazil.
[17] Departamento de Genética, Universidade Federal do Rio Grande do Sul, Porto Alegre, 90040-060, Brasil.
[18] Instituto Patagónico de Ciencias Sociales y Humanas, Centro Nacional Patagónico, CONICET, Puerto Madryn, U9129ACD, Argentina.
[19] Instituto de Diversidad y Evolución Austral (IDEAus), Centro Nacional Patagónico, CONICET, Puerto Madryn, Argentina.
[20] Instituto de Alta Investigación, Universidad de Tarapacá, Arica, Arica 1000000, Chile.
[21] Programa de Genetica Humana, ICBM, Facultad de Medicina, Universidad de Chile, Santiago, Arica 1000000, Chile.
[22] Melbourne Integrative Genomics, Schools of BioSciences and Mathematics & Statistics, University of Melbourne, Melbourne, VIC 3010, Australia.
[23] Ministry of Education Key Laboratory of Contemporary Anthropology and Collaborative Innovation Center of Genetics and Development, School of Life Sciences and Human Phenome Institute, Fudan University, Yangpu District, Shanghai, China.

Abstract

Adaptation following Out-of-Africa migrations is likely to have affected human pigmentation. Here, we developed a methodology to predict how quantitative measurements of skin (melanin index) and eye (iris luminance) evolved throughout time periods and geographical regions. This prediction methodology takes into account pseudo-haploidy, an important feature of the ancient genomes due to their sequencing at low-coverage. We proceeded by designing SNP sets for prediction of each trait and by training prediction models using the CANDELA cohort, a genotyped dataset from 5 Latin American countries including ~6,000 individuals with phenotypic measurements. For validation purposes, we first applied the prediction models to 2,803 non-phenotyped contemporary individuals. We have then applied the trained models to nearly 4,000 samples selected from the large public database of ancient genomes (Harvard Ancient Genome Database v44). Our results provide a fine characterization of the gradual pigmentation changes starting from 15,000 years ago and occurring mainly in Eurasia. Among other trends, we have detected a recent increase of melanin index in all North and Eastern Asian populations in the most recent time period (0-1500 CE), which correspond to values predicted in contemporary Northern populations. Besides, focusing on the SNPs involved in melanin index prediction, we also show a significant negative trend between the substitution effects of the derived alleles and the age of these alleles. Our study further strengthens the claims of convergent melanin evolution in the European and East Asian population and show that recent genetic variants impact the phenotypes to a greater extent.

3.2.1. Introduction

There is strong evidence that, as is the case in other mammals, pigmentation in humans has evolved under the influence of selection[1–3]. In particular, skin pigmentation

in native populations from around the world (i.e. prior to the large-scale population movements starting in the late 15th century) shows a correlation with latitude, an observation that has been interpreted as probably an adaptation to variable levels of exposure to UV radiation[4,5]. It has also been suggested that sexual selection could have played a role in pigmentation evolution, particularly for hair and eye colour[6]. Consistent with some form of selection playing a role in the evolution of human pigmentation, genome-wide scans, in contemporary populations as well as in ancient DNA data, have provided evidence for selective sweeps at loci including pigmentation genes[7]. Furthermore, some evidence has been provided of the action of polygenic selection for skin pigmentation, consistent with the greater polygenicity of skin, relative to hair and eye colour[8,9].

Genome-wide association studies (GWAS) have allowed the identification of genetic variants impacting on pigmentation variation in the general human population[8,10–14]. Initial studies concentrated in Europeans[8,10–12,15,16] and, subsequently, non-European populations have begun to be examined[17–20]. We recently conducted a GWAS for skin, hair and eye pigmentation in > 6,000 Latin Americans of mixed Native American, European and African ancestry[17]. This allowed us to replicate the effect of variants previously described in Europeans and as well as to identify a number of novel variants. Some of these novel variants probably originated in East Asia prior to the initial settlement of the Americas, reflecting the parallel evolution of lighter skin pigmentation in West and East Eurasia. The current availability of a restricted number of SNPs known to affect pigmentation in the general population enables the prediction of pigmentation traits using genotypes for individuals for which there is no phenotypic information. This is of considerable interest in forensics and can also be applied to non-contemporary DNA samples[21–23]. In the case of forensics, we have recently stressed that prediction accuracy is contingent on the extent to which the genetic background of the unknown samples is represented in data used to build the prediction models[9].

Here we endeavour to reconstruct the evolution of eye and skin pigmentation across the world and across time periods based on predictions from genetic data. For this, we adapt a prediction approach that we developed in Latin Americans, benefiting from the mixed Native American, European and African background of these individuals[17,24–27]. Although phenotypic inferences from ancient DNA data using prediction models optimized based on contemporary populations should be regarded with caution, we observe several compelling patterns consistent with independent evidence. We also observe a recent trend of increasing skin pigmentation and decreasing eye luminance in the Eastern Asia, Siberia and Central Asia.

3.2.2. Methods

Definition of sets of predictive SNPs

In a previous study, we have created the "CAN-S" and "CAN-E" SNP sets to respectively predict skin MI and eye Luminance in the CANDELA samples[9]. CAN-S and CAN-E consists of 120 and 56 SNPs respectively. We replicated a similar procedure to predict

pigmentation traits for the ancient samples, starting from these SNP sets. To use our prediction models on the ancient data and to obtain a robust phenotypic prediction, we retained the SNPs common to both CAN-E(S) and ancient samples (see next section). For the remaining SNPs (i.e. those not genotyped on ancient samples), we curated sets of proxy SNPs from our CANDELA database that found a match in the ancient data. Proxy SNPs are variants pertaining to the same LD block as one of the original SNPs; they were obtained using PLINK command --clump (with options -p1 0.9, -r2 0.50, -kb 250). Because of high LD, we expect them to provide similar prediction power to the SNPs in the original sets. In addition, to extend the prediction horizon beyond the contributing ancestries within CANDELA samples to global population we needed to account for the absent and underrepresented ancestries in the prediction model. Thus, from the literature, we additionally curated genetic variants independent from the SNPs already present in the SNP sets. Altogether, these three steps led to sets of 63 SNPs for MI and 39 SNPs for Luminance prediction on the ancient samples, named "aCAN-S" and "aCAN-E" respectively (details of the SNP sets are mentioned in Supplementary table 8 and 9). This procedure is widely detailed in Supplementary method S1 and S2.

As a result, the new aCAN-S set produced prediction R^2 0.412 (weighted Linear regression, weightage scheme is detailed in Supplementary method S3) where the original CAN-S (includes 120 SNPs) along with 5 basic covariates (Age, Sex, BMI, estimated European and sub-Saharan African Ancestry) as predictors produced R^2 0.45 on the same CANDELA samples. For Luminance, the new aCAN-E with 39 SNPs produced prediction R^2 0.555 whereas the original CAN-E (includes 56 SNPs) set along with the same covariates as above produced R^2 0.606.

The aCAN-S set directly included 5 of the top 10 SNPs from CAN-S and further 3 SNPs were accounted for through proxy SNPs (new SNPs located in the same LD blocks with the corresponding SNPs from CAN-S). For Luminance, aCAN-E contains 5 of the top 10 SNPs from CAN-E (directly 4 SNPs and one through proxy SNP).

Reference population

We used the CANDELA consortium data previously analysed for GWAS and prediction of pigmentation traits[9,17,24–27] as reference population to train the prediction models. This cohort comprises of pigmentation data (among others: skin Melanin Index – MI – and digital eye colour Luminance[9]) from five Latin American countries (Brazil, Colombia, Chile, Mexico and Peru) on more than 7,000 individuals (detailed Country-wise distribution provided in Supplementary Table 1). MI was measured on the skin generally unexposed to sunlight (by reflectometry) and Luminance was obtained from the iris pigmentation of the digital photographs of the individuals. As part of quality control, we removed all the samples outside the age range of 18 to 45 years. We also removed 8 pairs of closely related samples (pairwise IBD relationship close to 1). As an obvious check, the samples with missing information on MI or Luminance were removed from further analyses for the respective traits. These filters led to finally retain 6,549 samples for MI and 5,754 samples for Luminance analysis. Both the phenotypes were quantitative in nature; their distribution over the retained samples are: MI ranging from 20.34 to 80.67

with mean = 35.14 and SD = 5.7, Luminance ranging from 0.006 to 0.559 with mean = 0.148 and SD = 0.085.

The individuals in the CANDELA studies were genotyped using Illumina's Omni Express chip (approx. 700k SNPs after quality checks) and later imputed to ~11.1M autosomal variants (using the 1000 Genomes Phase 3 data as reference). All details over chip genotyping and imputation are provided in Adhikari et al[17].

Predicted ancient samples

We collected the data on the ancient *Homo Sapiens* samples from publicly accessible Harvard Ancient Genomes Database (https://reich.hms.harvard.edu/datasets - hereafter denoted as HAGDB, version V44 downloaded on 1st May 2021). This data includes archaic hominin (Denisova and Neanderthal) and ancient human samples mostly coded as pseudohaploids. Due to quality degradation and low coverage samples, nearly all the ancient DNA records are pseudohaploids (except 5 samples, which were coded as diploids). By "pseudohaploid", we mean that only one allele of each SNP was recorded. Therefore, the recorded alleles from consecutive SNPs were not expected to originate from the same chromosome. In contrast, the reference samples from CANDELA are densely genotyped and diploid.

We excluded 11 archaic hominin samples present in the data, as we focus only on ancient *Homo Sapiens*. We retained ancient human samples based on unique ID and whose predicted age ('Master ID' and 'Date mean in BP' columns respectively in annotation file v44.0_1240k_public.anno) is more than 450 years (counting from 1950 AD, see Article 2, Table 1). To further ensure to avoid all the samples having recent admixture, we removed the ones having estimated 'FullDate' from v44.0_1240k_public.anno going beyond 1500AD, which corresponds to the beginning of the modern age migrations. The 5,084 retained samples are spread across the globe, with continental Europe constituting more than 58% of the samples and continental Africa less than 2% (detailed distribution of the aDNA samples in Supplementary tables 3 and 4). We have grouped samples into 15 subcontinental areas following current administrative borders (with some rearrangements – see Supplementary Note 2 for details). This partitioning led to exclude 57 samples from Morocco, Canary Islands and Guatemala as well as samples which could not be mapped to any sub-continental geographic region.

The aDNA samples from the HAGDB contains 1,150,639 autosomal SNPs. As genotyping call rate per sample was low, we retained only the ones which had at least 10% of all the SNPs genotyped. Out of these samples, the average genotyping rate was 51.8% (median = 55.5%). As an additional check to obtain a robust prediction of the phenotypes, we removed the ancient samples having less than 5 SNPs in both the corresponding aCAN-E and aCAN-S SNP sets. We finally predicted the pigmentation traits on 3,597 ancient human samples (global distribution of the ancient samples has been presented in Supplementary figure S1).

Predicted contemporary samples

Besides ancient genomes, we have also predicted the pigmentation phenotypes from contemporary genomes for validation purposes. Two publicly-available databases (of diploid data) were used: CEPH-HGDP[28] and 1000 Genomes Phase 3[29] (henceforth

denoted as HGDP and 1000GP, respectively). We discarded some populations to avoid recent continental admixture (including admixed Americans – see retained populations in Supplementary Table 2), leaving 1901 samples from 1000GP and 929 samples from HGDP, distributed across 8 sub-continental ancestry regions and 70 populations.

We tested predictions under 2 modalities: diploid and pseudohaploid. For the pseudohaploid modality, we retrieved the data from the HAGDB release for 1000GP samples (as pseudohaploids genotypes of these samples were released along with HAGDB ancient samples) and generated pseudohaploid genotypes for the HGDP. To that end, we used the original diploid genotypes for HGDP samples (mentioned above) and randomly altered each heterozygotes to one of the homozygote and repeated the said procedure multiple times to remove the effect of randomisation.

Pseudohaploid prediction and diploid to pseudohaploid conversion

Because of the difference in the genotyping methods in the reference CANDELA samples (diploid) and ancient human data (pseudohaploid) we have thus devised the following algorithm to train a model on diploid samples that could then be applied to predict on the pseudohaploid samples.

For the diploid samples, we clustered together the ones called at the exact same SNPs, i.e. having the exact same set of SNPs present in the data. For each such cluster, we trained the models on the full CANDELA data selecting the exact same set of SNPs, so that trained models have the exact same variables as the samples to be predicted. Each time we trained a weighted Linear Regression model (higher weightage is given to under-represented ancestries; see the weightage scheme in Supplementary methods S3) on the CANDELA samples in a 10-fold cross-validation format where we use 90% data to train and hold out 10% data to calculate prediction R^2. This process (also explained in our previous article[9]) is repeated 10 times each time keeping aside one fold to calculate R^2. Therefore, we got 10 predictions of the corresponding phenotype for each diploid sample. We took the average of them as the final predicted value and proceeded with further analysis.

For the pseudohaploid samples, we added an intermediary step before training the models so that the CANDELA (both training and hold out part) data becomes more consistent with the aDNA samples. For all the CANDELA individuals having heterozygous genotypes we randomly assigned homozygous genotype for the reference or alternate allele. We repeated this random allocation 100 times to remove any bias caused by the randomisation. This forced homozygous data, which would be analogous to the pseudohaploid genotypes, is used as training data and the prediction is achieved as described here above for diploid samples. In both cases, we also made sure to align the reference alleles in the training and prediction data.

Correlations of predicted phenotypes with solar radiation and latitude

To assess the correctness of our prediction in contemporary humans, we correlated the predicted phenotypes in modern populations (1000GP and HGDP) with ultra-violet radiation (UVR) estimated at the places of origin of these populations and with the absolute latitude of these places. The places of origin of combined 73 populations

were obtained from 1000G and HGDP databases (Supplementary Table 2)[28,29]. We adjusted those of STU (Sri Lankan Tamil in the UK – 1000G) and ITU (Indian Telugu in the UK – 1000G) to match their places of origin (respectively North of Sri Lanka and Hyderabad in India) instead of their sampling locations (both London, UK). As UVR estimate, we interpolated insolation incident on a horizontal surface (from the 6th release (2008) of the NASA Surface meteorology and Solar Energy) at the place of origin of each population. This parameter returns the amount of kWh per square meter and per day, averaged over 22 years.

Correlations of derived allele age with allele alteration effect

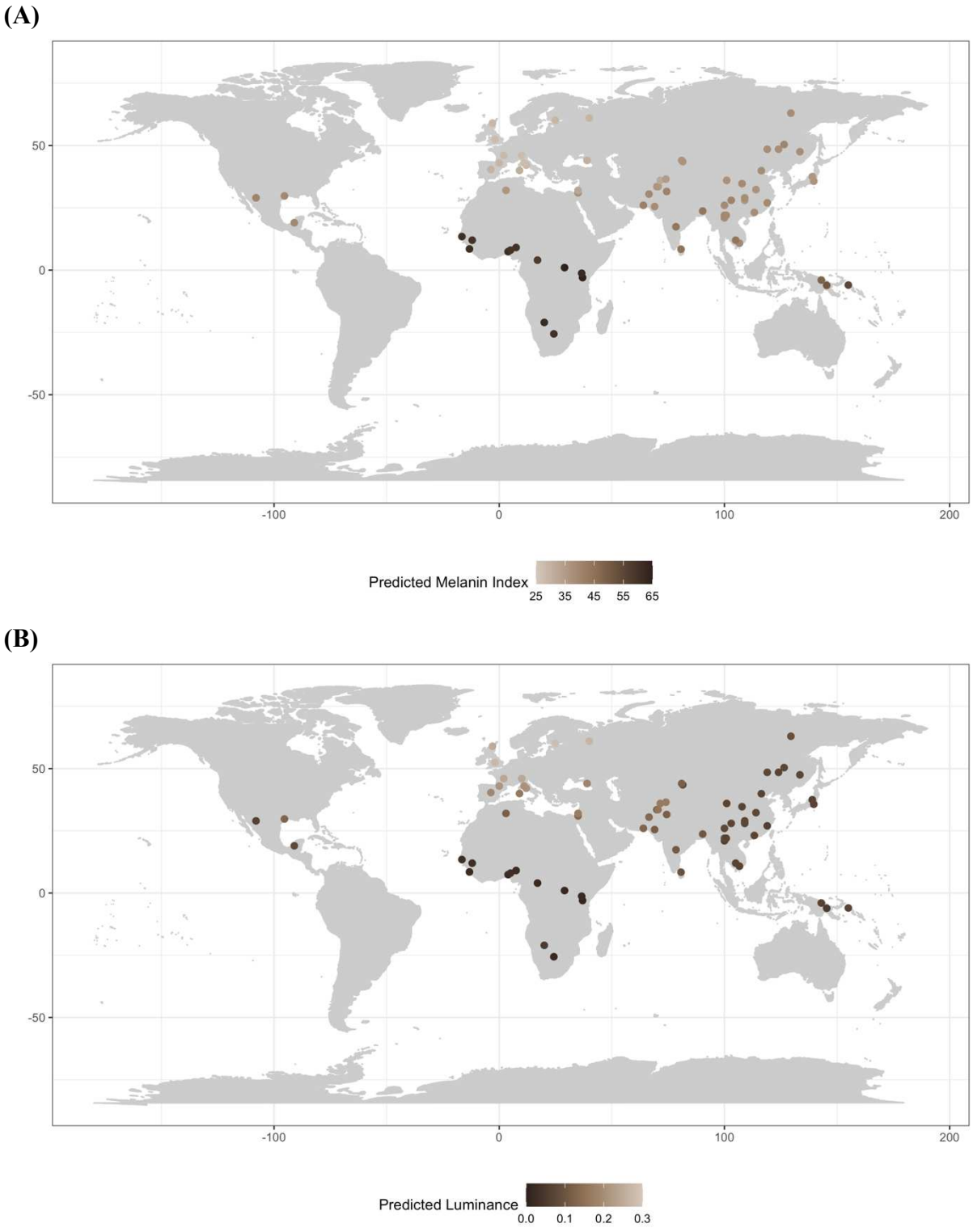
Human pigmentation has long been identified to be under selection pressure[7,30–32]. But most of the studies concentrated on the samples found in the European continent. We were interested to see the effect of the derived alleles on the broader human population. To that end, we obtained the effects (β) of the 63 SNPs from aCAN-S on MI and those of the 39 SNPs from aCAN-E on Luminance, from each of the 10-fold cross validation models trained on the CANDELA data (see above) and retained the average effect size for each SNP, aligned to the derived allele. Besides, we obtained two age estimates for these alleles. Firstly, we obtained empirical estimates of the derived allele age from the aDNA data, by recording the five earliest occurrences of the derived allele within the samples and as a point estimation we used the median of the five earliest instances (i.e. the third earliest instance of the derived allele – henceforth simply mentioned as the age of the derived allele). Secondly, we obtained the allelic age estimates obtained by the method of GEVA[33] (https://human.genome.dating). We computed the correlation of the allele effect on the phenotypes and the age of the derived alleles.

3.2.3. Results

3.2.3.1. Prediction of pigmentation in contemporary samples

We first sought to assess the validity of our prediction approach on contemporary samples. To that end, we predicted pigmentation for 2,830 geo-referenced population samples from around the world (combining data from the 1000 Genome[29] and CEPH-HGDP[28] projects, after discarding populations with evidence of recent inter-continental admixture).

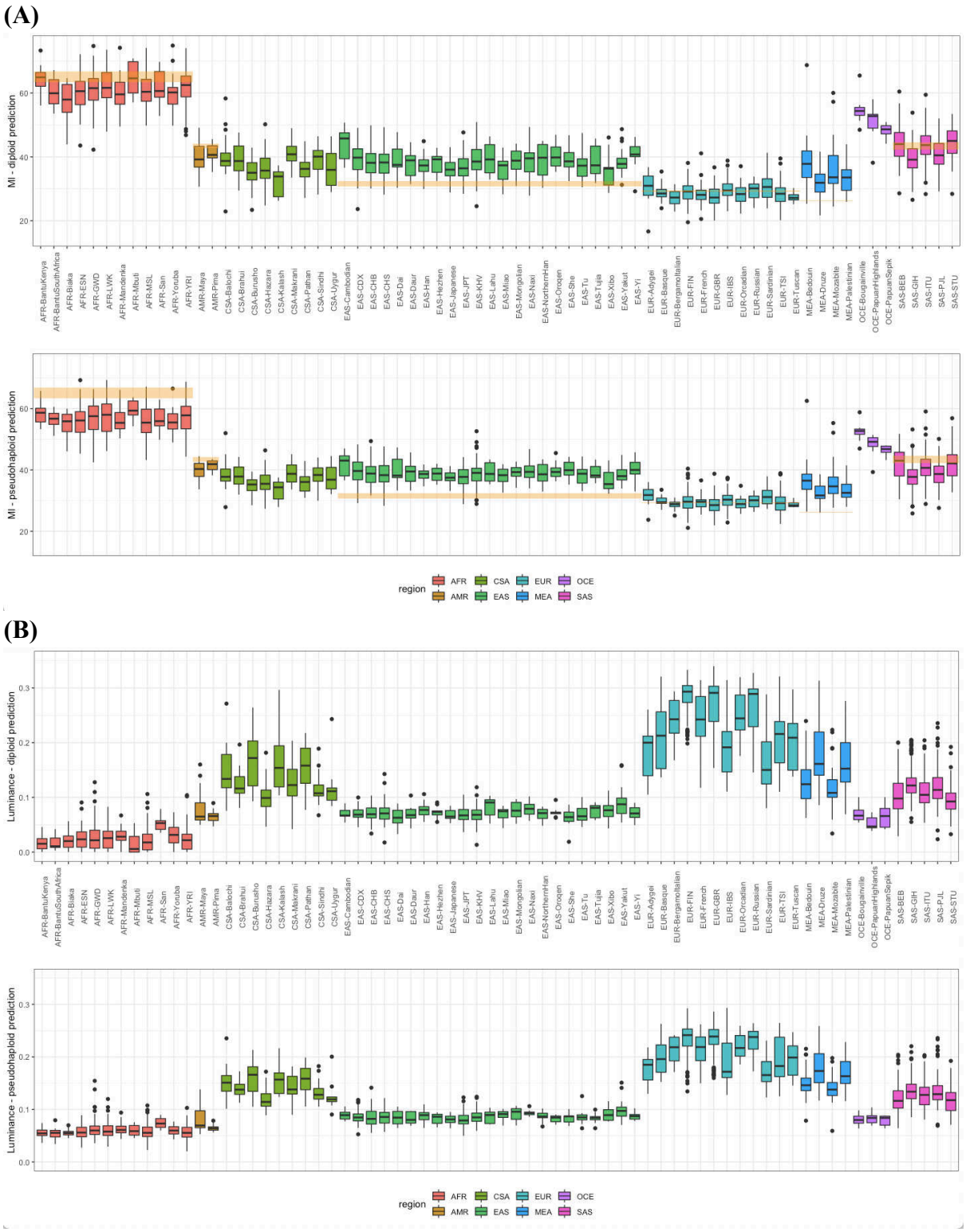
Since most currently available ancient DNA data is in haploid form prediction in the contemporary samples was performed under two modalities: (a) using diploid genotypes and (b) using pseudohaploid genotypes (randomly assigning any heterozygous genotype to a homozygous state). Article 2, Figure 1 shows the geographical distribution of mean predicted pigmentation of skin (Melanin Index, MI) and eyes (Luminance) using diploid data.



Article 2, Fig. 2 Prediction on the diploid contemporary datasets

(A) Skin Melanin Index and (B) Eye colour Luminance has been predicted for the 2,803 samples retained from 1000Genome Phase III and CEPH-HGDP (populations with recent admixture were excluded). We used the Linear Regression models (trained on diploid CANDELA data) to predict the pigmentation traits from diploid data.

Comparing different continents, the lowest and highest predicted MI were obtained for the European and sub-Saharan populations respectively. Predicted Luminance broadly followed an opposite trend across world populations, but with eye colour variation being essentially restricted to Europeans/Western Asians, as seen before. The phenotypic predictions obtained here are in line with direct observations of skin and eye pigmentation in world populations [21,34]. Within continents, population variation in mean predicted pigmentation is also consistent with observed regional variation in pigmentation (e.g. North vs South Europe or East Asia vs South Asia). These tendencies are also reflected in the box plots shown in top-panels of Article 2, Figure 2A and 2B (average predicted phenotypes for the population cohorts have been reported in the Supplementary Table 6). For MI (Article 2, Figure 2A), we also reported observed values for most subcontinental groups and found them well in-line with the predicted values of these groups. Article 2, Figure 2 shows the results of pigmentation prediction using pseudohaploid data. Similar geographic trends are seen as for predictions in diploid data, albeit with a wider overall range in the diploid data predictions (Supplementary Table 6). As a consequence, pseudohaploid predictions are less in line with direct population observations, compared to diploid predictions, especially for the population with more extreme values (i.e. Africans and Europeans).



Article 2, Fig. 3 Prediction comparison on the diploid and haploid genotypes using contemporary samples

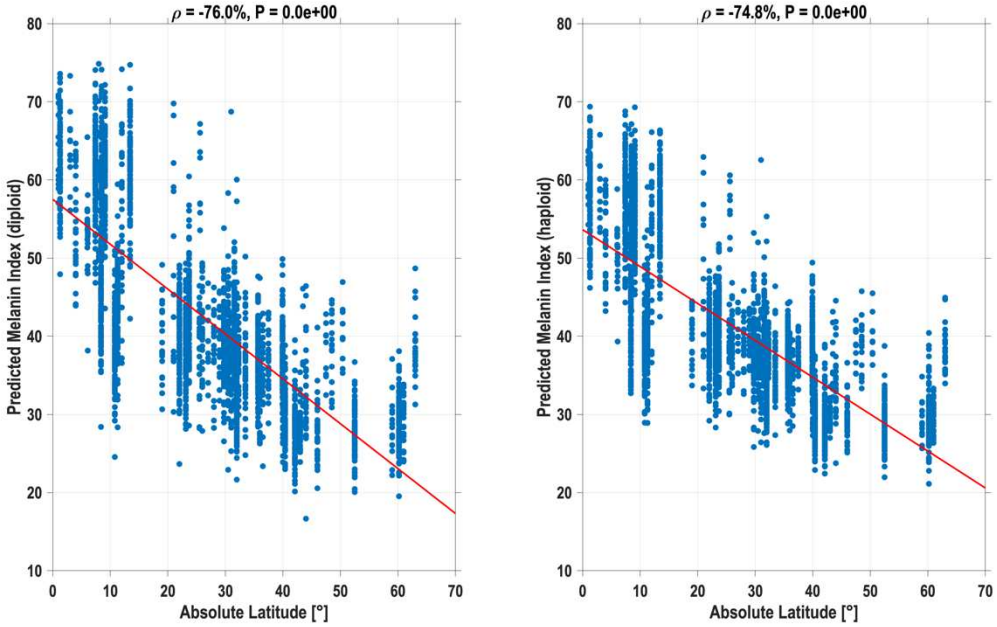
(A) Skin Melanin Index (MI) and (B) Eye colour Luminance (Luminance) respectively has been predicted for the 2,803 samples combined from 1000 Genomes Phase 3 and CEPH-HGDP studies (removing Latin American and African American samples to avoid recent admixture). For MI (A), yellow horizontal bars correspond to the (average±standard error) of observed values over a subcontinental group.

3.2.3.2. Correlation between predicted pigmentation and environmental variables

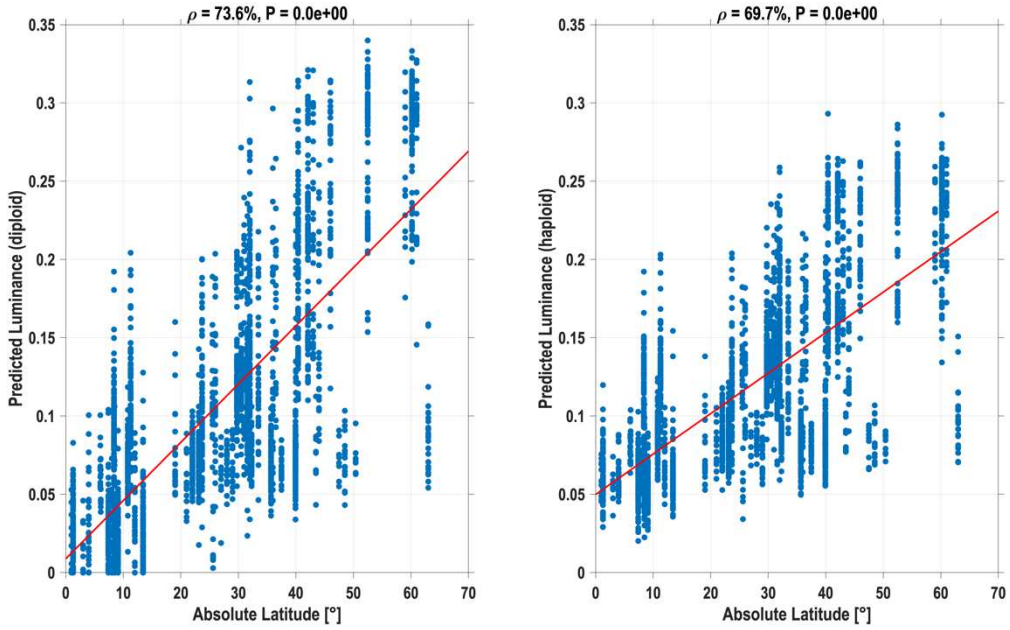
We next assessed the correlation between our phenotypic predictions and two environmental variables, absolute latitude and solar radiation. It has been shown[5,35,36] that these variables significantly correlate with mean observed pigmentation in human populations. Unlike the pigmentation phenotypes, these geographic variables are available for the 1000GP and HGDP samples. We obtained the geographic coordiantes for each of the 70 sampled populations from the annotation files of the 1000GP and HGDP. We then interpolated publicly available data (from https://eosweb.larc.nasa.gov/sse/) of solar radiation at these 70 locations.

We observe trends in pigmentation variation in relation to solar radiation and geographic location consistent with the literature (i.e. MI increasing/decreasing with solar radiation/absolute latitude and opposite trends for Luminance). Howerer the strengths of correlation (and P-value) widely varied considerably. Stronger (and more significant) correlations were observed with absolute latitude than with solar radiation, e.g. the lowest R2 was 48% for absolute latitude versus 26% for solar radiation (considering 1000GP and HGDP data together). Article 2, Figure 3 shows the correlations with absolute latitude (solar radiation correlations are shown in in Supplementary figure S2). Furthermore, we observed that diploid predictions yield stronger correlations than haploid predictions, consistent with the lower accuracy of the haploid predictions. As seen in Article 2, Figure 2 and the Y-axes of Article 2, Figure 3, the range of pseudohaploid predictions is more narrow; in line with the lower significance of correlations with those predictions. Finally, we observed stronger correlations with MI than with Luminance and in 1000GP data than on data from HGDP (Supplementary Table 5).

(A)



(B)



Article 2, Fig. 4 Correlation between absolute latitude and predicted pigmentation phenotypes

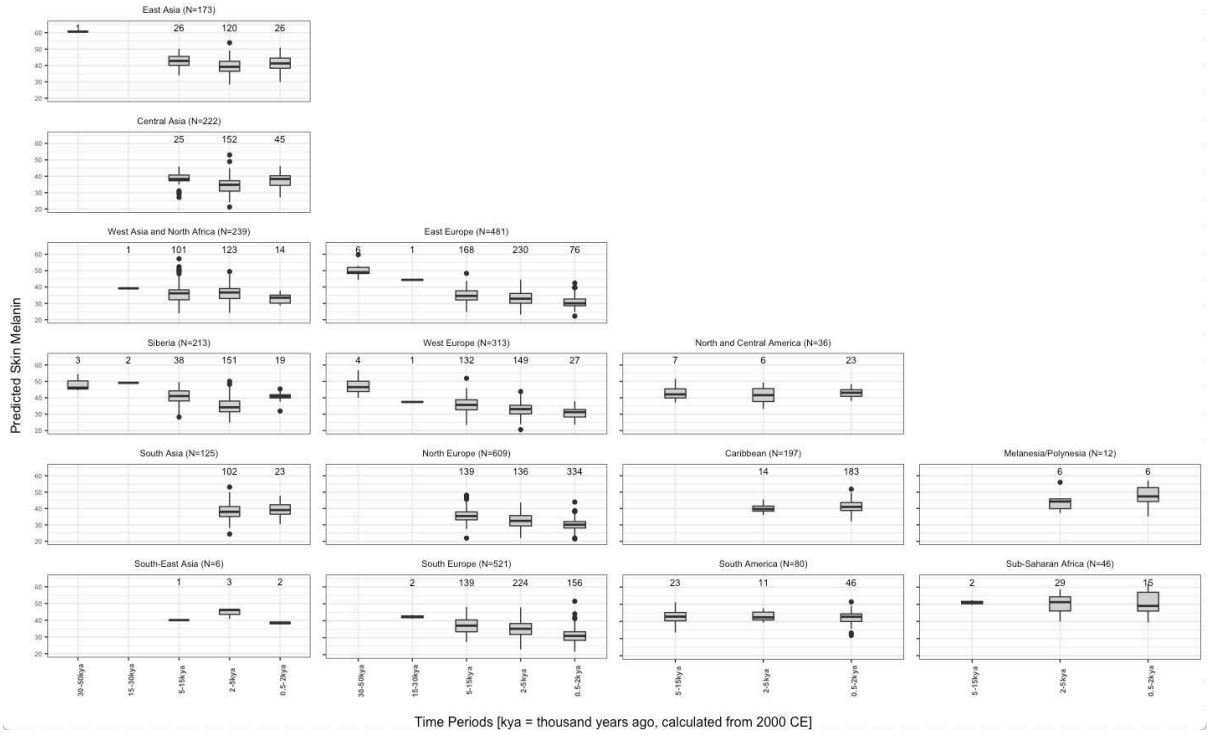
Correlation of latitude of the sampling location (absolute value) of the 1000GP and CEPH-HGDP studies with the predicted (A) skin Melanin Index and (B) eye colour Luminance. Prediction of the pigmentation phenotypes were based on both diploid and pseudohaploid genotypes (on left and right panels, respectively).

3.2.3.3. Prediction of pigmentation in ancient human samples

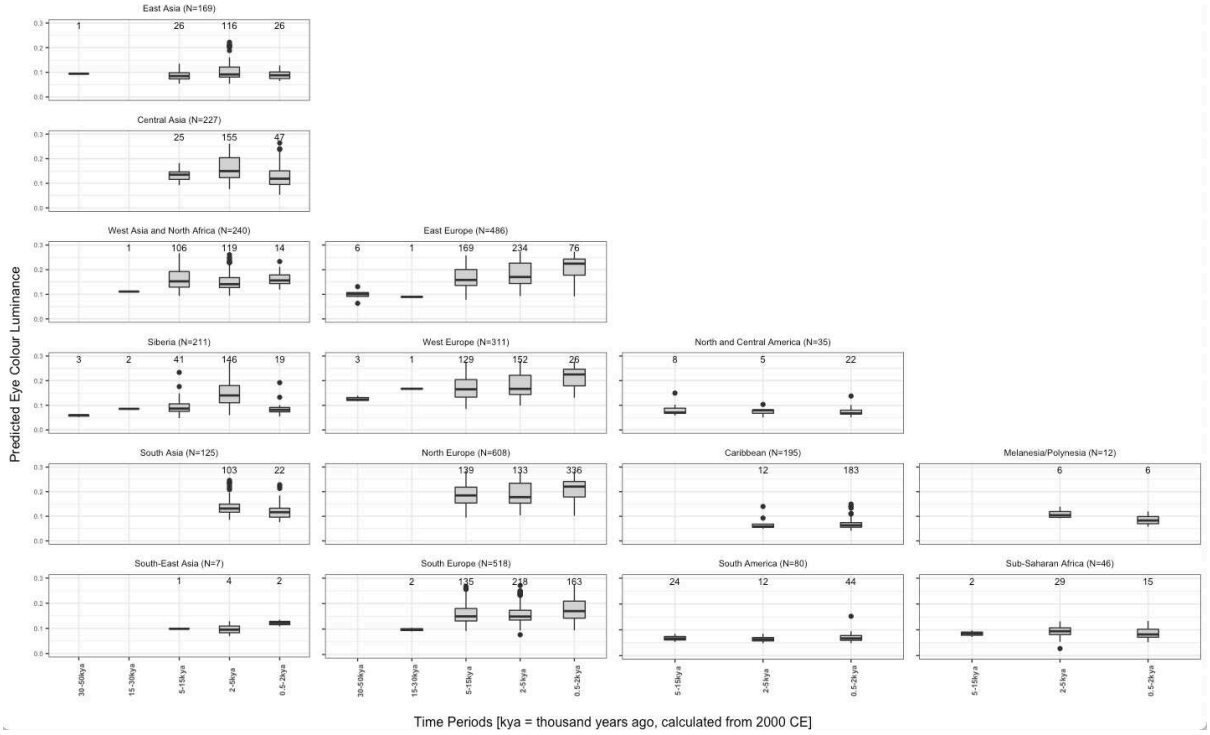
For prediction of pigmentation in ancient human samples we used the HAGDB database (Harvard Ancient Genomes Database, https://reich.hms.harvard.edu/datasets, version V44). We grouped the samples in the HAGDB dataset dated prior to 450 years ago into five time periods (30-50kya, 15-30kya, 5-15kya, 2-5kya and 0.5-2kya, kya = 1000 years ago, calculated from 2000AD) and fifteen geographic regions (6 subregions in Asia, 4 subregions in Europe, 3 in combined North and South America, sub-Saharan Africa and Melanesia/Polynesia) (Article 2, Table 1). Although the 10 Eurasian subregions have been grouped into five time periods in other parts of the world data is available for at most three of these time periods (the most recent, starting from 15,000 YA). We obtained average predicted pigmentation in each geographic region across time periods (Article 2, Figure 4). Ancient humans (30 to 50 kya) are predicted to have highly pigmented skin (Article 2, Figure 4). In Eurasia we observe that in more recent times there is a drop in predicted pigmentation, which occurs later and is less pronounced as one goes from West to East. Central Asia, East Asia and Siberia show a gradual decrease of MI until the penultimate time interval (2–5 kya) then skin pigmentation increases for all the regions in the most recent time interval considered.In the Americas, the oldest samples (from around the time of initial colonization of the continent, ~20 kya) have similar pigmentation levels to East Asian samples from the same time period. In North America we observe a drop in pigmentation for samples dated at <5 kya, similar to what predicted for East Asia. This drop in pigmentation is circumscribed to North America, the rest of the continent showing similar pigmentation levels as upon initial colonization. West Asia and North Africa show a gradual decrease in pigmentation up to until 1500 AD, similar to the four European regions where it follows a monotonic trend[37]. South Asia data is only available for the two most recent time periods and shows an increase in MI with time, similar to the trend seen in East and Central Asia and Siberia, as well as Melanesia/Polynesia. Overall only data from the Caribbean, South America and Sub-Saharan Africa show no evidence of a trend in pigmentation variation with time, but rather remain relatively constant across time periods (see Article 2, Figure 4A).

As seen in contemporary data, trends in Luminance predictions have a more restricted range compared to MI. Variation in Luminance is mainly observed in continental Europe and some parts of Central, North and Western Asia. Outside these regions, we do not observe any significant variation in predicted Luminance (see Article 2, Figure 4B). Predictions for West Asia and North Africa show a gradual increasing trend, as for Europe. For East and Central Asia and Siberia, there is a trend towards higher Luminance until the penultimate time period (2–5 kya), which has reversed in the most recent interval. This trend is similar to what has been predicted for MI in these regions, albeit in the opposite direction. For South Asian data, we see a small decreasing trend in the recent two intervals. This is similar to East and Central Asia and Siberia for the truncated time period.

(A)



(B)



Article 2, Fig. 5 Spatio-temporal prediction on the haploid ancient human samples

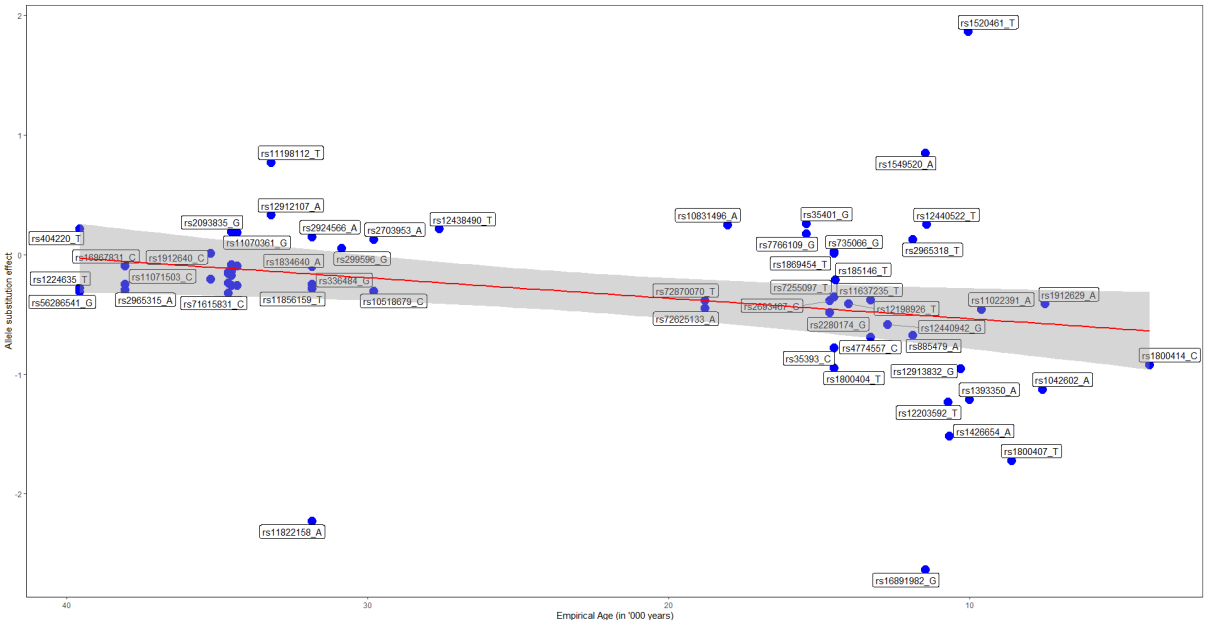
Skin Melanin Index (MI) and eye colour Luminance (Luminance) has been predicted for the 3,597 samples (Figure A and B respectively for MI and Luminance). We used pseudohaploidized CANDELA data to train our linear regression models and predicted on

the ancient samples. We divided the available aDNA data into 15 geographical regions and 5 time intervals (3 time intervals in case of lack of samples from the earlier intervals) to display the evolution trends of human pigmentation.

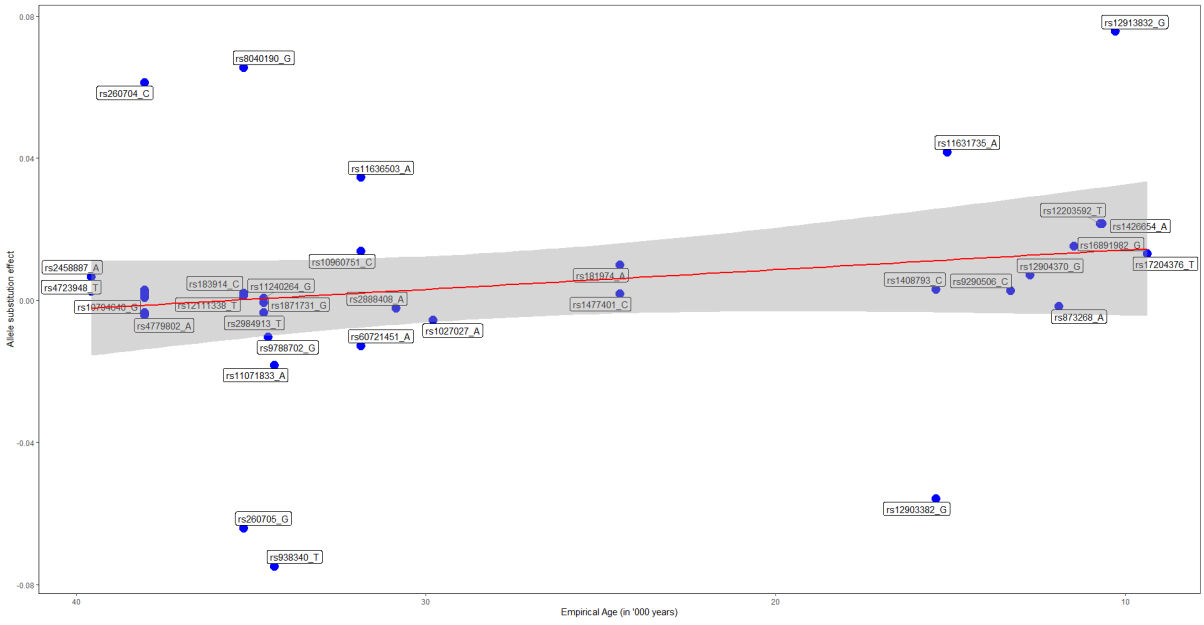
3.2.3.4. Allelic effects and age of pigmentation alleles

The predictions shown in Article 2, Figure 4 point to a relatively recent decrease in pigmentation in certain human populations. We examined the relationship between the phenotypic effect of each derived allele included in the pigmentation prediction model and the estimated age of that allele. We estimated allele age empirically (the oldest sample in which this allele is present) as well as using a populations genetic model (GEVA)[33] (these two allele age estimates are moderately correlated (for MI: r = 0.34, P-value 0.005, Luminance : 0.29, P-value 0.07; Supplementary table 7 and Supplementary figure S3)). We observe a significant correlation between allele age and allelic effect on skin and eye pigmentation (Article 2, Figure 5A for MI and 5B for Luminance), with a trend for younger alleles (<20 kya) to have a stronger phenotypic effect. This correlation is stronger for empirically estimated allele age than for the GEVA estimates. Only 2/11 of the young alleles (i.e. estimated age less than 20 kya) having a negative effect on eye pigmentation, compared to 10/25 of the older alleles (i.e. estimated age more than 30 kya) are associated with decreasing pigmentation levels, including some having relatively strong effects (Article 2, Figure 5). For MI, only 9 out of 32 young alleles (i.e. estimated age less than 20 kya) have a positive effect.

(A)



(B)



Article 2, Fig. 6 Allelic age compared to effect on pigmentation

The allele alteration effects captured from the prediction models trained on CANDELA samples are compared against the allelic age of the genetic variants present within the models. The age is calculated empirically from the ancient human samples in HAGDB. The earliest 5 instances of the derived alleles are kept in the plots and coefficients of the variants obtained from the linear regression models are regressed on the median of the earliest instances. (A) Skin Melanin Index and (B) Eye colour Luminance

3.2.4. Discussion

Inferences on ancient DNA data using prediction models trained on data from contemporary populations should be regarded with caution. The estimated effect sizes of the genetic variants derived from the modern sampling cohorts may not resemble the exact effect those had in the ancient population due to, for instance, gene-environmental interactions, epistatic effects[23] or variants not included in the prediction. Despite these limitations, it is remarkable that in the predictions performed here we observe patterns that match well established data from various other sources. Our prediction on the European continent shows a similar trend identified by the previous articles[7,23]. In a recent article[37], the authors show the West Eurasians were of intermediate complexion from HIrisPlex-S system, which is in line with our predicted range of MI. In our previous article[9], we have shown that a direct comparison with the categorical skin colour levels can be arranged if we consider 33 to 47 indices on continuous MI scale as a range for intermediate skin colour category where below 33 and over 47 should be considered as Pale and Dark complexions, respectively. Though due to lack of ancestral diversity in the HIrisPlex-S training data[21] and presence of heterozygous in the model is not ideal for the prediction on wide-spread ancient psudohaploid samples, we can still establish certain commonalities in the final prediction results. We have also shown that in the three most recent time intervals (15 kya onwards), the West Asian samples had significantly greater predicted skin pigmentation than the East European samples, just as reported in Lazaridis et al 2022[37].

Also the predicted trend for the Eastern Eurasian population show a parallel trend of skin lightening compared to the European cohorts, though due to a different selection force[20,23,38]. We have also shown that the first Americans did not go through a similar variation in skin or eye colour. Since the division of the Eastern Asians and Native Americans, the later cohort did not go through another selection pressure and pigmentation traits maintained similar levels to that from the before the split.

It has been indicated in the literature that latitude and solar radiation significantly relates to the pigmentation of the habitants of a certain region[4,5,35]. Since performance of our models could not be evaluated on the contemporary datasets (due to lack of phenotypic information), we tested the correlation of the predicted pigmentation phenotypes and solar radiation of the sampling sites. Our analysis shows strong correlation between the predicted phenotypes on the contemporary samples and the sampled locations (See Supplementary table 5). Also the studies on pigmentation phenotypes allowed us to gather information on the trait distribution (mostly for MI) on certain ancestry cohorts which we used as benchmarks for our prediction accuracies on the contemporary datasets (1000GP + HGDP). The broad level agreement of predicted phenotypes and average benchmark numbers sufficed to extend our prediction models to new samples. But at the same time, the availability of contemporary samples in both the modalities – diploid and pseudohaploid and subsequent agreement in the prediction outcome irrespective of the modality proved the efficacy of our newly devised algorithm.

Altogether, these results allowed us to apply our prediction models to non-phenotyped pseudohaploid ancient human samples.

It is worth mentioning that the said correlation of the solar radiation and predicted pigmentation takes a small hit for the high latitude samples EUR-FIN (Article 2, Figure 3) as we predict them to be of darker skin pigmentation not aligning to the decreasing trend along the latitudes, which is counter-intuitive. But it has been discussed in the previous articles[30] that these aboriginal population living in high latitude regions tend to be darker to protect themselves from the extended UV radiation during the long days and UV reflectance from the snow. Also their dietary plans fulfil the requirement for Vitamin D[39].

Though the prediction on the ancient samples should be examined with caution as the population structure of the different time intervals as well as CANDELA samples have never been considered and accounted in the models. However, please note that multiple genes present in our SNP sets have already been through selection for various ancestry cohorts (e.g. MC1R, OCA2, TYR, SLC24A5, SLC45A2, IRF4 for European, MC1R, OCA2 for Asian, OCA2 for Oceanian, SLC24A5, SLAC45A2 for African population)[30–32].

Also CANDELA samples have already been explained as a admixture of European, African and native American ancestries[24,25]. This puts us in an advantageous position in terms of the world ancestry coverage, still we did not deem this to be sufficient to predict on all the contemporary cohorts and as an extension not on the entire HAGDB. To mitigate this issue, we categorically selected the ancestry cohorts we miss in the CANDELA samples – east Asian and south Asian population. We also added African studies as original CANDELA samples represent only 6% sub-Saharan African ancestry.

At the same time, the prediction results are derived from a novel approach we undertook to counter the pseudohaploids available through the aDNA database. In Article 2, Figure 2, we show that prediction on the diploid and pseudohaploids follow a similar pattern. Additionally, strong correlation of the predicted phenotypes and latitude or solar radiation proves our diploid to haploid conversion method adapted in the prediction models works satisfactorily (see the correlation on Supplementary table 5). The detailed weightage method takes care of the ancestries scarcely represented in our CANDELA data. This results hints towards successful portability of our prediction models on the HAGDB datasets.

Article 2, Table 1 Number of ancient human samples examined for five different time periods and 15 geographic subregions

	50,000 to 30,000 years ago	30,000 to 15,000 years ago	15,000 to 5,000 years ago	5,000 to 2,000 years ago	2,000 to 500 years ago	**Total**
North and Central America	-	-	9	7	31	**47**
Carribean	-	-	-	14	198	**212**
South America	-	-	25	12	51	**88**
West Europe	4	1	148	171	29	**353**
North Europe	-	-	151	153	367	**671**
South Europe	-	2	149	237	168	**556**
East Europe	6	2	191	250	90	**539**
West Asia and North Africa	-	1	111	132	15	**259**
Central Asia	-	-	30	169	55	**254**
South Asia	-	-	-	106	25	**131**
South-East Asia	-	-	1	7	3	**11**
East Asia	1	-	29	124	28	**182**
Siberia	3	2	47	164	19	**235**
Sub-Saharian Africa	-	-	2	29	15	**46**
Melanesia/Polynesia	-	-	-	7	6	**13**
Total	**14**	**8**	**893**	**1582**	**1100**	**3597**

3.2.5. References

1. Pickrell, J. K. *et al.* Signals of recent positive selection in a worldwide sample of human populations. *Genome Res.* **19**, 826–837 (2009).
2. Voight, B. F., Kudaravalli, S., Wen, X. & Pritchard, J. K. A Map of Recent Positive Selection in the Human Genome. *PLoS Biol.* **4**, e72 (2006).
3. Mathieson, I. *et al.* Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* **528**, 499–503 (2015).
4. Jablonski, N. G. & Chaplin, G. The evolution of human skin coloration. *J. Hum. Evol.* **39**, 57–106 (2000).
5. Jablonski, N. G. & Chaplin, G. Human skin pigmentation as an adaptation to UV radiation. *Proc. Natl. Acad. Sci.* **107**, 8962–8968 (2010).
6. Jablonski, N. G. & Chaplin, G. The colours of humanity: the evolution of pigmentation in the human lineage. *Philos. Trans. R. Soc. B Biol. Sci.* **372**, 20160349 (2017).
7. Wilde, S. *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.* **111**, 4832–4837 (2014).
8. Candille, S. I. *et al.* Genome-Wide Association Studies of Quantitatively Measured Skin, Hair, and Eye Pigmentation in Four European Populations. *PLoS ONE* **7**, e48294 (2012).
9. Palmal, S. *et al.* Prediction of eye, hair and skin colour in Latin Americans. *Forensic Sci. Int. Genet.* **53**, 102517 (2021).
10. Kayser, M. *et al.* Three Genome-wide Association Studies and a Linkage Analysis Identify HERC2 as a Human Iris Color Gene. *Am. J. Hum. Genet.* **82**, 411–423 (2008).
11. Sulem, P. *et al.* Genetic determinants of hair, eye and skin pigmentation in Europeans. *Nat. Genet.* **39**, 1443–1452 (2007).
12. Sulem, P. *et al.* Two newly identified genetic determinants of pigmentation in Europeans. *Nat. Genet.* **40**, 835–837 (2008).
13. Liu, F. *et al.* Genetics of skin color variation in Europeans: genome-wide association studies with functional follow-up. *Hum. Genet.* **134**, 823–835 (2015).
14. The International Visible Trait Genetics Consortium *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**, 652–656 (2018).
15. Han, J. *et al.* A Genome-Wide Association Study Identifies Novel Alleles Associated with Hair Color and Skin Pigmentation. *PLoS Genet.* **4**, e1000074 (2008).
16. Liu, F. *et al.* Digital Quantification of Human Eye Color Highlights Genetic Association of Three New Loci. *PLoS Genet.* **6**, e1000934 (2010).
17. Adhikari, K. *et al.* A GWAS in Latin Americans highlights the convergent evolution of lighter skin pigmentation in Eurasia. *Nat. Commun.* **10**, 358 (2019).
18. Crawford, N. G. *et al.* Loci associated with skin pigmentation identified in African populations. *Science* **358**, eaan8433 (2017).

19. Martin, A. R. *et al.* An Unexpectedly Complex Architecture for Skin Pigmentation in Africans. *Cell* **171**, 1340-1353.e14 (2017).
20. Edwards, M. *et al.* Association of the OCA2 Polymorphism His615Arg with Melanin Content in East Asian Populations: Further Evidence of Convergent Evolution of Skin Pigmentation. *PLoS Genet.* **6**, e1000867 (2010).
21. Chaitanya, L. *et al.* The HIrisPlex-S system for eye, hair and skin colour prediction from DNA: Introduction and forensic developmental validation. *Forensic Sci. Int. Genet.* **35**, 123–135 (2018).
22. King, T. E. *et al.* Identification of the remains of King Richard III. *Nat. Commun.* **5**, 5631 (2014).
23. Ju, D. & Mathieson, I. The evolution of skin pigmentation-associated variation in West Eurasia. *Proc. Natl. Acad. Sci.* **118**, e2009227118 (2021).
24. Adhikari, K., Mendoza-Revilla, J., Chacón-Duque, J. C., Fuentes-Guajardo, M. & Ruiz-Linares, A. Admixture in Latin America. *Curr. Opin. Genet. Dev.* **41**, 106–114 (2016).
25. Ruiz-Linares, A. *et al.* Admixture in Latin America: Geographic Structure, Phenotypic Diversity and Self-Perception of Ancestry Based on 7,342 Individuals. *PLoS Genet.* **10**, e1004572 (2014).
26. Adhikari, K. *et al.* A genome-wide association scan implicates DCHS2, RUNX2, GLI3, PAX1 and EDAR in human facial variation. *Nat. Commun.* **7**, 11616 (2016).
27. Adhikari, K. *et al.* A genome-wide association scan in admixed Latin Americans identifies loci influencing facial and scalp hair features. *Nat. Commun.* **7**, 10815 (2016).
28. Bergström, A., Stringer, C., Hajdinjak, M., Scerri, E. M. L. & Skoglund, P. Origins of modern human ancestry. *Nature* **590**, 229–237 (2021).
29. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).
30. Deng, L. & Xu, S. Adaptation of human skin color in various populations. *Hereditas* **155**, 1 (2018).
31. Sturm, R. A. & Duffy, D. L. Human pigmentation genes under environmental selection. *Genome Biol.* **13**, 248 (2012).
32. Beleza, S. *et al.* The Timing of Pigmentation Lightening in Europeans. *Mol. Biol. Evol.* **30**, 24–35 (2013).
33. Albers, P. K. & McVean, G. Dating genomic variants and shared ancestry in population-scale sequencing data. *PLOS Biol.* **18**, e3000586 (2020).
34. Walsh, S. *et al.* The HIrisPlex system for simultaneous prediction of hair and eye colour from DNA. *Forensic Sci. Int. Genet.* **7**, 98–115 (2013).
35. Chaplin, G. Geographic distribution of environmental factors influencing human skin coloration. *Am. J. Phys. Anthropol.* **125**, 292–302 (2004).
36. Chaplin, G. & Jablonski, N. G. Vitamin D and the evolution of human depigmentation. *Am. J. Phys. Anthropol.* **139**, 451–461 (2009).
37. Lazaridis, I. *et al.* A genetic probe into the ancient and medieval history of Southern Europe and West Asia. *Science* **377**, 940–951 (2022).
38. Norton, H. L. *et al.* Genetic Evidence for the Convergent Evolution of Light Skin in Europeans and East Asians. *Mol. Biol. Evol.* **24**, 710–722 (2006).

39. Schæbel, L. K., Bonefeld-Jørgensen, E. C., Laurberg, P., Vestergaard, H. & Andersen, S. Vitamin D-rich marine Inuit diet and markers of inflammation – a population-based survey in Greenland. *J. Nutr. Sci.* **4**, e40 (2015).

3.3. Article 3: Transcriptome-wide association study reveals novel calcium channel gene influencing skin pigmentation in Latin Americans

This article is in preparation. Supplementary information is available in Appendix D. In this article's analyses my contributions are:

- Data preparation:
  - Phenotype quality control,
  - Summary data alignment from previous GWAS on CANDELA cohort
- TWAS pipeline implementation
  - Analysis of single tissue models,
  - Analysis of multi-tissue models
- Post TWAS analysis:
  - Gene annotations,
  - results visualization (Manhattan plots),
  - comparison plots
- Draft preparation and writing

Authors and affiliations

Sagnik Palmal[1*], Kaustubh Adhikari[2,3*], *Javier Mendoza-Revilla[4,5], Caio C. Silva de Cerqueira[6], Juan Camilo Chacón-Duque[7], Malena Hurtado[4], Valeria Villegas[4], Vanessa Granja[4], Claudia Jaramillo[8,9], William Arias[9], Rodrigo Barquera Lozano[10,11], Paola Everardo-Martínez[10], Jorge Gómez-Valdés[10], Hugo Villamil-Ramírez[12], Tábita Hünemeier[13], Virginia Ramallo[14,15], Maria-Laura Parolin[16], Rolando Gonzalez-José[15], Lavinia Schüler-Faccini[14], Maria-Cátira Bortolini[14], Victor Acuña-Alonzo[10], Samuel Canizales-Quinteros[12], Carla Gallo[4], Giovanni Poletti[4], Gabriel Bedoya[9], Francisco Rothhammer[17,18]*, *Desmond Tobin[19], Elisabeth Roider[20,21,22]*, Pierre Faux[1#], Andrés Ruiz-Linares[23,1,3#]

* Contributed equally to this work.

#Co-supervised this work

[1] UMR 7268 ADES, CNRS, Aix-Marseille Université, EFS, Faculté de Médecine Timone, Marseille, 13005, France.

[2] School of Mathematics and Statistics, Faculty of Science, Technology, Engineering and Mathematics, The Open University, Milton Keynes, MK7 6AA, United Kingdom.

[3] Department of Genetics, Evolution and Environment, and UCL Genetics Institute, University College London, London, WC1E 6BT, UK

[4]Laboratorios de Investigación y Desarrollo, Facultad de Ciencias y Filosofía, Universidad Peruana Cayetano Heredia, Lima, 31, Perú.

[5] Unit of Human Evolutionary Genetics, Institut Pasteur, Paris, 75015, France.
[6] Scientific Police of São Paulo State, Ourinhos-SP, 19900-109, Brazil.
[7] Division of Vertebrates and Anthropology, Department of Earth Sciences, Natural History Museum, London, SW7 5BD, UK.
[8] Department of Biotechnology, Kinnaird College for Women, 93 - Jail Road, Lahore, 54000, Pakistan.
[9] GENMOL (Genética Molecular), Universidad de Antioquia, Medellín, 5001000, Colombia.
[10] National Institute of Anthropology and History, Mexico City, 6600, Mexico.
[11] Department of Archaeogenetics, Max Planck Institute for the Science of Human History (MPI-SHH), Jena, 07745, Germany.
[12] Unidad de Genomica de Poblaciones Aplicada a la Salud, Facultad de Química, UNAM-Instituto Nacional de Medicina Genómica, Mexico City, 4510, Mexico.
[13] Departamento de Genética e Biologia Evolutiva, Instituto de Biociências, Universidade de São Paulo, São Paulo, SP, 05508-090, Brazil.
[14] Departamento de Genética, Universidade Federal do Rio Grande do Sul, Porto Alegre, 90040-060, Brasil.
[15] Instituto Patagónico de Ciencias Sociales y Humanas, Centro Nacional Patagónico, CONICET, Puerto Madryn, U9129ACD, Argentina.
[16] Instituto de Diversidad y Evolución Austral (IDEAus), Centro Nacional Patagónico, CONICET, Puerto Madryn, Argentina.
[17] Instituto de Alta Investigación, Universidad de Tarapacá, Arica, Arica 1000000, Chile.
[18] Programa de Genetica Humana, ICBM, Facultad de Medicina, Universidad de Chile, Santiago, Arica 1000000, Chile.
[19] UCD Charles Institute of Dermatology, University College Dublin, Dublin, Ireland.
[20] Cutaneous Biology Research Center, Department of Dermatology, Massachusetts General Hospital, Harvard Medical School, Charlestown, MA 02129, USA
[21] Department of Dermatology, University Hospital of Basel, 4031 Basel, Switzerland
[22] Department of Dermatology and Allergology, University of Szeged, 6720 Szeged, Hungary
[23] Ministry of Education Key Laboratory of Contemporary Anthropology and Collaborative Innovation Center of Genetics and Development, School of Life Sciences and Human Phenome Institute, Fudan University, Yangpu District, Shanghai, China.

Abstract

Human skin colour is a highly visible, variable and well-studied trait – hundreds of genetic variants that impact the phenotype have already been identified in the general population. Although Genome-wide association studies (GWAS) have highlighted several genomic regions, we are yet to identify all the causal genes and explain missing heritability. To better understand the underlying genetic pathways, we performed a transcriptome-wide association study (TWAS) based on genetic and phenotype data for >6,000 Latin American individuals with the gene expression data imputed from the Genotype-Tissue Expression (GTEx) project using the PrediXcan models. The method S-PrediXcan, based on GWAS summary statistics performed better than basic PrediXcan method (using genotype data). We further incorporated multiple cutaneous tissues in the summary based S-MultiXcan method and identified a novel calcium channel gene CACNA2D2 on chromosome 3 as a novel locus influencing skin pigmentation. We captured almost 84% of the genes identified in the single tissue based S-PrediXcan (Bonferroni-corrected significance level) through multi-tissue method S-MultiXcan and most of the genes identified were not detected in the previous GWA study performed on the same set of individuals. Combining gene expression data with our summary statistics information yielded better results when we accounted for multiple cutaneous tissues, further investigation on the sharing of those expression trait loci (eQTLs) will provide us with a better understanding of the genetic mechanism.

3.3.1. Introduction

With modern technological advances, we have more precise and high coverage genomic data available in the present. Genome-wide association studies (GWAS) have been

robustly successful in leveraging this data to associate genetic loci to complex phenotypes[1]. Although the identification of a causal variant from GWAS is not very straightforward in the presence of linkage disequilibrium (LD)[2], the trait-associated SNPs often coincide with the regulatory regions of the genes and tend to be enriched for expression quantitative trait loci (eQTLs)[3,4], highlighting the importance of genetically controlled gene expression in the complex human phenotypes. Additionally, the majority of the identified variants identified reside in the non-coding regions of the genome and hence does not have obvious functionality. Gene-prioritization methods bridge the gap between the GWAS discoveries and trait related genes through genetic regulatory mechanisms that influence complex phenotypes. One such method is transcriptome-wide association study (TWAS)[5], which sequentially imputes gene expression from genotype data and performs association tests to identify the trait associated genes leveraging phenotype data[6–10]. Another advantage of the TWAS method is the lower burden of multiple testing compared to GWAS approaches as the former aggregates genomic information into functionally relevant units, which can be mapped to genes instead of individual variants in the case of GWAS[6,10].

TWAS approaches with gene expression information from single tissue consider one trait associated tissue at a time, e.g. PrediXcan[6] whereas MultiXcan[11] accommodates shared eQTLs across multiple tissues. Moreover PrediXcan or MultiXcan are based on raw genotype data that were extended to summary statistics based versions – S-PrediXcan[7] and S-MultiXcan[11], similar to FUSION[10] which is capable of handling the summary statistics data as well. Considering the comparative study[3] of these methods, we decided to analyse the PrediXcan and MultiXcan (and their summary based approaches) for our skin colour phenotype.

Skin pigmentation is a highly researched and well documented phenotype, quite a few genomic regions affecting skin pigmentation have been identified in previous GWA studies [12–18]. Although most of the studies has concentrated on the European cohorts, which show lesser variation in skin pigmentation, several regions on the chromosomes 3, 5, 6, 9, 11, 14, 15, 16, 20 and X have been identified[18–20]. Recent studies focusing on the African population have also been able to identify novel associations[13,21]. Adhikari et al. 2019 has shown a convergent pattern of skin colour evolution in western and eastern Eurasia through a GWAS on admixed population[12]. Studies have implicated ~15 genes for skin colour variation in human, while ~171 genes have been identified for model organisms[13,22].

Despite these efforts, we still lack complete information on the expression of related genes and connected pathways[23]. To gain better understanding of the skin pigmentation and the genetic interplays within of the cutaneous tissues we used CANDELA consortium data [24,25] which has been utilised before to identify pigmentation related genetic loci through GWAS [12] and exploiting the GWAS findings in a forensic setup for prediction of

pigmentation traits [26]. We identified regions previously implicated through the GWA study as well as a novel calcium channel gene in the chromosomal region 3p21.

3.3.2. Methods

Study subjects

We analysed genetic and phenotypic data from CANDELA consortium, where data was collected from over 7,000 individuals from five Latin American countries. Volunteers provided their written informed consent. Ethics approval was obtained from: Universidad Nacional Autónoma de México (México), Universidad de Antioquia (Colombia), Universidad Perúana Cayetano Heredia (Perú), Universidad de Tarapacá (Chile), Universidade Federal do Rio Grande do Sul (Brazil) and University College London (UK).

Phenotyping

Skin pigmentation was quantitatively measured using DermaSpectrometer DSMEII reflectometer (Cortex Technology, Hadsund, Denmark). The measurements were taken from both inner arms and average of the readings were finally used as skin Melanin Index (MI). This scale measures the pigmentation level for an individual, darker skin tone corresponds to higher reading. We analyzed 6,357 individuals before applying quality control checks including 651 from Brazil, 1745 from Chile, 1507 from Colombia, 1207 from Mexico and 1247 from Peru. The average MI of the cohort was 34.99 (median MI = 34.49 and standard deviation = 5.36) whereas the range of MI was from 20 to 65. The average MI for each of the countries varied from 32.05 (Brazil) to 36.32 (Mexico). Further details on the sampling can be found in our previous article [12].

DNA genotyping and quality control

The individuals were genotyped using the Illumina HumanOmniExpress chip, which includes 730,525 SNPs. The basic QC checks included (using PLINK v1.9 [27]) removing individuals and SNPs with more than 5% missing value, genetic markers with less than 1% minor allele frequency (MAF) and sex mismatch from the records during interviews and estimated from X-chromosome heterozygosity. Finally, after the exclusions, 669,462 SNPs remained for further analysis. CANDELA individuals being highly admixed of European, Native American and sub-Saharan African ancestry, Hardy-Weinberg P values are inflated and therefore not a good exclusion criterion[28]. We therefore performed stringent software and biological checks on the variants – we checked for batch effects on genotyping samples where we obtained a consistency rate >= 0.999 in all the cases. To further check the genotyping accuracy, we selected few samples for high-coverage sequencing and independently compared the genotype calls, ending up with an accuracy >= 0.999.

SNP genotype imputation

The genotype data was further phased and imputed using SHAPEIT2 [29] and IMPUTE2 [30] respectively using 1000 Genomes Phase 3 data as reference for the untyped variants. The monomorphic positions for the Latin American samples (Colombia, Mexico and Puerto Rico) in 1000 Genome reference data were omitted resulting in 11,025,002 SNPs being imputed. 48,695 SNPs out of those imputed were removed owing to low imputation quality score ('info' score less than 0.4). Post exclusion, median 'info' score obtained from

IMPUTE2 for the remaining SNPs was 0.986, highlighting high imputation certainty. Genotype probabilities obtained from IMPUTE2 for each locus were converted into most probable genotypes leveraging PLINK V1.9[27] (limiting to variants with uncertainty less than 0.1). Imputed SNPs with MAF < 1% or uncalled genotypes >5% were excluded from further analysis. The details about the related exclusion has been elaborated in Adhikari et al. 2019 [12]. After the quality checks, the final genotype data contained 9,143,600 including 8,780,546 autosomal biallelic SNPs.

Sample filtering

From all the genotyped samples, we selected individuals with age between 18 to 45 years. Additionally, samples with missing age or reported age having large discordance to the age calculated from the recorded date of birth were also removed. As African ancestry is underrepresented in the CANDELA cohort, we removed samples having estimated African ancestry (computed using ADMIXTURE[31]) higher than 20%. Further analysing we found traces of East Asian ancestry from ADMIXTURE software, when provisioned for additional ancestries. We removed all the individuals with estimated East Asian ancestry more than 5%. Furthermore, within the remaining samples, we selected the highly related pairs (approximately up to third-degree relatedness, IBD more than 0.1) and removed one per pair. To obtain maximum samples for further analysis, we first removed the sample represented in maximum pairs and update the pairs in a recursive manner until each individual is part of maximum one pair each. Then, we removed the individual from each pair with more age difference to the average sampling age to obtain the final set of 6,143 samples for GWAS analysis.

GWAS procedure

The GWAS for skin Melanin Index was performed using multiple linear regression model implemented in PLINK v1.9[27]. The additive genetic effect model was used to test association on the set of imputed SNPs with age, sex and six leading genetic principal components (PCs) as covariates. The PCs were calculated based on the 6,143 individuals and 160,858 LD-pruned autosomal SNPs using PLINK v1.9[27]. The final number of PCs to be included in the model was based on the scree and PC scatter plots (more details in Adhikari et al 2019[12]).

Transcriptome data

To capture the genetic effect on tissue specific expression levels, Genotype-Tissue Expression (GTEx) consortium collects data on the genotype and expression readings from the same set of individuals [32]. They targeted to collect data from donors after their death on 54 distinct body sites initially, but for few of them the did not finally collect or collected but did not process [32]. At the time of our analysis, PredictDB (https://predictdb.org/) which stores the transcriptome prediction weights used information on 49 tissues [11]. The project at the initiation collected data from 237 post-mortem samples on the designated body locations and genotyped ~4.3 million variants which later increased to ~6.8 million after imputation excluding SNPs with less than 5% MAF. Although the PredictDB stores and provides weightage for 49 tissue types, we selected only 3 tissues which are, a priori, the most related to skin pigmentation, namely,

Skin - Not Sun Exposed (Suprapubic), Skin - Sun Exposed (Lower leg) and Cells Cultured Fibroblasts.

TWAS methods

A transcriptome association study is a two-step method which leverages the GWAS outputs with summary statistics or the individual genotype-phenotype data and estimated genetic variant effect on the gene expression. The first step is to estimate the gene expression of the individuals included in the study and in the later to find the gene-trait association from statistical models.

A. Gene expression imputation for CANDELA individuals

We have explored the TWAS methods – methods based on raw genotype: PrediXcan, MultiXcan and their corresponding methods based on summary statistics based methods: S-PrediXcan and S-MultiXcan[6,7,11] (we have explained the tissue information and genotype-phenotype data used for each of the methods in Article 3, Fig. 1). These methods predict the gene expression of the gene from the respective tissues based on the predictDB (https://predictdb.org/) estimates of the SNP effects on gene expression. The gene expression can be visualised as the weighted average of the SNPs represented for the certain gene in the certain tissue. It may be noted that, the set of SNPs may vary for the same gene across the tissues and all the genes are not present in all the tissues. PrediXcan and S-PrediXcan predicts the expression for a single tissue related to the phenotype, whereas, MultiXcan and S-MultiXcan are the multi-tissue models which take the predicted expression for each tissues affecting the phenotype and combine them to find the final vector of gene expression for the phenotype.

B. Gene-trait association analysis

The gene-trait association is obtained from the phenotype information and the predicted gene expression levels. For multi tissue models, to avoid collinearity, only the informative principal components are retained as described in Barbeira et al 2019[11]. To mitigate the burden of multiple testing, the association significance threshold was calculated separately for each of the tissues using Bonferroni's correction, as the number of genes annotated for the tissues varies.

For this analysis, as skin colour is controlled by three tissues among the tissues collected in the GTEx project - Skin not sun exposed (Suprapubic), Skin sun exposed (Lower leg) and Cells cultured (fibroblasts) [32], we aim to perform single-tissue analysis for each of the tissues individually and combine for all three tissues to perform the multi-tissue analysis. The authors of the MultiXcan method (and the summary based approach S-MultiXcan) has already pointed out that MultiXcan is able to capture more significantly associated genes compared to single tissue PrediXcan (and the summary based approach S-PrediXcan) if the phenotype is impacted by multiple tissues[11]. As per our assumption of 3 cutaneous tissues impacting skin colour, we expect to gather more information on gene functioning with MultiXcan (or the summary statistics based adaptation S-MultiXcan). We

have used the FastMan (https://github.com/kaustubhad/fastman) package to present the gene association results through Manhattan plots.

		SNP-level data used	
		Raw genotypes and phenotypes	Summary statistics
Tissue Used	Single	PrediXcan	S-PrediXcan
	Multiple, combined	MultiXcan	S-MultiXcan

Article 3, Fig. 1 Methods summary
Diagrammatic comparison between PrediXcan and MultiXcan methods and their summary statistics based adaptations

3.3.3. Results and discussion

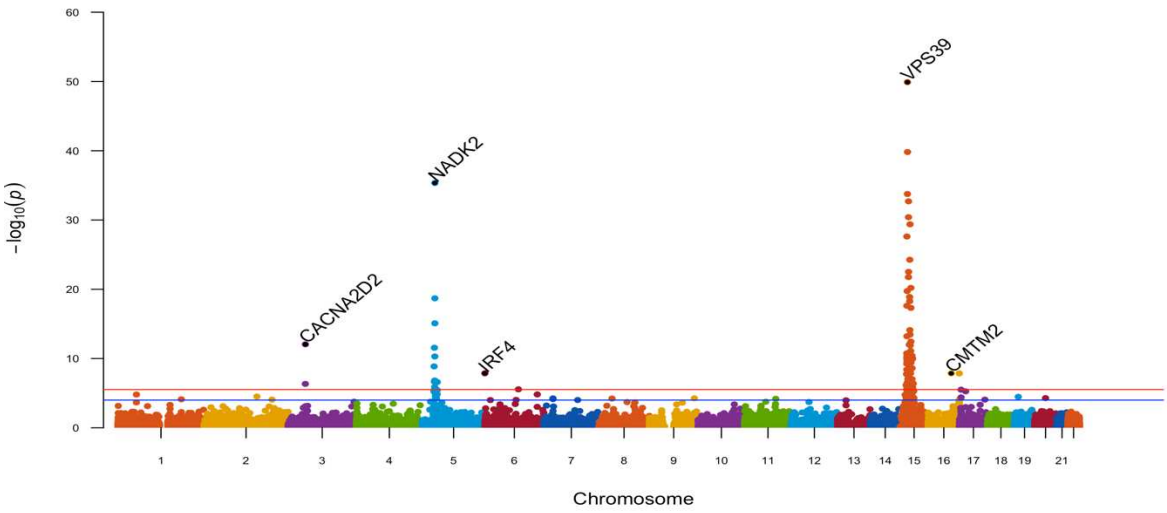
3.3.3.1. TWAS results

Transcriptome-wide association studies (TWAS) can be performed using the gene expression for a single tissue as well as multiple tissues. To implement single tissue method - PrediXcan[6], we leveraged the publicly available GTEx consortium data version8 [32,33] which contains information on the genotypes and gene expression form the same individuals. We are required to incorporate the genotypic information of our study individuals and the GTEx data to obtain the genetically regulated expression (GReX) levels for each of the genes in the respective tissues. Subsequently the variation of the expression levels for those tissues is equated to the variation of the skin melanin index to obtain trait-associated genes. We performed single-tissue PrediXcan analysis on 6,143 individuals for three cutaneous tissues - Cells cultured (fibroblasts), Skin not sun exposed (Suprapubic) and Skin sun exposed (Lower leg) and identified 37, 29 and 30 significantly associated genes, respectively, with Bonferroni correction significance threshold applied separately for each of the tissues (List of the significant genes are provided in Supplementary tables 1-3). Manhattan plots for the whole genome are presented in Supplementary figure S1 for each of the tissues.

Provided we had access to the summary statistics of the GWAS of skin melanin index from Adhikari et al 2019[12] on same individuals used in the PrediXcan mentioned above, we also

explored the S-PrediXcan[7] method. We identified genomic regions (detail gene list has been provided in the Supplementary tables 4-6) similar to the ones identified from PrediXcan, although the summary based approach often yielded stronger significance as this approach does a pre imputation for the SNPs present in the weightage database (namely PredictDB [6]), the Manhattan plots of the gene-trait association has been presented in the supplementary figure S2 and comparison of the genes which produced association p-value less than 1e-5 for at least one of the methods were illustrated in Supplementary figure S3. Going forward, we shall concentrate on the summary based approach as the association results remain unaltered in both methods where summary based approach reduce computational burden and provide more flexibility and higher association levels for most of the significant genes.

S-PrediXcan takes only one tissue into consideration while looking for trait-gene association. Nonetheless, traits where the eQTLs are shared across tissues and multiple tissues simultaneously influence a phenotype, Barbeira et al 2019[11] has demonstrated that combining the tissues perform better than individual ones and thus proposed S-MultiXcan[11]. We used transcriptomic data for the three cutaneous tissues mentioned above for skin colour analysis. We obtained 75 significantly associated genes spanning across 5 chromosomes from this multi-tissue approach. To apply Bonferroni corrected threshold of significance, we considered all the genes with appearance in at least one of the tissues (Bonferroni corrected threshold - pvalue less than 2.24e-6). Genomic regions on chromosomes 5, 6 and 15 have already been identified from the single tissue analyses, although it did not reach significance level for chromosome 16, TUBB3 (S-MultiXcan association pvalue = 8.6e-5, next to MC1R gene) previously has been associated with the skin tanning[34]. Additionally, we identified a novel calcium channel gene CACNA2D2 on chromosome 3 which turns out to be significant (S-MultiXcan pvalue 8.8e-13) only when three cutaneous tissues are considered simultaneously. Details of the full list of significant genes is provided in Supplementary Table 7 and the Manhattan plot of the genome is presented in Article 3, Fig. 2. When we compare the associated genes from the S-MultiXcan method, the list captures 84% (51 out of 61 unique genes) of the genes identified by the three individual tissue specific S-PrediXcan methods (the detailed gene comparison is presented in Supplementary table 8).



Article 3, Fig. 2 Manhattan plots for Skin Melanin Index using S-MultiXcan

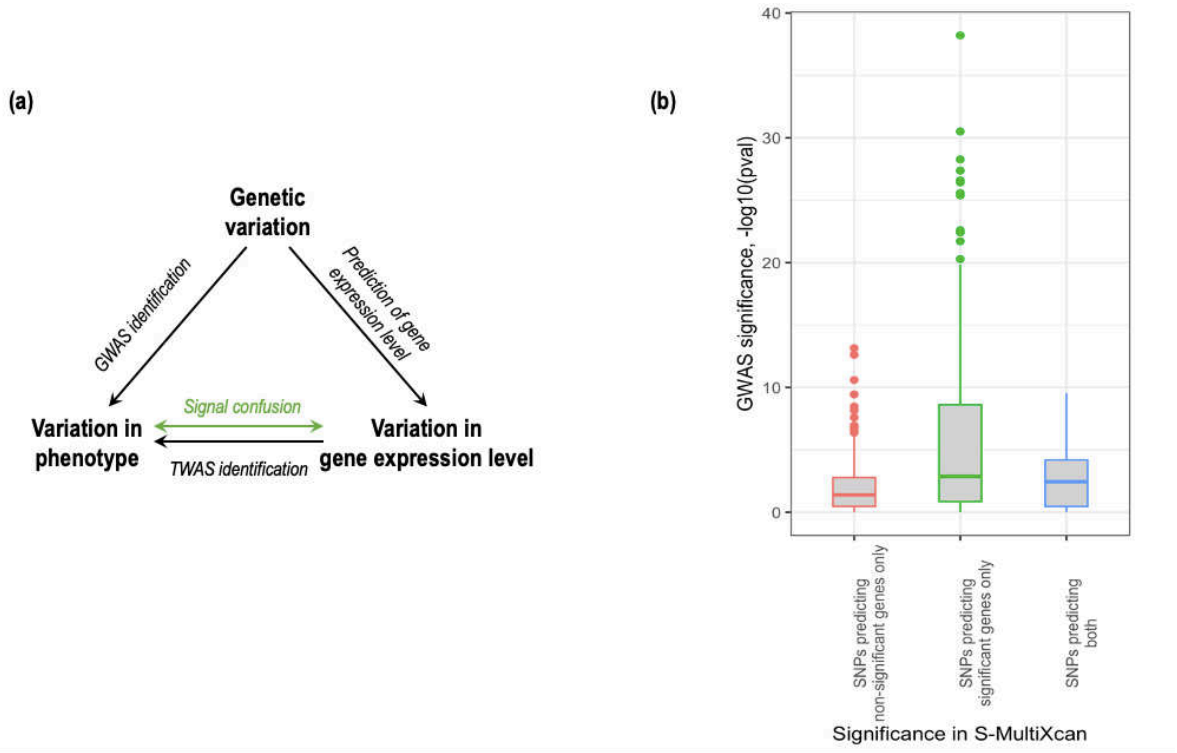
Predicted gene expression association for the genes in three cutaneous tissues based on the GWAS summary results of 6,143 samples from CANDELA study. Here the assigned position of a gene is the average of the start and end basepairs, according to the GRCh37 mapping. Red and blue lines represent significant and suggestive thresholds respectively.

3.3.3.2. Comparison of the TWAS and GWAS results

Our TWAS results from the S-MultiXcan[11] method is obtained leveraging the GWAS summary statistics from Adhikari et al 2019[12] on the same phenotype, which invoked our curiosity to compare the final set of associated genes. IRF4 gene on chromosomal region 6p25 has been identified through TWAS which is the only common entry in the two lists of significantly pigmentation associated genes (details of the gene lists are provided in the Supplementary tables 7 and 9) owing to the same intronic variant (rs12203592) being present in the PredictDB[6] as well as being reported in the GWAS analysis. One possible biological explanation for many mismatches in the gene lists could be that many of the SNPs identified in the GWAS (including the SLC genes on chromosome 5 and 15) are amino-acid changing variants: they are not expected to impact on transcription, thus should not appear as eQTLs and thus not in the TWAS[35]. These amino acid changes likely affect the function of the protein directly. As we have obtained a lot of eQTLs in chromosome 5, 15 and 16 regions in the S-MultiXcan analysis, this could relate to the multiplicity of independent association signals found in the GWAS. Some of the identified SNPs are amino acid changing variants, others are eQTLs.

Although most of the genes identified in GWAS were not significant in TWAS, we still observe peaks of significance for genes located in their neighbourhood (see chromosomes 5 and 15 in Article 3, Fig. 2). Genomic regions 5p13, 15q21 and 16q24 have been picked up by both GWAS and TWAS. We hypothesize that these TWAS peaks are signal confusions due to genetic variants affecting both the studied phenotype (hence significantly associated to that phenotype through GWAS) and the expression of neighbouring genes

(hence selected by PrediXcan – and related models – for prediction of the gene expression level). Thus, significant variations in gene expression level in these regions already associated to the phenotype GWAS would not have a causal effect on the phenotypic variations. Instead, it would come out significant because of a shared cause. We summarized this view in Article 3, Fig. 3a. To support it, we have retrieved GWAS significance of all the SNPs located in the two main GWAS/TWAS peaks (~9 Mb in 5p12 and 5p13 and ~20 Mb in 15q14, 15q15, 15q21 and 15q22, as indicated in the enlarged figures in Supplementary S4) and used by PrediXcan to predict expression level of genes in the vicinity. In Article 3, Fig. 3b, we classified these SNPs according to the TWAS significance of the gene(s) that they predict and plotted the distribution of the GWAS significance per class: genes significant in TWAS had their expression levels predicted by SNPs whose GWAS significance was a few orders of magnitude stronger than that of SNPs predicting non-significant TWAS genes.



Article 3, Fig. 3 Overlap of TWAS and GWAS signals

(a) *The same genetic variation (e.g. a SNP) may cause a phenotype variation through an effect on the protein by gene A (this effect being identified by GWAS) and a variation in the level of expression of gene B (making this genetic variation useful for predicting the expression of gene B). In turn, a variation in expression level of gene B may cause a phenotype variation; identifying such relation is the purpose of TWAS. However, when the same genetic variation has both effects, it may result in a TWAS false-positive signal, as both consequences of the same cause correlate (the green arrow).* ***(b)*** *GWAS significance of SNPs used by PrediXcan for prediction of gene expression level and located within two GWAS peaks (5p12-13 and 15q14-22). SNPs were classified according to the TWAS (non-)significance of the gene whose expression they predict.*

3.3.3.3. CACNA2D2 and human pigmentation

Here we present the novel gene CACNA2D2 in the genomic band 3p21 identified through the association of its predicted expression levels and melanin index, in three human skin tissues. This gene remained insignificant for each of the three tissues analysed separately in S-PrediXcan, however as we jointly analysed all of them, this calcium channel gene came out to be significantly associated. From our analysis, we are yet to ascertain the directional effect of the expression of the gene. In the single tissue analyses, we have observed expression of CACNA2D2 has a positive correlation with the Melanin Index in the Fibroblasts tissue, which is reversed in the remaining two tissues. To follow-up on our findings, experiments were performed on the mouse tissues; so far these experiments have produced contradictory inference on the direction, although the effect of the gene is visible in both cases. In one of the experiments when CACNA2D2 gene was silenced it resulted in increased melanin content compared to control group in the mouse skin tissue, whereas in a different setup when Amiodarone, which inhibits the functioning of the said gene was used in a controlled setup, melanin pigmentation decreased visibly (experimental outcomes has been attached in Supplementary figure S5).

We are yet to confirm the complete pathway but one possible connection can be lower requirement of calcium in the people of darker complexion.

Our skin pigmentation protects us from the harmful UVR, higher melanin in the skin absorbs more UVR and protects underlying skin tissues from its effects [36]. As solar radiation is highest near the equator and gradually decreases towards the poles, skin pigmentation follows a largely similar pattern[37]. But this reduction of penetrating UVR, especially in the areas with lower sun exposure, also affects vitamin D synthesis in the cutaneous tissues, which in turn affects the intestinal calcium absorption causing bone damage[38]. Additionally expression levels of CACNA2D2 has been associated in the pathways lung and breast cancer among others[39,40]. Prior knowledge on this gene therefore supports our finding, making it a relevant candidate for further investigations.

3.3.4. Conclusion

In conclusion, we performed the S-MultiXcan[11] analysis on the GWAS summary statistics obtained from a previous study on 6,143 individuals from Latin America[12]. The S-MultiXcan method has been proven to be more effective when there are multiple tissues acting on a single phenotype[11], which has been mirrored in our study as well. We have used three cutaneous tissues from the GTEx database[32] – cells cultured fibroblasts, skin not sun exposed suprapubic and skin sun exposed lower leg. We obtained meaningful association results for each of the three tissues and it improved when we combined all of them to perform the multi-tissue analysis. This goes on to show the combined impact of the tissues on our phenotype under consideration. Moreover in the single-tissue analysis, where we had access to both genotype and GWAS summary data, we obtained almost similar results using PrediXcan and S-PrediXcan methods. Mostly for practical reasons we explored the summary based approach when combining multiple tissues to identify gene-

trait associations. Our results have a small overlap with our previous GWAS mentioned above[12], we could replicate IRF4 on chromosome 6 for trait eQTLs, although the PrediXcan (alternatively S-PrediXcan) method replicated known regions from our GWAS in chromosome 5,6 and 15. Only when we combined three cutaneous tissues in the S-MultiXcan approach, we could identify new gene expression association with the phenotype skin Melanin Index -a calcium channel gene CACNA2D2 in genomic band 3p21 has been our novel identification, which has been previously associated for lower expression in cancer cells[39]. In our study, higher expression of the gene has been associated with lower melanin pigmentation.

Nonetheless, these methods (PrediXcan and MultiXcan, or the respective summary adaptations) do not help to identify a basis for colocalization of GWAS causal variants and eQTL causal variants[41]. Additionally the GTEx database suffers from the lack of ancestral diversity[6,33]. In the GWA studies using CANDELA cohort, we have observed significant associations for SNPs that are polymorphic only in Native Americans. Such SNPs are not used in the prediction models, as GTEx is primarily European. Once the limitations are removed by adding more data from diverse population, the statistical studies will get more accurate in the future.

3.3.5. References

1. Welter, D. et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 42, D1001–D1006 (2014).

2. Wainberg, M. et al. Opportunities and challenges for transcriptome-wide association studies. Nat. Genet. 51, 592–599 (2019).

3. Fryett, J. J., Inshaw, J., Morris, A. P. & Cordell, H. J. Comparison of methods for transcriptome imputation through application to two common complex diseases. Eur. J. Hum. Genet. 26, 1658–1667 (2018).

4. Nicolae, D. L. et al. Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS. PLoS Genet. 6, e1000888 (2010).

5. Hindorff, L. A. et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl. Acad. Sci. 106, 9362–9367 (2009).

6. GTEx Consortium et al. A gene-based association method for mapping traits using reference transcriptome data. Nat. Genet. 47, 1091–1098 (2015).

7. GTEx Consortium et al. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. Nat. Commun. 9, 1825 (2018).

8. Alzheimer's Disease Genetics Consortium, et al. A statistical framework for cross-tissue transcriptome-wide association analysis. Nat. Genet. 51, 568–576 (2019).

9. Pividori, M. et al. PhenomeXcan: Mapping the genome to the phenome through the transcriptome. Sci. Adv. 6, eaba2083 (2020).

10. Gusev, A. et al. Integrative approaches for large-scale transcriptome-wide association studies. Nat. Genet. 48, 245–252 (2016).

11. Barbeira, A. N. et al. Integrating predicted transcriptome from multiple tissues improves association detection. PLOS Genet. 15, e1007889 (2019).

12. Adhikari, K. et al. A GWAS in Latin Americans highlights the convergent evolution of lighter skin pigmentation in Eurasia. Nat. Commun. 10, 358 (2019).

13. Martin, A. R. et al. An Unexpectedly Complex Architecture for Skin Pigmentation in Africans. Cell 171, 1340-1353.e14 (2017).

14. Deng, L. & Xu, S. Adaptation of human skin color in various populations. Hereditas 155, 1 (2018).

15. Norton, H. L. et al. Genetic Evidence for the Convergent Evolution of Light Skin in Europeans and East Asians. Mol. Biol. Evol. 24, 710–722 (2006).

16. Jablonski, N. G. & Chaplin, G. The evolution of human skin coloration. J. Hum. Evol. 39, 57–106 (2000).

17. Jablonski, N. G. & Chaplin, G. Epidermal pigmentation in the human lineage is an adaptation to ultraviolet radiation. J. Hum. Evol. 65, 671–675 (2013).

18. Liu, F. et al. Genetics of skin color variation in Europeans: genome-wide association studies with functional follow-up. Hum. Genet. 134, 823–835 (2015).

19. Law, M. H. et al. Genome-Wide Association Shows that Pigmentation Genes Play a Role in Skin Aging. J. Invest. Dermatol. 137, 1887–1894 (2017).

20. Candille, S. I. et al. Genome-Wide Association Studies of Quantitatively Measured Skin, Hair, and Eye Pigmentation in Four European Populations. PLoS ONE 7, e48294 (2012).

21. Crawford, N. G. et al. Loci associated with skin pigmentation identified in African populations. Science 358, eaan8433 (2017).

22. Baxter, L. L., Watkins-Chow, D. E., Pavan, W. J. & Loftus, S. K. A curated gene list for expanding the horizons of pigmentation biology. Pigment Cell Melanoma Res. 32, 348–358 (2019).

23. Jablonski, N. G. The evolution of human skin pigmentation involved the interactions of genetic, environmental, and cultural variables. Pigment Cell Melanoma Res. pcmr.12976 (2021) doi:10.1111/pcmr.12976.

24. Ruiz-Linares, A. et al. Admixture in Latin America: Geographic Structure, Phenotypic Diversity and Self-Perception of Ancestry Based on 7,342 Individuals. PLoS Genet. 10, e1004572 (2014).

25. Adhikari, K., Mendoza-Revilla, J., Chacón-Duque, J. C., Fuentes-Guajardo, M. & Ruiz-Linares, A. Admixture in Latin America. Curr. Opin. Genet. Dev. 41, 106–114 (2016).

26. Palmal, S. et al. Prediction of eye, hair and skin colour in Latin Americans. Forensic Sci. Int. Genet. 53, 102517 (2021).

27. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience 4, 7 (2015).

28. Adhikari, K. et al. A genome-wide association scan implicates DCHS2, RUNX2, GLI3, PAX1 and EDAR in human facial variation. Nat. Commun. 7, 11616 (2016).

29. O'Connell, J. et al. A General Approach for Haplotype Phasing across the Full Spectrum of Relatedness. PLoS Genet. 10, e1004234 (2014).

30. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat. Genet. 44, 955–959 (2012).

31. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 19, 1655–1664 (2009).

32. The GTEx Consortium et al. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science 348, 648–660 (2015).

33. Lonsdale, J. et al. The Genotype-Tissue Expression (GTEx) project. Nat. Genet. 45, 580–585 (2013).

34. Herraiz, C. et al. Functional Characterization of MC1R-TUBB3 Intergenic Splice Variants of the Human Melanocortin 1 Receptor. PLOS ONE 10, e0144757 (2015).

35. Mostafavi, H., Spence, J. P., Naqvi, S. & Pritchard, J. K. Limited overlap of eQTLs and GWAS hits due to systematic differences in discovery. http://biorxiv.org/lookup/doi/10.1101/2022.05.07.491045 (2022) doi:10.1101/2022.05.07.491045.

36. Webb, A. et al. Colour Counts: Sunlight and Skin Type as Drivers of Vitamin D Deficiency at UK Latitudes. Nutrients 10, 457 (2018).

37. Chaplin, G. Geographic distribution of environmental factors influencing human skin coloration. Am. J. Phys. Anthropol. 125, 292–302 (2004).

38. Vieth, R. Weaker bones and white skin as adaptions to improve anthropological "fitness" for northern environments. Osteoporos. Int. 31, 617–624 (2020).

39. Carboni, G. L. et al. CACNA2D2-mediated apoptosis in NSCLC cells is associated with alterations of the intracellular calcium signaling and disruption of mitochondria membrane integrity. Oncogene 22, 615–626 (2003).

40. Gao, B. et al. Functional Properties of a New Voltage-dependent Calcium Channel α2δ Auxiliary Subunit Gene (CACNA2D2). J. Biol. Chem. 275, 12237–12242 (2000).

41. Hormozdiari, F. et al. Colocalization of GWAS and eQTL Signals Detects Target Genes. Am. J. Hum. Genet. 99, 1245–1260 (2016).

4. General discussion

4.1. Summary of the thesis

This thesis elaborates the usage of prediction approaches on human pigmentation phenotypes. Pigmentation – eye, hair and skin colour constitute a reasonable part of our physical appearance. As a consequence, they play a significant role in the forensic, medical and anthropological genetics and attracts scientific efforts to better understand the phenotypes[20,78,79,137,140]. Nonetheless, in most of the cases the study samples have been restricted to continental Europe[138]. Although eye and hair colour variation is mostly restricted to Europe, with some exception in West Asia or Polynesia[136], skin colour shows greater variation outside of the continent[187]. There have been attempts to capture the associated genetic variants for all the pigmentation phenotypes and although often the accuracy has surpassed an acceptable threshold within the modelling ancestry[20], global application of any of the SNP arrays remain a concern.

We utilised large CANDELA consortium data sampled from five Latin American countries – Brazil, Chile, Colombia, Mexico and Peru. Studies have argued about the recent admixture of European, Native American and African ancestries for the study individuals[28,223,260]. The 3-way admixture provides us a good platform to explore the existing state-of-the-art predictive SNP array- HIrisPlex-S[20]. Although the categorical eye and hair colour provides a good prediction accuracy, we realized skin colour prediction could improve further. In Chapter 3.1, we proposed our new SNP sets for all the phenotypes which improved skin colour prediction performance (38.4% prediction R^2 for HIrisPlex-S, 46% prediction R^2 for CAN-S) as CANDELA data has captured a wide range of skin melanin index (20 to 65). Moreover, owing to the large European ancestry proportion in the population, we also performed equally good to the existing SNP array systems for eye (89.14% prediction accuracy for HIrisPlex-S, 89.26% prediction accuracy for CAN-E) and hair colour (84.77% prediction accuracy for HIrisPlex-S, 85.67% prediction accuracy for CAN-H). To explain the result, we split the CANDELA population into varying European and native American ancestry buckets and high native ancestry buckets showed better performance for the CANDELA SNP sets where high European ancestry buckets had comparative performance for the quantitative traits (melanin index, eye colour hue, chroma and luminance). This highlights the lack of native American ancestry in the HIrisPlex-S affects its predictive power on the Latin American individuals.

In Chapter 3.2, we wanted to explore spatiotemporal trends of human pigmentation evolution. Therefore, we improved our eye colour luminance and skin melanin index prediction models to extend to the ancient DNA individuals. Additionally, we adapted CAN-E/S SNP sets for the sparse ancient data. We validated our adjusted models on the out-of-training samples from contemporary world population(1000 Genomes Phase 3[261] and CEPH-HGDP[262]). Our results showed strong coherence with the expected

pigmentation of the contemporary samples and a rational trend for regional variation[61]. We also used correlation of the predicted pigmentation levels with environmental variables known to affect the traits (e.g. solar radiation and latitude)[4]. Motivated by our model performance, we applied our prediction approach on the ancient data from Harvard Ancient Genomes Database. Our final predicted trends show a gradual skin lightening and increasing eye luminance for the Eurasian population in the last 15,000 years. In Siberia, Central and East Asia the trend is reversed in recent times, where melanin index is increasing in last two millenniums. For the continental America, we find that pigmentation is stagnated at the point of split from East Asia and Siberian population[263] for the last 15,000 years. Our study have produced concordant results with other recent studies representing specific historic ages or certain geographic regions within Eurasia[139,264]. However, prediction of the pigmentation evolution trends in the last 50,000 years has not been attempted before at a global scale.

We further explored gene expression prediction in Chapter 3.3 through Transcriptome-wide association studies (TWAS). We integrated the tissue gene expression datatsets from GTEx portal with the genotype data from CANDELA studies to impute the gene expression levels in our data and used skin melanin index as our phenotype. Single tissue based TWAS approach PrediXcan[259] and its summary statistics based adaptation summary-PrediXcan (S-PrediXcan)[265] produced similar results and replicated the known regions responsible for skin pigmentation from previous GWAS study[28]. Extending the analysis, we explored multi-tissue based summary approach S-MultiXcan[222] where we identified more genomic regions including a novel calcium channel gene CACNA2D2 on chromosome 3, whose expression levels has previously been associated with breast and lung cancer[266,267], among others.

4.2. Variation of prediction models and inference

In Chapter 3.1, we have explored the power of genetic predictors in collaboration with analytical methods. For quantitative traits we implemented linear and stepwise regression, whereas, for categorical traits we utilised multinomial logistic and ordinal regression in statistical models. For all the phenotypes we used three machine learning models – random forest (RF), extreme gradient boost (XGBoost) and artificial neural network (ANN), all of which can handle both quantitative and categorical variables. We observe that RF performs better than other models for categorical eye colour and most of the quantitative eye colour phenotypes. It is known that eye colour variation is largely captured by few very strong variants. RF through multiple trees is able to capture the variation, also the interaction of the HERC2-OCA2 variants[208] is captured implicitly in the model, although explicitly mentioning the interaction as a variable in the model should increase the accuracy of the models. This is because RF selects a proportion of the variants to build each tree. There have been scenarios when both HERC2 (rs12913832) and OCA2

variants were not selected in the same tree, hence, the interaction effect will not be captured. Explicitly declaring this interaction as a new variable avoids this situation and can potentially improve the prediction. It is worth mentioning that RF has been used in the literature for the purpose of identifying SNP-SNP interaction[90,91,268].

On the other side, skin pigmentation variation is determined by a large set of genetic variants and while building each tree, RF keeps few of the variables aside. While epistasis for skin colour has been reported[269], it affects skin colour to a lesser extent than eye or hair colour. Hence RF model loses some predictive power in the random assortment of the variables in the trees. Linear regression performs better in this particular case as it takes all the genetic variants in the model, so does not miss out on the potential predictive powers. At the same time lower impact of the specific SNP-SNP interactions does not harm linear regression. This highlights the importance of application of various predictive models with basic differences in the algorithm. Performance comparison of the models elaborates the intrinsic genetic structures of the pigmentation phenotypes.

4.3. TWAS finding

Genome-wide association studies (GWAS) have been applied on numerous phenotypes and it has successfully identified statistically associated SNPs with complex diseases and traits[98]. Nonetheless, there persists issues with interpretability and causality of those identified signals. Complexities of the linkage disequilibrium structure and context-dependent regulatory machinery in the genome prohibits us from a direct interpretation. Thus restricts the ability to work on the validation of downstream functionality and identification of therapeutics. In the last decade, large-scale consortiums, e.g. Genotype-Tissue Expression (GTEx) project have obtained data on tissue-specific gene expressions human. These datasets can be leveraged to dig deeper into the mechanisms of the cross-tissue transcriptional regulation and expedite expression quantitative trait loci (eQTL) discoveries. Moreover incorporating eQTL information into genetic association studies can be helpful to connect the SNPs, genes and complex phenotypes. Co-localization methods have been developed to identify eQTLs in the trait associated loci obtained from GWASs to pinpoint the candidate risk genes for complex phenotypes. Transcriptome-wide association studies (TWAS) are one such group of methods which have provided novel insights into the genetic architecture of the complex diseases and phenotypes from genotype-based imputations of gene expression. One of the reasons of their success is the capability to effectively assemble signals of multiple eQTLs with moderate effects. This also facilitates to reduce the multiple testing burden as TWAS tests for the genes rather the SNPs. Studies have shown that multi-tissue modelling can enhance the statistical power to discover eQTLs and gene-network studies[270].

Multi-tissue based TWAS approach (S-MultiXcan) on skin melanin index from CANDELA individuals identified novel gene expression association of CACNA2D2. We are already investigating the impact of CACNA2D2 mouse samples. It is evident from the experiments that silencing the calcium channel gene has been affecting the mice

pigmentation, although we are yet to ascertain the impact direction of the expression levels, which does not follow a unique direction for all three cutaneous tissues concerned. There are several other significantly associated genes discovered in Chapter 3.3 which merit more investigation in the future. We are yet to comprehend the complete genomic infrastructure affecting our skin colour, the follow-up analysis is going to help us gather more meaningful information. Nevertheless, one needs to be cautious about the associated gene list as statistical association does not imply causality. Wainberg et al. 2019[214] has discussed that non-causal genes can be apparently associated through correlation of imputed gene expressions at the same locus. This can be compared to non-causal SNPs being associated in GWAS through LD.

Hu et al. 2019[219] has discussed the limitations of TWAS approaches in identifying the relevant tissues as well as functional genes. When different genes share eQTLs from the same locus or from high LD, it might inflate the association of non-causal genes. Similarly when the eQTLs of a given gene are shared across different tissues or in LD, irrelevant tissues might become significantly associated[219]. This emphasises the importance of the relevant tissue identification, although for many complex phenotypes prior biological information remains unavailable. The existing TWAS methods have some limitations and are still evolving, however if the outputs are inferred with caution, they hold great potential for finding more causal relations.

4.4. Pigmentation evolution

Human pigmentation evolution is a complex process mostly shaped by UVR and its effects on vitamin availability[54]. Since the earliest Homo started to lose skin protection in terms of body hair and did not protect the bodies from harmful UVR using covering, skin became the effective layer of protection. Additionally, skin synthesises vitamin D_3 from the UVB radiation. Vitamin D_3 is essential ingredient for human growth, skeletal development and calcium absorption, lack of the vitamin can cause immobilisation and pelvic deformities, in extreme cases death[271]. For female, this is even more important as vitamin D_3 is required during pregnancy and lactation[4]. Some researchers consider melanin pigmentation an adaptive trait as UVR-induced sunburn or skin cancer threatens reproductive success[272] and the pigmentation genes are also impacting metabolism pathways[273], whereas, some argue that preferences for a particular skin tone in the partner caused the evolution of skin colour .

Eye colouration on the other hand does not show geographical patterns as human skin colour[56] and the evolution is more likely to be a combination of natural and sexual selection[141]. Dark iris colours protect from intraocular light scattering, which protects against strong sunlight and UVR, contrary to blue iris colours, which helps with intraocular light scattering in the regions with lower and seasonal sunlight[274]. Studies have also suggested a preference of the blue eyed males for blue eyed females for phenotypically-based assurance of paternity[57]. Recent studies have claimed that eye

colour evolution may have been caused by the pleiotropic effects of the favourable skin colour variants[141].

Our prediction from the aDNA data on skin and eye colour shows trends of pigmentation evolution on a global scale. Our predicted skin lightening in Eurasian regions nearly 15,000 ya aligns with the previous study[264]. It has already been claimed that the earliest human populating northern America came from North and North-Eastern Asia through the through Beringia via overland and coastal routes[275,276]. Our prediction shows identical skin melanin distribution in eastern Eurasia and American continent nearly 15,000-12,000 ya which resonates with the migration claim. This analysis also reflects the adaptation for vitamin D_3 synthesis in lower UVB regions, e.g. high latitude artic region, where inhabitants could acquire required vitamin through available foods and skin colour shows a higher than expected melanin level based on latitude[32]. Despite our model training data lacking global coverage, predicted phenotypes performed satisfactorily in contemporary global samples, additional data from the missing cohorts would potentially improve the prediction models. Nonetheless, our predicted inter-continental melanin indices coincide with the expected skin colour ordering. The same goes for the intra-continental ordering.

Being a predominantly European trait, our eye colour luminance prediction does not suffer from the lack of representation problem. We show an increase in luminance in Europe since 15,000 ya. Although luminance increased in eastern Asia around 5,000 ya, it is not as prominent as continental Europe. Our prediction of darker eye colour before 15,000 ya agrees with a recent study[139].

Varied pigmentation in the human society is a result of long history of migration, environmental adaptation to varying UVR and other factors, food and cultural adaptations and genetic interventions. In this thesis, we have painted a picture of pigmentation evolution over 50,000 years fragments of which agree with previous studies representing specific regions and time intervals.

4.5. Future perspectives

We are yet to fully comprehend the complete genomic architecture of the pigmentation traits which will help us elevate the prediction accuracy in the future and this can be extended beyond pigmentation traits. The relationship of the melanoma and skin colour is already established[277,278]. In Chapter 3.3, we have also identified gene expression associated with skin colour which has already been discovered in breast and lung cancer pathways. Once we improve our understanding of the pigmentation traits, that should enhance our prediction of other related traits as well.

We have also identified the genomic association for facial morphologies in a recent study[279] on the CANDELA cohort. In the future we could leverage this study to prepare a set of genetic markers to produce prediction models on facial features, following similar methods as we have described in Chapter 3.1 for pigmentation traits. Moreover, we can also extend this to predict evolutionary trends of the facial morphological features like

eye and skin pigmentation studied in Chapter 3.2. Once we achieve a satisfactory prediction setup, along with pigmentation, these traits will enhance forensic genetics to predict more appearance phenotypes.

Our models should be explored as a biometric verification tool in certain setups where genetic information is not present in the database but portrait photographs are available. The multi-trait prediction approach of the facial and pigmentation features should probabilistically identify the photograph of the sample donor, i.e. maximising the probability of matches of the photographs and DNA samples collected. This combined prediction method needs to go through a rigorous training process to deduce a weighted error distance metric, that would be able to optimise the predicted errors of each of the individual phenotypes.

Additionally, we can explore another avenue of combined predictive models for correlated phenotypes to improve current individual pigmentation prediction models. The correlation structure and pleiotropy of the three pigmentation traits are well-known[280,281] which, if exploited, can produce better prediction accuracy. Studies have already managed to implement statistical[281] and deep learning[282] models in this regard and claimed to achieve better results. In the deep learning model, the authors have also explored SNP-SNP interaction variables along with the individual SNPs and basic covariates. Interaction within the underlying genetic structures of the pigmentation traits are already studied - HERC2-OCA2 region on region 15q13 or MC1R variants on region 16q24 have been noted[65,208,269,283] in the literature for their epistatic effect. In the future we should exploit the known genetic structure of the pigmentation phenotypes in order to produce better predictive powers. Methods to explore two or higher order interactions might be more beneficial in this regard[90,91].

To better our skin colour prediction models, we should explore the additional TWAS findings in the model for their potential. Till now all the genetic variables in the prediction models of Chapter 3.1 are obtained exclusively from the previous GWAS study by Adhikari et al 2019. The gene expression associated genes (or their variants) should also improve our models and as an extension, we should explore eye and hair colour as well for the TWAS.

We should explore mixed linear models for pigmentation prediction in the future. These models are capable to handle the genetic relationship matrix (GRM) which is not feasible in the linear regression or machine learning models. In the current setup in Chapter 3.1, we are using only the top few principal components of the GRM, which provide us with the population structure of the study samples. Along with the genetic information of the individuals we want to predict, the information about the phenotypes of the related individuals in the study should potentially enhance the predictive models. Intuitively this scenario should improve our model performance as we'll not be discarding non-major components from the principal component analysis of the GRM.

4.6. Importance of ancestry inclusion

The previous GWAS on human pigmentation by Adhikari et al 2019[28] has been instrumental for this thesis. This was one of the earliest large cohort analyses on the admixed population of Latin America. The authors identified novel genes affecting the phenotypes and replicated several others. Although human pigmentation has been studied for a long time now, most of the genetic studies have concentrated on continental Europe. Eye and hair pigmentation are largely a European trait, but skin pigmentation shows more variation elsewhere. That leaves us with incomplete understanding of these pigmentation traits. Of late, studies on biobank level datasets and large-scale meta-analyses using GWAS summary results have tried to bridge the gap discovering more novel associations [191,204,205] to enhance our knowledge of the traits. Studies with large datasets are the new standard now, nonetheless we shall never fully appreciate the genetic functions of the traits unless representative studies from other ancestries emerge. Recent analyses on under-represented African population has produced noticeable results to improve our understanding of skin colour[197,198].

In Chapter 3.1 we have demonstrated that the models trained on a certain ancestry might be not performing in the same level for other ancestries, hence the online available models built in a certain ancestry might not be applicable to a global population, which points to the essential take away that each population needs an ancestry specific SNP array and predictive model and for a single model to predict on a global population needs global representation in the modelling sample.

In Chapter 3.2, we take on the aDNA data curated from all over the planet. This data represents more ancestries than the ones contributing for the CANDELA individuals. We tackled the issue with additional SNPs from the literature and an adaptive weighted linear regression model. In the ideal situation, we would add individuals from varied ancestries not contributing to CANDELA cohort. This would help us to identify new SNPs for pigmentation prediction as well as individuals to train our models. Although we have validated our prediction on the contemporary samples satisfactorily with our existing models, additional ancestries in the future will only enhance the global prediction.

Similarly, the tissue specific expression data we used from GTEx[259] database in Chapter 3.3, has representation from only European ancestries. This is a great study with enormous potential to capture the gene expression information in the human body. We expect them to include other ancestries in the future to elaborate the database which will definitely benefit the non-European sampling cohorts (e.g. CANDELA) to shed light on more genetic networks affecting other phenotypes as well. In the future we need to have advanced methods to incorporate the large datasets efficiently and studies are also making an effort in that direction[77,284], but until we have better representation from various population, we would still lack complete understanding of the phenotypes.

4.7. Concluding remarks

Therefore, CANDELA cohort was essential for the discoveries explained in this thesis. Although this population provides a 3-way ancestry admixture setup with an additional minor component from the East Asian ancestry, the Latin American population is underrepresented in the genetics studies. This unique scenario of population admixture challenges and enhances the knowledge of the pigmentation phenotypes which has mostly concentrated on European ancestry.

For pigmentation phenotype prediction, this cohort was crucial for preparation of new SNP sets and comparison with the existing assays. We further extended our methodology for prediction on the ancient human samples and provided an elaborate account of pigmentation evolution. The unique admixture profile of the cohort was key to the extensive coverage of the world population. This thesis identifies the potential of phenotype prediction utilising previous GWAS studies and provides guideline for new SNP set preparation. This can be extended to other visible phenotypes along with the pigmentation traits to have a more comprehensive understanding for individual identification.

Nonetheless, we should be cautious about the implication of the misuse of the predictive approaches, more so if we add the facial morphologies in the pipeline of human appearance prediction. Pigmentation traits, especially skin colour have been misused for specific isolated narratives by anthropologists, more so by the non-experts to highlight racism claims[125]. However, advanced methods now estimate ancestries from DNA samples, which counter the racism theory and have helped the appearance phenotypes to regain popularity[285].

To conclude, this study helps to navigate the existing genetic knowledge for predictive assignments and improves our understanding of the genetic architecture and predictive ability of the human pigmentation traits. This is the first instance of using transcriptomics data for pigmentation phenotypes within TWAS methods and the genotype-phenotype data from this cohort has led to identification of novel genetic regions. However, we are yet to achieve complete understanding – bigger datasets and more diverse studies in the future can improve our existing knowledge which will further benefit the predictive approaches.

References

1. Jablonski, N. G. *Living color: the biological and social meaning of skin color*. (University of California Press, 2014).

2. Slotten, R. A. *The heretic in Darwin's court: the life of Alfred Russel Wallace*. (Columbia University Press, 2004).

3. Benito-Martínez, S. *et al.* Research Techniques Made Simple: Cell Biology Methods for the Analysis of Pigmentation. *J. Invest. Dermatol.* **140**, 257-268.e8 (2020).

4. Jablonski, N. G. & Chaplin, G. The evolution of human skin coloration. *J. Hum. Evol.* **39**, 57–106 (2000).

5. Pavan, W. J. & Sturm, R. A. The Genetics of Human Skin and Hair Pigmentation. *Annu. Rev. Genomics Hum. Genet.* **20**, 41–72 (2019).

6. Hearing, V. J. Biogenesis of pigment granules: a sensitive way to regulate melanocyte function. *J. Dermatol. Sci.* **37**, 3–14 (2005).

7. Wasmeier, C., Hume, A. N., Bolasco, G. & Seabra, M. C. Melanosomes at a glance. *J. Cell Sci.* **121**, 3995–3999 (2008).

8. Raposo, G. & Marks, M. S. Melanosomes — dark organelles enlighten endosomal membrane transport. *Nat. Rev. Mol. Cell Biol.* **8**, 786–797 (2007).

9. Cichorek, M., Wachulska, M., Stasiewicz, A. & Tymińska, A. Skin melanocytes: biology and development. *Adv. Dermatol. Allergol.* **1**, 30–41 (2013).

10. Lin, J. Y. & Fisher, D. E. Melanocyte biology and skin pigmentation. *Nature* **445**, 843–850 (2007).

11. Sturm, R. A., Box, N. F. & Ramsay, M. Human pigmentation genetics: the difference is only skin deep. *BioEssays* **20**, 712–721 (1998).

12. Ando, H. *et al.* Melanosomes Are Transferred from Melanocytes to Keratinocytes through the Processes of Packaging, Release, Uptake, and Dispersion. *J. Invest. Dermatol.* **132**, 1222–1229 (2012).

13. Videira, I. F. dos S., Moura, D. F. L. & Magina, S. Mechanisms regulating melanogenesis*. *An. Bras. Dermatol.* **88**, 76–83 (2013).

14. Ortonne, J.-P. Melanins and Melanocytes: Form and Function. *Clin. Drug Investig.* **10**, 1–16 (1995).

15. Fu, C. *et al.* Roles of inflammation factors in melanogenesis (Review). *Mol. Med. Rep.* (2020) doi:10.3892/mmr.2020.10950.

16. Barsh, G. S. What Controls Variation in Human Skin Color? *PLoS Biol.* **1**, e27 (2003).

17. Swope, V. & Abdel-Malek, Z. MC1R: Front and Center in the Bright Side of Dark Eumelanin and DNA Repair. *Int. J. Mol. Sci.* **19**, 2667 (2018).

18. Del Bino, S. *et al.* Chemical analysis of constitutive pigmentation of human epidermis reveals constant eumelanin to pheomelanin ratio. *Pigment Cell Melanoma Res.* **28**, 707–717 (2015).

19. Thody, A. J. *et al.* Pheomelanin as well as Eumelanin Is Present in Human Epidermis. *J. Invest. Dermatol.* **97**, 340–344 (1991).

20. Chaitanya, L. *et al.* The HIrisPlex-S system for eye, hair and skin colour prediction from DNA: Introduction and forensic developmental validation. *Forensic Sci. Int. Genet.* **35**, 123–135 (2018).

21. Valverde, P., Healy, E., Jackson, I., Rees, J. L. & Thody, A. J. Variants of the melanocyte–stimulating hormone receptor gene are associated with red hair and fair skin in humans. *Nat. Genet.* **11**, 328–330 (1995).

22. Bastiaens, M. The melanocortin-1-receptor gene is the major freckle gene. *Hum. Mol. Genet.* **10**, 1701–1708 (2001).

23. Wolf Horrell, E. M., Boulanger, M. C. & D'Orazio, J. A. Melanocortin 1 Receptor: Structure, Function, and Regulation. *Front. Genet.* **7**, (2016).

24. White, D. & Rabago-Smith, M. Genotype–phenotype associations and human eye color. *J. Hum. Genet.* **56**, 5–7 (2011).

25. Del Bino, S. & Bernerd, F. Variations in skin colour and the biological consequences of ultraviolet radiation exposure. *Br. J. Dermatol.* **169**, 33–40 (2013).

26. Sturm, R. A. & Larsson, M. Genetics of human iris colour and patterns. *Pigment Cell Melanoma Res.* **22**, 544–562 (2009).

27. Prota, G., Hu, D.-N., Vincensi, M. R., McCORMICK, S. A. & Napolitano, A. Characterization of Melanins in Human Irides and Cultured Uveal Melanocytes From Eyes of Different Colors. *Exp. Eye Res.* **67**, 293–299 (1998).

28. Adhikari, K. *et al.* A GWAS in Latin Americans highlights the convergent evolution of lighter skin pigmentation in Eurasia. *Nat. Commun.* **10**, 358 (2019).

29. Ito, S. & Wakamatsu, K. Diversity of human hair pigmentation as studied by chemical analysis of eumelanin and pheomelanin: Human hair pigmentation. *J. Eur. Acad. Dermatol. Venereol.* **25**, 1369–1380 (2011).

30. Ito, S. & Fujita, K. Microanalysis of eumelanin and pheomelanin in hair and melanomas by chemical degradation and liquid chromatography. *Anal. Biochem.* **144**, 527–536 (1985).

31. Alaluf, S. *et al.* The Impact of Epidermal Melanin on Objective Measurements of Human Skin Colour. *PIGMENT CELL Res.* **15**, 119–126 (2002).

32. Deng, L. & Xu, S. Adaptation of human skin color in various populations. *Hereditas* **155**, 1 (2018).

33. Ito, S. & Wakamatsu, K. Quantitative Analysis of Eumelanin and Pheomelanin in Humans, Mice, and Other Animals: a Comparative Review. *Pigment Cell Res.* **16**, 523–531 (2003).

34. Napolitano, A., Panzella, L., Monfrecola, G. & d'Ischia, M. Pheomelanin-induced oxidative stress: bright and dark chemistry bridging red hair phenotype and melanoma. *Pigment Cell Melanoma Res.* **27**, 721–733 (2014).

35. Rosenblum, E. B., Hoekstra, H. E. & Nachman, M. W. ADAPTIVE REPTILE COLOR VARIATION AND THE EVOLUTION OF THE MCIR GENE. *Evolution* **58**, 1794–1808 (2004).

36. vonHoldt, B. M., Bailey, E. & Eizirik, E. Animal Pigmentation Genetics in Ecology, Evolution, and Domestication. *J. Hered.* **112**, 393–394 (2021).

37. Endler, J. A. A Predator's View of Animal Color Patterns. in *Evolutionary Biology* (eds. Hecht, M. K., Steere, W. C. & Wallace, B.) 319–364 (Springer US, 1978). doi:10.1007/978-1-4615-6956-5_5.

38. Endler, J. A. Natural Selection on Color Patterns in Poecilia reticulata. *Evolution* **34**, 76 (1980).

39. Wright, S. COLOR INHERITANCE IN MAMMALS. *J. Hered.* **8**, 224–235 (1917).

40. WRIGHT, S. COLOR INHERITANCE IN MAMMALSX., The Cat—Curious Association of Deafness with Blue-eyed White Color and of Femaleness with Tortoise-shelled Color, Long Known—Variations of Tiger Pattern Present Interesting Features. *J. Hered.* (1918) doi:10.1093/oxfordjournals.jhered.a111923.

41. Mills, M. G. & Patterson, L. B. Not just black and white: Pigment pattern development and evolution in vertebrates. *Semin. Cell Dev. Biol.* **20**, 72–81 (2009).

42. Protas, M. E. & Patel, N. H. Evolution of Coloration Patterns. *Annu. Rev. Cell Dev. Biol.* **24**, 425–446 (2008).

43. Baxter, L. L., Hou, L., Loftus, S. K. & Pavan, W. J. Spotlight on Spotted Mice: A Review of White Spotting Mouse Mutants and Associated Human Pigmentation Disorders. *Pigment Cell Res.* **17**, 215–224 (2004).

44. Caro, T. & Mallarino, R. Coloration in Mammals. *Trends Ecol. Evol.* **35**, 357–366 (2020).

45. Kamberov, Y. G. *et al.* Comparative evidence for the independent evolution of hair and sweat gland traits in primates. *J. Hum. Evol.* **125**, 99–105 (2018).

46. JABLONSKI, N. G. Human Skin Pigmentation as an Example of Adaptive Evolution. *Proc. Am. Philos. Soc.* **156**, 45–57 (2012).

47. Jablonski, N. G. The Evolution of Human Skin and Skin Color. *Annu. Rev. Anthropol.* **33**, 585–623 (2004).

48. Hart, P. H., Gorman, S. & Finlay-Jones, J. J. Modulation of the immune system by UV radiation: more than just the effects of vitamin D? *Nat. Rev. Immunol.* **11**, 584–596 (2011).

49. Jablonski, N. G. & Chaplin, G. Human skin pigmentation as an adaptation to UV radiation. *Proc. Natl. Acad. Sci.* **107**, 8962–8968 (2010).

50. Jablonski, N. G. & Chaplin, G. The colours of humanity: the evolution of pigmentation in the human lineage. *Philos. Trans. R. Soc. B Biol. Sci.* **372**, 20160349 (2017).

51. Sturm, R. A. & Duffy, D. L. Human pigmentation genes under environmental selection. *Genome Biol.* **13**, 248 (2012).

52. Relethford, J. H. Hemispheric difference in human skin color. *Am. J. Phys. Anthropol.* **104**, 449–457 (1997).

53. Candille, S. I. *et al.* Genome-Wide Association Studies of Quantitatively Measured Skin, Hair, and Eye Pigmentation in Four European Populations. *PLoS ONE* **7**, e48294 (2012).

54. Jablonski, N. G. The evolution of human skin pigmentation involved the interactions of genetic, environmental, and cultural variables. *Pigment Cell Melanoma Res.* pcmr.12976 (2021) doi:10.1111/pcmr.12976.

55. Wilde, S. *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.* **111**, 4832–4837 (2014).

56. Liu, F., Wen, B. & Kayser, M. Colorful DNA polymorphisms in humans. *Semin. Cell Dev. Biol.* **24**, 562–575 (2013).

57. Laeng, B., Mathisen, R. & Johnsen, J.-A. Why do blue-eyed men prefer women with the same eye color? *Behav. Ecol. Sociobiol.* **61**, 371–384 (2006).

58. Aoki, K. Sexual selection as a cause of human skin colour variation: Darwin's hypothesis revisited. *Ann. Hum. Biol.* **29**, 589–608 (2002).

59. Frost, P. Human Skin Color: A Possible Relationship Between Its Sexual Dimorphism and Its Social Perception. *Perspect. Biol. Med.* **32**, 38–58 (1988).

60. King, T. E. *et al.* Identification of the remains of King Richard III. *Nat. Commun.* **5**, 5631 (2014).

61. Beals, R. L. & Hoijer, H. An Introduction to Anthropology. *Macmilan N. Y.* (1965).

62. Loos, R. J. F. 15 years of genome-wide association studies and no signs of slowing down. *Nat. Commun.* **11**, 5900 (2020).

63. Uffelmann, E. *et al.* Genome-wide association studies. *Nat. Rev. Methods Primer* **1**, 59 (2021).

64. Wei, W.-H., Hemani, G. & Haley, C. S. Detecting epistasis in human complex traits. *Nat. Rev. Genet.* **15**, 722–733 (2014).

65. Mackay, T. F. C. Epistasis and quantitative traits: using model organisms to study gene–gene interactions. *Nat. Rev. Genet.* **15**, 22–33 (2014).

66. Grinberg, N. F., Orhobor, O. I. & King, R. D. An evaluation of machine-learning for predicting phenotype: studies in yeast, rice, and wheat. *Mach. Learn.* **109**, 251–277 (2020).

67. Aronson, S. J. & Rehm, H. L. Building the foundation for genomics in precision medicine. *Nature* **526**, 336–342 (2015).

68. Chatterjee, N., Shi, J. & García-Closas, M. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nat. Rev. Genet.* **17**, 392–406 (2016).

69. Torkamani, A., Wineinger, N. E. & Topol, E. J. The personal and clinical utility of polygenic risk scores. *Nat. Rev. Genet.* **19**, 581–590 (2018).

70. Ma, W. *et al.* A deep convolutional neural network approach for predicting phenotypes from genotypes. *Planta* **248**, 1307–1318 (2018).

71. Montesinos-López, A., Montesinos-López, O. A., Gianola, D., Crossa, J. & Hernández-Suárez, C. M. Multi-environment Genomic Prediction of Plant Traits Using Deep Learners With Dense Architecture. *G3 GenesGenomesGenetics* **8**, 3813–3828 (2018).

72. Montesinos-López, O. A. *et al.* New Deep Learning Genomic-Based Prediction Model for Multiple Traits with Binary, Ordinal, and Continuous Phenotypes. *G3 GenesGenomesGenetics* **9**, 1545–1556 (2019).

73. Lees, J. A. *et al.* Improved Prediction of Bacterial Genotype-Phenotype Associations Using Interpretable Pangenome-Spanning Regressions. *mBio* **11**, e01344-20 (2020).

74. Moradigaravand, D. *et al.* Prediction of antibiotic resistance in Escherichia coli from large-scale pan-genome data. *PLOS Comput. Biol.* **14**, e1006258 (2018).

75. Cui, T., Mekkaoui, K. E., Havulinna, A., Marttinen, P. & Kaski, S. *Improving Neural Networks for Genotype-Phenotype Prediction Using Published Summary Statistics*. http://biorxiv.org/lookup/doi/10.1101/2021.11.09.467937 (2021) doi:10.1101/2021.11.09.467937.

76. Liu, F. *et al.* Update on the predictability of tall stature from DNA markers in Europeans. *Forensic Sci. Int. Genet.* **42**, 8–13 (2019).

77. Mbatchou, J. *et al.* Computationally efficient whole-genome regression for quantitative and binary traits. *Nat. Genet.* **53**, 1097–1103 (2021).

78. Liu, F. *et al.* Eye color and the prediction of complex phenotypes from genotypes. *Curr. Biol.* **19**, R192–R193 (2009).

79. Branicki, W. *et al.* Model-based prediction of human hair color using DNA variants. *Hum. Genet.* **129**, 443–454 (2011).

80. Walsh, S. *et al.* The HIrisPlex system for simultaneous prediction of hair and eye colour from DNA. *Forensic Sci. Int. Genet.* **7**, 98–115 (2013).

81. Pirinen, M., Donnelly, P. & Spencer, C. C. A. Including known covariates can reduce power to detect genetic effects in case-control studies. *Nat. Genet.* **44**, 848–851 (2012).

82. Goddard, M. E., Wray, N. R., Verbyla, K. & Visscher, P. M. Estimating Effects and Making Predictions from Genome-Wide Marker Data. *Stat. Sci.* **24**, (2009).

83. Speed, D. & Balding, D. J. MultiBLUP: improved SNP-based prediction for complex traits. *Genome Res.* **24**, 1550–1557 (2014).

84. Wang, J. *et al.* Expanding the BLUP alphabet for genomic prediction adaptable to the genetic architectures of complex traits. *Heredity* **121**, 648–662 (2018).

85. Gualdrón Duarte, J. L. *et al.* Performances of Adaptive MultiBLUP, Bayesian regressions, and weighted-GBLUP approaches for genomic predictions in Belgian Blue beef cattle. *BMC Genomics* **21**, 545 (2020).

86. Montesinos-López, O. A. *et al.* A review of deep learning applications for genomic selection. *BMC Genomics* **22**, 19 (2021).

87. van Dijk, A. D. J., Kootstra, G., Kruijer, W. & de Ridder, D. Machine learning in plant science and plant breeding. *iScience* **24**, 101890 (2021).

88. de los Campos, G. *et al.* Predicting Quantitative Traits With Regression Models for Dense Molecular Markers and Pedigree. *Genetics* **182**, 375–385 (2009).

89. Lello, L. *et al.* Accurate Genomic Prediction of Human Height. *Genetics* **210**, 477–497 (2018).

90. Saha, S., Perrin, L., Röder, L., Brun, C. & Spinelli, L. Epi-MEIF: detecting higher order epistatic interactions for complex traits using mixed effect conditional inference forests. *Nucleic Acids Res.* gkac715 (2022) doi:10.1093/nar/gkac715.

91. Jiang, R., Tang, W., Wu, X. & Fu, W. A random forest approach to the detection of epistatic interactions in case-control studies. *BMC Bioinformatics* **10**, S65 (2009).

92. Li, B. *et al.* Genomic Prediction of Breeding Values Using a Subset of SNPs Identified by Three Machine Learning Methods. *Front. Genet.* **9**, 237 (2018).

93. Badré, A., Zhang, L., Muchero, W., Reynolds, J. C. & Pan, C. Deep neural network improves the estimation of polygenic risk scores for breast cancer. *J. Hum. Genet.* **66**, 359–369 (2021).

94. Gola, D., Erdmann, J., Müller-Myhsok, B., Schunkert, H. & König, I. R. Polygenic risk scores outperform machine learning methods in predicting coronary artery disease status. *Genet. Epidemiol.* **44**, 125–138 (2020).

95. Muneeb, M. & Henschel, A. Eye-color and Type-2 diabetes phenotype prediction from genotype data using deep learning methods. *BMC Bioinformatics* **22**, 198 (2021).

96. Zeng, S. *et al.* G2PDeep: a web-based deep-learning framework for quantitative phenotype prediction and discovery of genomic markers. *Nucleic Acids Res.* **49**, W228–W236 (2021).

97. Voigt, P. & von dem Bussche, A. *The EU General Data Protection Regulation (GDPR): A Practical Guide*. (Springer International Publishing : Imprint: Springer, 2017). doi:10.1007/978-3-319-57959-7.

98. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).

99. Lambert, S. A., Abraham, G. & Inouye, M. Towards clinical utility of polygenic risk scores. *Hum. Mol. Genet.* **28**, R133–R142 (2019).

100. Euesden, J., Lewis, C. M. & O'Reilly, P. F. PRSice: Polygenic Risk Score software. *Bioinformatics* **31**, 1466–1468 (2015).

101. Zhang, Q., Privé, F., Vilhjálmsson, B. & Speed, D. Improved genetic prediction of complex traits from individual-level data or summary statistics. *Nat. Commun.* **12**, 4192 (2021).

102. Privé, F., Arbel, J. & Vilhjálmsson, B. J. LDpred2: better, faster, stronger. *Bioinformatics* **36**, 5424–5431 (2021).

103. Márquez-Luna, C. *et al.* Incorporating functional priors improves polygenic prediction accuracy in UK Biobank and 23andMe data sets. *Nat. Commun.* **12**, 6052 (2021).

104. Khera, A. V. *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* **50**, 1219–1224 (2018).

105. Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C. A. & Smoller, J. W. Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nat. Commun.* **10**, 1776 (2019).

106. Lloyd-Jones, L. R. *et al.* Improved polygenic prediction by Bayesian multiple regression on summary statistics. *Nat. Commun.* **10**, 5086 (2019).

107. Zhu, X. & Stephens, M. Bayesian large-scale multiple regression with summary statistics from genome-wide association studies. *Ann. Appl. Stat.* **11**, (2017).

108. Kerminen, S. *et al.* Geographic Variation and Bias in the Polygenic Scores of Complex Diseases and Traits in Finland. *Am. J. Hum. Genet.* **104**, 1169–1181 (2019).

109. Zaidi, A. A. & Mathieson, I. Demographic history mediates the effect of stratification on polygenic scores. *eLife* **9**, e61548 (2020).

110. Gamulin, O. *et al.* Possibility of Human Gender Recognition Using Raman Spectra of Teeth. *Molecules* **26**, 3983 (2021).

111. Bruzek, J. & Murail, P. Methodology and Reliability of Sex Determination From the Skeleton. in *Forensic Anthropology and Medicine* (eds. Schmitt, A., Cunha, E. & Pinheiro, J.) 225–242 (Humana Press, 2006). doi:10.1007/978-1-59745-099-7_9.

112. Nagare, S., Chaudhari, R., Birangane, R. & Parkarwar, P. Sex determination in forensic identification, a review. *J. Forensic Dent. Sci.* **10**, 61 (2018).

113. Prabhat, M. *et al.* Computed tomography based forensic gender determination by measuring the size and volume of the maxillary sinuses. *J. Forensic Dent. Sci.* **8**, 40 (2016).

114. Secgin, Y., Oner, Z., Turan, M. & Oner, S. Gender prediction with parameters obtained from pelvis computed tomography images and decision tree algorithm. *Med. Sci. Int. Med. J.* **10**, 356 (2021).

115. Öner, S., Turan, M. & Öner, Z. Estimation of gender by using decision tree, a machine learning algorithm, with patellar measurements obtained from MDCT images. *Med. Rec.* (2021) doi:10.37990/medr.843451.

116. Atas, I. Human Gender Prediction Based on Deep Transfer Learning from Panoramic Radiograph Images. (2022) doi:10.48550/ARXIV.2205.09850.

117. Huynh, C., Brunelle, E., Halámková, L., Agudelo, J. & Halámek, J. Forensic Identification of Gender from Fingerprints. *Anal. Chem.* **87**, 11531–11536 (2015).

118. Rösing, F. W. *et al.* Recommendations for the forensic diagnosis of sex and age from skeletons. *HOMO* **58**, 75–89 (2007).

119. Santosh, K. C. *et al.* Machine Learning Techniques for Human Age and Gender Identification Based on Teeth X-Ray Images. *J. Healthc. Eng.* **2022**, 1–14 (2022).

120. Ruff, C. B. *et al.* Stature and body mass estimation from skeletal remains in the European Holocene. *Am. J. Phys. Anthropol.* **148**, 601–617 (2012).

121. Lee, S. *et al.* Estimation of stature from femur length measured using computed tomography after the analysis of three-dimensional characteristics of femur bone in Korean cadavers. *Int. J. Legal Med.* **131**, 1355–1362 (2017).

122. Petrovečki, V., Mayer, D., Šlaus, M., Strinović, D. & Škavić, J. Prediction of Stature Based on Radiographic Measurements of Cadaver Long Bones: A Study of the Croatian Population. *J. Forensic Sci.* **52**, 547–552 (2007).

123. Albanese, J., Tuck, A., Gomes, J. & Cardoso, H. F. V. An alternative approach for estimating stature from long bones that is not population- or group-specific. *Forensic Sci. Int.* **259**, 59–68 (2016).

124. Kayser, M. Forensic DNA Phenotyping: Predicting human appearance from crime scene material for investigative purposes. *Forensic Sci. Int. Genet.* **18**, 33–48 (2015).

125. Walsh, S. DNA Phenotyping: The prediction of human pigmentation traits from genetic data. (Erasmus University Rotterdam, 2013).

126. Ruiz, Y. *et al.* Further development of forensic eye color predictive tests. *Forensic Sci. Int. Genet.* **7**, 28–40 (2013).

127. Spichenok, O. *et al.* Prediction of eye and skin color in diverse populations using seven SNPs. *Forensic Sci. Int. Genet.* **5**, 472–478 (2011).

128. Hart, K. L. *et al.* Improved eye- and skin-color prediction based on 8 SNPs. *Croat. Med. J.* **54**, 248–256 (2013).

129. Allwood, J. S. & Harbison, S. SNP model development for the prediction of eye colour in New Zealand. *Forensic Sci. Int. Genet.* **7**, 444–452 (2013).

130. Walsh, S. *et al.* IrisPlex: A sensitive DNA tool for accurate prediction of blue and brown eye colour in the absence of ancestry information. *Forensic Sci. Int. Genet.* **5**, 170–180 (2011).

131. Yun, L., Gu, Y., Rajeevan, H. & Kidd, K. K. Application of six IrisPlex SNPs and comparison of two eye color prediction systems in diverse Eurasia populations. *Int. J. Legal Med.* **128**, 447–453 (2014).

132. Dario, P. *et al.* Assessment of IrisPlex-based multiplex for eye and skin color prediction with application to a Portuguese population. *Int. J. Legal Med.* **129**, 1191–1200 (2015).

133. Kastelic, V., Pośpiech, E., Draus-Barini, J., Branicki, W. & Drobnič, K. Prediction of eye color in the Slovenian population using the IrisPlex SNPs. *Croat. Med. J.* **54**, 381–386 (2013).

134. Purps, J., Geppert, M., Nagy, M. & Roewer, L. Evaluation of the IrisPlex eye colour prediction tool in a German population sample. *Forensic Sci. Int. Genet. Suppl. Ser.* **3**, e202–e203 (2011).

135. Dembinski, G. M. & Picard, C. J. Evaluation of the IrisPlex DNA-based eye color prediction assay in a United States population. *Forensic Sci. Int. Genet.* **9**, 111–117 (2014).

136. Walsh, S. *et al.* Developmental validation of the HIrisPlex system: DNA-based eye and hair colour prediction for forensic and anthropological usage. *Forensic Sci. Int. Genet.* **9**, 150–161 (2014).

137. Draus-Barini, J. *et al.* Bona fide colour: DNA prediction of human eye and hair colour from ancient and contemporary skeletal remains. *Investig. Genet.* **4**, 3 (2013).

138. Carratto, T. M. T. *et al.* Insights on hair, skin and eye color of ancient and contemporary Native Americans. *Forensic Sci. Int. Genet.* **48**, 102335 (2020).

139. Lazaridis, I. *et al.* A genetic probe into the ancient and medieval history of Southern Europe and West Asia. *Science* **377**, 940–951 (2022).

140. Breslin, K. *et al.* HIrisPlex-S system for eye, hair, and skin color prediction from DNA: Massively parallel sequencing solutions for two common forensically used platforms. *Forensic Sci. Int. Genet.* **43**, 102152 (2019).

141. Edwards, M. *et al.* 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**, 141–162 (2016).

142. Salvoro, C. *et al.* Performance of four models for eye color prediction in an Italian population sample. *Forensic Sci. Int. Genet.* **40**, 192–200 (2019).

143. Maroñas, O. *et al.* Development of a forensic skin colour predictive test. *Forensic Sci. Int. Genet.* **13**, 34–44 (2014).

144. Canales Serrano, A. Forensic DNA phenotyping: A promising tool to aid forensic investigation. Current situation. *Span. J. Leg. Med.* **46**, 183–190 (2020).

145. Cox, S. L. *et al.* Predicting skeletal stature using ancient DNA. *Am. J. Biol. Anthropol.* **177**, 162–174 (2022).

146. Thorne, A. G. & Wolpoff, M. H. The multiregional evolution of humans. *Sci. Am.* **266**, 76–83 (1992).

147. Eckhardt, R. B., Wolpoff, M. H. & Thorne, A. G. Multiregional Evolution. *Science* **262**, 973–974 (1993).

148. Disotell, T. R. Human evolution: Origins of modern humans still look recent. *Curr. Biol.* **9**, R647–R650 (1999).

149. Pagani, L. *et al.* Tracing the Route of Modern Humans out of Africa by Using 225 Human Genome Sequences from Ethiopians and Egyptians. *Am. J. Hum. Genet.* **96**, 986–991 (2015).

150. Cann, R. L., Stoneking, M. & Wilson, A. C. Mitochondrial DNA and human evolution. *Nature* **325**, 31–36 (1987).

151. Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nat. Genet.* **46**, 919–925 (2014).

152. López, S., Van Dorp, L. & Hellenthal, G. Human Dispersal Out of Africa: A Lasting Debate. *Evol. Bioinforma.* **11s2**, EBO.S33489 (2015).

153. Garrigan, D., Mobasher, Z., Kingan, S. B., Wilder, J. A. & Hammer, M. F. Deep Haplotype Divergence and Long-Range Linkage Disequilibrium at Xp21.1 Provide Evidence That Humans Descend From a Structured Ancestral Population. *Genetics* **170**, 1849–1856 (2005).

154. Plagnol, V. & Wall, J. D. Possible Ancestral Structure in Human Populations. *PLoS Genet.* **2**, e105 (2006).

155. Currat, M. & Excoffier, L. Modern Humans Did Not Admix with Neanderthals during Their Range Expansion into Europe. *PLoS Biol.* **2**, e421 (2004).

156. Green, R. E. *et al.* A Draft Sequence of the Neandertal Genome. *Science* **328**, 710–722 (2010).

157. Reich, D. *et al.* Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468**, 1053–1060 (2010).

158. Reich, D. *et al.* Denisova Admixture and the First Modern Human Dispersals into Southeast Asia and Oceania. *Am. J. Hum. Genet.* **89**, 516–528 (2011).

159. Eriksson, A. & Manica, A. Effect of ancient population structure on the degree of polymorphism shared between modern human populations and ancient hominins. *Proc. Natl. Acad. Sci.* **109**, 13956–13960 (2012).

160. Sankararaman, S., Patterson, N., Li, H., Pääbo, S. & Reich, D. The Date of Interbreeding between Neandertals and Modern Humans. *PLoS Genet.* **8**, e1002947 (2012).

161. Yang, M. A., Malaspinas, A.-S., Durand, E. Y. & Slatkin, M. Ancient Structure in Africa Unlikely to Explain Neanderthal and Non-African Genetic Similarity. *Mol. Biol. Evol.* **29**, 2987–2995 (2012).

162. Sabeti, P. C. *et al.* Positive Natural Selection in the Human Lineage. *Science* **312**, 1614–1620 (2006).

163. Nakagome, S. *et al.* Estimating the Ages of Selection Signals from Different Epochs in Human History. *Mol. Biol. Evol.* **33**, 657–669 (2016).

164. Field, Y. *et al.* Detection of human adaptation during the past 2000 years. *Science* **354**, 760–764 (2016).

165. Marciniak, S. & Perry, G. H. Harnessing ancient genomes to study the history of human adaptation. *Nat. Rev. Genet.* **18**, 659–674 (2017).

166. The International HapMap Consortium *et al.* Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**, 913–918 (2007).

167. Rocha, J. The Evolutionary History of Human Skin Pigmentation. *J. Mol. Evol.* **88**, 77–87 (2020).

168. Irving-Pease, E. K., Muktupavela, R., Dannemann, M. & Racimo, F. Quantitative Human Paleogenetics: What can Ancient DNA Tell us About Complex Trait Evolution? *Front. Genet.* **12**, 703541 (2021).

169. Martin, A. R. *et al.* Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *Am. J. Hum. Genet.* **100**, 635–649 (2017).

170. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).

171. Olalde, I. *et al.* Derived immune and ancestral pigmentation alleles in a 7,000-year-old Mesolithic European. *Nature* **507**, 225–228 (2014).

172. González-Fortes, G. *et al.* Paleogenomic Evidence for Multi-generational Mixing between Neolithic Farmers and Mesolithic Hunter-Gatherers in the Lower Danube Basin. *Curr. Biol.* **27**, 1801-1810.e10 (2017).

173. Ju, D. & Mathieson, I. The evolution of skin pigmentation-associated variation in West Eurasia. *Proc. Natl. Acad. Sci.* **118**, e2009227118 (2021).

174. The LifeLines Cohort Study *et al.* Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat. Genet.* **47**, 1114–1120 (2015).

175. O'Connor, L. J. *et al.* Extreme Polygenicity of Complex Traits Is Explained by Negative Selection. *Am. J. Hum. Genet.* **105**, 456–476 (2019).

176. Mathieson, I. *et al.* Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* **528**, 499–503 (2015).

177. Martiniano, R. *et al.* The population genomics of archaeological transition in west Iberia: Investigation of ancient substructure using imputation and haplotype-based methods. *PLOS Genet.* **13**, e1006852 (2017).

178. Cox, S. L., Ruff, C. B., Maier, R. M. & Mathieson, I. Genetic contributions to variation in human stature in prehistoric Europe. *Proc. Natl. Acad. Sci.* **116**, 21484–21492 (2019).

179. Burger, J. *et al.* Low Prevalence of Lactase Persistence in Bronze Age Europe Indicates Ongoing Strong Selection over the Last 3,000 Years. *Curr. Biol.* **30**, 4307-4315.e13 (2020).

180. Enattah, N. S. *et al.* Independent Introduction of Two Lactase-Persistence Alleles into Human Populations Reflects Different History of Adaptation to Milk Culture. *Am. J. Hum. Genet.* **82**, 57–72 (2008).

181. Gerbault, P. *et al.* Evolution of lactase persistence: an example of human niche construction. *Philos. Trans. R. Soc. B Biol. Sci.* **366**, 863–877 (2011).

182. Olalde, I. *et al.* The Beaker phenomenon and the genomic transformation of northwest Europe. *Nature* **555**, 190–196 (2018).

183. Allentoft, M. E. *et al.* Population genomics of Bronze Age Eurasia. *Nature* **522**, 167–172 (2015).

184. Furholt, M., Grossmann, R. & Szmyt, M. Corded Ware and Bell Beakers: A Practice-based Perspective on Local Communities, Transregional Interaction and Social Heterogeneity in Late Neolithic Europe. *Transitional Landsc.* 117–131 (2016).

185. Parra, E. J. Human pigmentation variation: Evolution, genetic basis, and implications for public health. *Am. J. Phys. Anthropol.* **134**, 85–105 (2007).

186. Stokowski, R. P. *et al.* A Genomewide Association Study of Skin Pigmentation in a South Asian Population. *Am. J. Hum. Genet.* **81**, 1119–1132 (2007).

187. Sulem, P. *et al.* Genetic determinants of hair, eye and skin pigmentation in Europeans. *Nat. Genet.* **39**, 1443–1452 (2007).

188. Kumar, S. & Newfeld, S. J. *Modern Genetic Analysis: Integrating Genes and Genomes . Second Edition. By* Anthony J F Griffiths, William M Gelbart, Richard C Lewontin, and Jeffrey H Miller. *New York: W. H. Freeman* . $75.00. xxi + 736 p; ill.; index. ISBN: 0–7167–4382–5. 2002. *Q. Rev. Biol.* **77**, 456–457 (2002).

189. Hurst, C. C. On the inheritance of eye-colour in man. *Proc. R. Soc. Lond. Ser. B Contain. Pap. Biol. Character* **80**, 85–96 (1908).

190. Mackey, D. A. What colour are your eyes? Teaching the genetics of eye colour & colour vision. Edridge Green Lecture RCOphth Annual Congress Glasgow May 2019. *Eye* **36**, 704–715 (2022).

191. Morgan, M. D. *et al.* Genome-wide study of hair colour in UK Biobank explains most of the SNP heritability. *Nat. Commun.* **9**, 5271 (2018).

192. Davenport, G. C. & Davenport, C. B. Heredity of Skin Pigmentation in Man. *Am. Nat.* **44**, 641–672 (1910).

193. Yang, Q. *et al.* Genetic analyses of oculocutaneous albinism types 1 and 2 with four novel mutations. *BMC Med. Genet.* **20**, 106 (2019).

194. Lee, S. T. *et al.* Diverse mutations of the P gene among African-Americans with type II (tyrosinase-positive) oculocutaneous albinism (OCA2). *Hum. Mol. Genet.* **3**, 2047–2051 (1994).

195. *Albinism in Africa: historical, geographic, medical, genetic, and psychosocial aspects*. (Academic Press, an imprint of Elsevier, 2018).

196. Lee, S. Y. *et al.* A new type of oculocutaneous albinism with a novel OCA2 mutation. *Yeungnam Univ. J. Med.* **38**, 160–164 (2021).

197. Crawford, N. G. *et al.* Loci associated with skin pigmentation identified in African populations. *Science* **358**, eaan8433 (2017).

198. Martin, A. R. *et al.* An Unexpectedly Complex Architecture for Skin Pigmentation in Africans. *Cell* **171**, 1340-1353.e14 (2017).

199. Eaton, K. *et al.* Association study confirms the role of two *OCA2* polymorphisms in normal skin pigmentation variation in East Asian populations: OCA2 POLYMORPHISMS AND MELANIN LEVELS IN EAST ASIA. *Am. J. Hum. Biol.* **27**, 520–525 (2015).

200. Norton, H. L. *et al.* Quantitative assessment of skin, hair, and iris variation in a diverse sample of individuals and associated genetic variation: QUANTITATIVE ASSESSMENT OF PIGMENTARY PHENOTYPE. *Am. J. Phys. Anthropol.* **160**, 570–581 (2016).

201. Yang, Z. *et al.* A Genetic Mechanism for Convergent Skin Lightening during Recent Human Evolution. *Mol. Biol. Evol.* **33**, 1177–1187 (2016).

202. Iliescu, F. M. *et al.* The influences of genes, the environment, and social factors on the evolution of skin color diversity in India. *Am. J. Hum. Biol.* **30**, e23170 (2018).

203. Jonnalagadda, M. *et al.* Identifying signatures of positive selection in pigmentation genes in two South Asian populations: JONNALAGADDA et al. *Am. J. Hum. Biol.* **29**, e23012 (2017).

204. Simcoe, M. *et al.* Genome-wide association study in almost 195,000 individuals identifies 50 previously unidentified genetic loci for eye color. *Sci. Adv.* **7**, eabd1239 (2021).

205. The International Visible Trait Genetics Consortium *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**, 652–656 (2018).

206. Visconti, A. *et al.* Genome-wide association study in 176,678 Europeans reveals genetic loci for tanning response to sun exposure. *Nat. Commun.* **9**, 1684 (2018).

207. Lona-Durazo, F. *et al.* Meta-analysis of GWA studies provides new insights on the genetic architecture of skin pigmentation in recently admixed populations. *BMC Genet.* **20**, 59 (2019).

208. Visser, M., Kayser, M. & 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**, 446–455 (2012).

209. Abdel-Malek, Z. A. *et al.* Melanocortins and the melanocortin 1 receptor, moving translationally towards melanoma prevention. *Arch. Biochem. Biophys.* **563**, 4–12 (2014).

210. Mitra, D. *et al.* An ultraviolet-radiation-independent pathway to melanoma carcinogenesis in the red hair/fair skin background. *Nature* **491**, 449–453 (2012).

211. Li, B. & Ritchie, M. D. From GWAS to Gene: Transcriptome-Wide Association Studies and Other Methods to Functionally Understand GWAS Discoveries. *Front. Genet.* **12**, 713230 (2021).

212. The GTEx Consortium *et al.* The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).

213. Maurano, M. T. *et al.* Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science* **337**, 1190–1195 (2012).

214. Wainberg, M. *et al.* Opportunities and challenges for transcriptome-wide association studies. *Nat. Genet.* **51**, 592–599 (2019).

215. Bhattacharya, A. *et al.* A framework for transcriptome-wide association studies in breast cancer in diverse study populations. *Genome Biol.* **21**, 42 (2020).

216. Dong, X. *et al.* A general framework for functionally informed set-based analysis: Application to a large-scale colorectal cancer study. *PLOS Genet.* **16**, e1008947 (2020).

217. Zhong, J. *et al.* A Transcriptome-Wide Association Study Identifies Novel Candidate Susceptibility Genes for Pancreatic Cancer. *JNCI J. Natl. Cancer Inst.* **112**, 1003–1012 (2020).

218. Tapia, A. L. *et al.* A large-scale transcriptome-wide association study (TWAS) of 10 blood cell phenotypes reveals complexities of TWAS fine-mapping. *Genet. Epidemiol.* **46**, 3–16 (2022).

219. Alzheimer's Disease Genetics Consortium, *et al.* A statistical framework for cross-tissue transcriptome-wide association analysis. *Nat. Genet.* **51**, 568–576 (2019).

220. Fryett, J. J., Inshaw, J., Morris, A. P. & Cordell, H. J. Comparison of methods for transcriptome imputation through application to two common complex diseases. *Eur. J. Hum. Genet.* **26**, 1658–1667 (2018).

221. Geoffroy, E., Gregga, I. & Wheeler, H. E. Population-Matched Transcriptome Prediction Increases TWAS Discovery and Replication Rate. *iScience* **23**, 101850 (2020).

222. Barbeira, A. N. *et al.* Integrating predicted transcriptome from multiple tissues improves association detection. *PLOS Genet.* **15**, e1007889 (2019).

223. Ruiz-Linares, A. *et al.* Admixture in Latin America: Geographic Structure, Phenotypic Diversity and Self-Perception of Ancestry Based on 7,342 Individuals. *PLoS Genet.* **10**, e1004572 (2014).

224. R Core Team. R: A Language and Environment for Statistical Computing. (2020).

225. Adhikari, K. *et al.* A genome-wide association scan in admixed Latin Americans identifies loci influencing facial and scalp hair features. *Nat. Commun.* **7**, 10815 (2016).

226. Schmid, P. Colour segmentation for the analysis of pigmented skin lesions. in *6th International Conference on Image Processing and its Applications* vol. 1997 688–692 (IEE, 1997).

227. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, 7 (2015).

228. O'Connell, J. *et al.* A General Approach for Haplotype Phasing across the Full Spectrum of Relatedness. *PLoS Genet.* **10**, e1004234 (2014).

229. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.* **44**, 955–959 (2012).

230. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).

231. Ruiz-Linares, A. *et al.* Admixture in Latin America: Geographic Structure, Phenotypic Diversity and Self-Perception of Ancestry Based on 7,342 Individuals. *PLoS Genet.* **10**, e1004572 (2014).

232. Li, S. *et al.* Detecting and correcting systematic variation in large-scale RNA sequencing data. *Nat. Biotechnol.* **32**, 888–895 (2014).

233. Smith, A. M. *et al.* Standard machine learning approaches outperform deep representation learning on phenotype prediction from transcriptomics data. *BMC Bioinformatics* **21**, 119 (2020).

234. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning*. (Springer New York Inc., 2008).

235. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).

236. Cordell, H. J. & Clayton, D. G. A Unified Stepwise Regression Procedure for Evaluating the Relative Effects of Polymorphisms within a Gene Using Case/Control or Family Data: Application to HLA in Type 1 Diabetes. *Am. J. Hum. Genet.* **70**, 124–141 (2002).

237. Knüppel, S. *et al.* Multi-locus stepwise regression: a haplotype-based algorithm for finding genetic associations applied to atopic dermatitis. *BMC Med. Genet.* **13**, 8 (2012).

238. Walsh, S. *et al.* Global skin colour prediction from DNA. *Hum. Genet.* **136**, 847–863 (2017).

239. Venables, W. N., Ripley, B. D. & Venables, W. N. *Modern applied statistics with S*. (Springer, 2002).

240. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (ACM, 2016). doi:10.1145/2939672.2939785.

241. Li, W., Yin, Y., Quan, X. & Zhang, H. Gene Expression Value Prediction Based on XGBoost Algorithm. *Front. Genet.* **10**, 1077 (2019).

242. Shen, C. *et al.* DLRAPom: a hybrid pipeline of Optimized XGBoost-guided integrative multiomics analysis for identifying targetable disease-related lncRNA–miRNA–mRNA regulatory axes. *Brief. Bioinform.* **23**, bbac046 (2022).

243. Sun, K. *et al.* Application of machine learning for ancestry inference using multi-InDel markers. *Forensic Sci. Int. Genet.* **59**, 102702 (2022).

244. Breiman, L. Random Forests. *Mach. Learn.* **45**, 5–32 (2001).

245. Sun, Y. V. *et al.* Classification of rheumatoid arthritis status with candidate gene and genome-wide single-nucleotide polymorphisms using random forests. *BMC Proc.* **1**, S62 (2007).

246. Díaz-Uriarte, R. & Alvarez de Andrés, S. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics* **7**, 3 (2006).

247. Ishwaran, H., Kogalur, U. B., Gorodeski, E. Z., Minn, A. J. & Lauer, M. S. High-Dimensional Variable Selection for Survival Data. *J. Am. Stat. Assoc.* **105**, 205–217 (2010).

248. Stephan, J., Stegle, O. & Beyer, A. A random forest approach to capture genetic effects in the presence of population structure. *Nat. Commun.* **6**, 7432 (2015).

249. Daoud, M. & Mayo, M. A survey of neural network-based cancer prediction models from microarray data. *Artif. Intell. Med.* **97**, 204–214 (2019).

250. Zaorska, K., Zawierucha, P. & Nowicki, M. Prediction of skin color, tanning and freckling from DNA in Polish population: linear regression, random forest and neural network approaches. *Hum. Genet.* **138**, 635–647 (2019).

251. Vidaki, A. *et al.* DNA methylation-based forensic age prediction using artificial neural networks and next generation sequencing. *Forensic Sci. Int. Genet.* **28**, 225–236 (2017).

252. Hanczar, B., Zehraoui, F., Issa, T. & Arles, M. Biological interpretation of deep neural network for phenotype prediction based on gene expression. *BMC Bioinformatics* **21**, 501 (2020).

253. Sheu, Y. Illuminating the Black Box: Interpreting Deep Neural Network Models for Psychiatric Research. *Front. Psychiatry* **11**, 551299 (2020).

254. Agarwal, R. *et al.* Neural Additive Models: Interpretable Machine Learning with Neural Nets. (2020) doi:10.48550/ARXIV.2004.13912.

255. Zou, H. & Hastie, T. Regularization and Variable Selection via the Elastic Net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67**, 301–320 (2005).

256. Yang, S. *et al.* Prioritizing genetic variants in GWAS with lasso using permutation-assisted tuning. *Bioinformatics* **36**, 3811–3817 (2020).

257. Waldmann, P., Mészáros, G., Gredler, B., Fuerst, C. & Sölkner, J. Evaluation of the lasso and the elastic net in genome-wide association studies. *Front. Genet.* **4**, (2013).

258. Papachristou, C., Ober, C. & Abney, M. A LASSO penalized regression approach for genome-wide association analyses using related individuals: application to the Genetic Analysis Workshop 19 simulated data. *BMC Proc.* **10**, 53 (2016).

259. GTEx Consortium *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* **47**, 1091–1098 (2015).

260. Adhikari, K., Mendoza-Revilla, J., Chacón-Duque, J. C., Fuentes-Guajardo, M. & Ruiz-Linares, A. Admixture in Latin America. *Curr. Opin. Genet. Dev.* **41**, 106–114 (2016).

261. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).

262. Bergström, A. *et al.* Insights into human genetic variation and population history from 929 diverse genomes. *Science* **367**, eaay5012 (2020).

263. Becerra-Valdivia, L. & Higham, T. The timing and effect of the earliest human arrivals in North America. *Nature* **584**, 93–97 (2020).

264. Beleza, S. *et al.* The Timing of Pigmentation Lightening in Europeans. *Mol. Biol. Evol.* **30**, 24–35 (2013).

265. GTEx Consortium *et al.* Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nat. Commun.* **9**, 1825 (2018).

266. Carboni, G. L. *et al.* CACNA2D2-mediated apoptosis in NSCLC cells is associated with alterations of the intracellular calcium signaling and disruption of mitochondria membrane integrity. *Oncogene* **22**, 615–626 (2003).

267. Gao, B. *et al.* Functional Properties of a New Voltage-dependent Calcium Channel α2δ Auxiliary Subunit Gene (CACNA2D2). *J. Biol. Chem.* **275**, 12237–12242 (2000).

268. Yoshida, M. & Koike, A. SNPInterForest: A new method for detecting epistatic interactions. *BMC Bioinformatics* **12**, 469 (2011).

269. Pośpiech, E. *et al.* The common occurrence of epistasis in the determination of human pigmentation and its impact on DNA-based pigmentation phenotype prediction. *Forensic Sci. Int. Genet.* **11**, 64–72 (2014).

270. Hore, V. *et al.* Tensor decomposition for multiple-tissue gene expression experiments. *Nat. Genet.* **48**, 1094–1100 (2016).

271. Neer, R. M. The evolutionary significance of vitamin D, skin pigment, and ultraviolet light. *Am. J. Phys. Anthropol.* **43**, 409–416 (1975).

272. Does the Melanin Pigment of Human Skin Have Adaptive Value?: An Essay in Human Ecology and the Evolution of Race. *Q. Rev. Biol.* **36**, 50–63 (1961).

273. Deol, M. S. Racial differences in pigmentation and natural selection. *Ann. Hum. Genet.* **38**, 501–503 (1975).

274. Higuchi, S., Motohashi, Y., Ishibashi, K. & Maeda, T. Influence of eye colors of Caucasians and Asians on suppression of melatonin secretion by light. *Am. J. Physiol.-Regul. Integr. Comp. Physiol.* **292**, R2352–R2356 (2007).

275. Nielsen, R. *et al.* Tracing the peopling of the world through genomics. *Nature* **541**, 302–310 (2017).

276. Moreno-Mayar, J. V. *et al.* Terminal Pleistocene Alaskan genome reveals first founding population of Native Americans. *Nature* **553**, 203–207 (2018).

277. Slominski, R. M., Zmijewski, M. A. & Slominski, A. T. The role of melanin pigment in melanoma. *Exp. Dermatol.* **24**, 258–259 (2015).

278. Sarna, M. *et al.* Cell elasticity is an important indicator of the metastatic phenotype of melanoma cells. *Exp. Dermatol.* **23**, 813–818 (2014).

279. Bonfante, B. *et al.* A GWAS in Latin Americans identifies novel face shape loci, implicating VPS13B and a Denisovan introgressed region in facial variation. *Sci. Adv.* **7**, eabc6160 (2021).

280. Flanagan, N. Pleiotropic effects of the melanocortin 1 receptor (MC1R) gene on human pigmentation. *Hum. Mol. Genet.* **9**, 2531–2537 (2000).

281. Chen, Y. *et al.* The impact of correlations between pigmentation phenotypes and underlying genotypes on genetic prediction of pigmentation traits. *Forensic Sci. Int. Genet.* **50**, 102395 (2021).

282. Alipanahi, B. *et al.* Accurate prediction of multiple pigmentation phenotypes. (2018).

283. Pośpiech, E., Draus-Barini, J., Kupiec, T., Wojas-Pelc, A. & Branicki, W. Gene–gene interactions contribute to eye colour variation in humans. *J. Hum. Genet.* **56**, 447–455 (2011).

284. Ko, S. *et al.* GWAS of longitudinal trajectories at biobank scale. *Am. J. Hum. Genet.* **109**, 433–445 (2022).

285. Fullwiley, D. The Biologistical Construction of Race: `Admixture' Technology and the New Genetic Medicine. *Soc. Stud. Sci.* **38**, 695–735 (2008).

Appendices

A. Article 1: Supplementary information

For references see references section in article 1 (3.1.5).

Supplementary note: Details on the construction of CAN-E/S/H SNP sets

Several We have established three SNP sets for prediction of one categorical (3-levels eye color) and three continuous (Hue, Chroma, Luminance) traits of eye color (CAN-E), of one continuous trait of skin pigmentation (melanin index – CAN-S) and of one categorical trait of hair color (3-levels hair color – CAN-H). The constitution of each one of these SNP sets was based on the genome-wide association results of the corresponding trait previously published in Adhikari et al. (2019). For the specific case of CAN-E, Luminance was retained as the reference traits among the four studied traits. Here below we detail the selection procedure. This procedure was independently applied for each of the three sets.

Pre-selection of SNPs. All SNPs suggestively associated to trait y in the discovery GWAS was pre-selected (p-value <1E-5). In this discovery GWAS, the SNP-phenotype association was independently tested for about 9M SNPs with a linear model involving 8 covariates (sex, age and 6 genetic PCs).

Computation of the predictive power of each SNP taken independently. For the ith pre-selected SNP, we computed its predictive power ρ_i^2 as the proportion of phenotypic variance explained by that SNP, thus $\rho_i^2 = \left(corr(\varepsilon_y, \varepsilon_g)\right)^2$, where ε_y and ε_g are the residuals of the linear regressions of the GWAS covariates ($X \equiv$ sex, age and 6 genetic PCs) respectively on the phenotype (y) and on the genotype of the ith SNP ($g_i \equiv$ the number of counts of the minor allele of each individual j, i.e. $g_{i,j} = \{0,1,2\}$).

Clustering of pre-selected SNPs in linkage disequilibrium (LD). Most often, a signal of association gathers various suggestive and/or significant SNPs from the same LD block. We clustered these SNPs into clumps using PLINK's function --clump with default parameters. This algorithm builds clumps of variants around an index SNP (i.e. the most significant SNP of a certain region) by gathering any variant that is either located max. 250 kb from the index SNP or has a LD R2 with that index SNP greater or equal than 0.50. The number of clumps obtained (N) were: 701 (CAN-E), 1,471 (CAN-S) and 207 (CAN-H).

Selection of one SNP per clump. We retained the SNP with highest predictive power within each clump.

Ranking of SNPs on their conditional predictive power. The SNP with highest predictive power is ranked first. Then, we iteratively compute the conditional predictive power of any SNP that is not yet ranked. This conditional predictive power is the proportion of phenotypic variance explained by a given (unranked) SNP accounting for the usual covariates (X) and the genotypes of the SNPs that are already ranked at this stage (M^k). In a more formal way, we apply the following iterative ranking to all SNPs but the one with highest predictive power.

$for\ k = 2\ \rightarrow N\ \{$

$M^k \equiv$ a matrix containing the genotypes of the $k-1$ SNPs already ranked

$v^k \equiv$ a vector containing the predictive power of the $N-k+1$ unranked SNPs, conditioned on the $k-1$ SNPs already ranked

$for\ i = k\ \rightarrow N\ \{$

$g_i \equiv$ the genotype of the ith unranked SNP

$\widehat{\beta_k^y} \equiv$ the estimates of the linear regression of $(X \quad M^k)$ on y

$\widehat{\beta_k^{g_i}} \equiv$ the estimates of the linear regression of $(X \quad M^k)$ on g_i

$\varepsilon_y = y - (X \quad M^k) \cdot \widehat{\beta_k^y}$

$\varepsilon_g = g_i - (X \quad M^k) \cdot \widehat{\beta_k^{g_i}}$

$v_i^k = \rho_i^2 = \left(corr(\varepsilon_y, \varepsilon_g)\right)^2$

$\}$

$j = argmax(v^k)$

The jth unranked SNP is ranked at rank k

$\}$

Determination of the set of SNPs with highest prediction accuracy. We retain the set Ω_i of SNPs ordered by rank of conditional predictive power ($\Omega_i = \{SNP_1, \cdots, SNP_i\}$) that has the highest prediction accuracy. Hence, the search for the optimal subset among selected SNPs is limited to a total of N subsets including SNPs by decreasing predictive power.

To that end, for each set Ω_i, we compute the prediction accuracy (R^2) of a Scenario 3 (see section *Prediction scenarios tested and approaches used, equation 3*) using RF and following the cross-validation scheme of our study (see Figure 1: the whole data is split in 10 folds and a model is trained on each leave-one-out). Before retaining the set Ω_i that has the highest prediction accuracy, the vector of prediction accuracies is locally averaged with a smoothing kernel in order to avoid potential local maxima. We eventually retain the set Ω_i for which i is minimal among those whose $R2_i \geq \max(R2) - 0.001$(see Supplementary Figure S5).

For the particular case of CAN-H, the trait is categorical (3-levels hair color). The optimal set Ω_i was thus determined for each of the three 1-versus-2 category and the largest one was eventually retained**.**

Supplementary note: Statistical and Machine learning models used for prediction

Multinomial Logistic Regression (MLR) is the generalization of the binary logistic regression to a multiclass problem. Where the probability of an individual belonging to a certain class is

$\Pr(Y_i = 1) = \frac{\exp(b_1.X_i)}{1+\sum_{k=1}^{K-1}\exp(b_k.X_i)}$,

$\Pr(Y_i = 2) = \frac{\exp(b_2.X_i)}{1+\sum_{k=1}^{K-1}\exp(b_k.X_i)}$,

…,

$\Pr(Y_i = (K-1)) = \frac{\exp(b_{K-1}.X_i)}{1+\sum_{k=1}^{K-1}\exp(b_k.X_i)}$, and

$\Pr(Y_i = K) = 1 - \Pr(Y_i = 1) - \Pr(Y_i = 2) - \ldots - \Pr(Y_i = (K-1))$,

b_k is the set of the regression coefficients associated with outcome k, is the set of independent variables and is the categorical phenotype of the individual i. The regression coefficients are estimated using an iterative algorithm and the final category predictions are derived using modal probabilities from MLR. In our analysis, k takes 3 values for both the categorical phenotypes - Blue/gray, Intermediate and Brown/black levels for eye color and blond, dark blond/light brown and brown/black for hair color and the independent variables are the basic covariates, estimated continental genetic ancestries (European and African) and SNPs from different SNP sets.

Linear Regression (LR) is a linear method to model the relationship of a quantitative dependent variable on a set of independent variables. A straight line is fitted across the datapoints that minimizes the error sum of squares and gives us the equation for prediction given the values of the independent variables. For the continuous traits (H, L, C, Melanin Index) we used LR for Scenario 0, 1 and 2 where the variables used each Scenario are already explained earlier.

Random Forest (RF) is an ensemble of multiple decision trees, where the modal voting is chosen as the prediction for a categorical variable or a mean of the predictions from each of the trees for a continuous variable. Biggest advantage of RF is that this method corrects for the tendency of decision trees to overfit using the hold out samples/ out of bag samples in the training process of the individual trees. For our prediction models, we used the R package 'randomForest' and tuned the trees using a grid search over ntree and mtry parameters. 'ntree' determines how many trees should be there in the forest whereas 'mtry' determines how many variables should be sampled before a new split. We took a parameter set (50, 100, 150, … , 750) for ntree and (3 to 40) for mtry. The best combination of ntree and mtry is selected from the validation data and applied on the test data to arrive at the accuracy metric. RF was also used for the selection of SNPs to be included in the CANDELA prediction set.

Extreme Gradient Boosting (XGBoost) is a sequential tree ensemble method where each new tree tries to explain the residual from the previous tree. The first tree of the series is modeled on the entire training data and each subsequent tree tries to make the model better. Thus this method can easily lead to overfitting. So we created a grid of hyperparameters where we took nrounds, early stopping rounds, max depth, eta, colsample by level, colsample by tree, subsample, min child weight and eval metric, alpha, gamma, lambda and 'gbtree' as booster to optimize the models. We took 'nrounds' values from 1 to 50 and 'early stopping rounds' values from (10, 15, 20) mandates the number of exact rounds the method will train if the accuracy does not improve on the validation set. Our objective was set as 'reg:linear' for the quantitative traits and 'multi:softmax' for categorical traits and evaluation metric was set as 'rmse' and 'mlogloss' respectively.

Neural Network (NN) is a circuit of neurons or directed graphs, that is composed of multiple layers of nodes, including one input and one output layer and a variety of hidden layers in between. Each node is connected to the previous or next or both layers through some edges, which carries a weight. The weights are optimized in a way to minimize the overall error rate of the network. We used 3 intermediate layers for all our models and for categorical traits the output layer had 3 neurons, one for each of the levels of both the traits (eye color and hair color) where for quantitative traits the output layer has only one neuron for the final prediction. The networks are compiled by a loss function and an optimization function. The loss functions were mean squared error for continuous traits and categorical cross-entropy for categorical traits and the optimization function was rmsprop. Activation functions also help us to shape the network. We used relu and tanh for continuous traits and additionally used softmax for categorical traits.

Ordinal regression (OrdinalR) is a type of regression analysis used for prediction of ordinal response variables. For ordinal regression we add the covariates and SNP information as our independent variables to the model where we assess the probability of each individual falling into one of the categories for eye color or hair color. Also the phenotypes are properly ordered to ensure there is a gradual change between the levels. Similar to MLR, an iterative procedure estimates the coefficients of the independent variables for the models and the final outcome is estimated as the category with highest probability estimated.

Stepwise Regression (StepR) is an iterative regression model which adds a new variable to the existing model or removes an existing variable from the model based on a prespecified criterion, which in this case is Akaike Information Criterion (AIC). This model gives us the precise set of variables needed to gain the optimum accuracy for the quantitative traits. At each step, this model checks for the best independent variable (from the set of remaining SNPs and covariates which are not already included in the model) to be added to the model only if it brings additional accuracy or removes any variable already added in the model that does not effectively increase the prediction

Supplementary note: Benchmark Prediction Strategies

For the categorical variables, there are certain prediction strategies which are considered as benchmark for comparison of any newly proposed prediction strategy. The common two strategies are:

1. A deterministic strategy: always selecting the largest category by proportion of samples, here called maxP (Maximum Proportion).
2. A random guessing strategy: randomly assigning a category with probability equal to the proportions of each of the levels, here called PropStrat (Proportional strategy).

A discussion of their characteristics, including how they are affected by a skewed trait distribution, is provided in the methods. Here we provide some further details explaining those observations.

For the categorical variables, we create confusion tables which are the cross-tables of the predicted levels of the categorical dependent variables against the observed levels. The accuracy gives us the proportion of the correctly classified individuals from the confusion table.

Accuracy = (True Positives + True Negatives) × 100 / (Total number of observations).

But along with accuracy, sensitivity is another metric that helps us to understand the separate predictability of each of the levels of the categorical variable. We calculate Sensitivity for each level from the confusion table as the percentage of correctly classified individuals for that level.

Sensitivity = (True Positives) × 100 / (True Positives + False Negatives).

Though maxP always dominates over PropStrat in terms of accuracy (proved below), sensitivity for the non-majority categories is better for PropStrat. For maxP, the sensitivity for the non-majority groups always remains at 0 and for the majority group at 1. Where for PropStrat, the sensitivity for each group is the proportion of that category.

We were motivated to add these two strategies as benchmark in the case of categorical variables with skewed distributions, as it is very easy to reach a certain accuracy with a biased strategy like maxP without actually making any effort for predictive modeling, but at the cost of sensitivity. Therefore, to get a sense of additional accuracy gain with the prediction models using the covariates and genetic information over these benchmark strategies, the corresponding accuracy levels are indicated on all the accuracy plots of categorical traits, and their accuracy values are provided in the Supplementary Tables.

Illustration with a binary trait:

Let us consider a classification setup of a binary variable with levels A and B with n_1 and n_2 samples respectively. After prediction, say, we get m_1 and m_2 samples predicted as A and B respectively, out of which n_{11} and n_{22} are identified correctly. The confusion table would look like the following:

Actual / Predicted	A	B	Row total (number of predicted samples)
A	n_{11}	n_{12}	m1
B	n_{21}	n_{22}	m2
Column total (number of samples)	n_1	n_2	Total = N

The accuracy of the prediction approach would be $(n_{11} + n_{22})/N$, where total number of samples $N = m_1 + m_2 = n_1 + n_2$.

Sensitivity will be n_{11}/n_1 and n_{22}/n_2 for A and B respectively.

Accuracy for the two benchmark strategies will be maxP = $\max(n_1/N, n_2/N)$ and propStrat = $(n_1/N)^2 + (n_2/N)^2$.

Sensitivity of the two categories in the case of maxP is $(n_1/n_1, 0/n_2)$ if $n_1 > n_2$ or $(0/n_1, n_2/n_2)$ if $n_1 < n_2$, so (1,0) or (0,1) depending on n_1 and n_2, whereas for propStrat it is $(n_1/N, n_2/N)$.

So for a very skewed distribution of the levels (where either n_1/N or n_2/N becomes very small and close to 0) then accuracy for maxP = $\max(n_1/N, n_2/N)$ also approaches 1. But for the minor level, the sensitivity remains to be 0. For propStrat accuracy also increases with a skewed distribution but it always remains less than that of maxP (proved below), whereas sensitivity always remains above zero.

Accuracy comparison between maxP and propStrat

Let us assume there are k levels of a categorical variable, where p_i denotes the proportion of the i-th level. So $\sum_{i=1}^{k} p_i = 1$.

Now according to the Proportional Strategy (propStrat), the correct guesses for each of the groups are proportional to the weightage of that particular group, since labels are predicted randomly with probabilities equal to group proportions. Hence overall accuracy is:

propStrat = $\sum_{i=1}^{k} p_i^2$

$\leq \sum_{i=1}^{k} p_i \cdot p_{max}$, where $p_{max} = \max_{1 \leq x \leq k} (p_i)$

$= p_{max} \cdot \sum_{i=1}^{k} p_i$

$= p_{max} \cdot 1 = \max_{1 \leq x \leq k} (p_i) = \text{maxP}$

So, we have proved that in terms of accuracy, maxP will always perform equal or better than propStrat. The equality holds when $p_{max} = p_i, for\ all\ i, 1 \leq i \leq k$, that is when all the levels have same frequency and then maxP = propStrat = 1/k.

Supplementary note: Diversity index

Gini-Simpson Index:

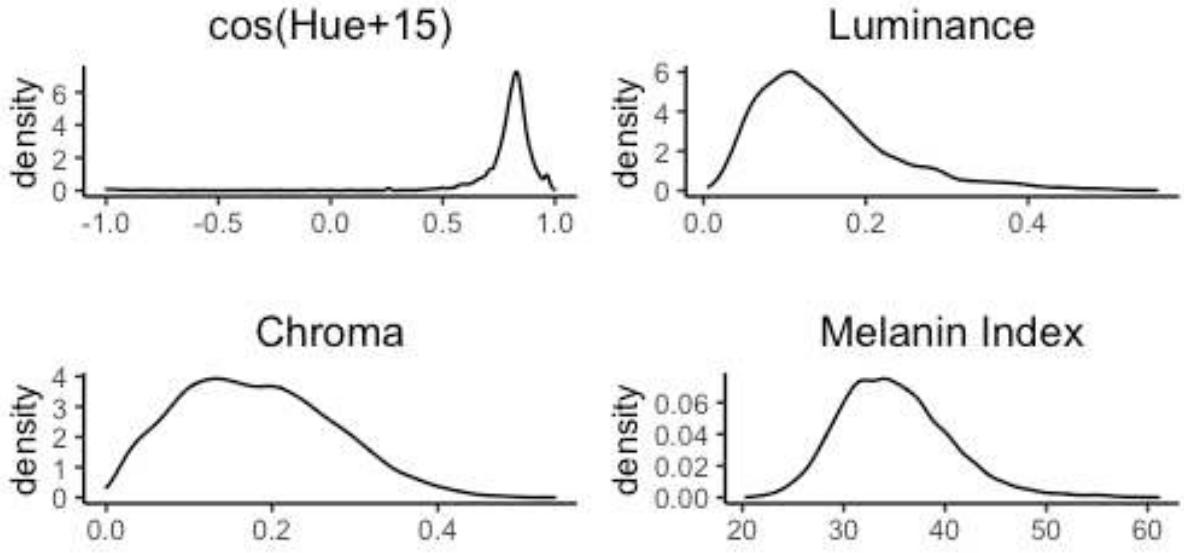
The diversity index (in Table S4) for categorical variables varies from 0 to 1, 0 meaning the data is completely alike and 1 meaning maximum variation within the categories where every observation belongs to a different category.

Gini-Simpson Index = 1 - $\sum_{i=1}^{R} p_i^2$, Where R is the number of levels of the variable and p_i is the sample proportion of level i.

Close observation reveals this diversity index is in fact the completement of the proportional strategy (propStrat) output, they adds up to 1 for each categorical variable. It also explains that higher the diversity, lesser chance of making a correct guess at random:

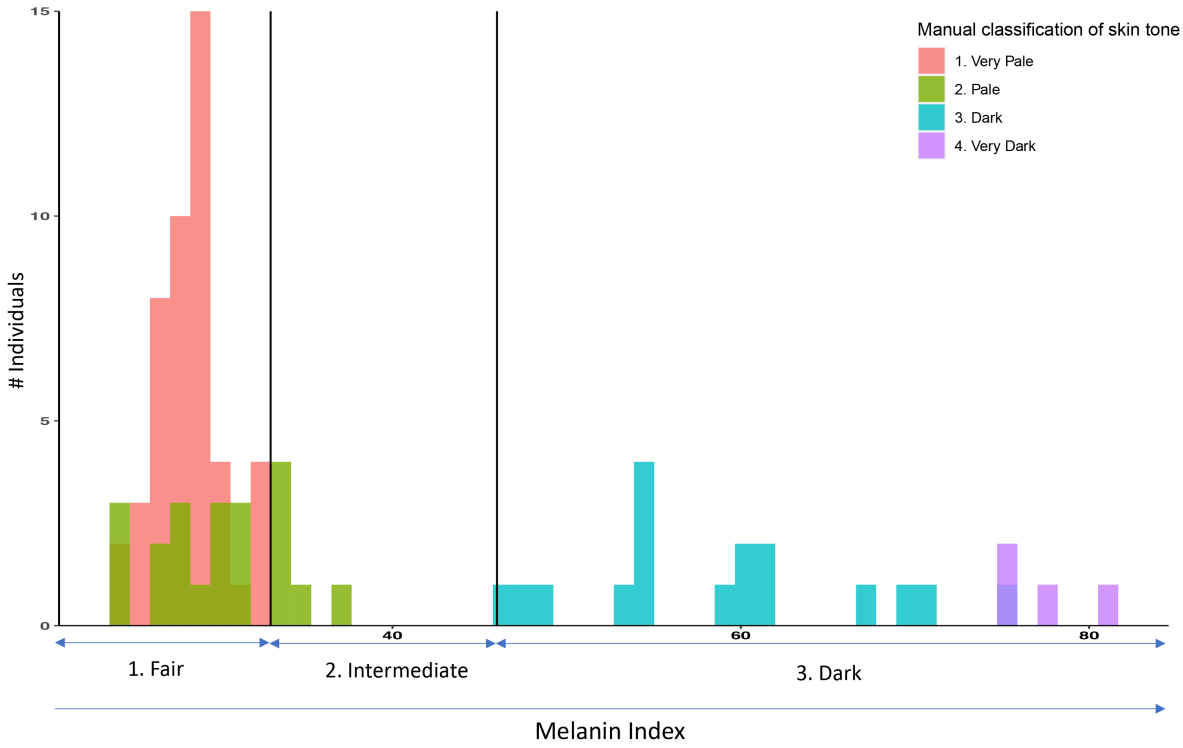
Supplementary figures

Supplementary figure S1: Frequency distribution of the quantitative traits



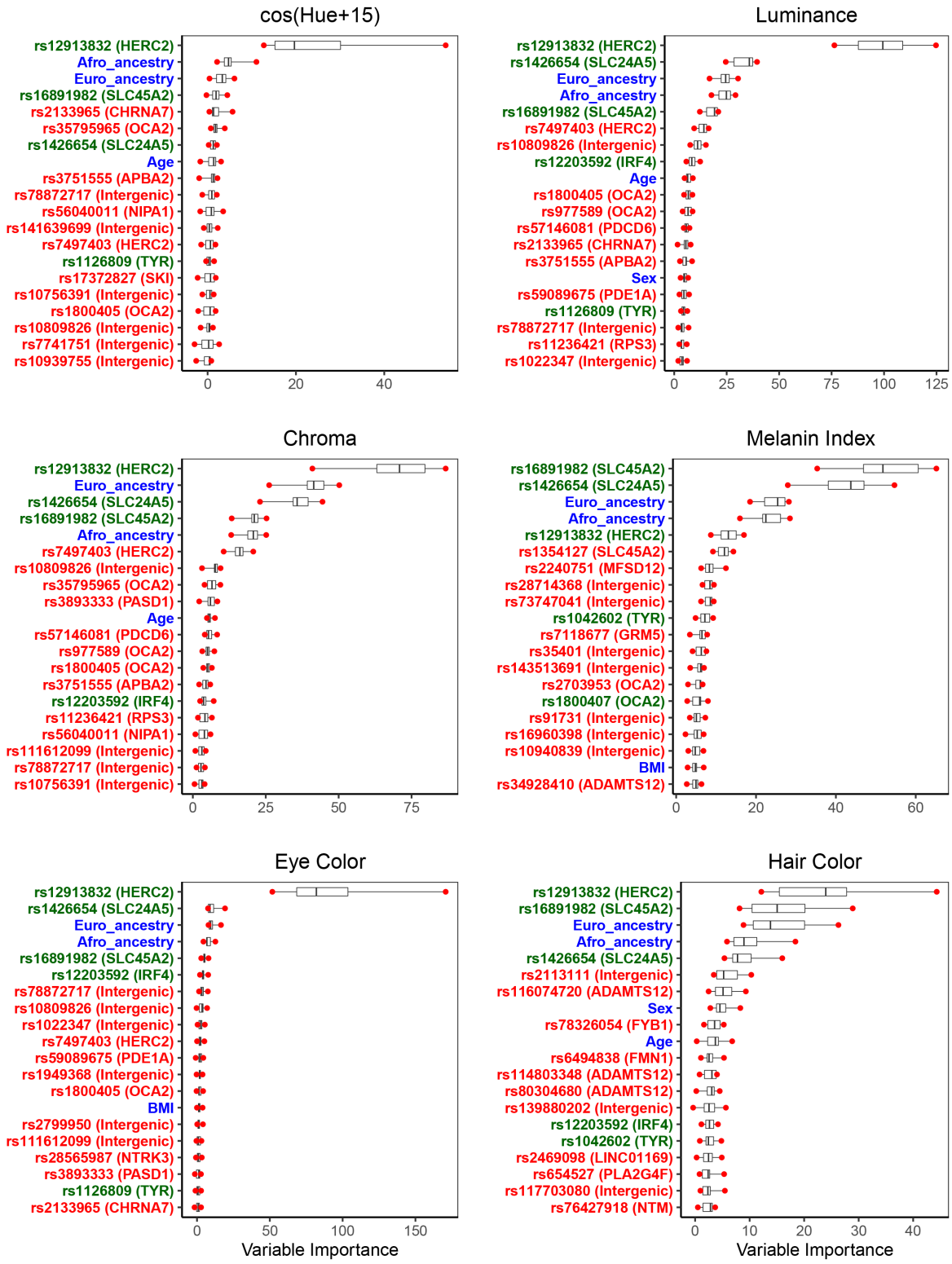
Supplementary figure S2: Melanin index thresholds defining categorical skin colour

We created two ethnicity clusters using 100% European ancestry individuals (70 individuals) and individuals with African ancestry higher than native or European ancestry (23 individuals) with individuals having estimated genetic ancestry and profile pictures available and we manually ranked them in very pale, pale for the European cluster and dark and very dark from African cluster. Both clusters also included cases when the rater failed to classify any individual in any of these groups in particular. We plot the histograms of Melanin Index for all 4 groups of individuals (very pale, pale, dark and very dark). Finally, we identified the threshold of very pale and pale as our Melanin Index cut-off between fair and intermediate categories, also the lowest melanin index of the individuals in the dark category as the cut-off between intermediate and dark categories at 33 and 47 respectively.



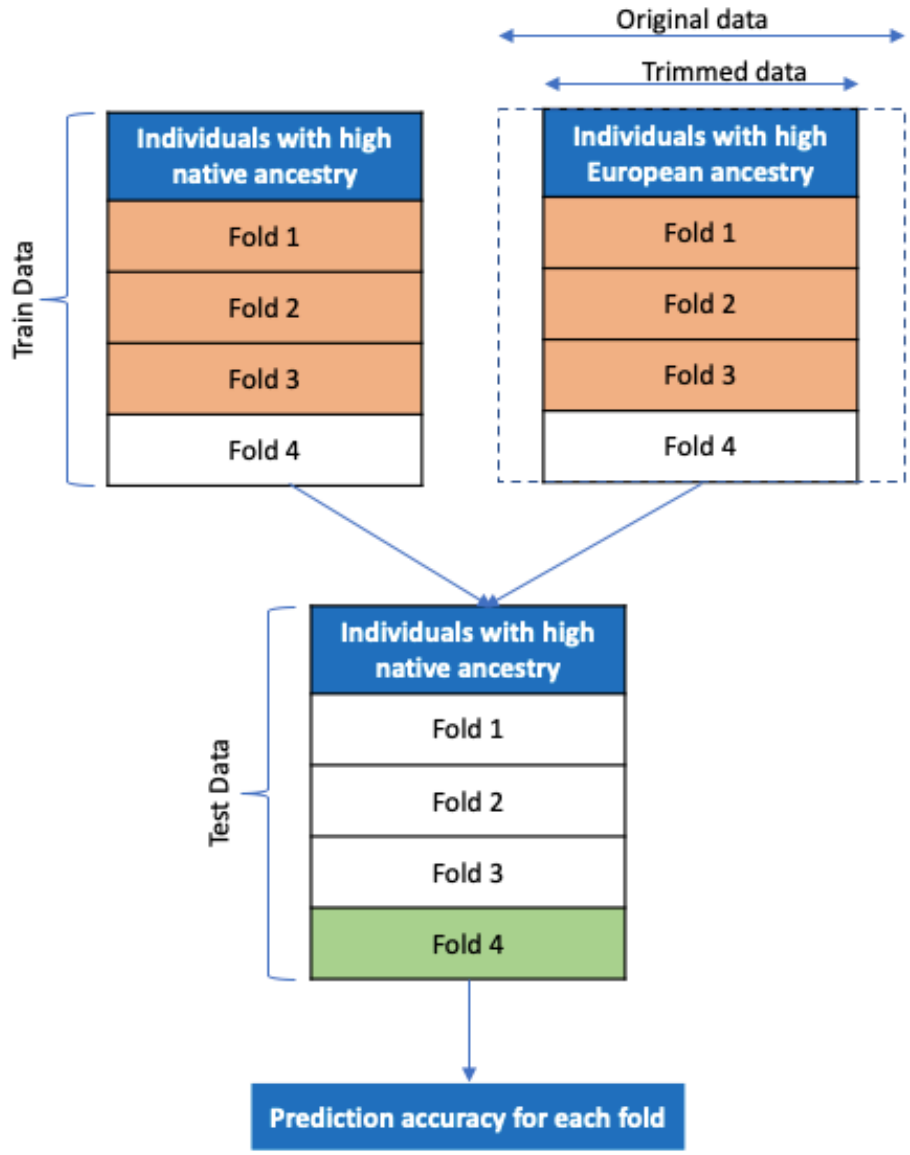
Supplementary figure S3: Variable importance and SNP sharing across sets

We derived the variable importance from the random forest models in Equation 3 using CAN sets and retained the top 20 variables in each case (SNPs and covariates combined). The SNPs which are present also in the HIrisPlex-S assay has been denoted by green, SNPs present only in the CAN sets are denoted by red and covariates are highlighted in blue. We have used increase in mean squared error when a certain variable is dropped as our variable importance metric. The full list of CAN SNPs with their variable importance is provided in Supplementary table S4.

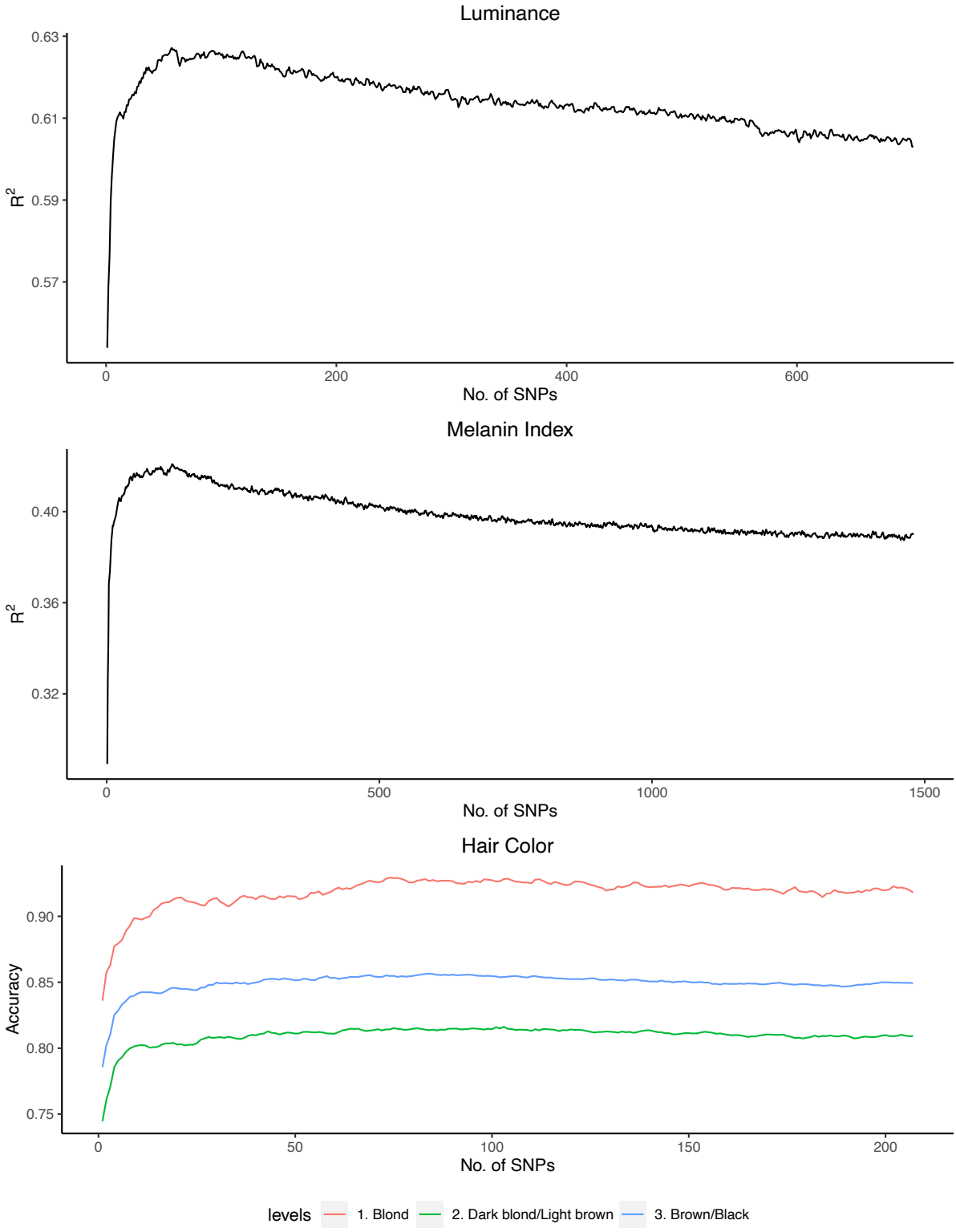

cos(Hue+15)
rs12913832 (HERC2)
Afro_ancestry
Euro_ancestry
rs16891982 (SLC45A2)
rs2133965 (CHRNA7)
rs35795965 (OCA2)
rs1426654 (SLC24A5)
Age
rs3751555 (APBA2)
rs78872717 (Intergenic)
rs56040011 (NIPA1)
rs141639699 (Intergenic)
rs7497403 (HERC2)
rs1126809 (TYR)
rs17372827 (SKI)
rs10756391 (Intergenic)
rs1800405 (OCA2)
rs10809826 (Intergenic)
rs7741751 (Intergenic)
rs10939755 (Intergenic)
0
20
40
Luminance
rs12913832 (HERC2)
rs1426654 (SLC24A5)
Euro_ancestry
Afro_ancestry
rs16891982 (SLC45A2)
rs7497403 (HERC2)
rs10809826 (Intergenic)
rs12203592 (IRF4)
Age
rs1800405 (OCA2)
rs977589 (OCA2)
rs57146081 (PDCD6)
rs2133965 (CHRNA7)
rs3751555 (APBA2)
Sex
rs59089675 (PDE1A)
rs1126809 (TYR)
rs78872717 (Intergenic)
rs11236421 (RPS3)
rs1022347 (Intergenic)
0
25
50
75
100
125
Chroma
rs12913832 (HERC2)
Euro_ancestry
rs1426654 (SLC24A5)
rs16891982 (SLC45A2)
Afro_ancestry
rs7497403 (HERC2)
rs10809826 (Intergenic)
rs35795965 (OCA2)
rs3893333 (PASD1)
Age
rs57146081 (PDCD6)
rs977589 (OCA2)
rs1800405 (OCA2)
rs3751555 (APBA2)
rs12203592 (IRF4)
rs11236421 (RPS3)
rs56040011 (NIPA1)
rs111612099 (Intergenic)
rs78872717 (Intergenic)
rs10756391 (Intergenic)
0
25
50
75
Melanin Index
rs16891982 (SLC45A2)
rs1426654 (SLC24A5)
Euro_ancestry
Afro_ancestry
rs12913832 (HERC2)
rs1354127 (SLC45A2)
rs2240751 (MFSD12)
rs28714368 (Intergenic)
rs73747041 (Intergenic)
rs1042602 (TYR)
rs7118677 (GRM5)
rs35401 (Intergenic)
rs143513691 (Intergenic)
rs2703953 (OCA2)
rs1800407 (OCA2)
rs91731 (Intergenic)
rs16960398 (Intergenic)
rs10940839 (Intergenic)
BMI
rs34928410 (ADAMTS12)
0
20
40
60
Eye Color
rs12913832 (HERC2)
rs1426654 (SLC24A5)
Euro_ancestry
Afro_ancestry
rs16891982 (SLC45A2)
rs12203592 (IRF4)
rs78872717 (Intergenic)
rs10809826 (Intergenic)
rs1022347 (Intergenic)
rs7497403 (HERC2)
rs59089675 (PDE1A)
rs1949368 (Intergenic)
rs1800405 (OCA2)
BMI
rs2799950 (Intergenic)
rs111612099 (Intergenic)
rs28565987 (NTRK3)
rs3893333 (PASD1)
rs1126809 (TYR)
rs2133965 (CHRNA7)
0
50
100
150
Variable Importance
Hair Color
rs12913832 (HERC2)
rs16891982 (SLC45A2)
Euro_ancestry
Afro_ancestry
rs1426654 (SLC24A5)
rs2113111 (Intergenic)
rs116074720 (ADAMTS12)
Sex
rs78326054 (FYB1)
Age
rs6494838 (FMN1)
rs114803348 (ADAMTS12)
rs80304680 (ADAMTS12)
rs139880202 (Intergenic)
rs12203592 (IRF4)
rs1042602 (TYR)
rs2469098 (LINC01169)
rs654527 (PLA2G4F)
rs117703080 (Intergenic)
rs76427918 (NTM)
0
10
20
30
40
Variable Importance

Supplementary figure S4: Modeling approach for ancestry portability.

The high European cohort data is trimmed to match the size of the high native cohort. Then both the datasets are divided in 4 equal folds, we train the random forest models on any 3 folds from one of the ancestry datasets in the training data and test the prediction accuracy on the remaining fold of the native ancestry data. So the models are trained on equal sized data from any of the extreme ancestry cohorts but the test data always remains to be highly native (for example, here fold 1 ,2 and 3 have been used to train the two models on the high native ancestry data and high European ancestry data – orange blocks and both the models are tested on the fold 4 – green bar of the high native ancestry data).



Supplementary figure S5: SNP selection to construct the CAN-E/H/S sets from SM1



Supplementary tables

Supplementary tables can be found in the following location: https://drive.google.com/drive/folders/12xahSRiEon1SvBlB6JqFrQUy_A4CEIbe?usp=sharing

Supplementary table S1: List of all SNPs included in the HIrisPlex-S set
See article1_supplementary_tables_1-9.xlsx, sheet 'Table S1'

Supplementary table S2: List of all SNPs included in the CAN-E/H/S SNP sets
See article1_supplementary_tables_1-9.xlsx, sheet 'Table S2'.

Supplementary table S3: Prediction accuracy numbers
See article1_supplementary _tables_1-9.xlsx, sheet 'Table S3'.

Supplementary table S4: Variable Importance in prediction for the covariates and CAN-E/H/S SNPs included in the pigmentation prediction models
See article1_supplementary _tables_1-9.xlsx, sheet 'Table S4'.

Supplementary table S5: Ancestry pools distribution
See article1_supplementary _tables_1-9.xlsx, sheet 'Table S5'.

Supplementary table S6: Prediction accuracy from reduced models
See article1_supplementary _tables_1-9.xlsx, sheet 'Table S6'.

Supplementary table S7: Variable importance of the top 10, 20, 30 SNPs from Random Forest model for all 6 phenotypes
See article1_supplementary _tables_1-9.xlsx, sheet 'Table S7'.

Supplementary table S8: Prediction accuracy for ancestry pools
See article1_supplementary _tables_1-9.xlsx, sheet 'Table S8'.

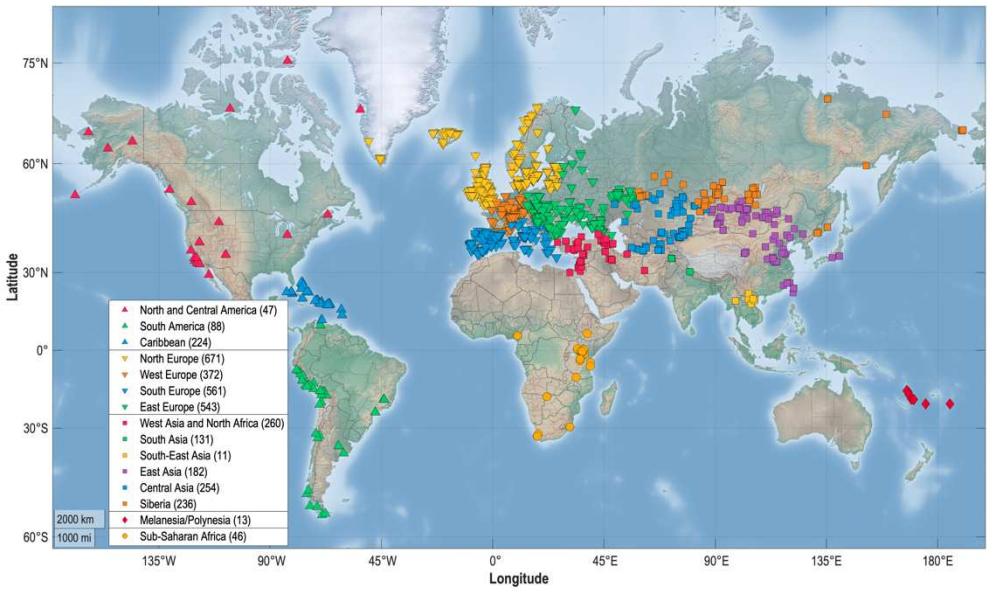
Supplementary table S9: Variability across ancestry pools
See article1_supplementary _tables_1-9.xlsx, sheet 'Table S9'.

B. Article 2: Supplementary information

For references see references section in article 2 (3.2.5).

Supplementary Figures

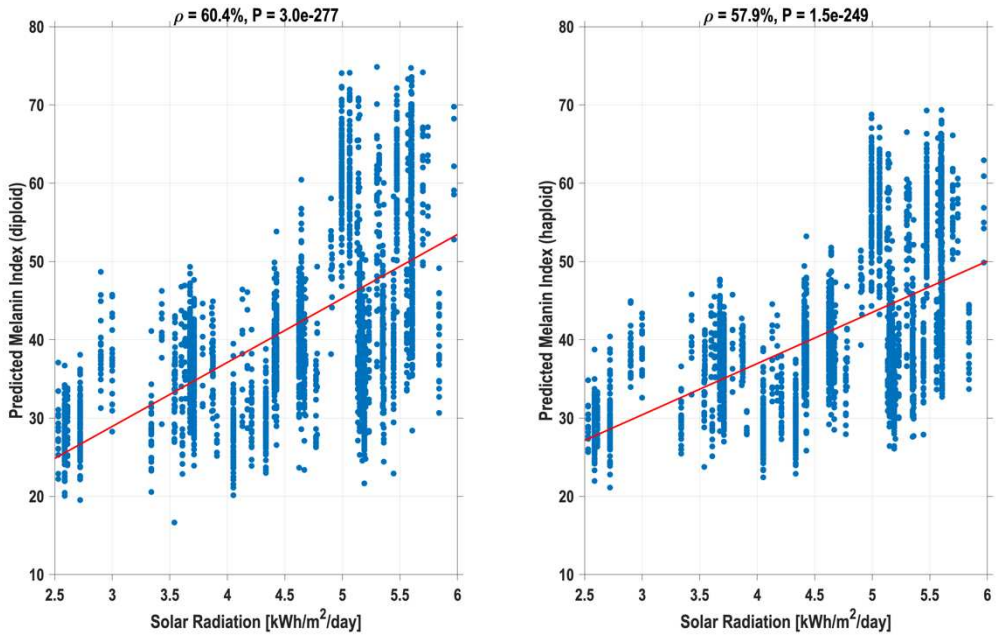
Supplementary Figure S1: Distribution of the ancient samples from HAGDB, the samples were divided into 15 geographic subregions



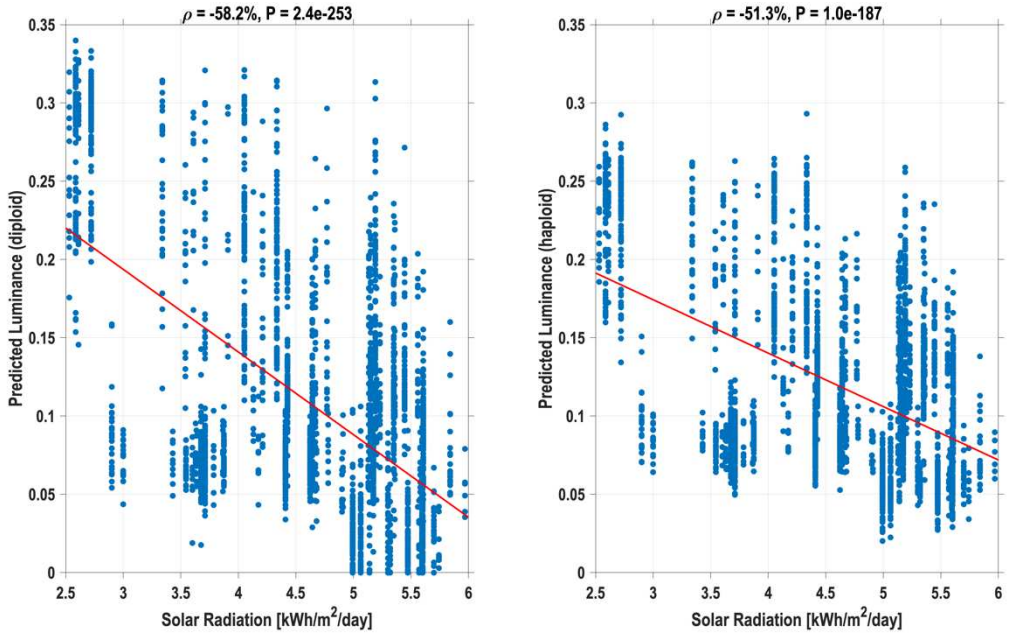
Supplementary Figure S2: Correlation between Solar Radiation and predicted pigmentation phenotypes

Correlation of insolation on a horizontal surface (in kWh/m^2/day) of the sampling location of the 1000GP and CEPH-HGDP studies with the predicted **(A)** skin Melanin Index and **(B)** eye colour Luminance. Prediction of the pigmentation phenotypes were based on both diploid and pseudohaploid genotypes (on left and right panels, respectively).

(A)



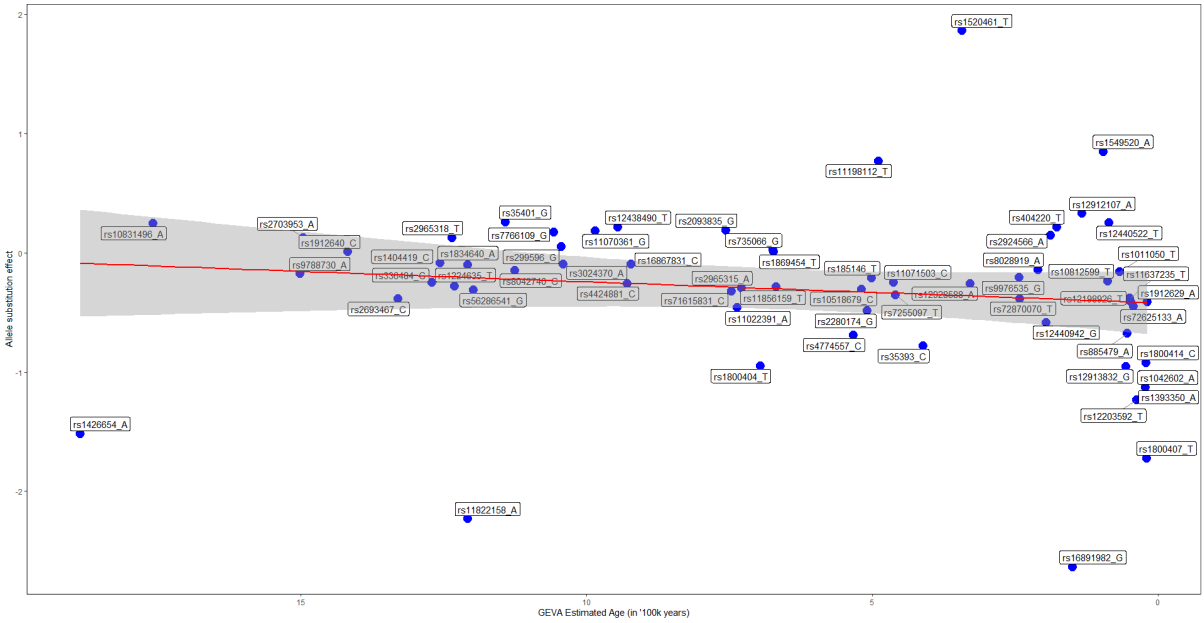
(B)



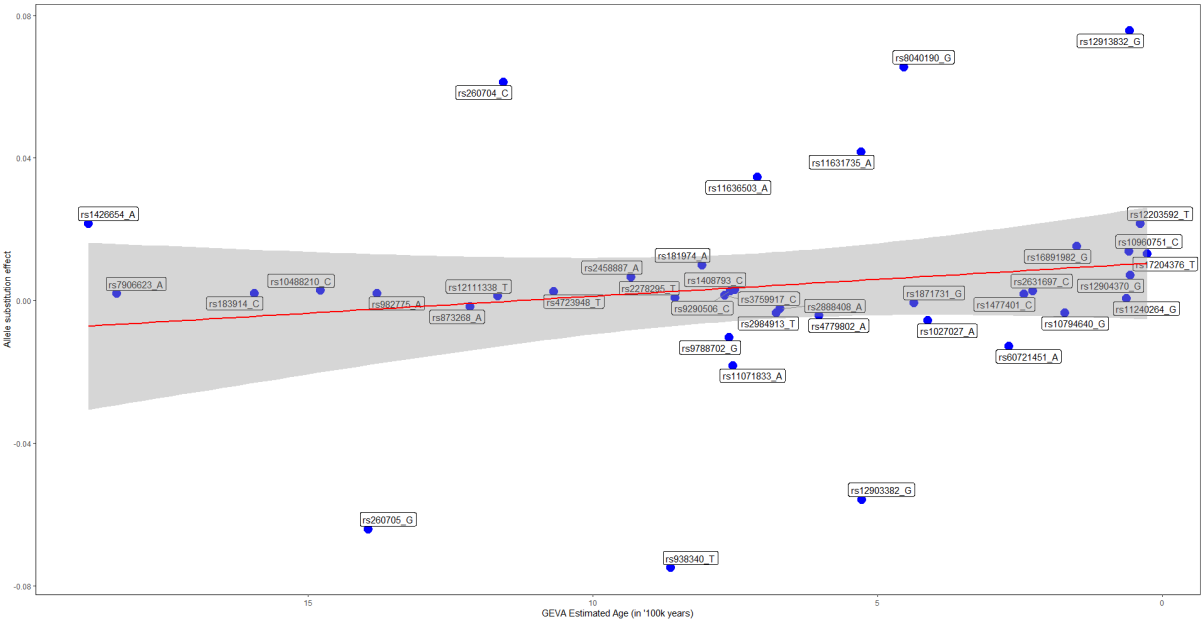
Supplementary Figure S3: Allelic age compared to effect on pigmentation

The allele alteration effects captured from the prediction models trained on CANDELA samples are compared against the allelic age of the genetic variants present within the models. The age estimates obtained from the GEVA (https://human.genome.dating/) portal are regressed on the coefficients of the variants obtained from the linear regression models. **(A)** Skin Melanin Index and **(B)** Eye colour Luminance.

(A)



(B)



Supplementary Notes

Supplementary Note S1: Details on the construction of the SNP sets for MI and Luminance prediction on the ancient samples.

well To predict pigmentation phenotypes (Skin Melanin Index and Eye colour Luminance) in the CANDELA samples, we have already created SNP sets CAN-S (120 SNPs) and CAN-E (56 SNPs) respectively in our previous article[6]. But all the SNPs from the SNPsets were not present in the HAGDB dataset. Only 27 SNPs for MI and 9 SNPs for Luminance were found to be a direct match. But missing out on majority of the SNPs caused a significant drop of the prediction R^2. So we created new SNP sets to predict the phenotypes of the aDNA samples – we took an approach similar to the previous one with the aim to maximise the overlapping of the SNP sets.

First we created a common base of the SNPs from aDNA data merge (HAGDB: https://reich.hms.harvard.edu/datasets) and our CANDELA imputed datasets (11,153,711 autosomal SNPs). Out of 1,150,639 autosomal SNPs from the aDNA genotype data from HAGDB, 995,859 SNPs were found to be common to CANDELA sets.

a. For all the SNPs in the common SNP base we calculated partial correlation for the phenotype and the corresponding SNPs, conditioning on the basic covariates. This partial correlation refers the correlation between the corresponding phenotype and the SNP when the effects of the covariates are removed from both.

b. From our CAN-S set we found 27 SNPs to be directly matched in the common SNP base. For CAN-E, there were 9 common SNPs.

c. For rest of the SNPs in CAN-E/S sets we checked all the SNPs in high LD for the respective SNPs. We created clump sets on our CANDELA data for the SNPs with default PLINK settings. For each of the SNPs not finding a direct match in the common SNP base (i.e. 93 SNPs for CAN-S and 47 SNPs for CAN-E), we checked the clump sets and tried to replace the candidate SNPs from CAN-E/S with a proxy from the corresponding clump set. In case of multiple matches, we chose the one with highest partial correlation mentioned in *b*. We found 13 SNPs from CAN-S and 6 from CAN-E to have an alternative candidate from the clump sets.

d. For the next set of SNPs to be added, we selected a list of SNPs from the higher partial R^2.

e. Now when we got a handful of candidate SNPs we selected only the SNP with the highest partial R^2 and added to 'Sex' in the covariate list and selected the next SNP having the highest partial R^2 from the remaining SNPs when conditioned on the covariate list. Next we added this recent SNP to the covariate list and checked for the next SNP based on the partial R^2 and gradually increasing our covariate list by one at each step.

f. When we finally reached the level of 200 top SNPs, all ordered on the basis of conditional predictive power given the SNPs with higher predictive power already added in the list we calculated the cumulative prediction R^2 sequentially adding one SNP at a time. So in the i-th level the Linear Regression model incorporates the top i SNPs and regressed on the phenotype melanin index. We plotted the cumulative R^2 for all the 200 SNPs.

g. We created our final aDNA CAN-S SNP set based on the top i SNPs from the ordered list leveraging the cumulative plot explained in *g* with the criterion $R^2(i)>=\max(R^2)*0.999$. Finally this new set was obtained with 41 SNPs producing prediction R^2 0.395 where CAN-S with 120 SNPs and basic covariates (Age, Sex, BMI, estimated Euroean and African Ancestry) was producing R^2 0.45 on the same CANDELA samples.

As a final result, out of the top 10 SNPs from CAN-S, we could replicate 8 SNPs in the aDNA CAN-S SNP set. 5 of them were directly present, and for 3 other SNPs, we produced a proxy SNP from the clumped files.

Supplementary Notes S2: Accommodate the ancestries not represented in CANDELA

For the CANDELA consortium study, individuals have been recruited from 5 latin American countries and it has been shown that this demography has been part of recent admixture of European, Native American and African ancestry. This makes the dataset a good starting point for the prediction of the aDNA samples collected across globe. But there are certain ancestries which are underrepresented in the CANDELA study, e.g. East Asian or South Asian ancestries.

To account for the remining ancestries, we selectively added SNPs affecting skin pigmentation in those population. For Luminance, we did not include any additional SNP from the literature, as it has been studied widely that eye colour is a european trait. Outside of the continent, the trait does not show significant variation. European ancestry has the highest contribution in the CANDELA data, so we are well placed to capture eye colour luminance variation in aDNA samples.

We elaborate our procedure below:

1. We gathered all the SNPs found to be associated with skin melanin index (association pvalue less than $5e^{-8}$. We made a combined list of unique SNPs. The SNP list then was pruned to avoid linkage disequilibrium. 1000Genome Phase III data for the corresponding ancestry was taken as the reference to create the LD blocks.
2. For the remaining SNPs, we checked for LD with the SNPs already present in the CANDELA MI SNPset. The SNPs in the new list found to be in LD with any of the existing SNPs were removed, as we can assume their effect will be taken care by the one already present in our SNPset.
3. For the rest of the SNPs we searched our CANDELA data if they are present in the geotype or imputation files. In case there is a mismatch, we also created a LD block with r^2 threshold 0.8 and checked if there can be a proxy SNP available in our CANDELA set.

We finally obtained 9 SNPs for East Asian cohort, 5 SNPs for South Asian cohorts and 11 SNPs for African cohort (22 unique and new SNPs all combined).

When we extended our admixture analysis to allow new ancestries in the CANDELA samples, we found East Asian ancestry to be present as a 4th but insignificant component. We exploited the East Asian and African ancestry present in our data to better train our weighted linear regression model for the smaller admixture components.

Supplementary Notes S3: Ancestry Weighted Linear Regression

As previously discussed, CANDELA samples are 3-way admixed between European, Native American and African ancestries. If we allow for another component in the ADMIXTURE software we see a statistically insignificant East Asian ancestry component. As we are aiming to predict the ancient human samples from potentially varied ancestral candidates and to compensate the lower representation of the African and east Asian ancestry in our CANDELA data, we assigned higher weights to the samples representing the minor ancestries. Our weighting algorithm is elaborated below:

1. We divided the samples based on their European and Native American ancestries – the main ancestry components. We created bins with width 10% of either ancestries and combined bins to create buckets of samples of samples with increasing European ancestry and decreasing native ancestry or vice versa.

So the 10 bins along the European ancestry were 0-10%, 10-20%, 20-30%, … ,90-100%. Similar 10 bins were created for Native American ancestry as well. Finally both ancestries were combined and the as the combined proportion of the ancestries should lie below 100%, the buckets surpassing the combined ancestry threshold were left empty.

2. All the buckets in the process having less than 3 samples were clubbed to the nearest bucket with lower sample size. This was done to avoid extremely high weights given to smaller buckets, as the final weightage was a function of the inverse of the number of samples in a certain bucket.
3. As mentioned before, we encountered a small east Asian component in the CANDELA samples. To magnify their ancestral contribution, we recalculated the ancestry contribution of each of the samples in CANDELA with East Asian ancestry as a 4th component and created bins with ancestry width 10% based on the same. Because of scarcity of samples, we combined all the samples with more than 30% East Asian ancestry in a single bucket and assigned a new weightage based on the inverse of the samples in each of the bins.
4. The maximum of Step2 and Step3 was assigned as the final weightage for the samples.

The weightage method helps us to navigate the insufficient representation from the other ancestries not part of the main continental ancestries in the CANDELA data, as already evidenced in Figure2.

.

Supplementary Tables

Supplementary tables can be found in the following location: https://drive.google.com/drive/folders/12xahSRiEon1SvBlB6JqFrQUy_A4CEIbe?usp=sharing

Supplementary Table S1: CANDELA samples used for the prediction of Ancient samples

See article2_supplementary_tables_1-10.xlsx, sheet 'Supplementary Table S1'.

Supplementary Table S2: Distribution of aDNA samples though space and time

See article2_supplementary_tables_1-10.xlsx, sheet 'Supplementary Table S2'.

Supplementary Table S3: Spatial coordinates, sample size and solar radiation values of modern public datasets

See article2_supplementary_tables_1-10.xlsx, sheet 'Supplementary Table S3'.

Supplementary Table S4: Sampling locations of the aDNA samples

See article2_supplementary_tables_1-10.xlsx, sheet 'Supplementary Table S4'.

Supplementary Table S5: Final aDNA samples retained

See article2_supplementary_tables_1-10.xlsx, sheet 'Supplementary Table S5'.

Supplementary Table S6: Correlations and squared correlations between predicted traits and solar radiations and latitude

See article2_supplementary_tables_1-10.xlsx, sheet 'Supplementary Table S6'.

Supplementary Table S7: Predicted Melanin Index and Eye Colour Luminance of the contemporary samples - both diploid and pseudohaploid modalities

See article2_supplementary_tables_1-10.xlsx, sheet 'Supplementary Table S7'.

Supplementary Table S8: Correlation of the allele alteration effect and the age of the derived allele

See article2_supplementary_tables_1-10.xlsx, sheet 'Supplementary Table S8'.

Supplementary Table S9: Description of the aCAN-E SNP set

See article2_supplementary_tables_1-10.xlsx, sheet 'Supplementary Table S9'.

Supplementary Table S10: Description of the aCAN-S SNP set

See article2_supplementary_tables_1-10.xlsx, sheet 'Supplementary Table S10'.

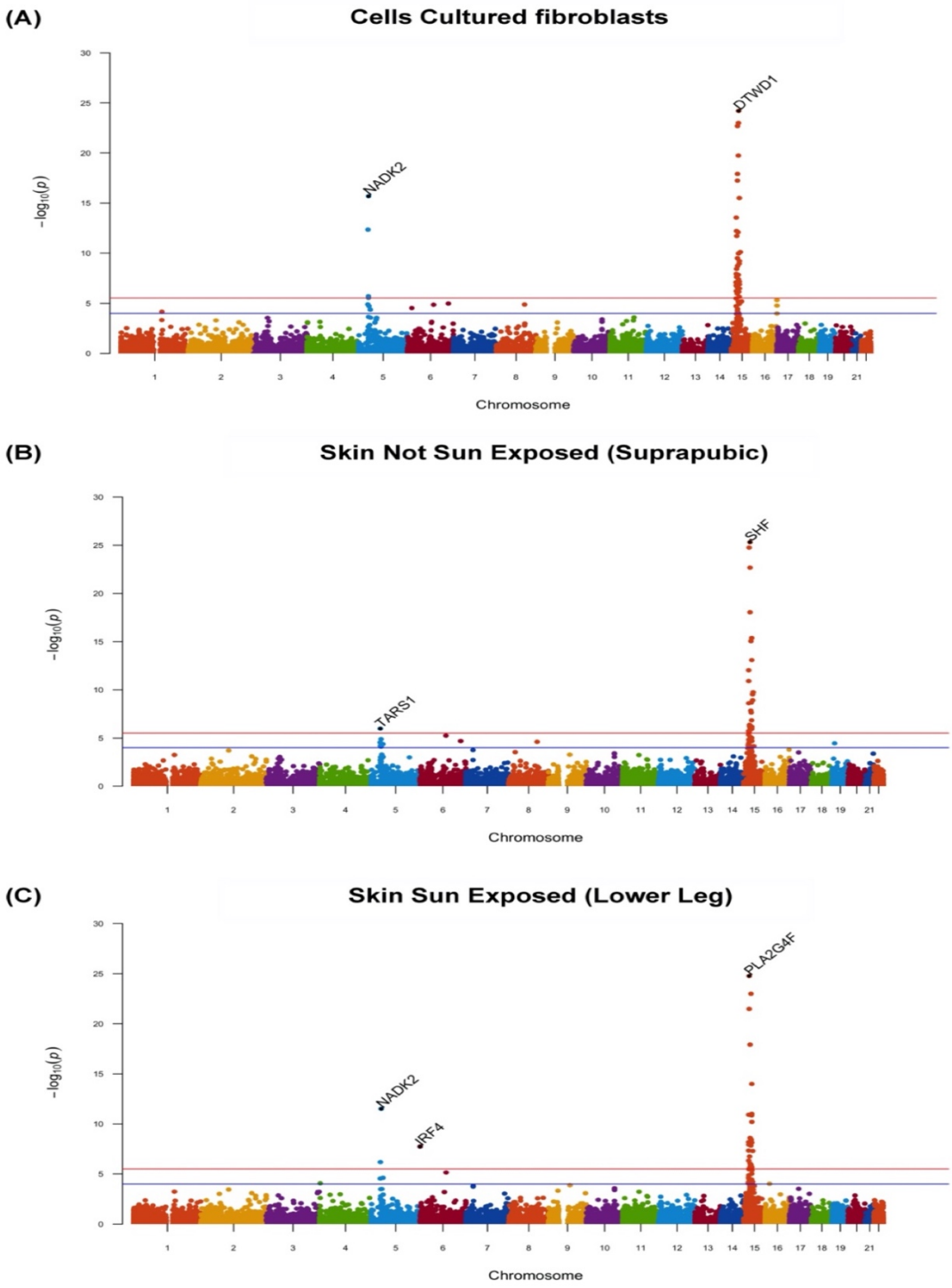
C. Article 3: Supplementary information

For references see references section in article 3 (3.3.5).

Supplementary Figures

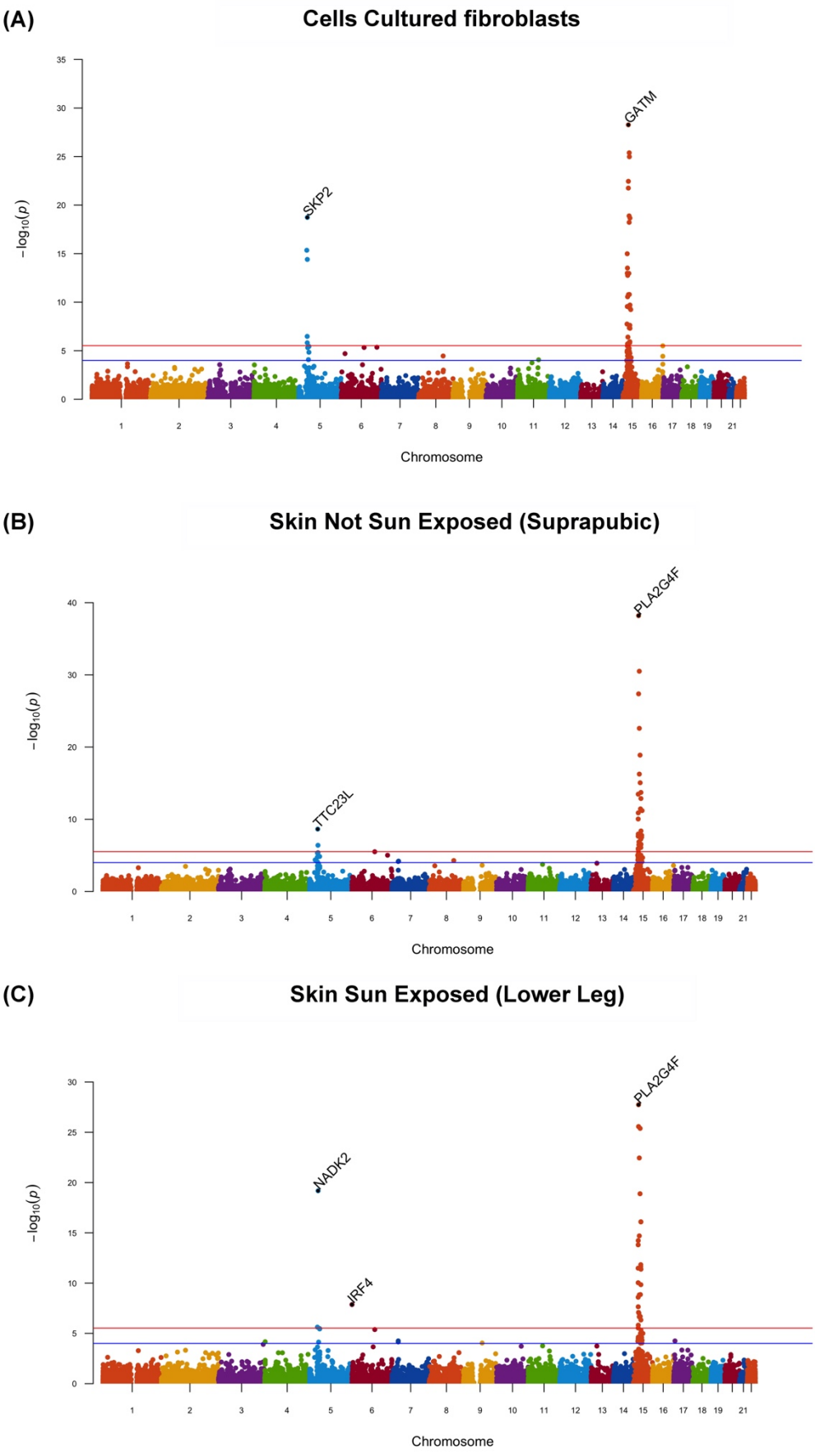
Supplementary Figure S1: Manhattan plots for Skin Melanin Index using PrediXcan

Predicted gene expression association for the skin pigmentation using 6,143 individuals from CANDELA study in tissue **(A)** Cells cultured fibroblasts **(B)** Skin not sun exposed suprapubic **(C**) Skin sun exposed lower leg. Here the assigned position of a gene is the average of the start and end basepairs, according to the GRCh37 mapping. Red and blue lines represent significant and suggestive thresholds respectively


(A)
Cells Cultured fibroblasts
DTWD1
NADK2
$-\log_{10}(p)$
Chromosome
(B)
Skin Not Sun Exposed (Suprapubic)
SHF
TARS1
$-\log_{10}(p)$
Chromosome
(C)
Skin Sun Exposed (Lower Leg)
PLA2G4F
NADK2
IRF4
$-\log_{10}(p)$
Chromosome

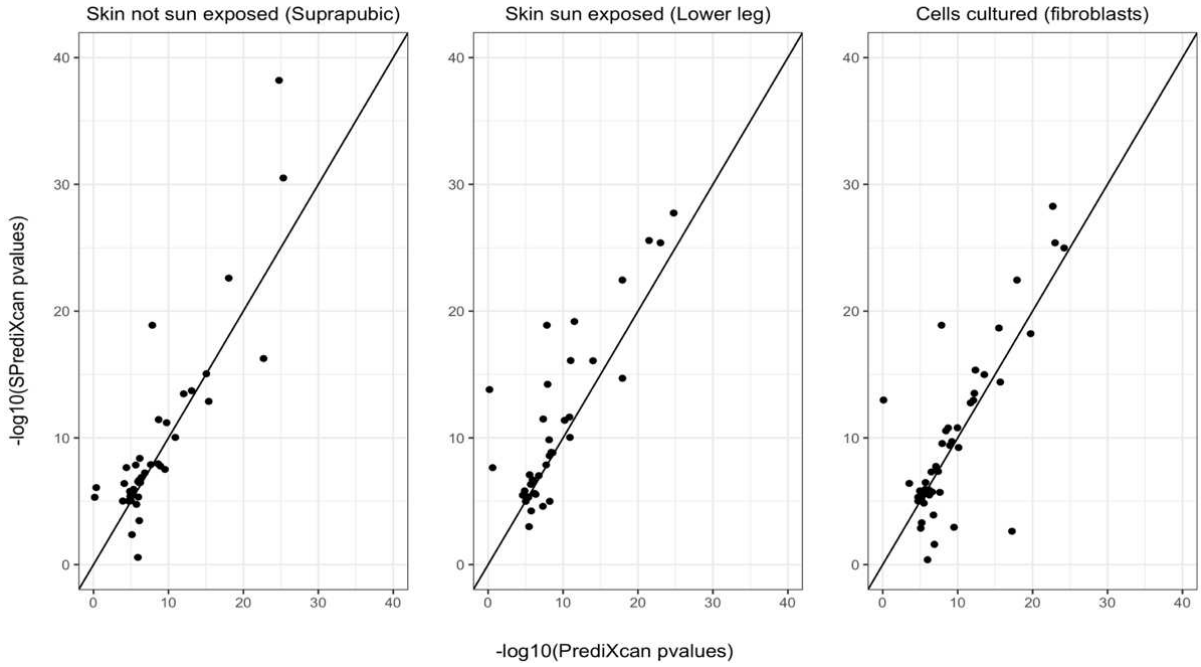
Supplementary Figure S2: Manhattan plots for Skin Melanin Index using S-PrediXcan on the Skin Melanin Index GWAS results on CANDELA samples

Predicted gene expression association for the genes in tissue **(A)** Cells cultured fibroblasts **(B)** Skin not sun exposed suprapubic **(C)** Skin sun exposed lower leg. Here the assigned position of a gene is the average of the start and end basepairs, according to the GRCh37 mapping.


(A)
Cells Cultured fibroblasts
SKP2
GATM
$-\log_{10}(p)$
Chromosome
(B)
Skin Not Sun Exposed (Suprapubic)
TTC23L
PLA2G4F
$-\log_{10}(p)$
Chromosome
(C)
Skin Sun Exposed (Lower Leg)
NADK2
IRF4
PLA2G4F
$-\log_{10}(p)$
Chromosome

Supplementary Figure S3: Comparison of the results from PrediXcan using individual data and S-PrediXcan using GWAS summary data

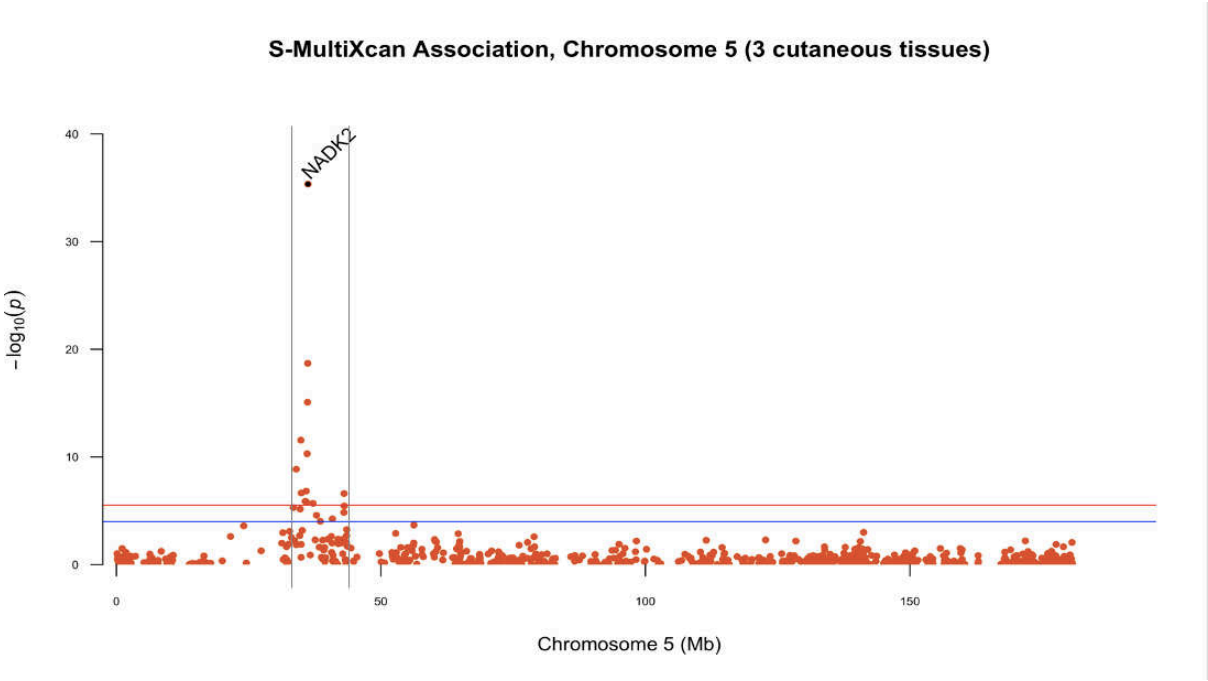
The comparison between the gene association reports from PrediXcan which uses raw genotype-phenotype data from CANDELA and Summary-PrediXcan which uses the GWAS summary data. The previous GWAS results from Adhikari et al 2019 have been used for the summary method. Here only the genes having p-values less than 1e-5 from at least one of the methods have been plotted. The diagonal line represents the X=Y straight line, hence the genes on the line had exact same p-values obtained from both the methods using raw data and summary data



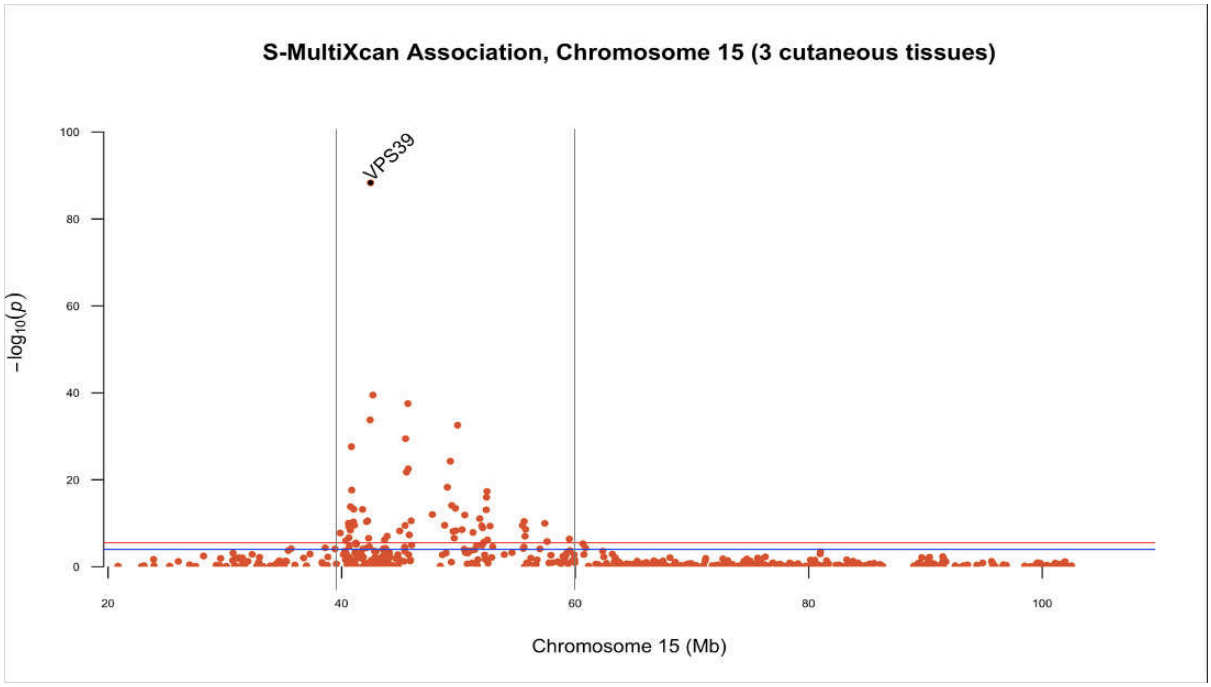
Supplementary Figure S4: Manhattan plot of TWAS association

The regions around the peaks for each of the chromosomes **(A)** 5 and **(B)** 15 has been marked by lines

(A)



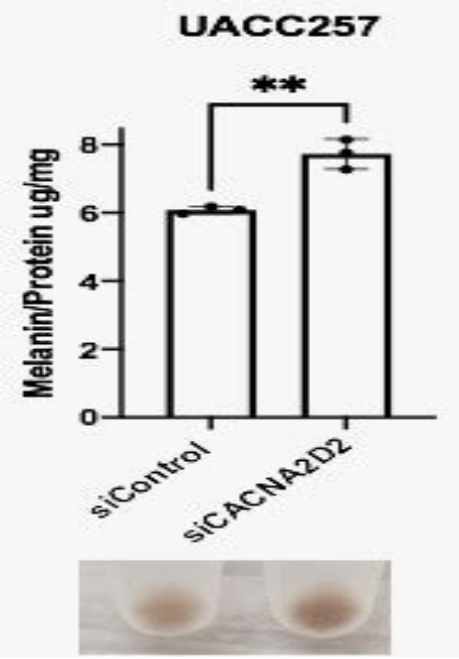
(B)



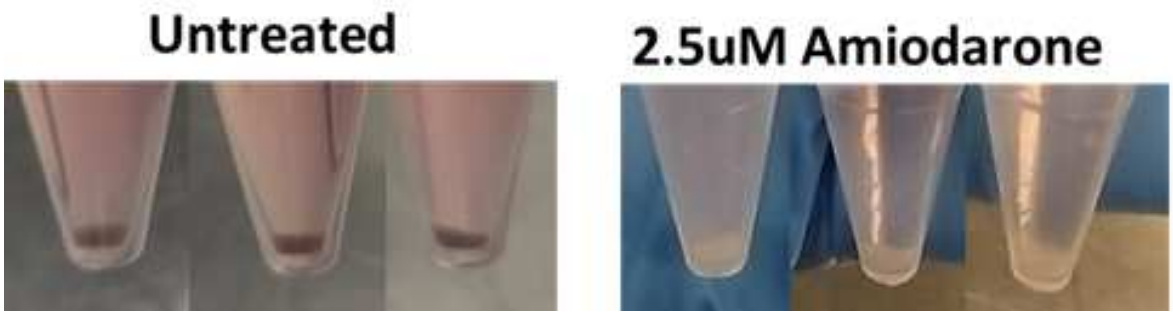
Supplementary Figures S5: Contradictory experimental outcomes of the treatment on CACNA2D2 gene

(A) Silencing the CACNA2D2 gene in the mice produces more pigmentation (B) Skin tissues treated with 2.5uM Amiodarone, which blocks the activity of CACNA2D2, decreases melanin content.

(A)



(B)



Supplementary Tables

Supplementary tables can be found in the following location: https://drive.google.com/drive/folders/12xahSRiEon1SvBlB6JqFrQUy_A4CEIbe?usp=sharing

Supplementary Table S1: PrediXcan results for the tissue Cells cultured fibroblasts

See article3_supplementary_tables_1-9.xlsx, sheet 'Supplementary Table S1'.

Supplementary Table S2: PrediXcan results for the tissue Skin sun exposed lower leg

See article3_supplementary_tables_1-9.xlsx, sheet 'Supplementary Table S2'.

Supplementary Table S3: PrediXcan results for the tissue Skin not sun exposed suprapubic

See article3_supplementary_tables_1-9.xlsx, sheet 'Supplementary Table S3'.

Supplementary Table S4: S-PrediXcan results for the tissue Cells cultured fibroblasts

See article3_supplementary_tables_1-9.xlsx, sheet 'Supplementary Table S4'.

Supplementary Table S5: S-PrediXcan results for the tissue Skin sun exposed lower leg

See article3_supplementary_tables_1-9.xlsx, sheet 'Supplementary Table S5'.

Supplementary Table S6: S-PrediXcan results for the tissue Skin not sun exposed suprapubic

See article3_supplementary_tables_1-9.xlsx, sheet 'Supplementary Table S6'.

Supplementary Table S7: S-MultiXcan results for the tissue Skin sun exposed lower leg, Skin not sun exposed suprapubic and Cells cultured fibroblasts

See article3_supplementary_tables_1-9.xlsx, sheet 'Supplementary Table S7'.

Supplementary Table S8: S-MultiXcan results using 3 tissues from GTEx database compared to individual tissue based S-PrediXcan result

See article3_supplementary_tables_1-9.xlsx, sheet 'Supplementary Table S8'.

Supplementary Table S9: S-MultiXcan results using 3 tissues from GTEx database compared with previous GWAS results

See article3_supplementary_tables_1-9.xlsx, sheet 'Supplementary Table S9'.