ENHANCING OUR GENETIC KNOWLEDGE OF HUMAN IRIS PIGMENTATION AND FACIAL MORPHOLOGY

by

Ryan Eller

A Thesis

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

Department of Biology at IUPUI

Indianapolis, Indiana

December 2019

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Susan Walsh, Chair
School of Science

Dr. Nicolas Berbari
School of Science

Dr. Christopher Lapish
School of Science

Dr. Christine Picard
School of Science

Dr. Randall Roper
School of Science

Approved by:
Dr. Theodore Cummins

To all my teachers, academic, personal, and spiritual

nanos gigantum humeris insidentes

ACKNOWLEDGMENTS

Reflecting on my collegiate career and life in general it’s hard to thank just about everyone who helped me get to where I am now. There are so many people to acknowledge that it becomes a herculean task to try and list them on a single sheet of paper.

With that being said, I would like to acknowledge my parents who provided all the love support, and patience (so much patience) to help me hurdle any challenge life threw my way.

I would like to thank all my teachers from elementary to graduate school for helping me build the foundation of knowledge on which this thesis rests.

I would like to thank all the members of my committee who have helped guide my research and provide invaluable feedback.

I would also like to thank Dr. Randall Roper, Dr. David Skalnik, and Dr. Simon Atkinson for convincing me to come to IUPUI and for providing a lab home for me during my first semester.

My research would certainly not be possible without the support of the IUPUI Biology department, the National Institute of Justice, all the study participants, and too many collaborators to count from places as distant as California, Pennsylvania, Australia, and Belgium.

I also want to acknowledge all the members of the Walsh lab that I’ve gotten to meet. Many people say that work is not work if you love what you do. That is certainly true, but the same can be said about getting to work with such an amazing group of people.

And last but certainly not least I would like to thank Dr. Susan Walsh for letting me join her lab. Going on this journey into the world of bioinformatics has been full of joy (when the computers cooperate) and frustration (when the computers deserve to be tossed out a window), and there is no one else I’d rather be on this journey with.

Truly, thank you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

The biological underpinnings that control iris pigmentation and facial morphology are two areas of research that over the last decade are becoming more thoroughly investigated due to the increased affordability of genotyping and advances in technology allowing for more advanced analysis techniques. Despite the ease of access to the data and the tools required to perform iris pigmentation and facial morphological studies, there are still numerous challenges researchers must overcome when exploring the genetics of these complex phenotypes. Some of these challenges include difficulty in working with the bioinformatic programs designed to analyze genetic associations, the inability to define a phenotype that captures the true nature of these traits, and analysis techniques that fail to model complex gene-gene interactions and their effect on a phenotype or phenotypes of interest.

In this body of work, I attempted to address these challenges by designing a bioinformatic pipeline, *Odyssey*, that bridges the communication gaps between various data preparation programs and the programs that analyze genomic data. With this program, genome-wide association studies (GWAS) could be conducted in a quicker, more efficient, and easier manner. I also redefined iris color as a quantitative measurement of pre-defined color classes. In this way it is possible to define and quantify the unique and intricate mixtures of color, which allows for the identification of known and novel variants that affect individual iris color. I also improved upon current prediction models by developing a neural network model capable of predicting a quantitative output to four pre-defined classes; blue/grey, light brown (hazel), perceived green, and dark brown. I examined the effects of defining a simple facial morphology phenotype that more accurately captures the lower face and jaw shape. I then analyzed this phenotype via a GWAS and found several novel variants that may be associated with a square and diamond shaped face. Lastly, I demonstrated that structural equation modeling can be used in combination with traditional GWAS to examine interactions amongst associated variants, which unearths potential biological relationships that impact the multifaceted phenotype of facial morphology.

CHAPTER 1. INTRODUCTION

An individual's unique physical appearance comprised of body morphology and pigmentation is one of the many aspects that defines one's identity. It is therefore no surprise that much research has been dedicated to studying why humans have such an appearance. Pigmentation, specifically iris pigmentation, is one phenotype that has drawn a lot of attention due to its forensic[1] and anthropologic[2] applications. Numerous iris color studies thus far have focused on understanding categorical iris color[3–6], and they have largely succeeded in identifying genes such as *HERC2*, *OCA2*, and *TYR* that are highly influential in determining blue vs brown eye color. However, while these color extremes have been studied extensively, and are capable of being predicted with a high degree of accuracy[7–10], there is a notable gap in our genetic understanding of the intermediary colors such as green and hazel[11–13]. Thus, a better comprehension of what genetically causes intermediate iris color will not only fill the genetic gaps, but it will also lead to more complete prediction models.

Human facial morphology is another phenotypic area of study that has been researched extensively in disease-based phenotypes of the face such as cleft lip and palate[14–18] and Down syndrome[19–23], but notably less in non-diseased facial variation. Over the past several years, more interest and research has been focused toward understanding these minor variations of the face. However, unlike the simple categorical phenotyping that accompanied early iris phenotyping, the majority of facial research has been based on Euclidean distances between landmarks[24–27]. More recently, as technology allows us to better reconstruct a 3D face from hand-held cameras, 3D modeling has helped improve facial phenotyping allowing for more advanced analyses to capture holistic views of facial phenotypes instead of simple distance measurements. One of the more successful adaptations of using 3D facial meshes to model and analyze facial morphology comes from Claes et al., 2018, where they were able to identify 38 loci related to facial morphology[28,29]. It is also important to note that these facial genome-wide association studies (GWAS), and GWAS in general, are adapted to specifically analyze independent variant effects on a phenotype. Because of this, potential variant interactions or pathway contributions are often left unexplored.

Therefore, the research presented here attempts to fill in the gaps pertaining to human iris color and facial morphology. I also present a tool that may help other researchers accelerate their own work by making bioinformatic tools more easily accessible to those without a computer programming background.

1.1 SQL Servers

Often researchers find it necessary to work with and archive vast amounts of data for research. From storing small output files from an analysis to large images or genomic data, the success of research is partially dependent upon data management and organization. However, there are many methods of storing data from a folder of excel or word documents to databases created via Microsoft Access to large production servers like SQL Server (Microsoft, Redmond, WA), MySQL (Oracle, Redwood City, CA), and Oracle Database (Oracle, Redwood City, CA). The choice of implementation method is often tied to the analysis that is being conducted, the level of experience of the user, and how much effort the user wishes to setup the data repository. Storing data via Microsoft Excel and Word are quickly implemented but are not designed to store large amounts of data since they are not as easily managed. Conversely, production servers take longer to setup, but once they are established allow for easy manipulation of terabyte size amounts of data. As production servers are designed for small to large corporations, they also come with a suite of analysis tools that allow for quick in-place (i.e. the data does not have to be extracted first) dataset analysis. However, despite these options, problems arise when scientists exceed the limits of easy-to-implement solutions. The problems that result from attempting to store large amounts of data not maintained by a database can range from disorganized data that cannot be extracted quickly to loss of data due to file corruption. The ability to maintain organizational schemes between lab personnel also grows to be a challenge when a server-managed database is not being used. The three main advantages of using SQL server are 1) organization, 2) ease of manipulation, and 3) security. Relational databases are especially useful since the lab collects many types of data (i.e. spectrophotometry, questionnaires, and pictures) that relate to a certain study participant. Since relational databases are setup via numerous primary and secondary keys and allow for many data entry rules, this forces data being entered to be organized. For example, rules exist so that no two people may have the same questionnaire number or that number fields cannot contain non-numerical data. Second, SQL servers are indexed, which allows for their fast lookup speeds. For example, with SQL it is possible to extract a single picture out of a group of 21,000 pictures in less than 10 seconds. Since SQL Servers are designed for enterprise, they have enterprise level security features, which helps keeps sensitive subject data secure. Another benefit of the lab's SQL server is that it is easily portable allowing for fast and easy data sharing between collaborators.

1.2 Genome Wide Association Studies and META Analyses

Genome-wide-association studies or GWAS are a form of observational study that attempts to draw correlations between a phenotype of interest and genotypic variants that are normally scattered throughout the genome (i.e. genome wide). To perform a GWAS, a researcher must have access to both phenotypic data and its corresponding genotypic data. Prior to analysis however, multiple factors must be investigated in order to control for a variety of confounders. First, a researcher may want to increase his or her genomic coverage since chances are the genotypic data is genotyped on a SNP array due to cost restraints. To increase coverage, the computationally demanding task of phasing and imputing the genotyped data must be performed in tandem with a sequenced reference genome such as the 1000 Genomes Project[30]. Through this process the non-genotyped "gaps" in the researcher's SNP array will be filled with predicted imputed data. Imputation quality metrics such as IMPUTE's[31,32] INFO or Minimach's[33] R^2 must also be consulted in order to remove any poorly imputed variants. Once the phenotypes and genotypes have been assembled, the researcher must address a variety of quality control (QC) measures prior to GWAS. One confounder is ancestry, which can be assessed visually through a principal component analysis (PCA) or non-visually via programs, such as Eigensoft[34], that automatically removes ancestral outliers. As one of the GWAS assumptions is having a group of independent non-related individuals, familial relatedness (e.g. mother-daughter relationships) must also be assessed and corrected for. Other QC steps, such as assessing for genotype and individual missingness rates, Hardy-Weinberg Equilibrium (HWE), excessive heterozygosity rates, and minor allele frequencies (MAF), should also be analyzed prior to performing GWAS. However, once all confounders have been addressed, a basic linear or logistic-based regression GWAS can be performed depending on whether the phenotype of interest is quantitative or categorical, respectively. GWAS results are often displayed in the form of Manhattan plots (See Figure 1.1) which displays the variant being tested (x-axis by chromosome and position) versus the reported p-value (y-axis on a $-\log_{10}$ scale). Since the test analyzes anywhere from several hundred thousand to millions of variants at one time, multiple testing corrections, such as the family wise error rate (FWER) correction of Bonferroni[35] (visualized by the black line occurring at -log$_{10}$ ~ 7.5 in Figure 1.1), or the more lenient false discovery rate (FDR) correction of Benjamini-Hochberg[36] (visualized as the fainter dashed line occurring at -log$_{10}$ ~ 5 in Figure 1.1) must be performed in order to reduce the number of false positives. Alternatively, more advanced GWAS analyses such as the linear-mixed model (LMM), which natively creates a genetic relatedness model that corrects for cryptic and familiar relatedness prior to running a GWAS may also be performed. Lastly, META analyses may be performed in cases where a GWAS is not powerful enough to detect associated variants and the researcher wishes to combine GWAS together. However, what is unique about META analyses is that it only requires a significance value (e.g. a p-value), an effect size (e.g. an odds-ratio or beta), a sample size, and a variant ID. In this way, researchers who do not wish, or cannot share their raw genotypic or phenotypic data, can safely share their analysis output with a fellow researcher to perform the combined analysis. Luckily a variety of free bioinformatic tools exist that can perform a variety of QC, ancestry and relatedness analyses, and eventually the GWAS itself, such as PLINK[37,38], GCTA[39], and GEMMA[40]. However, these tools often have their own unique syntax, as well as their own input and output requirements. Simply put, bioinformatic tools do not make it easy for a researcher to output a file in one program and enter it into another for a follow-up analysis.

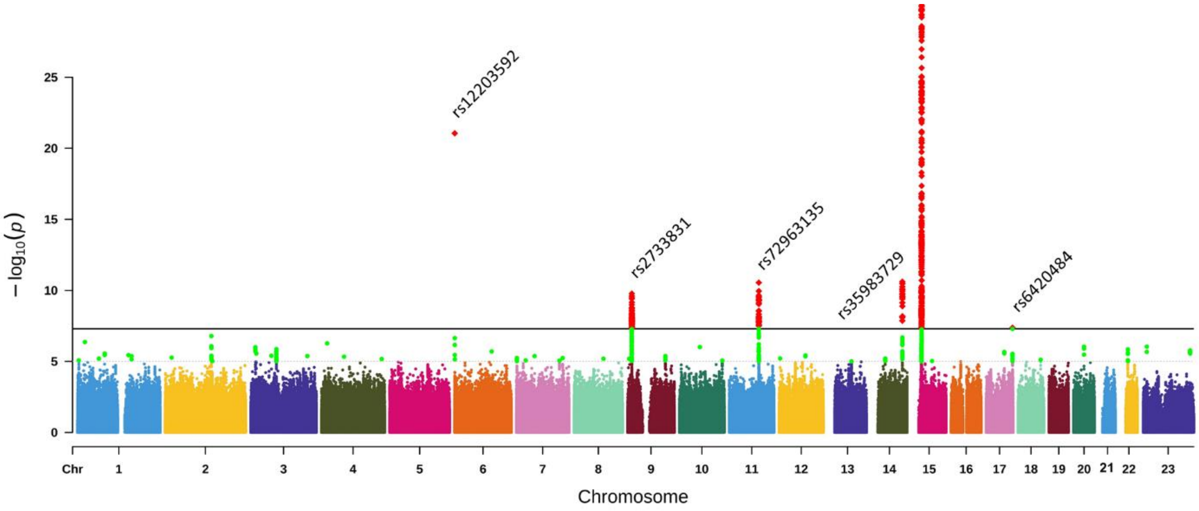


Figure 1.1. An Example Manhattan Plot

1.3 Iris Pigmentation – A Biological Background

Human pigmentation is highly visible and one of the most variable traits seen between individuals and populations. Much of this variation is due to an individual's geography and ancestry, since pigmentation has largely been seen to serve as a protection mechanism against UV damage, most notable with skin color[41] and even iris color[42]. As pigmentation is highly integrated into various biological, anthropological, and forensic-based applications, understanding its complex composition on a molecular and genetic level is very useful.

On the molecular level, iris color is the result of the pigment melanin[43]. Melanin pigments are produced within specialized organelles called melanosomes (See Figure 1.2, which is based on the figures from Wasmeier et al., 2008 and Scherer et al., 2010[43,44]), which are responsible for synthesizing, storing, and transporting the pigment via dendrites[43]. The final destination of these pigments varies depending on the location, with melanosomes being deposited into the iris stroma or the retinal pigment epithelial of the eyes.

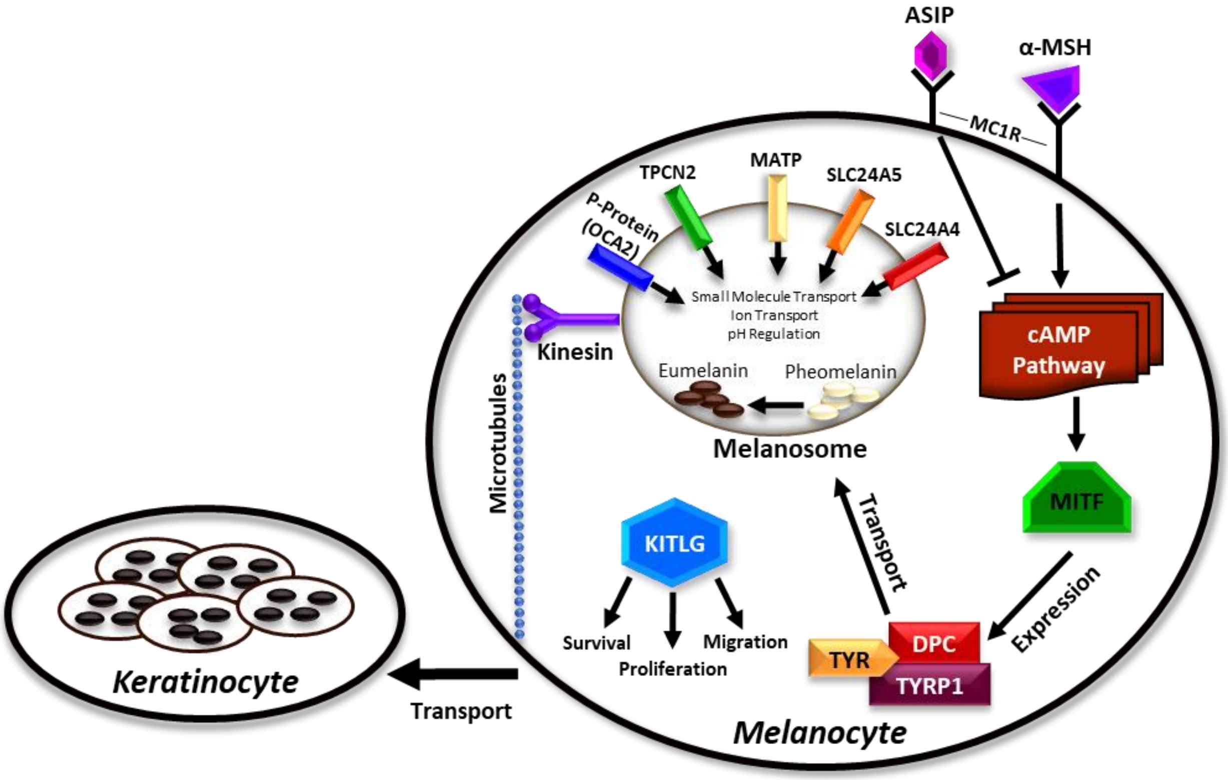


Figure 1.2. Melanin Production Pathway

There are also two different types of melanin; eumelanin, a dark brown polymer and pheomelanin, a red yellow polymer. Both pigments are derivates of the amino acid tyrosine with both forming the intermediate DOPA-quinone via the enzyme Tyrosinase[45,46]. At this stage, if cysteine is present then pheomelanin is formed; if not, eumelanin is produced (See Figure 1.3, which is based on the figure from Horrell et al., 2015[47]).

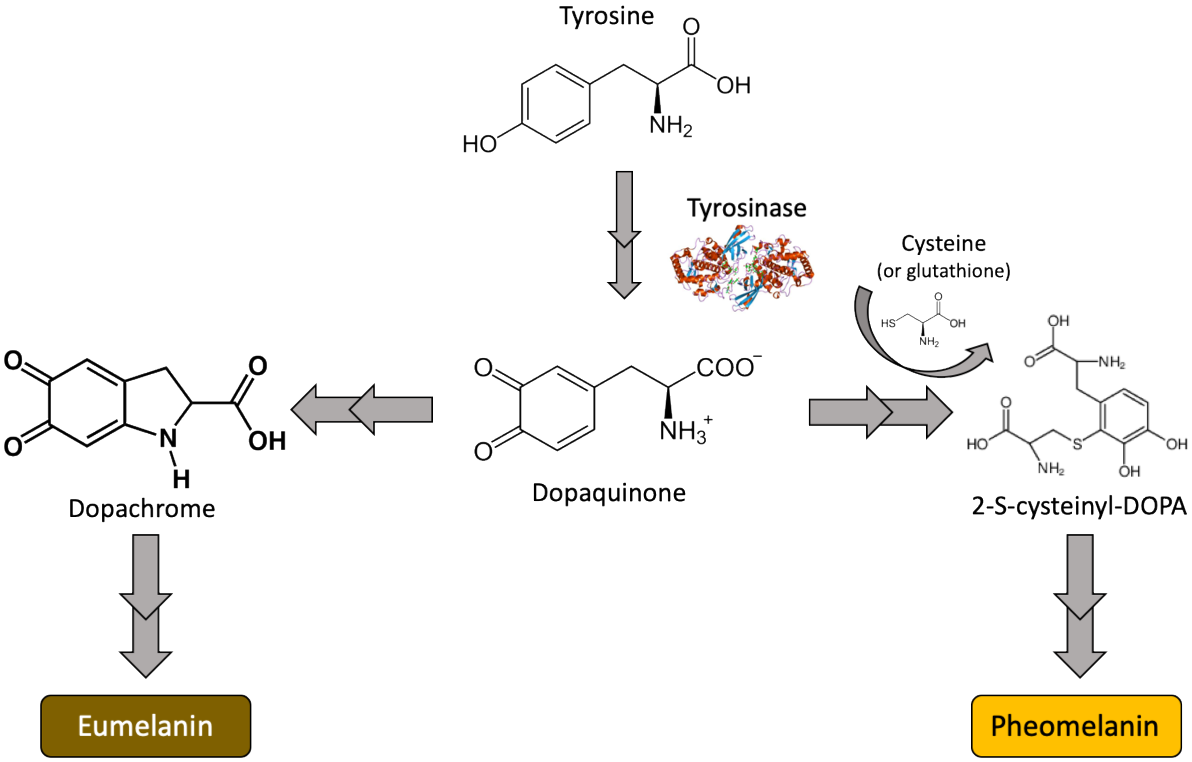


Figure 1.3. Chemical Differences Between Eumelanin and Pheomelanin

Genetically, a number of factors are responsible for the production, transport, and distribution of pigment, which ultimately give rise to pigmentation. In general, developmental genes such as stem cell factor (*SCF*) and its receptor *KIT*, endothelin 3 (*ET3*) and its endothelin B receptor (*EDNBR*), and several cadherin genes are responsible for the development of melanocytes[44,48]. Melanocortin receptors, specifically MC1R, are located on the surface of melanosomes and either increase the production of melanin, if exposed to the alpha-melanoctye-stimulating hormone (alpha-MSH) or decrease production if exposed to its ASIP antagonist[44,46]. This in turn either stimulates or suppressed the microphthalmia-associated transcription factor (*MITF*)[49], which either induces or silences the production of tyrosinase (TYR), the protein responsible for synthesizing melanin[50], and tyrosinase related proteins (TRYP). Other intermembrane proteins that are involved in melanosome survivability include SLC24A5, OCA2, and MATP, whose roles include maintaining the stability of the melanosome through regulating pH and transporting small molecules (See Figure 1.2)[44].

In the iris, several cells are responsible for the eyes visual appearance; uveal melanocytes derived from the neural crest, melanocytes in the iris pigment epithelium (IPE) derived from the neuroectoderm, and clump cells, which are thought to be of histiocytic origin[51–53] (See Figure 1.4[54]). While all cells contribute to giving eyes their unique color, the uveal melanocytes play the largest role. Uveal melanocytes, located primarily in the anterior border layer of the stroma[55], deposit pigments into the iris stroma (See Figure 1.5, which is a cartoon based on the light micrograph found in *Clinical Anatomy and Physiology of the Visual System*[56]). In comparison to melanocytes in the IPE, uveal melanocytes produce less pigment, and also produce both eumelanin and pheomelanin, whereas IPE melanocytes produce mainly eumelanin[52,57]. More importantly, the amount and type of pigment produced in the iris stroma varies, with higher melanin deposition and a larger eumelanin to pheomelanin ratio giving rise to darker color irises, and lower melanin content and/or smaller ratios yielding lighter colors[57]. Playing a more minor role in the coloration of the iris is the iris pigment epithelium, located on the posterior surface of the iris[52]. This epithelium is heavily pigmented with mostly eumelanin content, regardless of iris color. The IPE serves primarily as a protector of the retina by absorbing excess light, thereby minimizing UV damage through eumelanin's ability to reduce free radicals and, in particular, reactive oxygen species[58]. While the IPE always contains a large amount of eumelanin regardless of visible iris color, the notable exception is in individuals affected by Oculocutaneous albinism, in which the IPE is devoid of pigment, sometimes causing a pink tint to the iris from the blood vessels in the retina[59]. Lastly, clump cells (see Figure 1.5), which do not produce pigment, but rather phagocytose melanin, have been hypothesized to scavenge free pigment within the iris and are normally located near the sphincter muscle anterior of the IPE[55,56]. These cells nominally affect iris color, but can still be seen as clumps of color near the pupillary zone (i.e. the part of the iris that is nearest the pupil) or where the iris attaches to the ciliary body (i.e. near the periphery of the iris)[60].

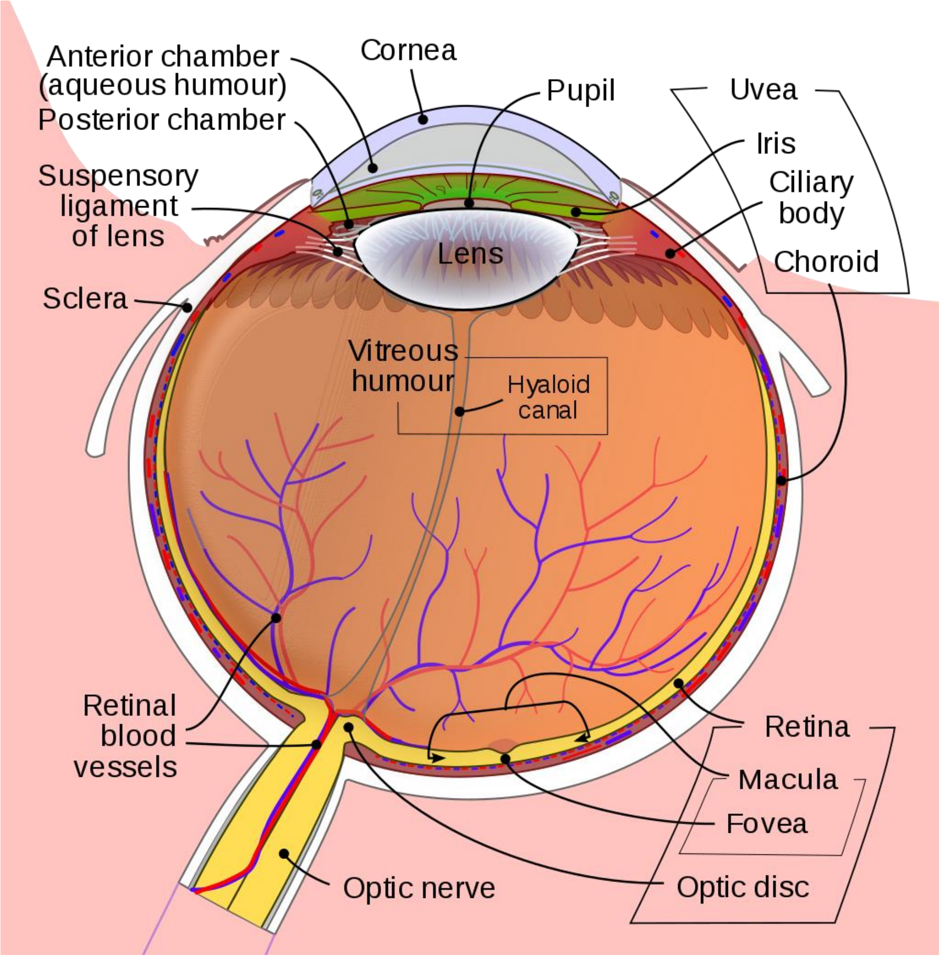


Figure 1.4. Iris Anatomy

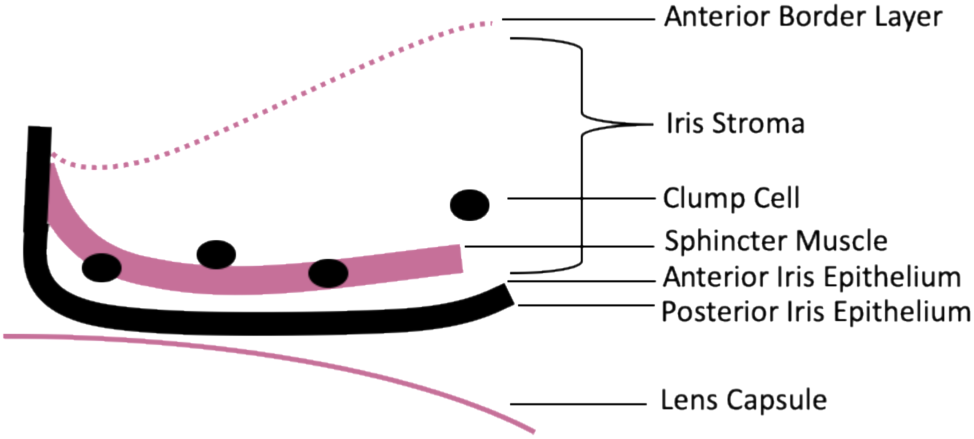


Figure 1.5. Layers of the Iris

1.4 Iris Color Phenotyping and Prediction

As our knowledge of iris pigmentation develops through tools such as GWAS, so too do our prediction tools. Initially, GWAS were performed on categorical colors that were defined as blue, intermediate, and brown[3–6]. Through these initial studies, it was found that the primary switch in blue vs brown iris color was *HERC2* through its regulation of *OCA2* with later studies identifying other iris color influencing genes such as *TYR*, *TYRP1*, *IRF4*, *SLC2A4*, and *SLC45A2*[11,61–64]. Using these identified variants, Bayesian[10] and non-Bayesian[7–9] prediction tools were designed, which have the capability of predicting certain iris colors with a high degree of accuracy. For example, the IrisPlex model is capable of predicting blue and brown iris color quite well, with an area under the receiver operating curve (AUC) of 0.91 for blue and 0.93 for brown eyes[11]. However, for the more intermediate colors, prediction AUC drops to 0.73 due to a lack of knowledge surrounding intermediate iris colors[11]. Therefore, a more in-depth look at the wide-ranging phenotype that is currently classified as 'intermediate' iris color is warranted. However, intermediate iris colors are often highly diverse and thus, a novel method of quantifying iris color may first need to be developed. Moving from categorical color to a more realistic quantitative measurement is needed to more thoroughly measure these intermediate colors. Quantitative measurement techniques developed and presented in Liu et al., 2010 and Wollstein et al., 2017 (See Figure 1.6, which demonstrates our use of the software used in Wollstein et al., 2017), used in tandem with traditional GWAS may be the solution to identifying novel variants associated with intermediate color that have eluded iris pigmentation studies thus far[11,65].

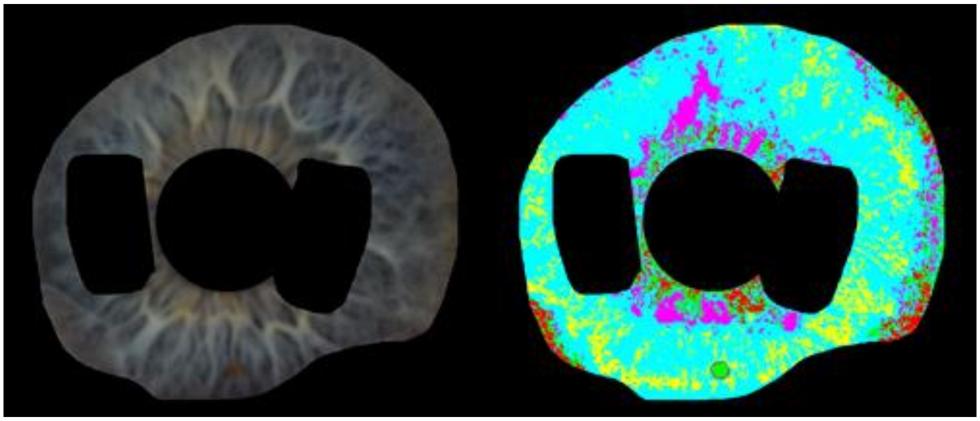


Figure 1.6. Quantitatively Measuring Iris Color. An example of quantitatively measuring iris color where perceived blue areas of an actual iris image (left) is represented on the classified image (right) in blue, perceived green/yellow in pink, light brown in green, dark brown in red, and crypts (i.e. a lack of color) in yellow

1.5 Normal Variation in Facial Morphology

It has already been concluded from twin studies that facial morphologies are based, to a large extent, on genetic composition due to their high rates of heritability[24,66]. However, with any genetic interaction that influences a phenotype, environment also has to be taken into consideration[67,68]. Age, sex, and ancestral origins can also influence facial morphologies, demonstrated in Williams and Slice, 2010, where facial shape was found to be variable in the orbits, zygomatic arches, and maxillary alveolar process[69]. The difficulties of identifying a gene's interaction on a complex and quantitative phenotype such as facial structure, amid other factors that contribute to the final phenotype, are but one reason why GWAS on facial morphology are few and far between. Yet, new technology such as next generation sequencing, advanced statistical software, and better methods of phenotyping face shape, combined with a better understanding of the genes that are involved in the determination of facial structure are allowing these genetic association studies to become more effective at finding increased numbers of facial morphology variants.

Since 2010 an increasing amount of studies have shown promise in associating genes/SNPs with facial structure amid a cohort of confounding variables (such as age, sex, and ancestral origin). Many of these first studies rely primarily on facial landmarking and then measuring the distance between said landmarks to define a phenotype (See Figure 1.7). Studies such as Paternoster et al., 2012 used principal components and 3D distances of facial features in a sample of Europeans in order to determine genetic-phenotypic associations. Paternoster was able to identify the relation between rs7559271 in *PAX3*, a gene known to be associated with Waardenburg syndrome Type I, and nasion to midenocanthion distance[70]. In 2012, a study was published by Liu et al. who also used 2D facial landmarking but supplemented that method with 3D MRI scans in order to better define the landmark in a three-dimensional space. By assessing Europeans, Liu et al., 2012 was able to determine that intronic SNPs of *PRDM16*, *PAX3*, *COL17A1*, *C5orf50*, and *TP63* were associated with facial morphology[24]. The researchers also noted that *PAX3* was one of the six genes associated with Waardenburg syndrome[25], and that variations in *PAX3* were associated with broad nasal root and an increase in the distance between corners of the eyes[24]. Mutations in *TP63*, a gene that encodes a transcription factor that helps to regulate developmental signaling and epithelial morphogenesis[71], was found to cause facial defects such as Ectrodactyly-ectodermal dysplasia-cleft lip/palate[72,73]. Variations in *TP63* were also associated with the distance between eyeballs[24].

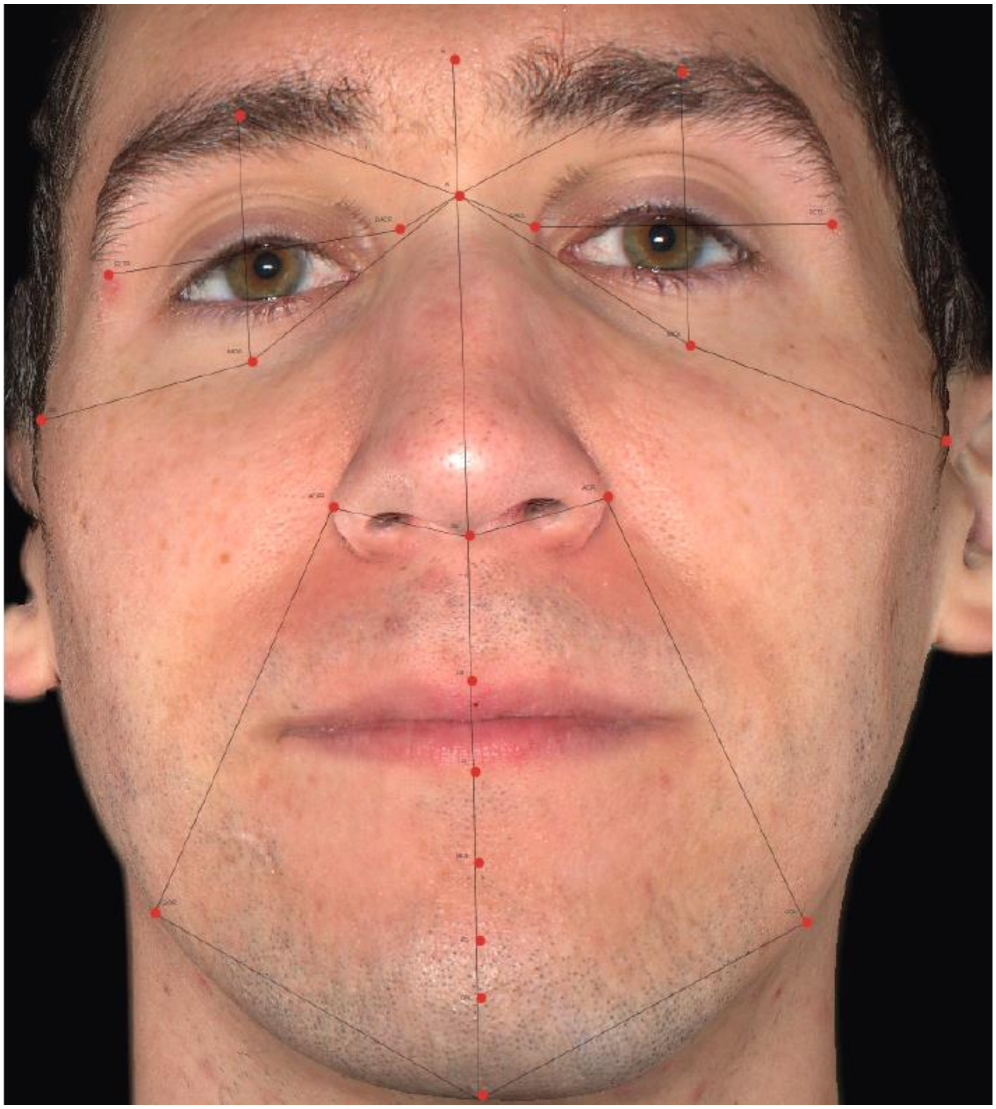


Figure 1.7. An Example of Human Facial Landmarking

Peng et al., 2013 performed a similar GWAS except the sample population was predominantly of Asian ancestral origin. Similarly, their genes of interest were selected from previous studies that had identified genes associated with diseased phenotypes such as Laron, Pfeiffer, and Kallmann syndrome. The researchers specifically identified SNPs from *GHR*, *FGFR1*, and *IRF6*[74]. Variants in *ENPP1* was found by Ermakov et al., 2010 to be essential in bone physiology and were associated with upper facial height in Chuvashians[75]. *GHR* was found to be correlated with mandible shape in Japanese and Chinese individuals[76–78], *FGFR1* is a genetic marker associated with the cephalic index in multiple populations[79], and *IRF6* has already been discussed as having an association with the formation of non-syndromic cleft lip and palate[16–18]. Using these SNPs and the 3D imagery data collected from the participants within the dataset, the researchers attempted to correlate SNPs with distances between 15 landmark distances. The authors do find several correlations between their SNPs of interest and facial morphology, particularly rs642961 (*IRF6*) and how its variants may contribute to more protrusive and thicker lips. In 2016, Adhikari et al. used 3D morphology measurements and a Euclidean distance phenotyping approach on South Americans to identify variants within *DCHS2*, *RUNX2*, *GLI3*, and *PAX1* that influence nose morphology, while they identified a variant within *EDAR* that influences chin protrusion[27].

Thus far, facial morphology GWAS have operated under the assumption that variants contribute to distances between two points, however, in 2018 Claes et al. attempted to model facial morphology in a different manner. Instead of measuring distances between points, they captured facial variation within a section of the face and performed analyses on those landmark correlations. Briefly, the researchers captured 3D photographs of faces and systematically, mapped their dense mesh landmarks to a facial average of 2329 individuals and in the process created quasi-landmarks (i.e. landmarks created from mapping an actual face onto a 'mean' face). These mapped faces were then normalized using a generalized Procrustes analysis to make the face symmetric and corrected for position and orientation. Facial landmarks were then analyzed via 3D correlation using an RV coefficient[80], which resulted in a squared similarity matrix that explains the correlations between each landmark versus all neighboring landmarks. This correlation matrix was input into a 5-level hierarchical spectral clustering procedure, which clustered groupings of landmarks with high 3D correlations into 63 total masks on five separate layers (see Figure 1.8). After normalizing each segment using a generalized Procrustes analysis, each segment, with its unique quasi-landmarks, were subjected to a principal component analysis, to determine the significant components contributing to that particular facial shape[28]. As a result, the group was able to identify 38 loci associated with facial morphology, of which 15 were replicated in a separate European cohort. Four loci were novel, while the eleven others were found to have literature supporting their connection to cranio-facial morphology.

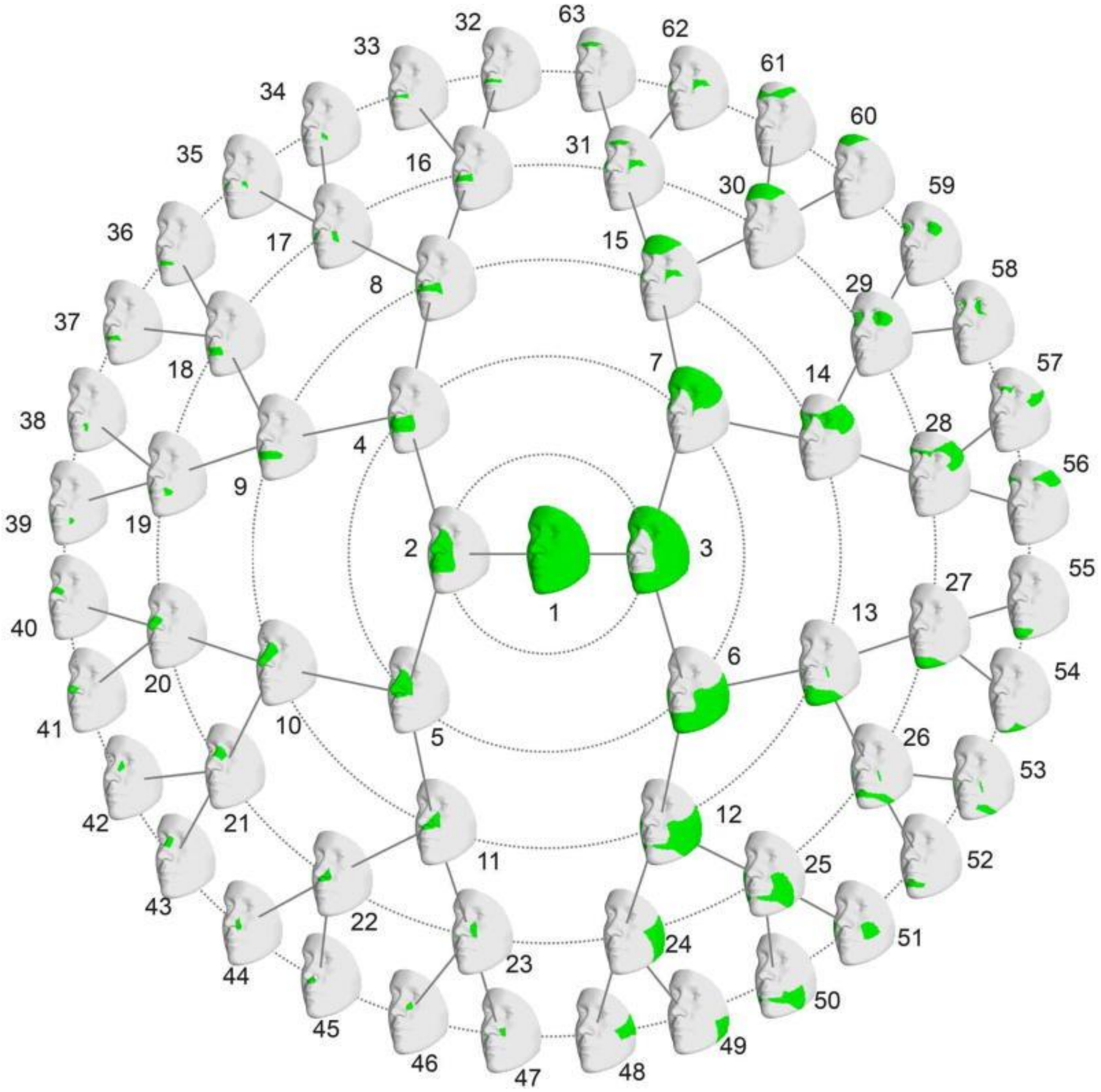


Figure 1.8. Hierarchical Clustering of Face Shape. Facial segmentation was performed on a dataset of N = 2329. First, a squared similarity matrix was constructed based on a landmark and all other neighboring 3D landmarks. Subsequently, a 5-level hierarchical spectral clustering was performed on this matrix that resulted in 63 total masks

By examining the progression of these studies, it is possible to see that researchers skipped a categorical approach to facial phenotyping, as was seen with the first iris color GWAS, and moved straight into a quantitative measure. By and large, these distance measurement GWAS have yielded plenty of genes that have later been found to be associated with facial morphology including *PAX3*, *PRDM16*, and *C5orf50*. However, as facial morphology continues to evolve and these old forms of measurements lose their effectiveness due to a lack of discovered novel SNPs, our methods of phenotyping this complex phenotype must evolve as well.

1.6 Canonical Correlation (Cancorr) and Cancorr Implementation in GWAS

A canonical correlation is a multivariate analysis of correlation, where the user wishes to analyze multiple X variants with multiple Y variants. The analysis has two parts: 1) variable grouping into latent variables (i.e. canonical variates or CV) where $CV_{X1} = a_1x_1 + a_2x_2 + a_3x_3 + \dots a_nx_n$ and $CV_{Y1} = b_1y_1 + b_2y_2 + b_3y_3 + \dots b_my_m$ and, 2) maximizing the correlation between canonical variates by adjusting the weights ($a_1 \dots a_n$ and $b_1 \dots b_n$). The output is termed a canonical pair which summarizes the relationship between subgroupings of the X and Y variables. Multiple canonical pairs are possible since different sub-groupings of either the X and or Y variables may be correlated with the Y or X subgroupings respectively. As a result, users may be able to explore higher dimensional correlations between several groups (e.g. X vs Y variables) in terms of the significance of the canonical pair and their respective variable standardized loadings, which gives insight into the relative effect or correlations one variable has with variables on the same size of the equation (e.g. the effect a_1x_1 has on $CV_{X1} = a_1x_1 + a_2x_2 + a_3x_3 + \dots a_nx_n$) and the opposite (e.g. the effect b_1y_1 has on $CV_{X1} = a_1x_1 + a_2x_2 + a_3x_3 + \dots a_nx_n$).

The implementation of canonical correlation to be used with a GWAS analysis stems from the idea that each SNP can be associated with multiple univariates instead of a single univariable that is normally conducted in a GWAS. In the analysis from Claes et al., 2018[28], an individual SNP (i.e. $CV_{SNP} = a_1SNP$) was input into a canonical correlation with a grouping of significant principal components found via the hierarchical spectral clustering and Procrustes normalization so that $CV_{PC} = b_1PC_1 + b_2PC_2 + b_3PC_3 + \dots b_mPC_m$. When analyzed, this yields the following general equation: $CV_{SNP} = CV_{PC}$ for each canonical pair. These identified CV_{PC}'s, which are simply a subset combination of PC's, can then be used as a projection target for the originally calculated PC's, rendering a single latent variable that can be used to perform a linear model based GWAS. While this method is robust it does require a substantial amount of computing power since multiple regressions are being calculated on each variant in the GWAS. In Claes et al., 2018, Matlab 2016b was used to perform the calculation, and while they do not provide such details, it would have required, at a minimum, a powerful cluster computer to perform such analyses in parallel, and expertise in resource management.

1.7 Structural Equation Modeling

Structural Equation Modeling (SEM) is a form of causal modeling that attempts to define the cause-effect relationship between observed and unobserved variables. Mathematically, SEM's are a combination of a measurement model which is constructed via confirmatory factor analysis, and the structural model that utilizes path analysis. Once the full SEM model is specified, parameter estimation is conducted by comparing the real covariance matrix between parameters and the estimated matrix created by numerical maximization. For this model, numerical maximization was carried out via maximum likelihood (ML) estimation. When the model converges on a solution, a host of fit indices are populated, which indicates the strength of how well the SEM models the data. If the fit indices indicate a good fitting model, then the parameters estimated (i.e. the regression weights) in the model may be interpreted. Here I use SEM to simultaneously model the polygenic and covariate genotypic effects of facial phenotypes. The analysis condenses the multidimensional facial phenotype from many principal components down to a single univariate phenotype via built-in dimensional reduction. While SEM's are traditionally used to analyze a relatively small number of measured indicators and latent factors with complex interactions on questionnaire datasets, this analysis sought to analyze and rank the effects of multiple genotypes on a phenotype. The output of which was to understand which genetic variants best explain the variance observed within each facial segment. In addition, most SEM's can be built visually with graphical statistical packages such as the AMOS SPSS module[81], however, for larger analyses that require hundreds of interactions, inputting the model in a graphical format becomes impractical. Thus, more robust packages such as the R-based Lavaan[82] package or the Mplus[83] software that allow for command line model input become more useful. As eluded to earlier with canonical correlation, the amount of computational power required to perform SEM's on GWAS-sized genotypic and phenotypic data requires an exceptional amount of resources that normally exceeds the limits of a high-powered desktop computer.

1.8 Research Aims

The aims of the research that will be presented here attempt to better our understanding of the genes that impact iris pigmentation and its prediction. It will be shown that a more thorough and quantitative phenotyping approach, especially regarding the intermediate iris colors of hazel and perceived green, is better suited to identify variants contributing to these phenotypes through GWAS analyses. It will also be demonstrated that the future of iris pigmentation prediction may lie in a quantitative model instead of a categorical model of prediction. This research will also provide a better understanding of facial morphology by performing a simple face/jaw shape GWAS, in addition to using more advanced multivariate statistical modeling to identify new variants and gene interaction networks that are active in facial morphology segments.

CHAPTER 2. ODYSSEY: A SEMI-AUTOMATED PIPELINE FOR PHASING, IMPUTATION, AND ANALYSIS OF GENOME-WIDE GENETIC DATA

Eller, R. J., Janga, S. C. & Walsh, S. Odyssey: a semi-automated pipeline for phasing, imputation, and analysis of genome-wide genetic data. *BMC bioinformatics* 20, 364 (2019).

2.1 Introduction

Genome-wide association studies (GWAS) have grown in popularity thanks to the increased availability of genome-wide data and sequence information. GWAS, while successful at identifying candidate variants, has also been aided by imputation methods[31,84,85] that increase coverage and allow for increased sensitivity. Imputation is often performed to fill in the genomic gaps, increase statistical power, and to "standardize" datasets so they can be combined with others that are genotyped with different arrays[86]. Over the last few decades several reference datasets have become available to use for imputation, such as the international HapMap Project in 2003[87], the 1000 Genome Project in 2015,[30] the Haplotype Reference Consortium (HRC) in 2016[88], and the more recently announced All of US Research Program currently being conducted by the NIH in 2018[89]. Increasing the number and diversity of reference panels allow for increased flexibility in how imputation is performed for a particular sample set. Van Rheenen et al., 2016 has shown that custom reference panels that combine an existing reference panel with sequence data collected from a subset of individuals within their analysis cohort may also increase imputation accuracy[90].

Current imputation options include the popular free-to-use imputation servers such as the Michigan Imputation Server (https://imputationserver.sph.umich.edu/) and the Sanger Imputation Server (https://imputation.sanger.ac.uk/), which provide an online solution to imputation. There are also offline solutions such as the Michigan Imputation Server Docker[33], and imputation packages such as Python's Genipe[91]. The strengths of online solutions are that they normally require no setup and are easy to use. However, a major drawback is that they require data to be sent off-site (albeit via a secure SFTP or SCP connection in the cases of the Sanger and Michigan Imputation Servers), which may or may not be possible for a researcher due to ethical or legal constraints. As with most online servers users may need to sit in a queue before their job is run, and users are often restricted by analysis options, such as the choice of phasing/imputation programs as well as the reference panels the sites support. At the time of writing, both the Sanger and Michigan Imputation Servers support three main panels: the HRC, 1000 Genomes (Phase 3), and the CAAPA African American Panel[92]. Sanger also provides access to the UK10K dataset, which is currently unsupported by the Michigan Imputation Server. It is important to note that apart from Sanger's option of imputing with a combined UK10K dataset and 1000 Genomes Reference panel, the online solutions do not give much flexibility if the user wishes to combine several reference panels or integrate collected data into a custom reference set to enrich the imputation. Users must then opt for offline solutions, such as the Michigan Imputation Server Docker image and Python Packages such as Genipe, that do not require data to be sent offsite and provide considerably more flexibility in the imputation analysis as they allow custom reference datasets to be installed. However, an issue of using offline solutions is that they need to be configured by the user, which may not be straightforward due to the many programs these pipelines require as well as their interconnected library dependencies.

While imputation is the main goal of all these platforms, it is imperative that data must be formatted properly before submitting it through phasing and imputation. Furthermore, quality control measures should be enacted to achieve the highest possible imputation accuracy. While a researcher should know what quality control measures they would like to use for imputation, there are inconsistencies between different programs and their default settings. The established online imputation servers perform some filters for minor allele frequency, duplicated variant detection, and VCF integrity, but most of the data cleanup is left to the user. While this data prep must be done offline, most of these platforms provide a thorough walk-through on how to implement these steps. It is also worth noting that the offline docker solution, which is similar to the Michigan Imputation Server, provides guidance with quality control but like its online counterpart does not perform it automatically. Thus, the responsibility of proper data preparation falls largely on the user and their ability to find acceptable imputation quality control thresholds, such as those found in Manolio et al., 2007[93]. In addition to cleaning the data, the user is expected to provide data in a compatible format for the imputation workflow, which is normally a VCF (for the Sanger and Michigan Imputation Servers and docker image) or a PLINK .bed/.bim/.fam (for the Genipe package). While most commercial genome array software, such as Illumina's Bead Studio or

Affymetrix's Power Tools, perform these conversions, the user must still rectify any genome array compatibility issues, such as remapping incompatible sample data to the same genome build used by the imputation reference panel.

Genome-wide association studies that use probabilistic imputation data or dosage data require a considerable number of programs for the analysis to be run. Currently, only PLINK 2.0[37] and SNPTEST[31,86,94] are capable of performing such analyses, short of writing a custom script. It is important to note that additional programs accept dosage data, but subsequently hard-call (i.e. probabilistically round) genotypes that are used in downstream analysis. While analyzing hard-called data is a valid strategy, data is ultimately being altered and may alter study outcomes. In addition to having few analysis programs that can analyze dosage data, it is often cumbersome and time consuming to input data into these programs. Dosage data often needs to be concatenated or merged (as imputation is normally done in segments) and then converted into a format accepted by PLINK 2.0 or SNPTEST in a manner that does not alter the data as previously described. Further complicating the matter of compatibility is the continual evolution of dosage data formats, such as Oxford's .bgen and PLINK's .pgen, since programs may not accept both file formats or even certain version iterations of a particular filetype. Due to the aforementioned issues, the transition between imputation and data analysis is the largest hurdle to analyzing imputation data and is probably an area in the largest need of improvement in imputation analysis workflows.

Admixture considerations while performing a GWAS lie primarily in performing a separate stratified analysis using ancestry informative programs such as the model-based Admixture[95] or PCA-based Eigensoft[34] programs, and therefore require knowledge of these additional programs to account for population stratification prior to GWAS analyses. Of course, another option is to perform the analysis via a program that supports a linear mixed model (LMM) and therefore does not require pre-ancestry testing, such as BOLT-LMM[96] or GEMMA[40], which takes ancestral interactions into account during the association analyses[97]. However, this may not be the desired algorithm of choice for most GWAS.

While much effort is expended on performing the analyses, it is essential to remember that dissemination of the results in an easy to understand manner is equally as important. Result condensation and visualization via charts, graphs, and summary tables is therefore important in any imputation analysis workflow. Advanced R plotting packages, such as Plotly[98], allow close integration with association analyses, providing users with interactive Manhattan and Quantile-Quantile (QQ) plots that give an overview of the GWAS results. Plotly data visualizations are also invaluable when assessing admixture-based PCA plots since the plots are often three-dimensional and more easily to interpret as dynamic images. At present, incorporation of data visualization into a GWAS pipeline is not present on any previously published workflows.

Genipe is one of the first to successfully integrate many of the imputation and GWAS workflow steps, as described above, into a single, easy-to-use package. The Python package is designed to facilitate the transfer of data through phasing, imputation, and various analyses using a variety of program dependencies such as PLINK, SHAPEIT, IMPUTE, and Python analysis packages as well as various custom analysis scripts. Similar to other imputation platforms, Genipe lacks built-in pre-imputation quality control measures, instead outsourcing quality control to the user via recommendations in the user manual. In addition, the program gives the option of running logistic and linear analyses, but fails to assess sample admixture, which would require the user to refer to external admixture analysis programs prior to running these analyses. However, Genipe does give the option of running an LMM, which historically has shown more success than naive logistic and linear analyses for admixed samples[99]. In addition, the program does not provide ways to visualize the association results, which would have provided a nice complement to its large repertoire of analysis options. Finally, while it is easy to setup Genipe’s Python-based framework, it does require the user to manually install and configure several of its dependencies.

Essentially, it would be beneficial from a time and resource perspective to have an imputation solution that can leverage the easy setup of online imputation servers with the flexibility of local imputation packages. Being able to control the workflow’s options and automations steps from a single configuration file would also be an advantage over programs that require the user to refer to a lengthy user-manual describing the necessary flags needed to implement a program feature. Here we describe a flexible and easy-to-use local pipeline that not only phases and imputes data, but also automates data preparation, organization, quality control, admixture and association analysis, and visualization of genome-wide data. This pipeline was designed to be compatible with all major operating systems and is also scalable, having the ability to leverage the computational power of HPS, facilitating parallelization and reducing GWAS run time from start to finish.

2.2 Methods

Several obstacles of many pipelines that contain multiple dependencies is portability, compatibility, and in the case of this resource intensive process, scalability. Odyssey attempts to address each of these issues by utilizing Singularity[100], which is similar to the commonly used Docker container solution (https://www.docker.com/). All of Odyssey's dependencies save two (IMPUTE4 due to licensing restrictions and GNU-Parallel due to technical limitations), are packaged into a Singularity container, which is contained within the Odyssey Github repository. Therefore, running Odyssey is as easy as installing Singularity on the host system allowing for increased portability. Since Singularity can be run on all major operating systems including Linux, MacOS, and Windows, this allows Odyssey to be compatible on the same systems. Unlike Docker, Singularity was created with High-Performance Systems in mind, and thanks to its unique handing of user security settings, is employed on many HPS around the world allowing Odyssey to scale from small desktops to large cluster computing systems.

Odyssey is primarily a collection of Bash and R scripts housed within a Github repository that are controlled by a single configuration text file. Researchers who wish to use the main functions of Odyssey would thus only need to interact with a single file that contains all the "flags" that affect how Odyssey behaves. Whereas other programs are controlled via command line by specifying flags and their subsequent options, Odyssey, explicitly states its options (as well as a small description of its purpose), which partially eliminates the need to refer to a user manual.

Odyssey also relies on a set of dependent programs, which are all installed and configured (save IMPUTE4 and GNU-Parallel) on startup, to perform the pipeline's main tasks. These bioinformatic programs include PLINK[37] and BCFTools[101] to perform quality control and analysis, SHAPEIT2[102] and Eagle2[103] for phasing, IMPUTE4[31,32] and Minimac4[33] for imputation, SNPTEST[31,86,94] for post-imputation quality control reporting, R as a platform for visualization and population stratification analysis, and GNU-Parallel[104] for increasing throughput. The pipeline is divided into the following main steps (see Figure 2.1).

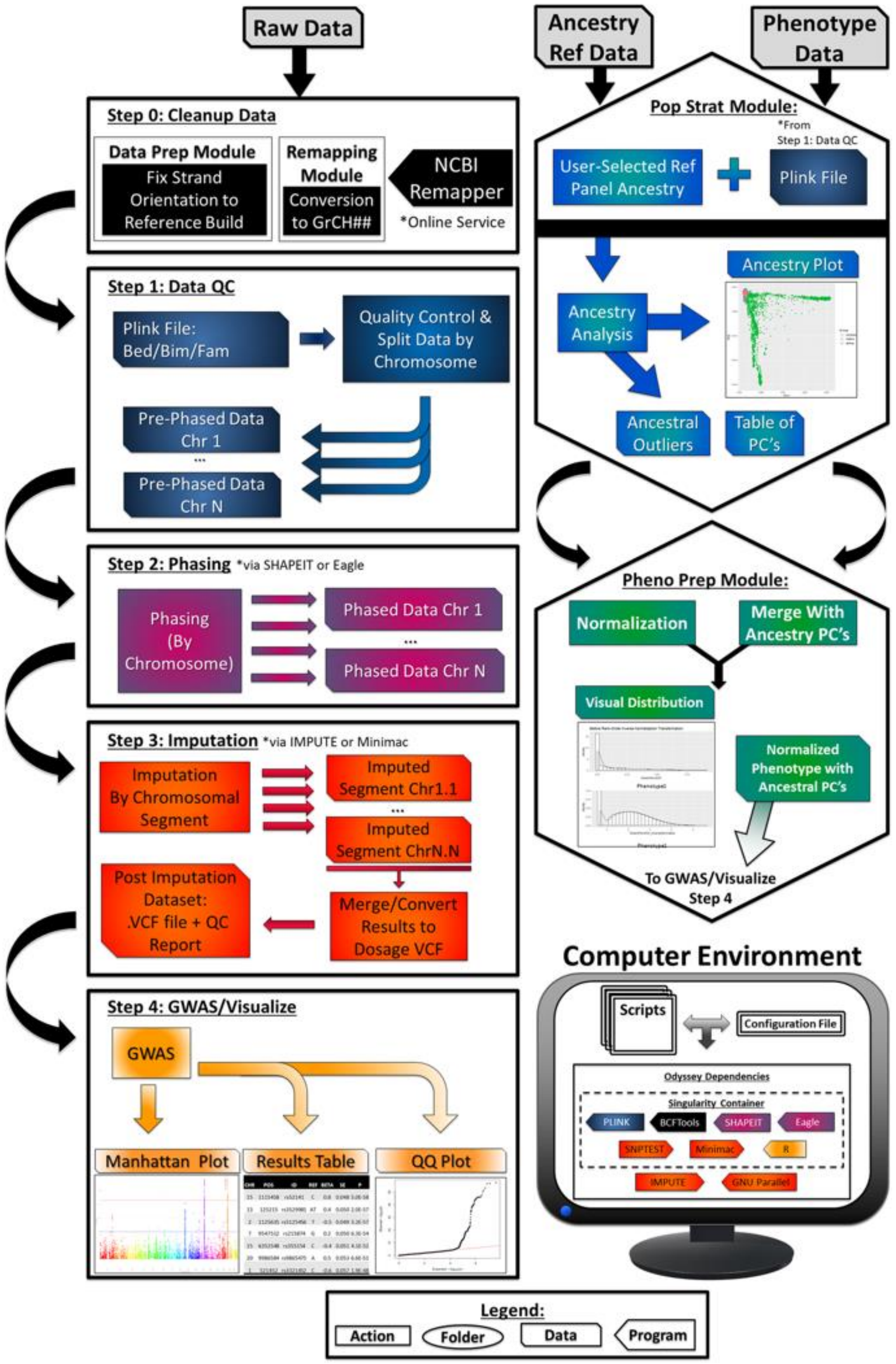


Figure 2.1. Odyssey Workflow. Odyssey performs 4 steps after data cleanup: Pre-Imputation Quality Control, Phasing, Imputation, and GWAS Analysis. Data can be easily removed from the pipeline at the ends of each major step. A Population Stratification and Phenotype Prep Module are provided, which assists in the removal of ancestral backgrounds deemed unwanted though a PCA-based approach and normalizing phenotypes

Step 0 provides a range of data cleanup options designed to take genotype data from a sequencer and prepare it for imputation and downstream analysis. The input criteria for Odyssey is a PLINK .bed/.bim/.fam. While there are a range of genotyping platforms, Illumina and Affymetrix were used as a starting point for which there are tools (i.e. BeadStudio with the PLINK Plugin, and Affymetrix Power Tools respectively) to convert raw array data into PLINK format. The Remapping Module in Step 0 gives the option of remapping input data to the genome build used in the imputation reference panel by utilizing NCBI's Coordinate Remapping Service (https://www.ncbi.nlm.nih.gov/genome/tools/remap). The Data Prep Module provides the option of using BCFTool's "fixref" plugin to correct strand orientation errors on the input data so that it matches a given reference dataset, which helps improve imputation as well as reducing the chance of getting an imputation error downstream. Both modules within Step 0 are optional and may be used if needed.

Step 1 calculates quality control metrics (including missingness, minor allele frequency, relatedness, etc.) with PLINK, and visualizes the data to better inform the use of the nature of the dataset. In addition, Odyssey provides the option to filter out variants that do not pass the default thresholds, which while set based on current practices[93], can be modified from Odyssey's configuration file. Quality controlled data is separated by chromosome and sent to Step 2 where it is phased with either SHAPEIT or Eagle, depending on user preference. Like most other imputation pipelines, Odyssey supports the phasing, imputation, and analysis of the X chromosome. At the end of Step 2, an internal check is performed to determine whether all chromosomes were phased properly. If a chromosome failed imputation, Odyssey displays which chromosome failed, why it failed (by returning the phasing error message), and can be set to re-phase the offending chromosome/s.

In Step 3 phased chromosomal data is imputed with IMPUTE or Minimac, depending on user preference, in chromosomal segments to ease the computational burden of imputation. Following imputation, another error check, similar to the error check following Step 2, is performed to check for imputation errors and provides guidance on fixing offending segments. Once all the chromosomal segments are imputed, a post-imputation quality control check is run, where poorly imputed variants are filtered out based on a user-specified IMPUTE INFO or Minimac R2 metric.

The resulting files are converted and merged into a dosage VCF-4.3 with PLINK and BCFTools, which can be loaded into most major analysis programs including PLINK and SNPTEST. Odyssey also provides a "Custom Reference Panel Creator Module", which semi-autonomously takes several user-provided reference panels (in .hap and .legend formats, which can be created from running VCF or PLINK files through SHAPEIT) and merges them together via IMPUTE to create a custom imputation reference panel. In this way users are not limited to using the default 1000 Genome Phase 3 Reference Panel that is downloaded and can thus use Odyssey to tailor their imputation runs to their own data.

Step 4 uses the dosage data calculated in Step 3 in addition to a user-provided phenotype file to perform a GWAS using PLINK, whose results are parsed, analyzed, and visualized in R via a summarized table, a Quantile-Quantile plot, and an interactive Manhattan plot using several R packages. In addition, a Population Stratification Module can be run prior to performing the GWAS, which visualizes the ancestral background of cohort individuals. Then, users can either incorporate this ancestry information into the final GWAS as a covariate or exclude individuals who lie outside of an acceptable ancestral background. This exclusion method is accomplished via an Eigensoft-like method[34] in which a reference set (e.g. the 1000 Genomes reference data) is combined with cohort data in a Principal Component Analysis to establish an X-dimensional centroid that identifies the ancestry the user wishes to retain. Outliers that fall outside of the X-dimensional centroid are determined based on a specified standard deviation or inter quartile range cutoff. Unlike Eigensoft, the exclusion method performed by Odyssey only occurs once as opposed to Eigensoft's iterative exclusion method.

Since imputation creates many files, Odyssey organizes all the data by grouping it into 6 folders (one folder for each step including a summary project folder that contains project meta data collected from each step) and provides a single dosage VCF.gz output that can be manipulated and viewed with programs such as PLINK, SNPTEST, or BCFTools. Odyssey also provides support for archiving multiple imputation runs and GWAS analyses since data is organized in the 6 folders within discrete "Project" directories. In this way a user may run multiple GWAS analyses or Imputation runs without worrying about data being overwritten. As an added benefit these modularized projects allow the user to zip and extract data at the end of each step. In this way, raw project data or the summarized results folder can be easily shared with collaborators and even integrated within their Odyssey pipeline for further analysis.

2.3 Results

Odyssey provides a user manual (See Appendix A), a tutorial, and a publicly available HGDP dataset[105] (http://www.hagsc.org/hgdp/files.html) to illustrate a sample workflow for new users. Benchmarking was conducted on Indiana University's large memory HPS, Carbonate. Carbonate contains 72 Lenovo NeXtScale nx360 M5 server compute nodes containing 12-core Intel Xeon E5–2680 v3 CPUs and 256 GB of RAM, in addition to 8 large-memory compute nodes containing 512 GB of RAM. RAM and CPU usage metrics were collected using the collectl utility (http://collectl.sourceforge.net/). To provide a baseline estimate of the resources needed by Odyssey for an imputation job, benchmarking was conducted using 3 CPU's when applicable.

The Human Genome Diversity Project (HGDP) dataset of 940 individuals with 542 K genetic markers (after quality control) was used in a SHAPEIT-IMPUTE and Eagle-Minimac workflow to show Odyssey's performance metrics. A breakdown of these benchmarks for each step can be found in Appendix B Tables B1 and B2, in addition to real-time analyses in Appendix B Figures B1-B11. To summarize, all 940 individuals were cleaned, pre-imputation quality-controlled, phased, imputed, post-imputation quality controlled, analyzed (by performing a linear regression on the dosage data and randomly generated phenotypic data), and visualized within 8 h when using SHAPEIT-IMPUTE and within 3 h when using the Eagle-Minimac workflow. Performing the optional Population Stratification add-in using the HGDP dataset and the 1000 Genomes reference set to remove non-European individuals took approximately 20 min. One of the major steps, imputation, using the 1000 Genomes Phase 3 reference panel provided by IMPUTE (https://mathgen.stats.ox.ac.uk/impute/1000GP_Phase3.html), imputed approximately 40 M (post-QC) genotypes from 542 k input genotypes in approximately 20 min by running the SHAPEIT-IMPUTE workflow on Carbonate's hyperthreaded Xeon E5–2680 CPU's, which performed 100 to 200 concurrent jobs. Conversely, when running the Eagle-Minimac workflow on the same hardware, using the same input genotypes, and a 1000 Genomes Phase 3 reference panel provided by the Minimac4 website (https://genome.sph.umich.edu/wiki/Minimac4), imputation took 45 min and 25.4 M (post-QC) variants were imputed. Therefore, in this comparison, although the choice of the Eagle-Minimac workflow was faster, the total number of variants available post QC for GWAS was only 64% of the total variants available when implementing the SHAPEIT-IMPUTE workflow under a set 0.3 INFO score threshold. This disparity could be due to the fact that imputation quality control cutoffs need to be adjusted when using alternative imputation programs and that reference panels are curated differently (e.g. some variants may be taken out of a reference panel to simplify the imputation analysis). These factors are all important considerations when choosing a workflow to help maximize the effectiveness of an imputation analysis. However, when all these aspects are held equal as shown by Liu et al., 2015[106], the accuracy differences between imputation workflows, specifically IMPUTE v Minimac, are small.

While a direct analysis with the popular online solutions, such as the Sanger Imputation Server, could not be easily measured (due to the randomness of queue wait times), in general a small dataset (N~ 900 with 550 K markers) could be submitted to Sanger and returned within similar time frames. This is expected due to the underlying programs that runs Odyssey, the Sanger and Michigan Imputation Server, and Genipe are similar, if not identical, and thus have similar time and resource requirements. Thus, in general the speed of the analysis will primarily depend on the hardware available to the user. While the runtime of the analyses will be similar, the setup time of these pipelines vary depending on the amount of data prep, the configuration of the imputation solutions, and the input of imputation options. Odyssey attempts to minimize setup time by employing modules that streamlines the data prep process, utilizing Singularity, which minimizes the time needed to configure the pipeline, and using a configuration file, which centralizes control of the pipeline and minimizes the need to constantly refer to a reference manual to lookup program options.

In the future, Odyssey's capabilities will be further improved via implementation into domain-specific language (DSL) implicit frameworks such as Snakemake[107] and by continuing to explore routes to optimize the pipeline to save time and space.

2.4 Conclusion

In conclusion, Odyssey allows quick and easier access to genome imputation by scientists who seek a local pipeline that is easy to setup, offers the flexibility to accommodate highly customizable analyses, and accommodates those who may not be allowed to outsource data to imputation servers. Odyssey attempts to take the best parts of the previous local and cloud imputation solutions and combine them into a portable, compatible, and scalable pipeline that offers a default simple analysis option for those wanting a simple analysis, or a highly customizable advanced analysis options for those looking for more complex analysis. Using modular and portable project directories, Odyssey is built to maximize collaborations as project data and results may be ported from one research group to another. Ultimately, Odyssey condenses a difficult workflow into a fast and easy-to-use pipeline that will benefit and complement biologists working with big data from multiple admixed cohorts.

CHAPTER 3. QUANTITATIVE IRIS COLOR ASSOCIATION & PREDICTION

3.1 Introduction

Human iris color, and human pigmentation in general, has been a widely studied topic over the last few years. Multiple groups have tried to identify causal variants that affect melanin production and all related steps that ultimately result in the deposition of pigment to the area of interest: iris, hair follicle, and dermis. Initially, iris color GWAS's chose to phenotype eye color on the basis of a 3-point color scale: blue, intermediate, and brown[3–6] out of its simplistic and easy implementation. Because of these studies, conducted primarily in Europeans, it was found that *HERC2* and *OCA2* are most informative for blue and brown iris phenotypes followed by other genes such as *TYR, TYRP1, IRF4, SLC2A4*, and *SLC45A2*[11,61–64]. However, as phenotyping with categorical color was able to find variants that most effected the color extremes, blue and brown, the analysis setup was unable to discern variants that had a profound effect on intermediate hues such as hazel and perceived green. Therefore, as the capabilities of categorical phenotyping were quickly becoming exhausted, quantitative iris color phenotyping has more recently been explored. Studies such as Liu et al., 2010[108] measured hue, saturation, and value (HSV) from digital photography in order to measure iris color. Other researchers such as Beleza et al., 2013 and Candille et al., 2012 performed similar analyses with digital photography, albeit using their own defined color scales[109,110]. However, one disadvantage of all these approaches is that the measurements either fails to capture the entire color content of the eye, such as in the analysis of Liu et al., 2010, or color is simplified to a custom scale and averaged over the entire iris[109,110]. Ultimately, phenotyping iris color quantitatively may provide the specificity that is needed to ascertain variants that contribute to intermediate pigmentation. For example, in 2017 Wollstein et al., used digital photography and a support vector machine algorithm to digitally categorize each pixel within a cropped iris image as having eumelanin, pheomelanin, or a lack of pigment based on training[65]. To test against other phenotypic quantitation methods, the researchers quantified iris color using their phenotypic method as well as the methods outlined in several other papers[108,109,111,112] and compared the GWAS effect strength of known pigmentation variants. As a result, the researchers found that their method yielded greater power due to their quantified phenotype being calculated per pixel and not an average over the entire iris. Therefore, by using more detailed and sophisticated phenotyping methods to detect these hard-to-find variants that may affect intermediate colors, it may be possible to ascertain if the variant has predictive capabilities. This in turn may supplement existing pigmentation prediction systems such as HIrisPlex-S[7–9], and in particular, help increase the accuracy of the IrisPlex iris color prediction model, improving its accuracy for the currently hard-to-predict intermediate category. Currently the model is capable of predicting blue and brown iris color with an area under the receiver operating curve (AUC) of 0.91 for blue and 0.93 for brown eyes[11]. However, for the more intermediate colors, prediction AUC drops to 0.73 due to a lack of genetic knowledge surrounding intermediate iris color[11]. Likewise, akin to phenotyping, perhaps it is now time to attempt quantitative color prediction from DNA.

The aim of this research was to expand upon and improve the quantitative iris color phenotyping first proposed in Frudakis, et al., 2008[62] and advanced in Wollstein et al. 2017 to determine whether the collection of quantitative color from a class phenotype (e.g. determining the percentage of blue content within an iris) increases the power of a GWAS to identify novel variants. Second, this research used known pigment variants to construct a model that could quantitatively predict a categorical color based on the same quantitative scale.

3.2 Materials & Methods

3.2.1 IUPUI Dataset Collection and Organization

To prepare for the various analyses that would be conducted on the dataset collected by the Walsh Lab at IUPUI, I prepared, setup, and currently manage a SQL server that houses all the lab's phenotypic and genotypic data. SQL Server data repositories are very efficient and safe, and thus an instance of Microsoft SQL Server was deemed the most appropriate for storing the lab's data. The capacity of the collection data, the importance of being able to quickly extract the data for analyses and securing the data to protect privacy were all factors that contributed to the SQL Server set up. Utilizing the SQL language for setup, deployment, and maintenance, the server was set up on the lab's computers and over 6600 records already collected via questionnaire, spectrophotometer, and camera were optimized and input into the server. A script protocol for importation and extraction in C# was generated to enter and extract data quickly for use in analyses. The data was secured on the computer level, via full volume encryption, server level via service key, as well as on the database level via a database key and auto-expiring certificates. Backups are automatically performed on in-lab RAID1 drives, and also off-site to the University's Scholarly Data Archive.

In order to make the data collection process quicker, more efficient, and less error prone, an automated online version of our questionnaire was set up and its link placed on our website. This minimizes erroneous responses by study participants and reduces the chance of error of lab-personnel copying a printed questionnaire's responses into the computer. The questionnaire also contains a variety of validation options resulting in real-time validation as participants are filling out study questions, which further increases the quality of our questionnaire responses.

3.2.2 Samples & Genotyping

3.2.2.1 IUPUI Dataset

The IUPUI dataset includes 2D images and genotypic data on 3528 individuals collected from individuals recruited in Ireland, Lebanon, Indianapolis, IN and Twinsburg, OH (IUPUI IRB 1409306349). Demographic distributions of the data can be found in Appendix C. Participant's self-reported information on various physical characteristics including age and ancestry was also obtained at the time of the collection. Individuals who were below 18 years of age were included if they had a parent or legal guardian's signature. No restrictions were placed on the recruitment of participants. Genotyping was performed by the University of Chicago's DNA Sequencing and Genotyping Facility (Chicago, IL) using Illumina's Infinium Multi-Ethnic Global-8 v1 array (Illumina, San Diego, California USA) consisting of 1.78M genome-wide markers.

3.2.2.2 Penn State University (PSU) Dataset

The PSU dataset includes three cohorts: Axiom, Euro180, and FEMMES. Collectively 2D images and genotypes of the participants from these cohorts were recruited through several studies at the Pennsylvania State University and sampled at the following locations: Urbana-Champaign, IL (PSU IRB 13103); New York, NY (PSU IRB 45727); Cincinnati, OH (UC IRB 2015-3073); Twinsburg, OH (PSU IRB 2503); State College, PA (PSU IRB 44929 and 4320); Austin, TX (PSU IRB 44929); and San Antonio, TX (PSU IRB 1278). Participants self-reported information on age, ethnicity, ancestry, and body characteristics, and data were gathered on height and weight. Individuals were excluded from the analysis if they were below 18 years of age and if they reported a personal history of significant trauma or facial surgery, or any medical condition that might alter the structure of the face. No restriction on ancestry or ethnicity was imposed during recruitment. PSU participants were genotyped on a variety of arrays as explained below in Table 3.1.

Table 3.1. Sample and Genotype Counts of PSU Cohorts

Cohort Name	Number of Individuals	Variants Genotyped	Genotype Array
Axiom	925	112K	Affymetrix Axiom Custom Array
Euro180	176	317K	Affymetrix Exome Array
FEMMES	176	518K	23andMe v4

3.2.2.3 University of Toronto Dataset

Between 2012 and 2014, 1465 participants of diverse ancestries were recruited for a research study on human pigmentation variation. Among these, there were 624 healthy volunteers of European ancestry. All 624 participants ranged between 18 and 35 years of age and were recruited using online and print advertisements directed towards the University of Toronto student community. A personal questionnaire was administered to each participant to determine their age, sex, self-described eye color and whether or not they had been diagnosed with any pigmentation- related diseases or disorders. Individuals were categorized as European if their four grandparents originated in any country in Europe, other than Turkey. When information about the grandparents was not known, the self-described ancestry of both parents was used to assess biogeographical ancestry. This study was approved by the University of Toronto Research and Ethics Board (Protocol Reference #27015), and all participants were required to provide written informed consent.

A photograph of each participant’s right eye was taken using a Miles Research Professional Iris Camera (Miles Research, United States). This camera consists of a Fujifilm Finepix S3 Pro DSLR

12-megapixel camera body attached to a 105-mm Nikkor lens. A biometric coaxial cable was used to deliver light to the iris at a constant light temperature to maintain color and brightness fidelity and reduce the impact of ambient light. All photographs were taken with an ISO of 200, a shutter speed of 1/125” and an aperture of f19. Photographs were initially acquired in RAW format and later converted to JPEG format using Adobe Camera Raw in Adobe Photoshop CS5 (Adobe Systems Incorporated, United States).

A 2-ml saliva sample was obtained from each participant using the Oragene DNA (OG-500) collection kit (DNA Genotek, Canada). All participants were instructed not to eat, drink or smoke for at least 30 minutes prior to obtaining the sample to ensure maximal sample purity. DNA was isolated from each sample using the protocol provided by DNA Genotek and eluted in 500 ml of TE (10 mM Tris- HCl, 1 mM EDTA, pH 8.0) Buffer. Genotyping was done using the Multi-Ethnic Global Array (MEGA) chip (Illumina Inc., San Diego, California, U.S.A.) at the Clinical Genomics Centre (Mount Sinai Hospital, Toronto, Ontario, Canada) using standard protocols. Four samples were included as blind duplicates, and the concordance rate was in all samples higher than 99.99%.

3.2.2.4 QIMR Berghofer Medical Research Institute Dataset

Participants from the QIMR dataset were genotyped on the Illumina Human610-Quad and Core+Exome SNP chips. These samples were genotyped in the context of a larger genome-wide association project that resulted in the genotyping of 28,028 individuals using the Illumina 317, 370, 610, 660, Core+Exome, PsychChip, Omni2.5 and OmniExpress SNP chips which included data from twins, their siblings and their parents. As these samples were genotyped in the context of a larger project, the data were integrated with the larger QIMR genotype project and the data were checked for pedigree, sex and Mendelian errors and for non-European ancestry. As the QIMR genotyping project included data from the multiple chip sets, to avoid introducing bias to the imputed data, individuals genotyped on the Human Hap Illumina chips (the 317, 370, 610, 660K chips) were imputed separately from those genotyped on the Omni chips (the Core+Exome, PsychChip, Omni2.5 and OmniExpress chips). All participants, and where appropriate their parent or guardian, gave informed consent, and all studies were approved by the QIMR Berghofer Human Research Ethics Committee. All SNPs are described using the dbSNP ID according to human reference assembly GRCh37.p13. Close-up images of study participants' irises were taken using a DSLR.

3.2.3 Phenotype Processing and Color Training/Quantitation

In total 21,555 left and right 2D iris images were sent to IUPUI where they were processed, quantified and either returned to collaborators for analysis or analyzed at IUPUI. Image processing included selecting the highest quality left and right iris image (if there were duplicates), selecting cropping out artifacts such as eyelashes, flashes, the pupil, and converting the JPG to PNG format. Images that were too blurry or had substandard lighting were dropped from the analysis. Images that displayed a non-standard white balance were also corrected by identifying all images that came from the same collection period and using the auto-white balance tool on Adobe Photoshop. The tool was used on an approximately 1cm by 1cm area of the sclera of five people within the cohort to establish a stable white balance correction adjustment, which was automatically applied to everyone in the cohort.

The training for quantitation of categorical color classes came from a Matlab implemented algorithm described elsewhere[65] and a built in-house Java program by the paper author, that will be referred to as IrisQuanter. Briefly, quantitative measurements of quantitative color were obtained in two steps: training of the program to recognize types of color classes (categories) and quantitation of iris images based on said color classes.

First iris color was segmented into classes of blue, perceived green/yellow, light brown, and dark brown. The crypt class was also created as a negative selector since the shadows that were cast within the crypts were reporting as dark brown. Thus, when the total color plus crypt percentages of an iris were calculated, the crypt percentage was subtracted from the total normalizing for the lack of color (i.e. shadows) that crypts cause. Each of these colors (and one crypt class) were trained by making a color training file (See Appendix D) that contained a range of pigments observed in real-life iris images obtained from participants in the IUPUI and Australian datasets. Approximately twenty irises containing significant amounts of the color being trained were used in the creation of each training image. These training images were then edited with Python and the Data Structures, Matplotlib, Pillow, and NumPy libraries[113–116] to remove pixels that did not fall into the hue, saturation, and value (HSV) upper and lower limits that had been set for each color category (See Figure 3.1). The aim of this filtering approach was to create color training images that did not have pixels with overlapping HSV's in more than one color training file in order to reduce the training error of the IrisQuanter. A screenshot of the IrisQuanter performing its color training can be seen in Figure 3.2 and the final HSV color space boundaries for each training category can be seen in Figure 3.3.

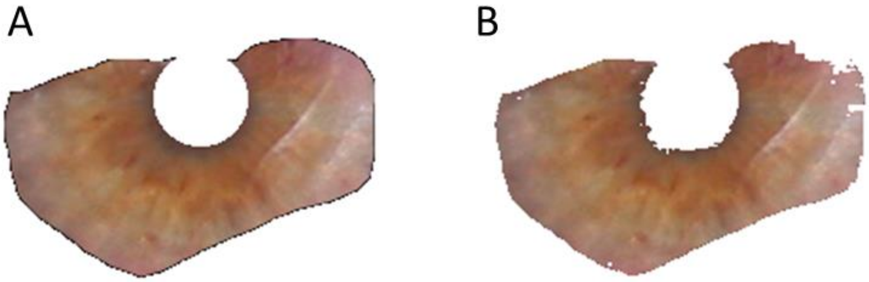


Figure 3.1. HSV Filtering. A) A light brown iris image prior to HSV filtering for light brown pixels. B) The same light brown iris image post HSV filtering. Notice how the darker areas near the pupil and the lighter areas in the upper right-hand corner have been trimmed

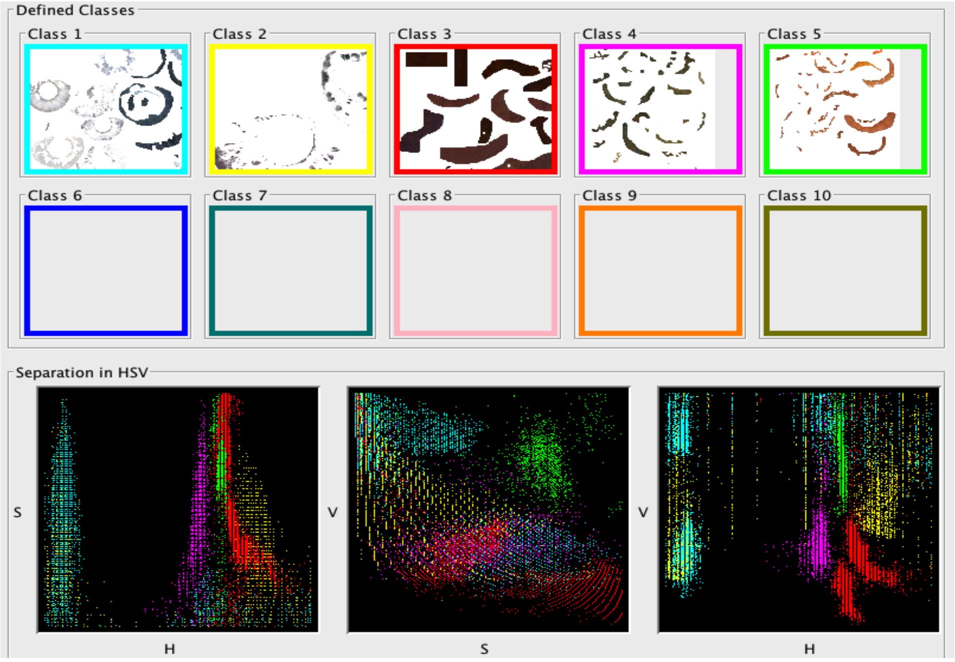


Figure 3.2. IrisQuanter Training. The IrisQuanter uses the defined classes (i.e. the color/crypt classes) from the training files we created from real irises. At the bottom the pre-HSV filtered HSV separations can be seen allowing for model tuning

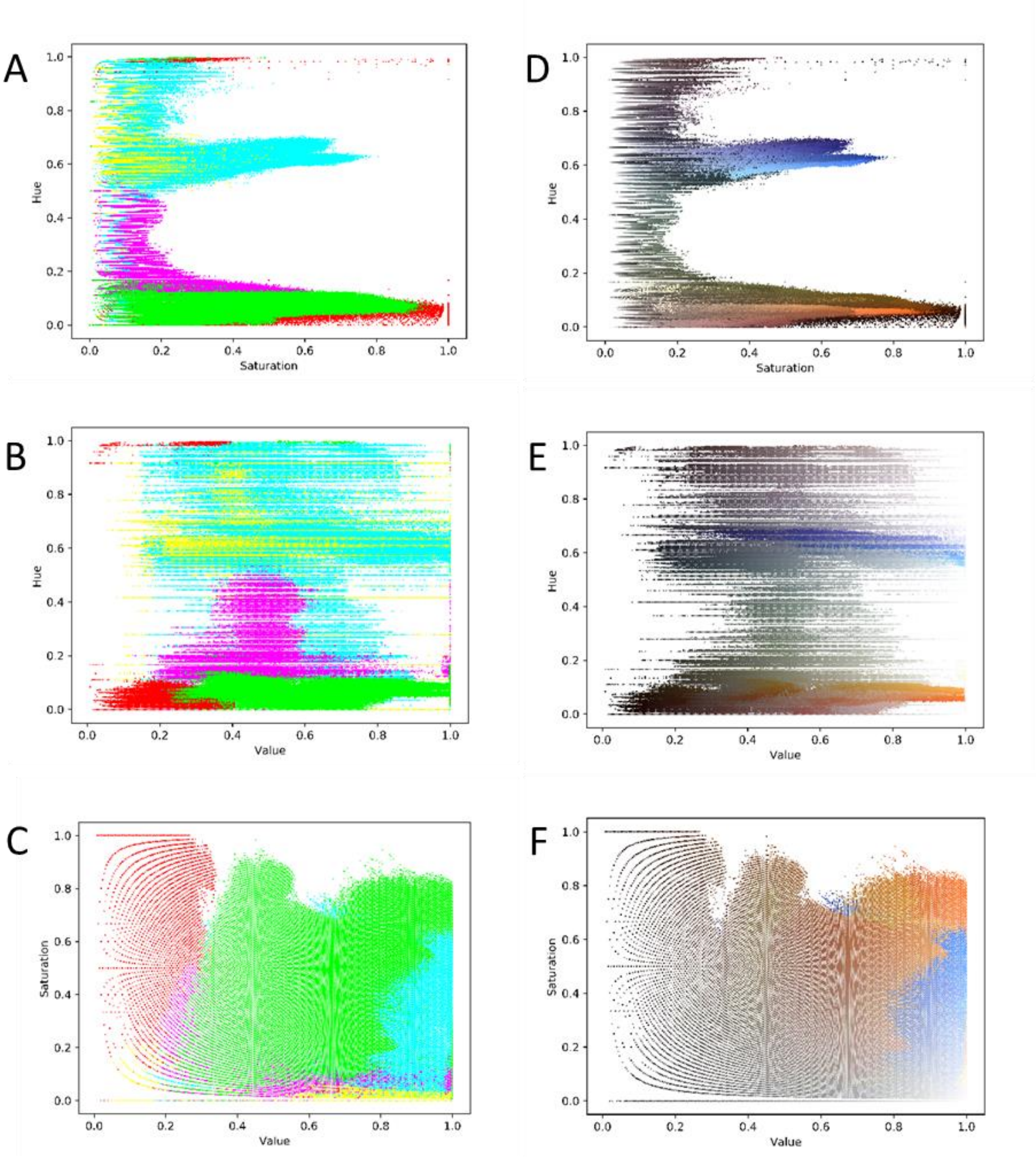


Figure 3.3. HSV Color Space Boundaries for Color Classes. Each pixel contained with the five color/crypt training images are plotted. A-C illustrates the separation between color categories on the basis of A) saturation v hue, B) value v hue, and C) value v saturation. Blue points represent the blue color category, pink represents the green color category, green represents light brown, red represents dark brown, and yellow represents crypts. D-F are identical to A-C except instead of illustrating the boundaries of each color class, each point is HSV colored

Once the color trainings had been established and all the iris photos were processed, the IrisQuanter was given the 2D left and right irises pictures of all the datasets. Quantified image results were returned in visual format to assess for accuracy and performance as well as a tabulated output which listed the pixel counts for each of the trained categories (see Figure 3.4).

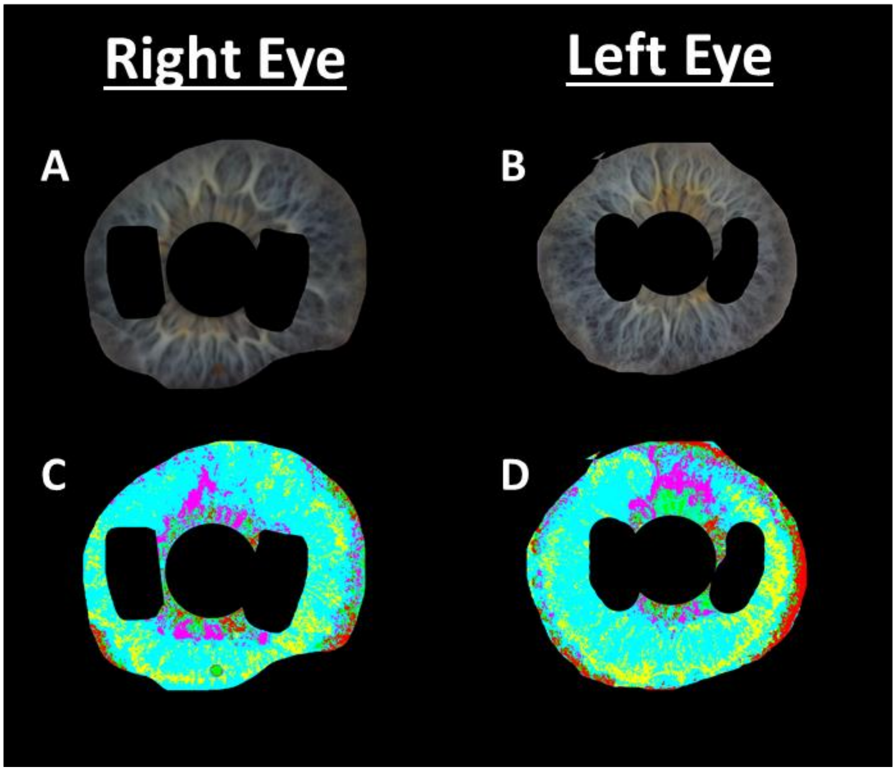


Figure 3.4. Visual Output of IrisQuanter. A-B) The processed left and right iris images are shown before quantitation and C-D) after quantitation. Note that the color categories for “blue defined” iris pixels are blue, “green defined” are pink, “light brown defined” are green, and “dark brown defined” are red. “Crypt defined” are also shown in yellow

To normalize for the variation in picture sizes and cropping, the pixel counts for each of the color classes, and not the crypt category, was summed and divided by the color classes to return the percent color of the image. The left and right iris color percentages were also averaged to give the final color class percentage for the individual. These percentages were then visualized and found to be non-normal. To correct for this, an inverse rank-order normalization was performed which drastically improves the normality of the dataset for all iris percentages (See Figure 3.5).

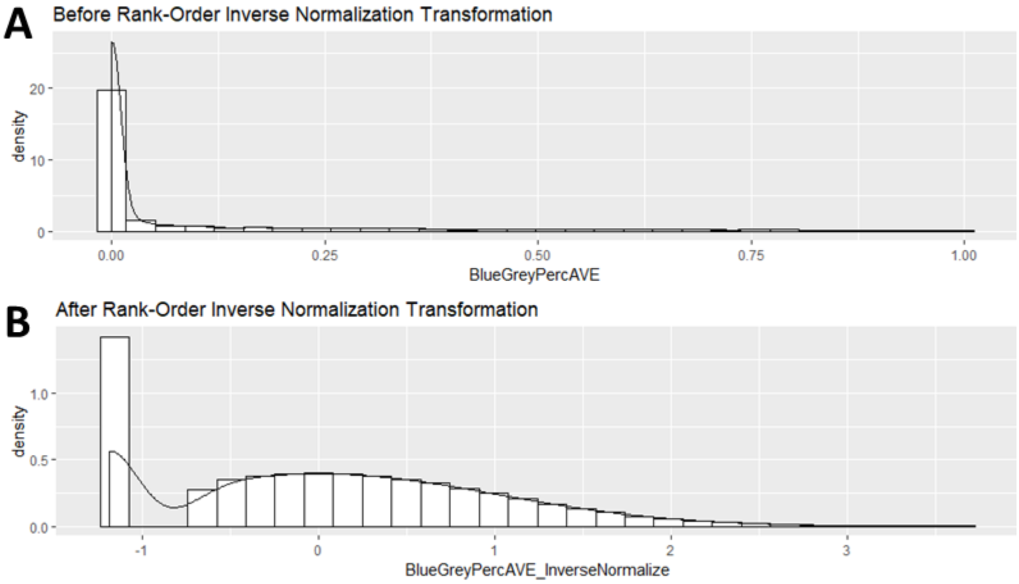


Figure 3.5. Inverse Rank-Order Transformation on Iris Phenotypes. A) Blue/Grey average percentage of the left and right irises prior to normalization and B) after inverse rank-order normalization. Note that the peak on the left side of the graph is caused by values tied for zero percent

We also performed a principal component analysis of all four-color percentages followed by a factor analysis of the eigenvectors to yield a single latent color factor, which encompasses all four classes. This latent factor was also analyzed for normality and corrected, due to a departure from normal (albeit less of a departure than the individual color percentages) via the same inverse rank order normalization method.

3.2.4 Ancestry Analysis

For the IUPUI and dataset, pre-phased quality controlled genotyped variants underwent a filter for Hardy-Weinberg Equilibrium ($p < 0.00001$) and were merged with the 1000 Genomes Phase 3 reference dataset. Variants that were in common between the datasets were assessed for linkage disequilibrium and then pruned using a 1500 kb window, 50 bp step size, and a 0.4 R^2 threshold yielding 194K variants. This pruned dataset which contained 2504 individuals from the 1000G reference in addition to the IUPUI individuals were used in a principal component analysis to construct an ancestry space. Using the eigenvalues who were found to explain more than 5% of the total amount of variance, an X-dimensional centroid was created from 503 individuals who were deemed by 1000G to be of European ancestry. This in term created a “European centroid” (See Figure 3.6). If participants were 3 standard deviation away from the centroid, they were deemed to not be of European descent and were subsequently dropped. As a result, 1081 individuals were removed from the IUPUI dataset. Ancestry analysis on the PSU dataset was performed the same way resulting in the removal of 719 individuals (705 from Axiom, 1 from Euro180, and 14 from FEMMES).

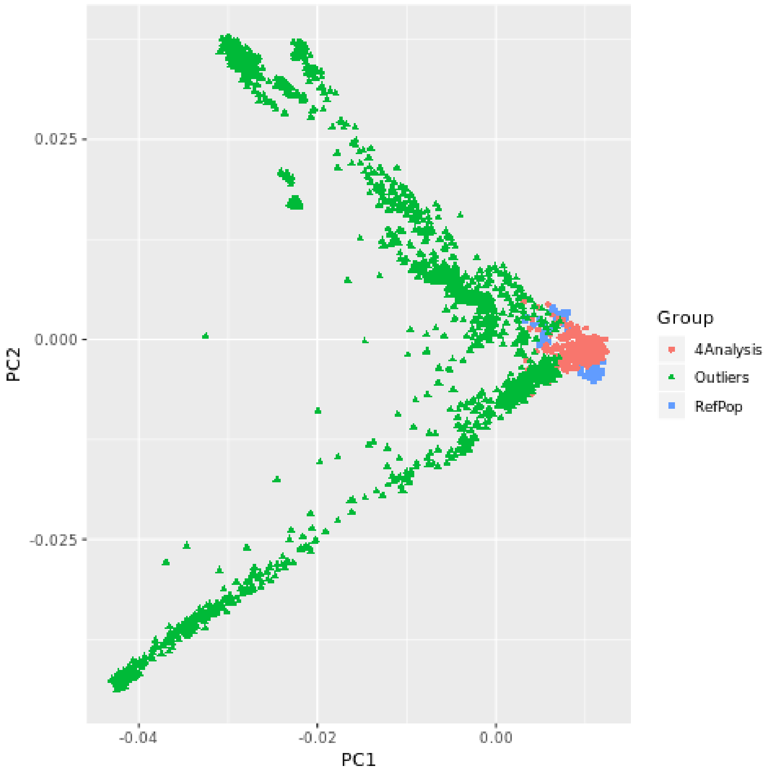


Figure 3.6. Analyzing Population Stratification via PCA. 6032 individuals were plotted, including 3528 IUPUI individuals of unknown ancestry and 2504 1000 Genome individuals of known ancestry, including 503 Europeans. The 503 reference Europeans (shown in blue) create the European centroid around which individuals are either dropped (shown in green as outliers) or kept for subsequent analysis (illustrated in pink)

3.2.5 Genotype Quality Control (QC), Imputation, and GWAS Analysis

The following are a summary of the quality controls, imputation settings, and GWAS analyses that were conducted for each study’s dataset. A summary of the post-imputation QC statistics as well as the surviving variants and individuals used for the GWAS analyses can be seen in Table 3.2.

Table 3.2. Iris Color Post-Imputation Quality Control

Dataset	Individuals	Variants Imputed	Imputation INFO Score Cutoff	Analysis Method*
IUPUI	2119	6.3M	0.7	LMM (GCTA)
Axiom	177	948K	0.7	LMM (GCTA)
Euro180	155	6.5M	0.7	LMM (GCTA)
Femmes	162	7.0M	0.7	LMM (GCTA)
Toronto	545	8.9M	0.8	LM (SNPTEST)
QIMR	3740	7.5M	0.7	LMM (GEMMA)

**Note: Mixed Linear Model (LMM) and Linear Model (LM)*

3.2.5.1 QC, Imputation, and GWAS Analysis: IUPUI and PSU

Quality control practices used to prepare the IUPUI and PSU datasets for imputation were performed using the GRCh37 (hg19) genome assembly[117]. Quality control included filtering out individuals who had more than 5% of their genome missing, variants that were missing in more than 5% of the dataset, individuals who were missing either phenotypic or genotypic data, and related individuals (i.e. identity by descent greater than 0.1875, or the halfway point between second and third degree relatives). Individuals who had unusually high heterozygosity (+/-3 standard deviations), were also excluded. In total 3528 IUPUI, 919 Axiom, 176 Euro180, and 176 Femmes individuals were selected for imputation via the Odyssey pipeline (See Chapter 2), using the SHAPEIT2[102] and IMPUTE4[31,32] workflow to phase and impute, respectively. The Haplotype Reference Consortium[88] reference panel was used for both phasing and imputation. After imputation, variants were filtered base on the imputation quality control INFO metric (INFO score > 0.7). Prior to GWAS, the dataset was pruned based on SNP missingness (missingness < 5%), minor allele frequency (MAF > 1%), and Hardy-Weinberg equilibrium ($p < 1 \times 10^{-6}$).

After converting the dosage data to hard-calls due to programming compatibility, linear mixed model GWAS were performed on 6.3M variants using the GCTA software[39,118]. Specifically, a genetic relatedness matrix (GRM) was first created separately for the autosomes and X chromosome using the imputed variants, which then informed the linear mixed model (LMM) of any residual cryptic relatedness between samples not filtered out in the ancestry analysis and identity by descent QC. LMM’s were conducted separately on autosomes and the X chromosome for the five color-based normalized phenotypes. In addition, covariates of age and sex were included in the IUPUI and Axiom LMM's. Age was the only covariate in the Euro180 and Femmes LMM's since all individuals were females. The separate autosome and X chromosome LMM results were combined following the analysis and analyzed.

3.2.5.2 QC, Imputation, and GWAS Analysis: University of Toronto

For the University of Toronto's dataset, our collaborator performed QC steps to remove samples and markers, according to the following criteria. Sample QC: 1) removal of samples with missing call rates < 0.9; 2) removal of samples that were outliers in Principal Component Analysis (PCA) plots; 3) removal of samples with sex discrepancies; 4) removal of samples that were outliers for heterozygosity; and 5) removal of related individuals (pi-hat > 0.2). Marker QC: 1) removal of markers with genotype call rate < 0.95; 2) removal of markers with Hardy-Weinberg p-values < 10^{-6}; 3) removal of Insertion/Deletion (Indel) markers; 4) removal of markers with allele frequencies < 0.01; 5) removal of markers not present in the 1000 Genomes reference panel, or that did not match on chromosome, position and alleles; 6) removal of A/T or G/C SNPs with MAF > 40% in the 1000 Genomes European reference sample; and 7) removal of SNPs with allele frequency differences > 20% between the study sample and the 1000 Genomes European reference sample. After these QC steps, we retained 545 samples and 561,400 markers. The samples were then phased using the program SHAPEIT2[119] and imputed using the Sanger Imputation Service, using the Positional Burrows-Wheeler Transform (PBWT) algorithm[120], and the samples of the 1000 Genomes Phase 3 as reference haplotypes[30]. The final number of variants after imputation and a filtering step for INFO > 0.3 was approximately 8.9 million.

GWAS were run on the SNPTEST software[31,86,94] using an additive model and the genotype dosages ('expected' method), for each of the five phenotypes. We ran the GWAS conditioning for sex and the first four principal components. Age was not correlated with any of the phenotypes. Afterwards, we filtered the SNPTEST results by removing SNPs with INFO < 0.8.

3.2.5.3 QC, Imputation, and GWAS Analysis: QIMR

Our collaborators imputed individuals with the Haplotype Reference Consortium (HRC.1.1) using a set of SNPs common to the first-generation genotyping platforms (N ~ 278,000). Imputation was performed on the Michigan Imputation Server using the SHAPEIT/minimac Pipeline[33]. Genotype data were screened for genotyping quality (GenCall < 0.7), SNP and individual call rates (< 0.95), HWE failure ($P < 1 \times 10^{-6}$) and MAF (< 0.01). After phenotype and genotype quality control, data were available for 7,624,941 SNPs and 2,361 participants.

GWAS were run on the GEMMA software[40,121] using a LMM for each of the five inverse rank-order normalized phenotypes. Covariates of age, sex, and five principal components were also included in the model.

3.2.6 IUPUI, PSU, Toronto, and QIMR META Analysis

A meta-analysis was carried out with the METAL program[122] using Stouffer's method[123] on the IUPUI, PSU (Axiom, Euro180, and Femmes), Toronto, and QIMR GWAS results on each of the five iris color phenotypes. As alternative genotyping platforms use different naming schemes, the variants were renamed by chromosome, position, and minor allele. The genotypes were coded additively based on the presence of the minor allele and were consistent between cohorts.

3.2.7 Iris Prediction Modeling Using Neural Networks

A feedforward neural network model that attempts to predict quantitative measurements of iris color classes (e.g. a model with a prediction of 80% blue, 15% green, and 5% light brown eye color) was created. 981 variants that have a known connection to pigment were collected from various sources (See Appendix E) and extracted from the IUPUI imputed dataset[124]. Of the 981 variants identified 875 overlapped with the contents of the imputed dataset. After excluding variants that were fixed ($MAF > 1 \times 10^{-7}$) and those that were linked assessed via a Plink LD pruning procedure using a 1500 kb window, a 50 bp step size, and a 0.4 R^2 threshold, 527 variants were left for feature selection. Feature/variant selection was performed in a manner that is similar to the variant ranking method as outlined in section 5.2.5. Briefly, 527 variants were broken into four equally sized groups and input into a SEM, composed of a single latent factor being comprised of eye color percentages. Variants were then regressed onto (i.e. are predictors of) the latent variable (See Figure 3.7).

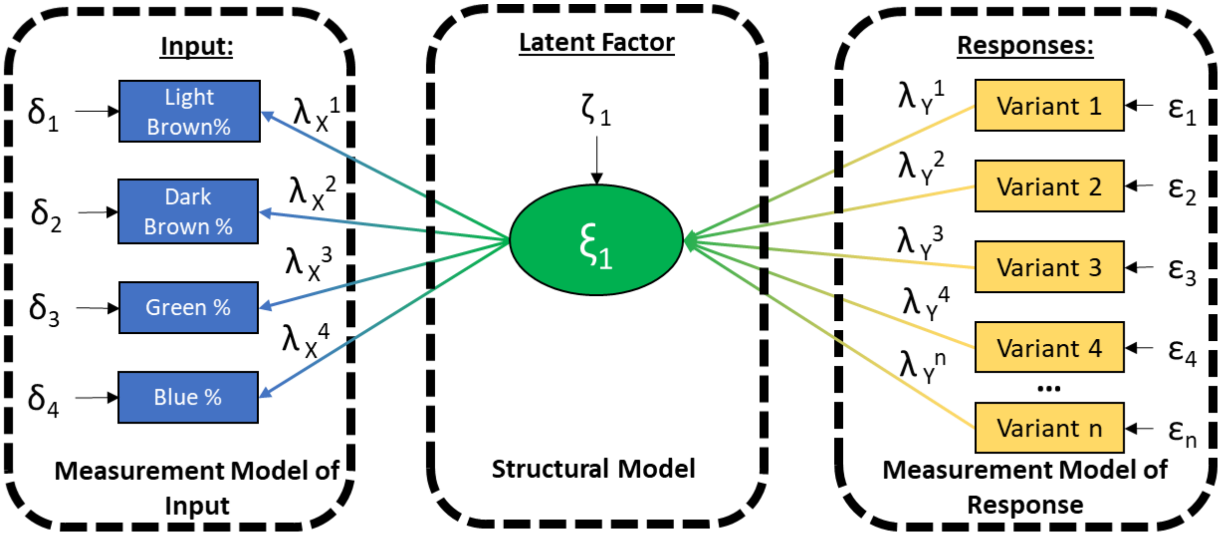
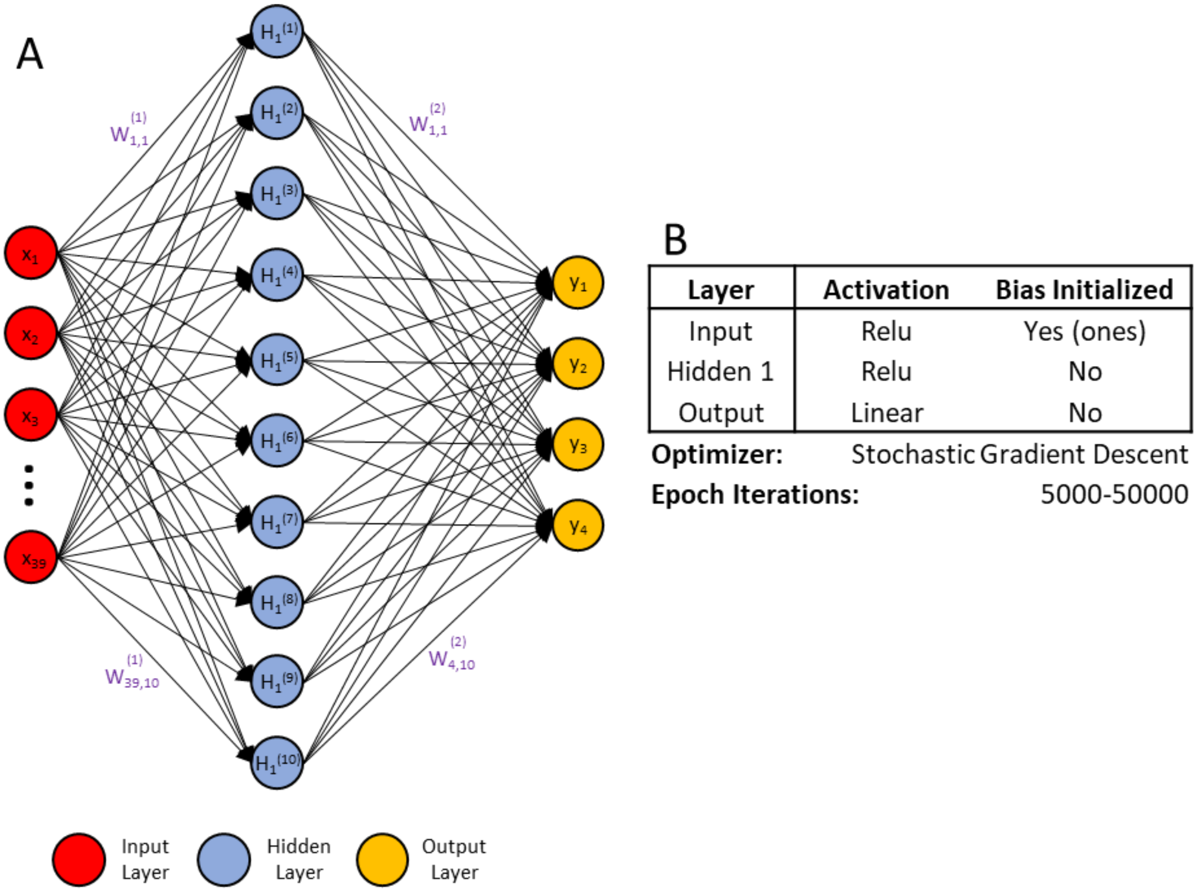


Figure 3.7. A Visual Representation of the SEM Used for Iris Pigmentation Feature Selection. The SEM is composed of iris color percentage exogenous variables (Input) that comprise the unobserved latent factor that explains the relationship between all iris colors (ξ1). Groupings of the 527 variants (Responses) are input into the model which are regressed onto the latent factor

Four SEMs, one for each grouping of variants were created separately due to computational resource limitations. Variants that were found to be nominally influential in explaining the latent factor (assessed by examining the significance of the variants regression weight at a threshold of $p < 0.2$) were passed to the second 'round' of SEM creation in which variants again were assessed, except with a more strict criteria of $p < 0.05$. This process was iteratively done, explained more in depth in section 5.2.5, until the most influential variants that contribute to iris color were identified. This list was comprised of 39 variants (See Appendix F) that were passed to a neural network model.

A neural network model was created in Python using the Tensorflow backend, Keras frontend, CUDA GPU acceleration on a Nvidia RTX2070, and the Scikit-learn, numPy, and Pandas package libraries[113,116,125–128]. The model was created with three layers that contained 39 input nodes (one for each input variant), ten hidden nodes, and four output nodes (one for each predicted color) respectively. Model specifications can be seen in Figure 3.8. Several models were created with this structure. First a ten-fold cross validated model was performed using a training to test ratio of 80-20 for 3276 individuals from the IUPUI database and run for 5000 epochs. A second model was created to serve as the 'upper estimate' model that contained all 3276 for training and was tested on 5 randomly selected individuals who were held out and run for 50000 epochs. Accuracy was measured in the same way between both models and consisted of measurements on mean squared error (MSE), root mean squared error (RMSE), and R^2. Mean absolute error (MAE) was also calculated for the entire model as well as each individual iris color.



Layer	Activation	Bias Initialized
Input	Relu	Yes (ones)
Hidden 1	Relu	No
Output	Linear	No

Optimizer: Stochastic Gradient Descent

Epoch Iterations: 5000-50000

Figure 3.8. Diagram of the Quantitative Iris Color Prediction Neural Network. A) The model is comprised of 39 input nodes (red), 10 hidden nodes (blue), and 4 output nodes (yellow), resulting in 430 interconnected backpropagated regression weights (purple; only four are shown for simplicity). B) Activation, optimizer, and epoch iterations of the neural network

3.3 Results & Discussion

3.3.1 Identification of Known and Novel Hits Regarding Iris Pigment

The META analysis consisting of 6898 individuals between six cohorts on the five phenotypes tested (See Tables 3.3 and 3.4) returned several known and expected loci as evident by the large peaks occurring on chromosomes 15 (*HERC2*), 14 (*SLC24A4*), 11 (*TYR*), 9 (*TYRP1*), and 6 (*IRF4*).

However, there were 3 potentially novel variants found within the META. First, within the blue META analysis (See Figure 3.9), known hits can be seen on chromosomes 6, 9, 11, 14, and 15 as mentioned above. However, one novel significant hit was found within chromosome 1, within the Axiom, Euro180, Femmes, Toronto, and QIMR datasets. Rs3820285 is a missense SNP located within *CELA3A* that has a 2.4% allelic frequency in Europeans. While *CELA3A* itself may not appear to have any relation to melanin formation, as the protein is a serine protease and has known to serve a digestive function in the intestine[129], the variant is found adjacent to numerous candidate cis-regulatory elements as identified by Encode[130,131] as well as 784bp downstream of an enhancer (ENSR00000921031) that is actively enhanced in keratinocytes. While there is no evidence to support the function of this enhancer, it is interesting that 16.5kb downstream of the enhancer is a gene, *CDC42*, that has been implicated in actin reformation and formation of filopodia[132,133]. While there are several methods for melanosomes to migrate from melanosomes to keratinocytes, one method is melanosome transfer via filopodia[134]. As Singh et al., 2010 demonstrates, constitutively active *CDC42* results in the increased melanosome transfer from melanocytes to keratinocytes[135]. Furthermore, rs3820285 is in relatively high LD (R^2 = 0.69) with rs182609273 (See Figure 3.10) which has shown evidence of histone deacetylase binding (found through ENCODE ChIP-seq studies). Therefore, while the initial correlation of rs3820285 to pigmentation may be weak, the variant may be in an area worth exploring due to its dense concentration of regulatory elements.

Table 3.3. Iris META Analysis Results

GWAS Pheno	RSID	META P	Axiom		Euro		Femmes		IUPUI		Toronto		QIMR	
			B	P	B	P	B	P	B	P	B	P	B	P
Blue	rs916977	6.4E-323	0.92	4.1E-11	-0.67	6.0E-10	-0.99	1.8E-11	-1.00	9.9E-123	-0.82	4.6E-29	-0.94	1.8E-175
Blue	rs12203592	1.1E-20	0.11	4.9E-01	NA	NA	0.28	5.0E-02	0.28	3.9E-13	0.29	1.3E-04	0.16	5.7E-07
Blue	rs1393350	5.9E-19	-0.04	7.6E-01	0.23	3.2E-02	0.04	7.6E-01	0.21	2.1E-09	0.06	1.9E-03	0.17	2.7E-09
Blue	rs35983729	1.3E-16	-0.34	3.7E-03	0.18	8.0E-02	0.04	6.7E-01	0.12	1.4E-04	0.16	5.9E-03	0.16	4.4E-10
Blue	rs2762457	5.8E-10	NA	NA	0.23	1.9E-02	0.22	4.7E-02	NA	NA	0.05	4.3E-01	0.15	2.6E-08
Blue	rs3820285	2.8E-08	0.02	9.7E-01	-1.33	5.8E-04	-0.12	6.9E-01	NA	NA	-0.26	2.1E-01	-0.38	3.1E-07
Green	rs12913832	5.7E-128	NA	NA	NA	NA	-0.41	5.8E-04	-0.70	1.8E-127	-1.03	7.1E-60	NA	NA
Green	rs12896471	5.7E-10	-0.15	2.1E-01	-0.05	6.1E-01	0.03	7.5E-01	0.07	4.4E-03	0.06	4.1E-01	0.16	3.7E-10
Green	rs77373930	4.2E-08	NA	NA	-0.05	8.7E-01	0.69	1.7E-01	NA	NA	0.98	2.0E-03	0.50	3.4E-06
Light Brown	rs1667394	1.8E-230	-0.50	1.2E-03	-0.18	1.2E-01	-1.08	1.5E-14	-0.72	3.2E-82	-0.56	1.2E-139	-0.86	1.9E-11
Light Brown	rs941799	2.9E-16	-0.27	3.7E-02	-0.23	4.2E-02	-0.08	3.8E-01	-0.10	4.2E-04	-0.19	1.3E-03	-0.16	6.5E-10
Light Brown	rs72963135	2.2E-12	NA	NA	-0.1	4.1E-01	0.08	5.2E-01	-0.17	7.4E-08	-0.13	9.0E-02	-0.14	2.2E-06
Light Brown	rs12203592	6.1E-11	0.12	4.8E-01	NA	NA	-0.24	8.0E-02	-0.19	4.4E-08	-0.18	2.0E-02	-0.12	3.7E-04
Dark Brown	rs1667394	2.8E-183	-0.85	5.7E-10	-0.68	7.9E-10	-0.84	3.5E-09	-0.30	-3.6E-20	-0.78	1.9E-24	-0.85	3.7E-159
Dark Brown	rs12203592	8.8E-22	-0.17	2.7E-01	NA	NA	-0.24	8.8E-02	-0.25	2.0E-16	-0.31	6.1E-05	-0.14	6.2E-06
Dark Brown	rs35983729	2.5E-11	0.09	4.3E-01	0.14	1.7E-01	0.00	9.9E-01	0.07	2.0E-03	0.08	1.6E-01	0.14	8.3E-09
Dark Brown	rs72963135	2.9E-11	NA	NA	-0.02	8.4E-01	0.00	1.0E+00	-0.14	1.2E-06	-0.17	1.2E-02	-0.12	1.7E-05
Dark Brown	rs2733831	1.7E-10	0.18	1.2E-01	0.2	5.0E-02	0.11	3.2E-01	0.10	2.0E-04	0.07	2.5E-01	0.12	4.4E-06
Dark Brown	rs6420484	4.2E-08	NA	NA	NA	NA	NA	NA	0.08	1.4E-03	0.14	2.4E-02	0.11	8.5E-06
Factor Reduced	rs1667394	4.8E-192	0.88	3.1E-12	0.7	4.5E-11	0.53	3.2E-11	0.59	2.1E-131	0.9	2.5E-31	0.73	8.4E-199
Factor Reduced	rs12203592	8.9E-12	0.24	1.1E-01	NA	NA	0.17	2.2E-01	0.23	1.7E-13	0.13	9.5E-02	0.08	2.0E-02
Factor Reduced	rs77373930	3.7E-08	NA	NA	0.06	8.7E-01	0.51	2.9E-01	NA	NA	0.84	7.3E-03	0.52	1.4E-06

*Note: B (beta); P (p-value)

Table 3.4. Iris META Analysis Datamining Results

GWAS Pheno	RSID	Chr	Position*	META P	Associated Gene	Eur Allele Freq
Blue	rs916977	15	28268218	6.4E-323	HERC2	0.24
Blue	rs12203592	6	396321	1.1E-20	IRF4	0.12
Blue	rs1393350	11	89277878	5.9E-19	TYR	0.24
Blue	rs35983729	14	92321417	1.3E-16	SLC24A4	0.43
Blue	rs2762457	9	12689313	5.8E-10	TYRP1	0.41
Blue	rs3820285	1	22009784	2.8E-08	CELA3A	0.02
Green	rs12913832	15	28120472	5.7E-128	HERC2	0.36
Green	rs12896471	14	92307559	5.7E-10	SLC24A4	0.43
Green	rs77373930	2	65970440	4.2E-08	AC007389.1	0.02
Light Brown	rs1667394	15	28285036	1.8E-230	HERC2	0.24
Light Brown	rs941799	14	92310481	2.9E-16	SLC24A4	0.43
Light Brown	rs72963135	11	89188404	2.2E-12	TYR	0.24
Light Brown	rs12203592	6	396321	6.1E-11	IRF4	0.12
Dark Brown	rs1667394	15	28285036	2.8E-183	HERC2	0.24
Dark Brown	rs12203592	6	396321	8.8E-22	IRF4	0.12
Dark Brown	rs35983729	14	92321417	2.5E-11	SLC24A4	0.43
Dark Brown	rs72963135	11	89188404	2.9E-11	TYR	0.24
Dark Brown	rs2733831	9	12703484	1.7E-10	TRYP1	0.41
Dark Brown	rs6420484	17	81,645,371	4.2E-08	TSPAN10; NPLOC4	0.36
Factor Reduced	rs1667394	15	28530182	4.8E-192	HERC2	0.24
Factor Reduced	rs12203592	6	396321	8.9E-12	IRF4	0.12
Factor Reduced	rs77373930	2	65970440	3.7E-08	AC007389.1	0.02

*Note: RSID position is based on hg19

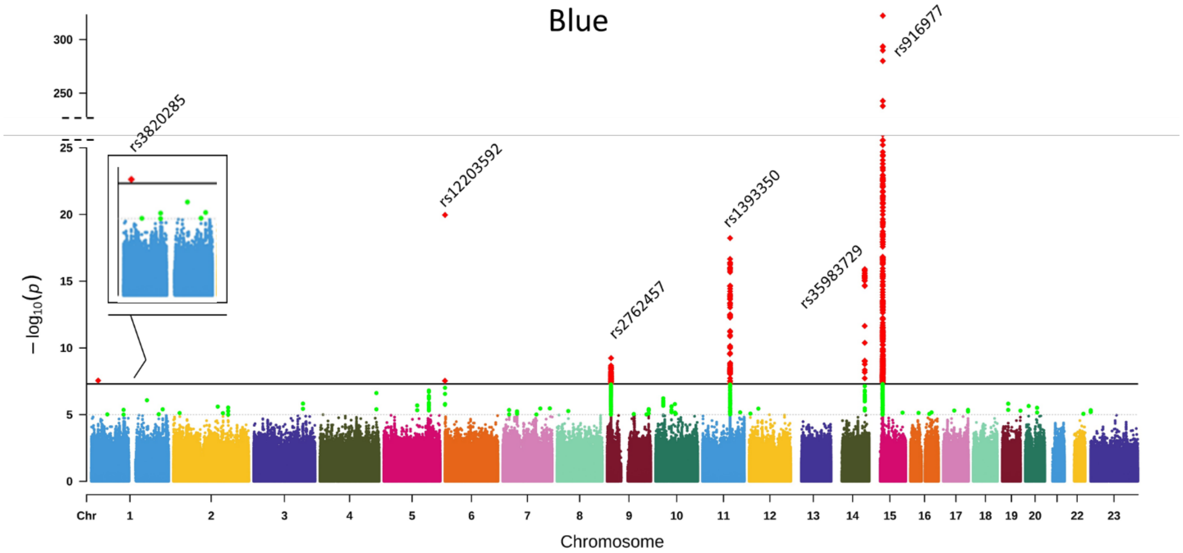


Figure 3.9. Iris META Manhattan Plots for Blue Irises. Near genome-wide significant variants ($p < 1 \times 10^{-5}$) are illustrated in green and genome-wide significant variants ($p < 5 \times 10^{-8}$) are shown in red

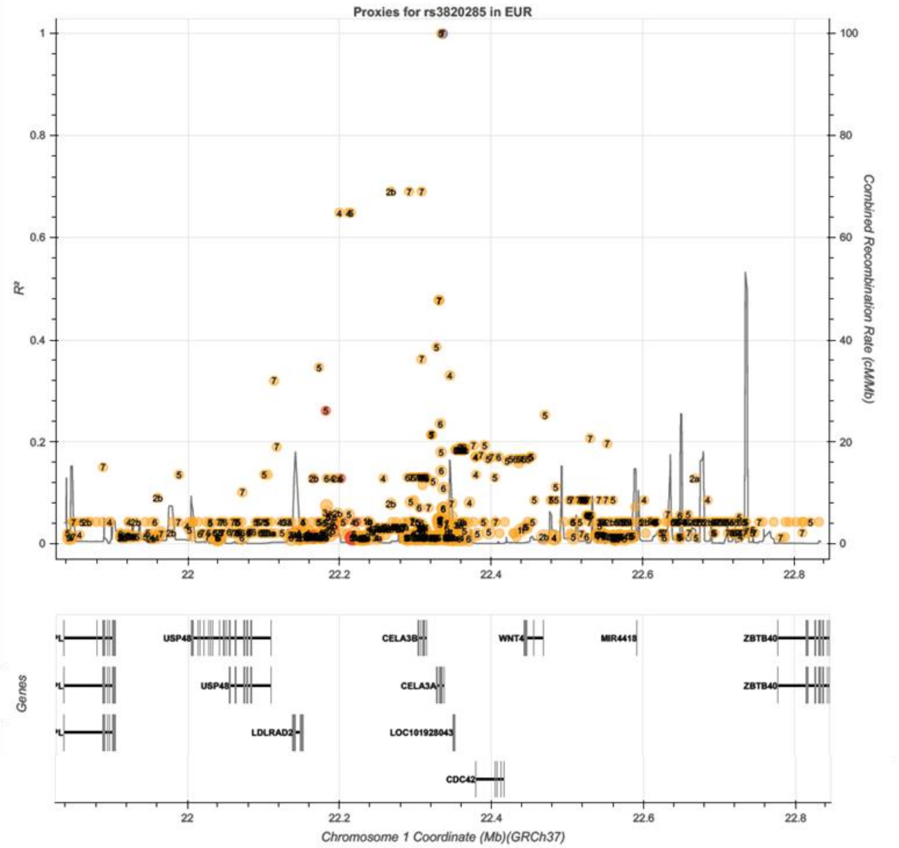


Figure 3.10. LD Plot of Rs3820285. Rs3820285 is located at R^2 = 1

Another potential variant, rs77373930, first appeared to be a false positive on the green META GWAS on chromosome 2 due to it only being found in four of the six datasets and because aside from it being in the middle of a lncRNA, it is in a gene desert. Rs77373930 is an intronic variant located within *AC007389.1*, a lncRNA (See Figure 3.11 A). However, as rs77373930 was also found on the Factor Reduced META GWAS (See Figure 3.11 B) a more thorough investigation was performed. As lncRNA often have regulatory roles, the lncRNA was searched on FANTOM CAT[136] to attempt to elucidate a function. Unfortunately, the FANTOM query did not have any records of co-expressed mRNA and suggested that the lncRNA was most expressed in the middle temporal gyrus and likely associate with obesity[137]. A query of the closet adjacent coding gene, *MEIS1* (~460kb upstream) found that knockouts of the gene in mouse models resulted in embryos that had partially duplicated retinas and smaller than average lenses[138]. Thus, while the connection to *MEIS1* is weak, the area may still be worth exploring.

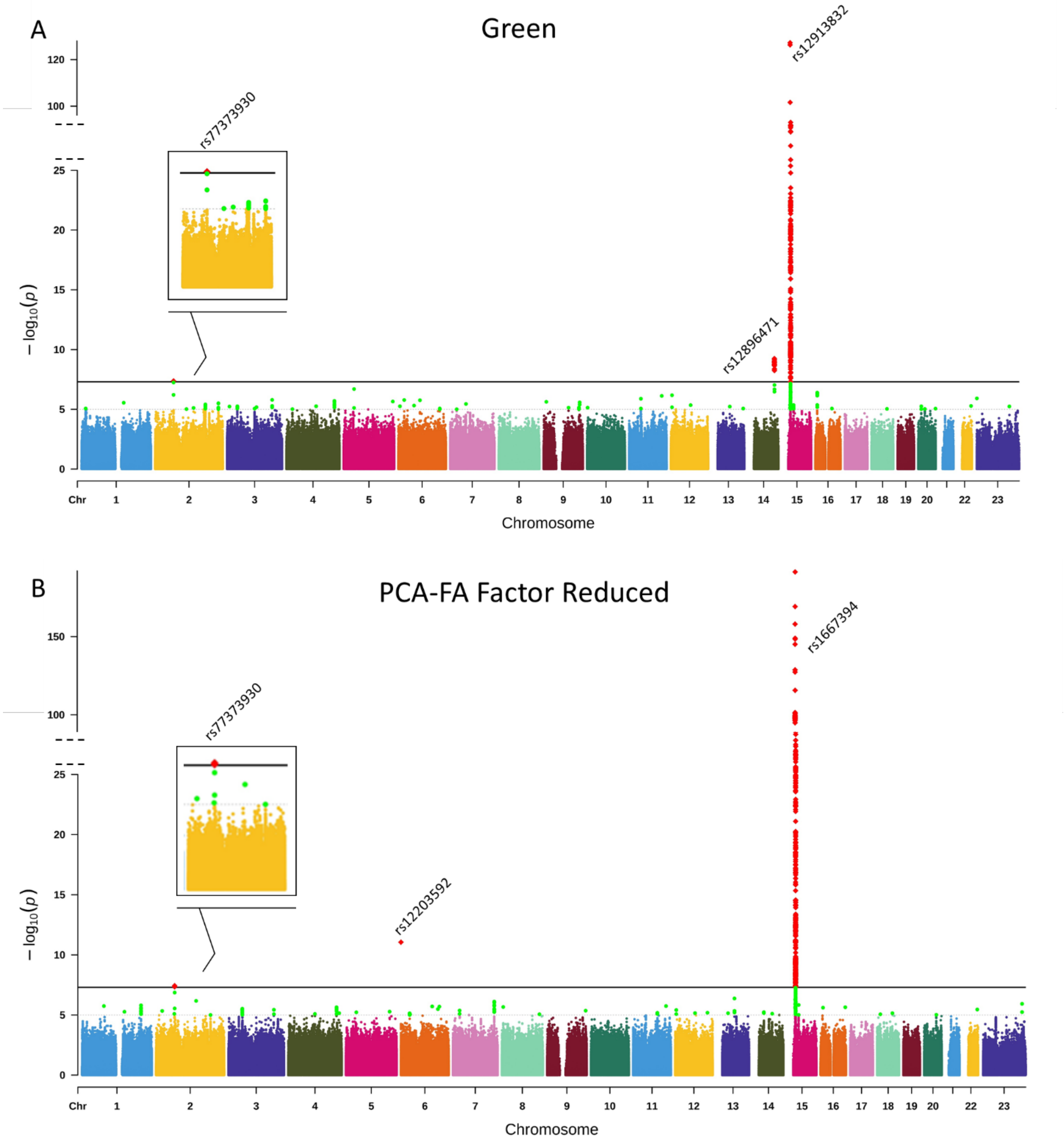


Figure 3.11. Iris META Manhattan Plots for Green and PCA-FA Phenotypes. A) Green Iris META and B) PCA-FA Factor Reduced META Manhattan plots. Near genome-wide significant variants ($p < 1 \times 10^{-5}$) are illustrated in green and genome-wide significant variants ($p < 5 \times 10^{-8}$) are shown in red

Finally, there was a significant hit on the dark brown META GWAS on chromosome 17 (See Figure 3.12). Rs6420484 which was only found in the IUPUI and QIMR cohorts, is an intronic variant that lies at the intersection of *NPLOC4* and *TSPAN10*. Importantly, variants in this area, such as rs9894429, have been found previously in eye GWAS and are in slight LD ($R^2 = 0.46$) with rs6420484[108]. However, in Liu et al., 2010, it is important to note that the researchers were quantifying eye color via averaging hue, saturation, and value measurements. While rs6420484 is located in *NPLOC4* and *TSPAN10* it is also in LD with several variants that are in neighboring *PDE6G* (See Figure 3.13). *PDE6G* encodes the gamma subunit of cyclic GMP-phosphodiesterase, is expressed in rod photoreceptors in the eye, and when mutated has also been shown to be involved in the development of Retinitis Pigmentosa, a disorder that results in abnormal pigment production in the retina[139,140]. No novel variants were found running the light brown META (See Figure 3.14), but the hits on chromosome 6 (*IRF4*), 11 (*TYR*), 14 (*SLC24A4*), and 15 (*HERC2*), were found in previous GWAS analyses[11,61–64].

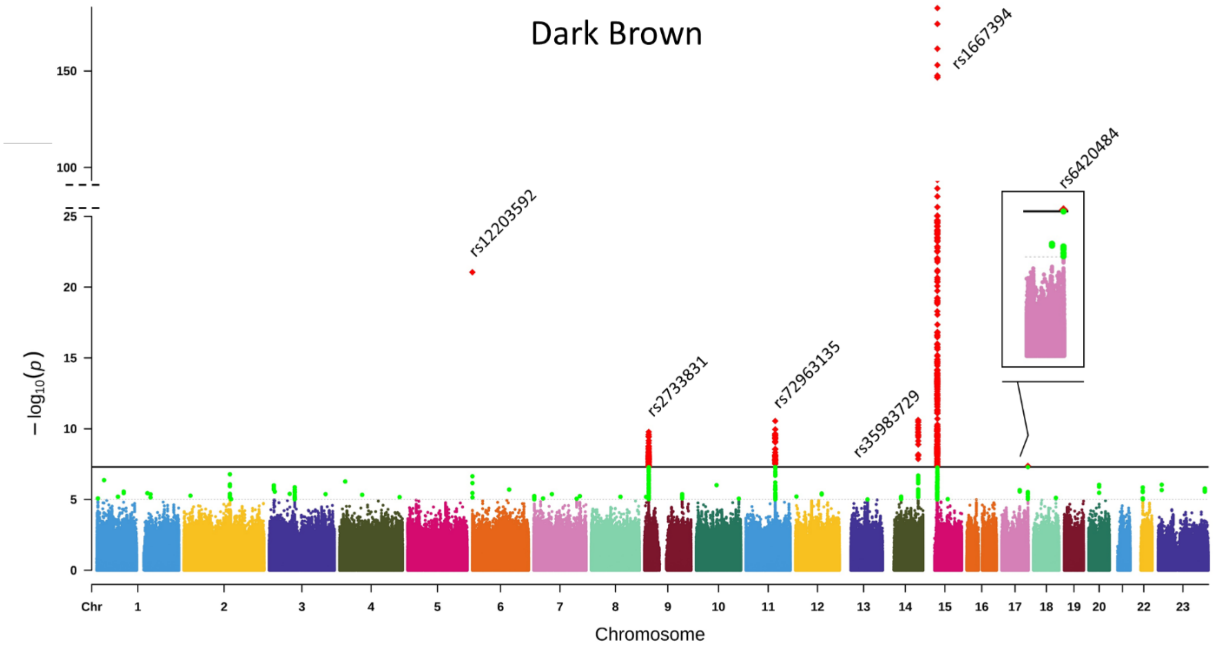


Figure 3.12. Iris META Manhattan Plots for Dark Brown Irises. Near genome-wide significant variants ($p < 1 \times 10^{-5}$) are illustrated in green and genome-wide significant variants ($p < 5 \times 10^{-8}$) are shown in red

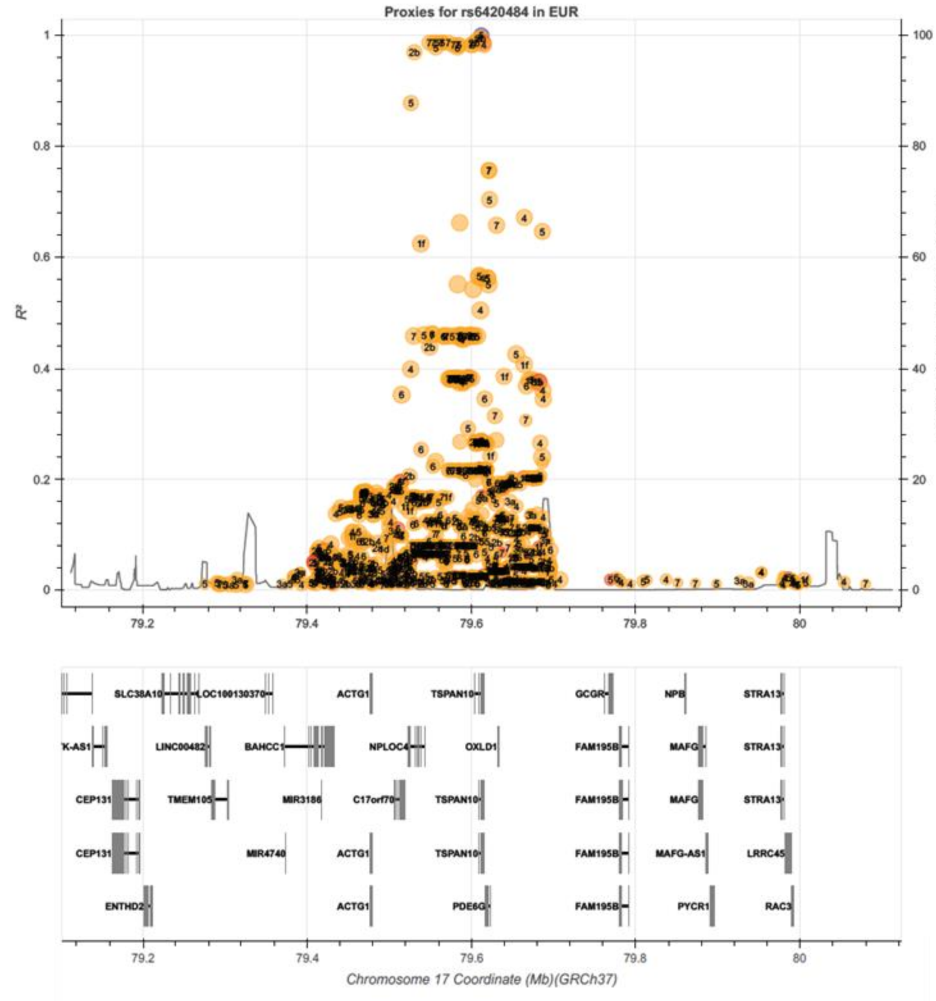


Figure 3.13. LD Plot of rs6420484. Rs6420484 is shown in blue

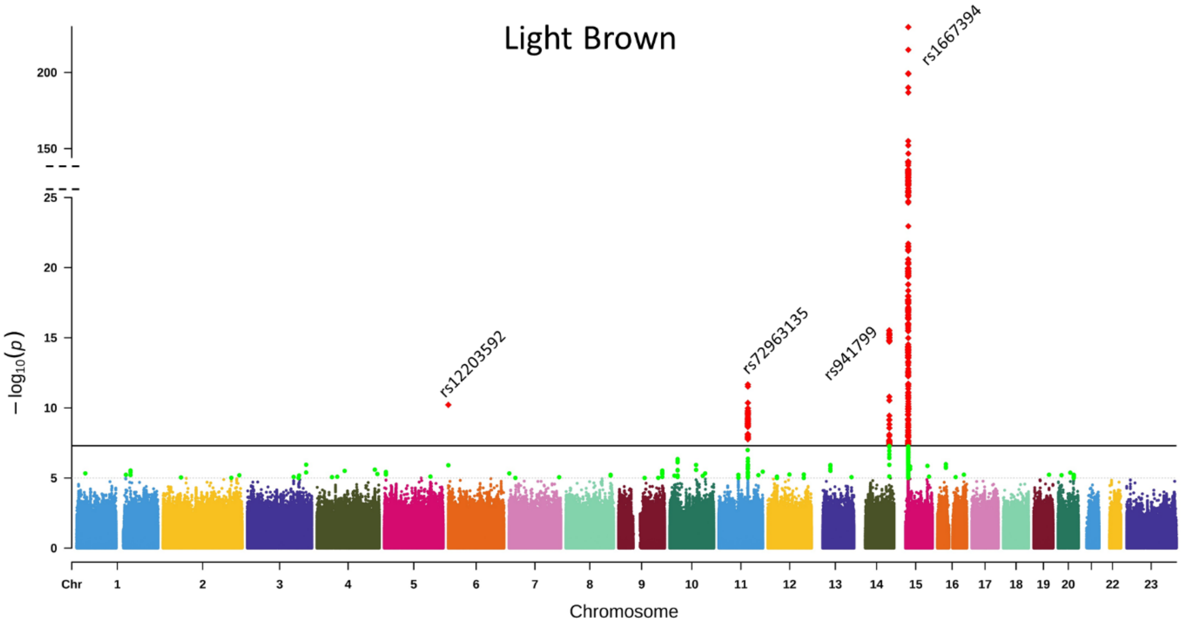


Figure 3.14. Iris META Manhattan Plots for Light Brown Irises. Near genome-wide significant variants ($p < 1 \times 10^{-5}$) are illustrated in green and genome-wide significant variants ($p < 5 \times 10^{-8}$) are shown in red

3.3.2 Evaluation of Quantitative Categorical Iris Color Prediction Model

Initially, we planned to build a prediction model based on the known iris pigmentation variants found in the literature in addition to the novel hits that we would find in the aforementioned GWAS. However, due to the weak strength of the novel hits found, we elected to proceed with building an iris prediction model using already known pigmentation variants to predict a quantitative output based on our color class phenotype scale. Therefore, after setting up the three-layer neural network, the model's performance was assessed through statistical measurements evaluating accuracy as well as 'visually' evaluating the model to assess how applicable it may be in forensic or anthropologic scenarios. The first neural network model, which will be referred to as the cross-validated (CV) model, was generated to assess performance via 10 cross-validations (80% training and 20% testing) of the full global cohort of 3265 individuals. Its performance metrics can be seen in Figure 3.15 and Table 3.5. These models were able to achieve a mean absolute error (MAE) of 13.57% +/- 0.36% overall across all eye color predictions. Individually the model can predict the percent quantity of blue irises with 84.4% (+/- 0.06%) accuracy, 95.0%

(+/-0.4%) for green, 82.3% (+/- 0.5%) for light brown, and 84.0% (+/- 0.8%) accuracy for dark brown. The overall RMSE was 0.1927 (+/- 0.0044). The full set of training individuals was then utilized to generate a full version of the model, denoted as the Full model, using the same parameters as the cross-validated model apart from increasing the epochs to 50000 from 5000, which as Figure 3.15 demonstrates, was negligible to increasing model performance. This model achieved a MAE of 11.5% overall across all eye color predictions (See Table 3.5).

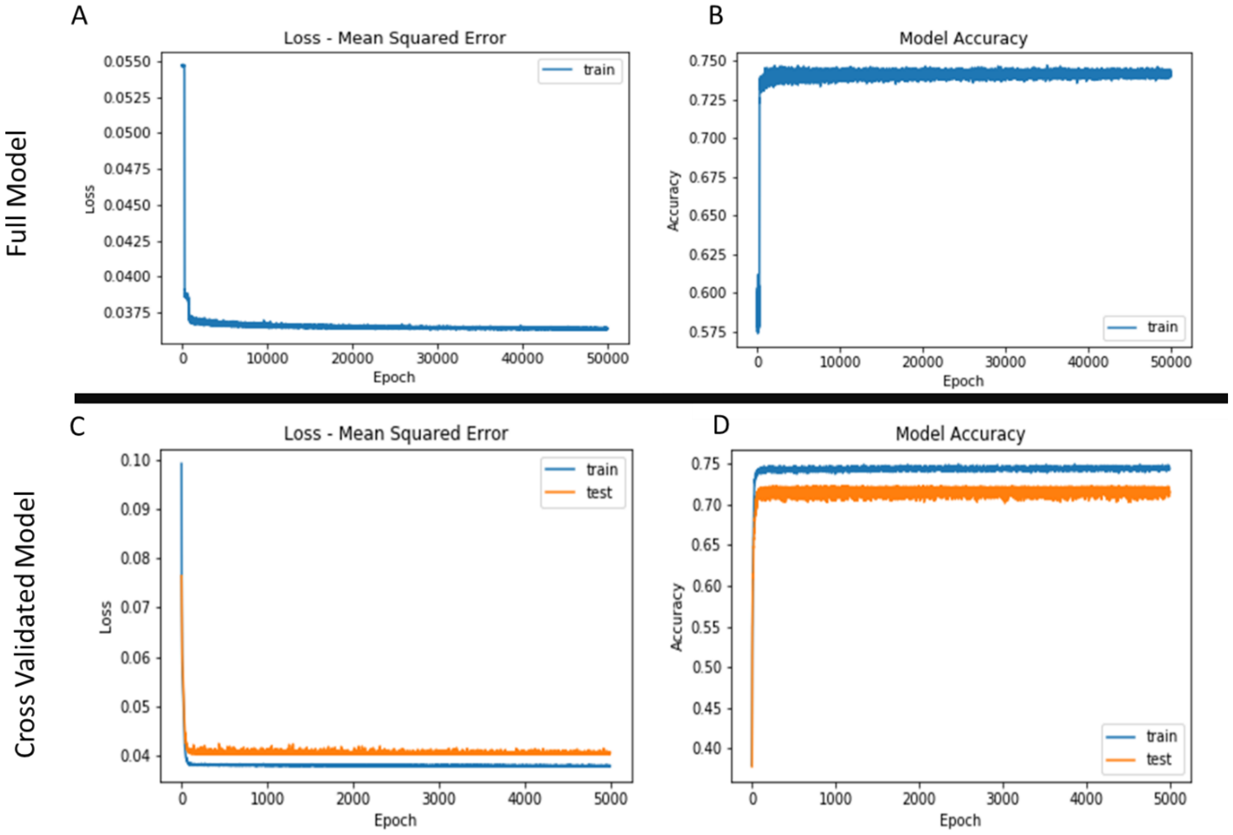


Figure 3.15. Neural Network Model Performance Per Epoch. Full Model A) loss and B) accuracy are shown per epoch. Ten-fold cross validated model C) loss and D) accuracy are also shown between the training (blue) and testing (orange) set

Finally, the Full model was used to visually inspect its prediction with that of the actual iris image. However, as the model only outputs percentages of color categories, the top five closest quantitated training iris images to that of the prediction were extracted. To accomplish this, we took the model prediction (i.e. a series of 4 color percentages) and then created a difference matrix between the prediction and the actual quantified percentages of the 3265 individuals that were in the testing set. By summing the errors between each prediction and actual color type the five smallest total errors were extracted, which allowed us to output the most similar quantified iris image. By observing three of these ‘predicted’ images next to the actual image, the model’s performance could be visually assessed (See Figure 3.16). While evaluating the model, it was possible to see that it was able to predict the general color of the iris from the five images that were visually tested. This visual check supports the cross-validation metrics that the model is able to quantitatively predict categorical eye color with a decent accuracy. The model does struggle with green eyes and also irises that have a combination of blue and brown. As one final benchmark the Full model was compared against the Irisplex model[7–9]. In the comparison, six variants were input as instructed on the HIrisPlex-S website (https://hirisplex.erasmusmc.nl/) for each of the test subjects. The Full model seemed to predict similarly to that of the IrisPlex model (see Cat Prediction in Table 3.6). Irisplex was better at predicting the iris color extremes of blue and brown better, as seen by its blue and brown predictions of individuals #2, #4, and #5. On the other hand, the neural network correctly predicted #2 and #4 to have primarily blue and dark brown iris color, but incorrectly predicting #5 as having more of an intermediate phenotype. However, the neural network appears to better predict intermediate colors with a higher accuracy especially in individual #1 while also predicting the correct lighter shade of light brown in individual #3.

Table 3.5. CV and Full Model Performance Metrics

	CV		Full*
Performance Metrics	**Mean**	**Stdev**	**Mean**
MSE	0.037	0.002	0.023
RMSE	0.198	0.004	0.151
R2	0.427	0.015	0.554
MAE Overall	0.136	0.004	0.115
MAE Blue	0.156	0.006	0.167
MAE Green	0.050	0.004	0.051
MAE Light Brown	0.177	0.005	0.185
MAE Dark Brown	0.160	0.008	0.058

* Note: Full Model Performance Metrics were carried out on a testing set of 5. Thus, they are reported for completeness, but true model performance should be assessed using the Cross-Validation Model

Table 3.6. Comparison of the Neural Network Model and Irisplex

	Blue			Light Brown		Intermed	Green		Dark Brown			Cat Prediction	
ID	**Actual**	**NN P**	**IrisPlex P**	**Actual**	**NN P**	**IrisPlex P**	**Actual**	**NN P**	**Actual**	**NN P**	**IrisPlex P**	**NN P**	**IrisPlexP**
1	31.7%	41.8%	0.884	57.7%	33.5%	0.073	3.0%	10.7%	7.6%	14.1%	0.044	Intermed	Blue
2	94.5%	69.0%	0.948	1.1%	13.0%	0.038	0.2%	7.6%	4.2%	10.4%	0.014	Blue	Blue
3	1.5%	5.2%	0.050	83.2%	61.9%	0.114	0.0%	1.4%	15.3%	31.5%	0.836	Intermed	Brown
4	3.0%	20.2%	0.000	20.9%	19.6%	0.070	0.0%	-0.1%	76.1%	60.3%	0.993	Brown	Brown
5	74.7%	40.9%	0.884	4.7%	34.1%	0.073	15.8%	10.8%	4.9%	14.2%	0.044	Intermed	Blue

* Note: Neural Network (NN), P (Percentage of color – NN; Probability of Color – Irisplex)

**Note: Cat Prediction is a categorical prediction in which the highest percentage of color is reported as being the iris color

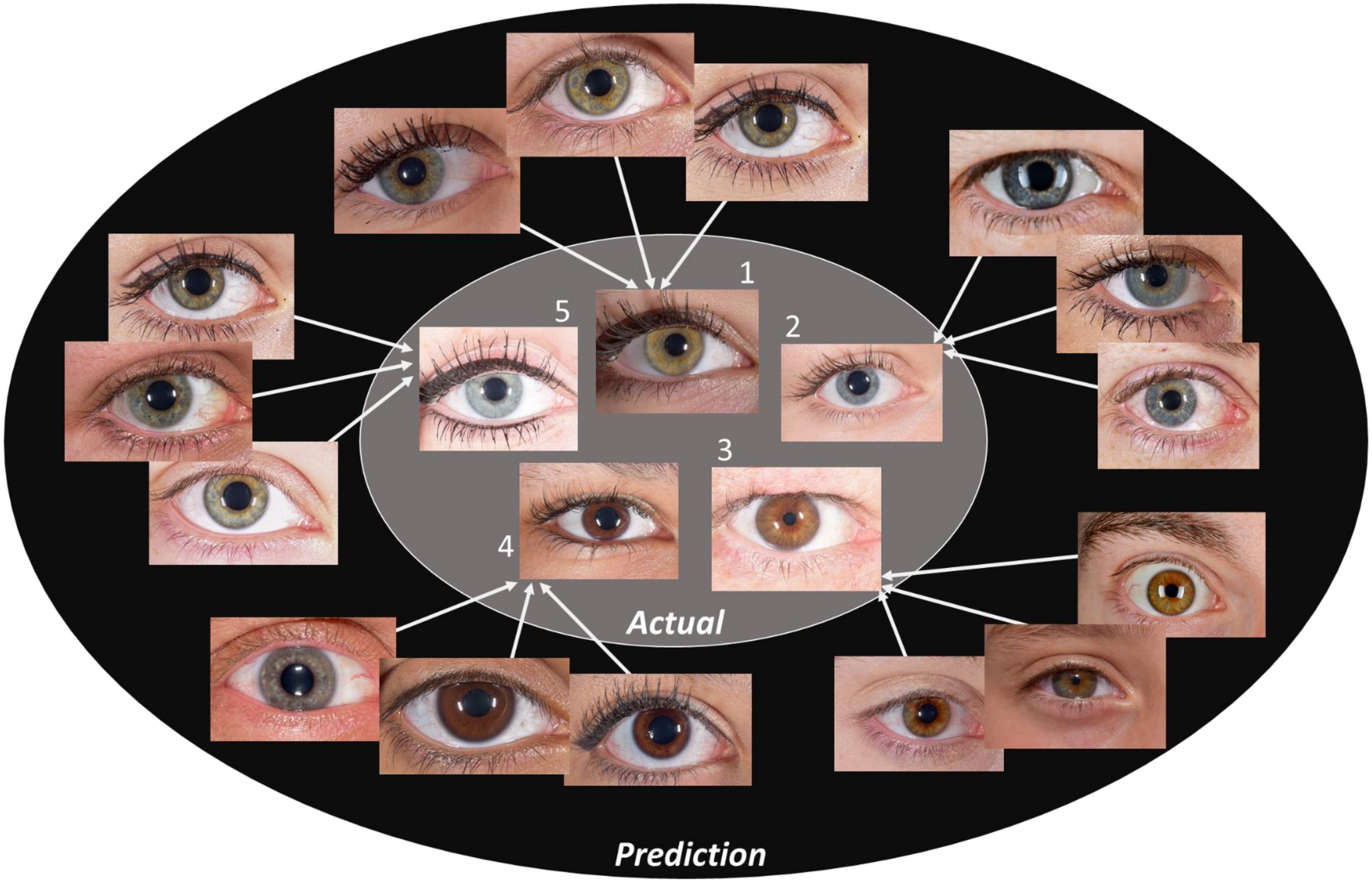


Figure 3.16. Visual Quantitative Prediction of Categorical Color. Five irises of known color (Actual) were input into the Full Model. A prediction of blue, green, light brown, and dark brown iris percentage were returned. These predicted values were then compared against quantitated percentages from the IUPUI testing dataset. Testing individuals with quantitated percentages closest to the prediction (calculated by summing the errors between all four-color categories) were output (prediction) to give a visual representation of the model's prediction. Note that the ID's of the individuals in the table match the figure

3.4 Conclusion

In conclusion, the META analysis conducted on nearly 6900 individuals has identified both known variants that have shown to contribute to iris color as well as some potentially new variants that may warrant follow-up analysis via replication studies, functional studies, or both. I have also, for the first time, produced a model that is able to predict quantitatively pre-set categorical color measurements, which can be converted to a physical image.

CHAPTER 4. FINDING NEW VARIANTS ASSOCIATED WITH CATEGORICAL FACIAL (MANDIBLE) DEFINITIONS

4.1 Introduction

Human facial morphology has been widely understood through the course of twin and disease-based studies that have explored the genetic influence of diseases such as cleft lip and palate, Down syndrome, and Prader-Willi syndrome. Through these studies, facial morphology has been shown to be a highly heritable trait[24,66]. However, until recently, knowledge of the genetic factors that influence non-diseased phenotypes has been limited. One reason for this is likely due to the face's multifaceted and complex structure. Nearly all of the facial morphology studies thus far have attempted to quantify facial shape using Euclidean distance measurements or a derivative of a distance metric either on individual points or on groupings of points[24–27]. Several studies have attempted to construct a more wholistic phenotype by combining either 2D or 3D landmarks to create face shapes whose pair-wise distance can be calculated to yield a distance phenotype[74,141]. To date, this has proved successful in identifying numerous variants and genes associated with the facial phenotype, such as *PAX3*, *PRDM16*, and *C5orf50*. However, as facial phenotypes are the result of complex embryologic developmental it should not be surprising that assessing single variant associations with distances calculated between two facial landmarks or even groupings of landmarks may limit the potential for discovery. A more recent approach by Claes et al., 2018 and distributed in White et al., 2019 examines a more holistic and quantitative view of facial morphology, in which a 3D facial mesh is broken up into highly correlated segments via a data-driven hierarchical spectral clustering methodology. These segments are subsequently analyzed individually (i.e. there are no Euclidean distance measurements) via a generalized Procrustes analysis, Principal Component analysis, a canonical correlation analysis, and ultimately a general linear model based GWAS [28,29].

Amid all these complex facial phenotyping methods, many standard approaches to GWAS's conducted on other phenotypes began with a simplified phenotype, which evolved into something more complex. For example, 'early' iris color GWAS's phenotyped eye color on the basis of a three-point color scale: blue, intermediate, and brown[3–6]. Quantitative iris color phenotyping was adopted only after exhausting the capabilities of the categorical approach. However, no facial

GWAS has been conducted on a simple categorical face shape that explores genetic contributions to the face holistically or to the jawline. Thus, the aim of this research was to analyze whether a simplified categorical phenotype of face and jaw shape could identify novel face-shape variants that may have been missed by the more complex facial phenotyping methods previously applied.

4.2 Materials & Methods

4.2.1 Samples & Genotyping

4.2.1.1 IUPUI Dataset

Refer to section 3.2.2.1 IUPUI Dataset Collection and Organization.

4.2.1.2 Penn State University (PSU) Dataset

The PSU dataset partitioned for this study includes 1760 individuals that contained 2D images and genotypes. Additional details of participant recruitment can be found in Section 3.2.2.2. PSU participants were genotyped by 23andMe on the v3, (900K SNPS) and v4 arrays (600K SNPS) (Mountain View, CA).

4.2.2 Phenotyping for Categorical Analysis

Study participants from both the IUPUI and PSU dataset were independently graded by three lab members into four broad facial categories; oval, round, square, and diamond (See Figure 4.1). Phenotypes were chosen based on face shape phenotypes commonly referred to by plastic surgeons[142,143]. Of particular note, we combined the ‘heart’ and ‘diamond’ face shape into just ‘diamond’ for grading simplicity. To train the raters to achieve high inter-rater correlation, three raters cycled through a set of approximately 50 images and discussed the features of each face that best fit the grading criteria. Once an understanding of the criteria was established, the raters independently rated all the PSU and IUPUI facial images. Any picture in which all three graders disagreed on a facial category, was dropped from the analysis, while pictures that had two or more raters in agreement were kept, with the agreed classification being assigned to that picture.

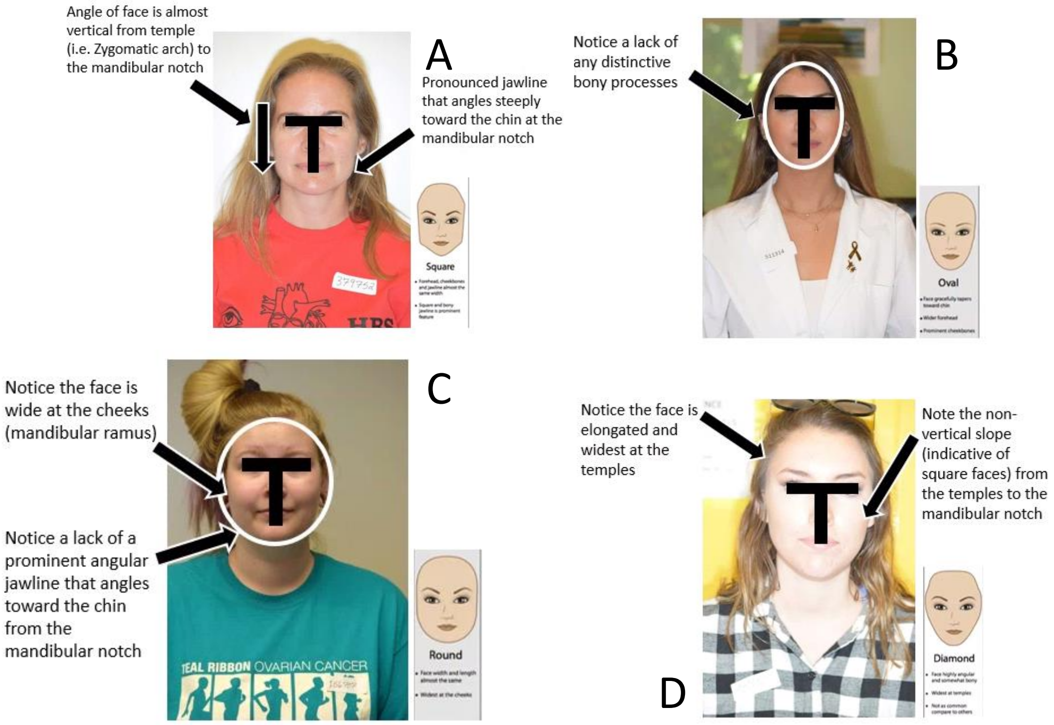

Angle of face is almost vertical from temple (i.e. Zygomatic arch) to the mandibular notch
A
Pronounced jawline that angles steeply toward the chin at the mandibular notch
Square
Forehead, cheekbones and jawline almost the same width
Square and bony jawline is prominent feature
Notice a lack of any distinctive bony processes
B
Oval
Face gracefully tapers toward chin
Wider forehead
Prominent cheekbones
C
Notice the face is wide at the cheeks (mandibular ramus)
Notice a lack of a prominent angular jawline that angles toward the chin from the mandibular notch
Round
Face width and length almost the same
Widest at the cheeks
Notice the face is elongated and widest at the temples
Note the non-vertical slope (indicative of square faces) from the temples to the mandibular notch
Diamond
Face highly angular and somewhat bony
Widest at temples
Not as common compare to others
D

Figure 4.1. Categorical Facial Shape Evaluation Criteria. Criteria is shown for A) square, B) Oval, C) Round, and D) diamond facial shapes. Each rater had a face model and a real-life facial example to refer back to when making ratings

4.2.3 Quality Control and Imputation

Quality control practices used to prepare the IUPUI dataset for imputation were performed using the GRCh37 (hg19) genome assembly[117]. Quality control included filtering out individuals who had more than 5% of their genome missing, variants that were missing in more than 5% of the dataset, individuals whose phenotypic data did not match their genotypic data, and related individuals (i.e. identity by descent greater than 0.1875 or the halfway point between second and third degree relatives). Individuals who had unusually high heterozygosity (+/-3 standard deviations), were also excluded. In total 3528 individuals were sent for imputation via the *Odyssey* pipeline (See Chapter 2), using SHAPEIT2[119] and IMPUTE4[84] to phase and impute respectively. A custom reference genome combining the Haplotype Reference Consortium[88] and the 1000 Genomes Phase 3[30] were used for both phasing and imputation, which resulted in the imputation of 43.9M variants. After imputation, variants were filtered based on the imputation quality control INFO metric (INFO score > 0.3). Prior to GWAS, the dataset was pruned based on SNP missingness (missingness < 5%), minor allele frequency (MAF > 1%), and Hardy-Weinberg equilibrium ($p < 1 \times 10^{-5}$) to yield approximately 6.7M variants for analysis.

As the PSU dataset was collected from multiple arrays, imputation was performed by our collaborator at PSU separately for each platform and then combined following Verma, et al. 2014[144]. For each dataset, standard data cleaning and quality assurance practices were performed based on the GRCh37 (hg19) genome assembly[145]. The genotypes were "harmonized" with 1000 Genomes Project (1000G) Phase 3[30] using Genotype Harmonizer[146] with a window size of 200 SNPs, a minimum of 10 variants, and alignment based on minor allele frequency (--mafAlign 0.1). This program was also used to filter out ambiguous SNPs, update the SNP id, and update the reference allele as needed, all in reference to the 1000G Phase 3 genotypes. After genotype harmonization, additional QC metrics, such as relatedness and ancestry analyses were performed. Prior to GWAS, the dataset was again pruned based on SNP missingness (missingness < 50%), minor allele frequency (MAF > 1%), and Hardy-Weinberg equilibrium ($p < 1 \times 10^{-6}$) to yield approximately 9.4M variants for analysis.

4.2.4 Ancestry Analysis

Ancestry analysis for the IUPUI dataset was performed in an identical manner as explained in Section 3.2.4. 1081 individuals were removed in this fashion.

For the PSU dataset from the post-imputation merged datasets, individuals containing primarily European ancestry were determined by projecting them into a principal component (PC) space constructed using the 1000G Phase 3 dataset. To do this, all indels, multi-allelic SNPs, and SNPs with MAF ≤ 0.1 in both the 1000G dataset and the PSU dataset were excluded. Those variants that were common to the 1000G and the merged dataset were used in the projection. On this list of variants, linkage disequilibrium (LD) pruning (50 bp window, 5 bp step size, 0.2 correlation threshold) was iteratively performed on the 1000G dataset until no variants were excluded. This LD-pruned list (n = 461,372 SNPs) was then used in a principal component analysis (PCA) to construct a population structure space based upon the 1000G project and projected the dataset onto that PCA space to obtain the ancestry axes of the dataset. Once in a combined PC space, Euclidean distance was calculated between all participants and the 1000G samples. Using a k-th nearest neighbor algorithm, the five nearest 1000G sample neighbors for each US participant was identified. The most common 1000G population label (e.g. CEU, GIH, YRI) from these five nearest neighbors was then assigned to the participant. Participants with the 1000G European population labels of CEU, TSI, FIN, GBR, and IBS were then selected for analysis.

4.2.5 Genome Wide Association Study (GWAS) and META Analysis

A series of categorical and one multinomial GWAS on the five categorical facial shapes of oval, round, diamond, square, and a combination of oval and round were performed at both the IUPUI and PSU sites using their respective data. After ancestry exclusion and other forms of quality control on these datasets, the total number of individuals were 1642, and 2149, respectively. In the IUPUI cohort a logistic model was created that was adjusted for age and sex. An additional genotype missing filter was performed to exclude any genotypes that were missing in more than 5% of the dataset as well as filters for HWE ($p < 0.00001$) and minor allele frequency (MAF > 1%). The PSU cohort was analyzed in an identical manner by our collaborator with the only difference being the inclusion of BMI as an additional covariate.

A meta-analysis was carried out with the METAL program[122] using Stouffer's method[123] on the PSU and IUPUI result for each of the 5 categorical facial phenotypes (See Table 4.1). As different genotyping platforms can use alternative naming schemes, the variants were renamed by chromosome, position, and minor allele. The genotypes were coded additively based on the presence of the minor allele and were consistent between cohorts.

Table 4.1. Categorical Facial Shape META Analysis Metadata. A breakdown of the META analysis listing the two META datasets by variants tested, cases (the facial shape being tested), and controls (the face shapes not being tested)

Phenotype	IUPUI META Dataset 1			PSU META Dataset 1		
	Variants	Cases	Controls	Variants	Cases	Controls
Diamond v Others	6.7M	227	1735	9.4M	277	1365
Square v Others	6.7M	530	1432	9.4M	322	1320
Round v Others	6.7M	626	1336	9.4M	268	1374
Oval v Others	6.7M	579	1383	9.4M	775	867
R/O v D/S*	6.7M	757	1205	9.4M	599	1043

**Note: Round/Oval (R/O) were considered cases and Diamond/Square (D/S) were considered controls*

4.3 Results & Discussion

Due to a lack in power, as evident by numerous signals which failed to reach the significance level of the family-wise error rate (FWER) adjusted threshold of 5 x 10^{-8}, a META analysis (n = 3791) consisting of a combination of the results generated from the IUPUI dataset (n = 2149) and the PSU dataset (n = 1642) was conducted. It is evident that the Round, Oval, and Oval/Round versus Square/Diamond analysis did not return any significant signals as illustrated in Figures 4.2 and 4.3.

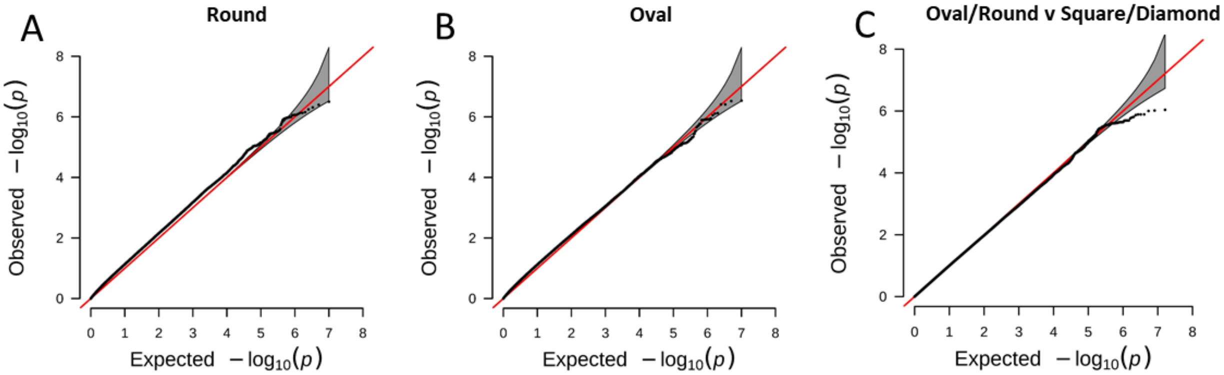


Figure 4.2. Quantile-Quantile Plots for the Round, Oval, and Oval/Round Face Shape Analyses. The META analysis results of A) Round, B) Oval, and C) Oval/Round v Square/Diamond are shown. The distance of the point from the red line, y = x, measures the deviation of the distribution from the -$\log_{10}$(p) that was expected under a gaussian distribution vs what was observed. The grey shaded area is the 95% confidence interval, and points in red indicate significant p-values (p < 5 x 10^{-8})

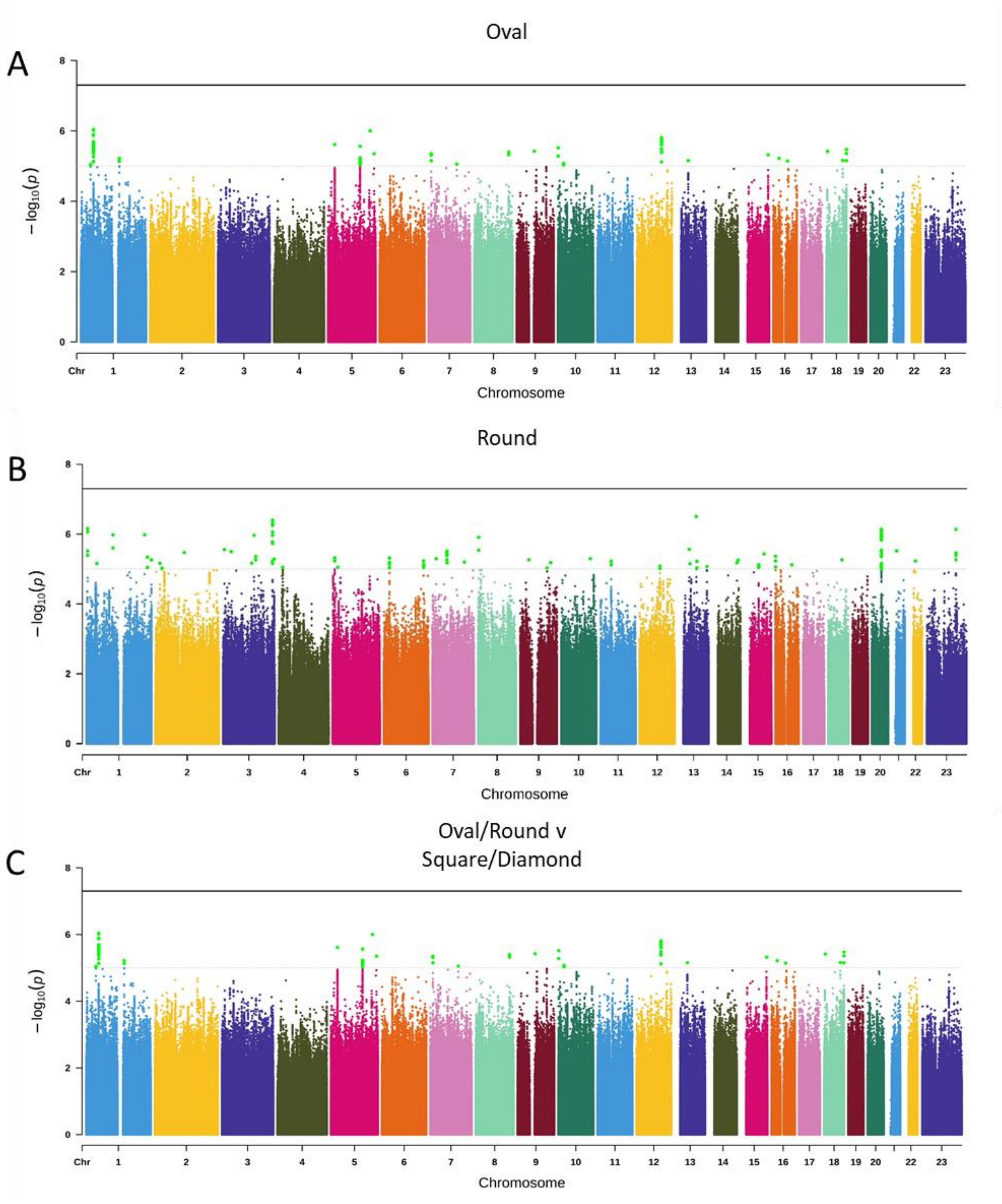


Figure 4.3. Manhattan Plots for the Round, Oval, and Oval/Round Face Shape Analyses. Manhattan plots of A) Oval, B) Round, C) Oval/Round v Square/Diamond. Near genome-wide significant variants ($p < 1 \times 10^{-5}$) are illustrated in green and genome-wide significant variants ($p < 5 \times 10^{-8}$) are shown in red

However, we did find several signals, illustrated in Figure 4.4 and reported in Tables 4.2 and 4.3, that may warrant a more in-depth look. The first variant seen in a cluster of three hits within the square META GWAS on chromosome 2 is rs187236608, an intron variant located within *STON1*. *STON1* encodes one of the two human homologs of the *Drosophilia melanogaster* Stoned B protein, which is involved in the formation of components within endocytic machinery[147]. While functionally, there does not seem to be much evidence that *STON1* plays a substantial role in bone development, while searching the literature it was found that *STON1* was formerly associated in a GWAS that reduced 276 facial linear distances using factor analysis[141]. However, in that study the variant was rs76889437, it was near significant ($p = 3 \times 10^{-6}$), and it was also suggested to be associated with *STON1*. Linkage disequilibrium (LD) analysis shows that rs187236608 is in high LD ($r^2 = 1$) with variants found on the neighboring *LHCGR* gene (See Figure 4.5).

Drawing on the connection of this peak signal with *LHCGR*, the second variant of the trio cluster in the square face analysis is rs111431304, which is located 106kb downstream of rs187236608 and is found within an intron on *LHCGR*. This gene encodes for a lutropin-choriogonadotropic hormone receptor[148]. Mutations of this receptor have led to Leydig Cell Hypoplasia Type I, due to the receptor being an integral part of Leydig cell development and, by extension, testosterone[149]. Leydig Cell Hypoplasia has been known to retard pubertal development and delayed bone maturation[150] in young males as well as contribute to bone health and strength in older males[151]. The literature also suggests that females may be affected by alterations of the *LHCGR* gene. Yarram et al., 2003 specifically demonstrate that mice knockout (KO) models of the luteinizing hormone receptor results in decreased bone mineral density, caused by a reduction in bone formation or an increase in bone resorption in both males and female mice[152].

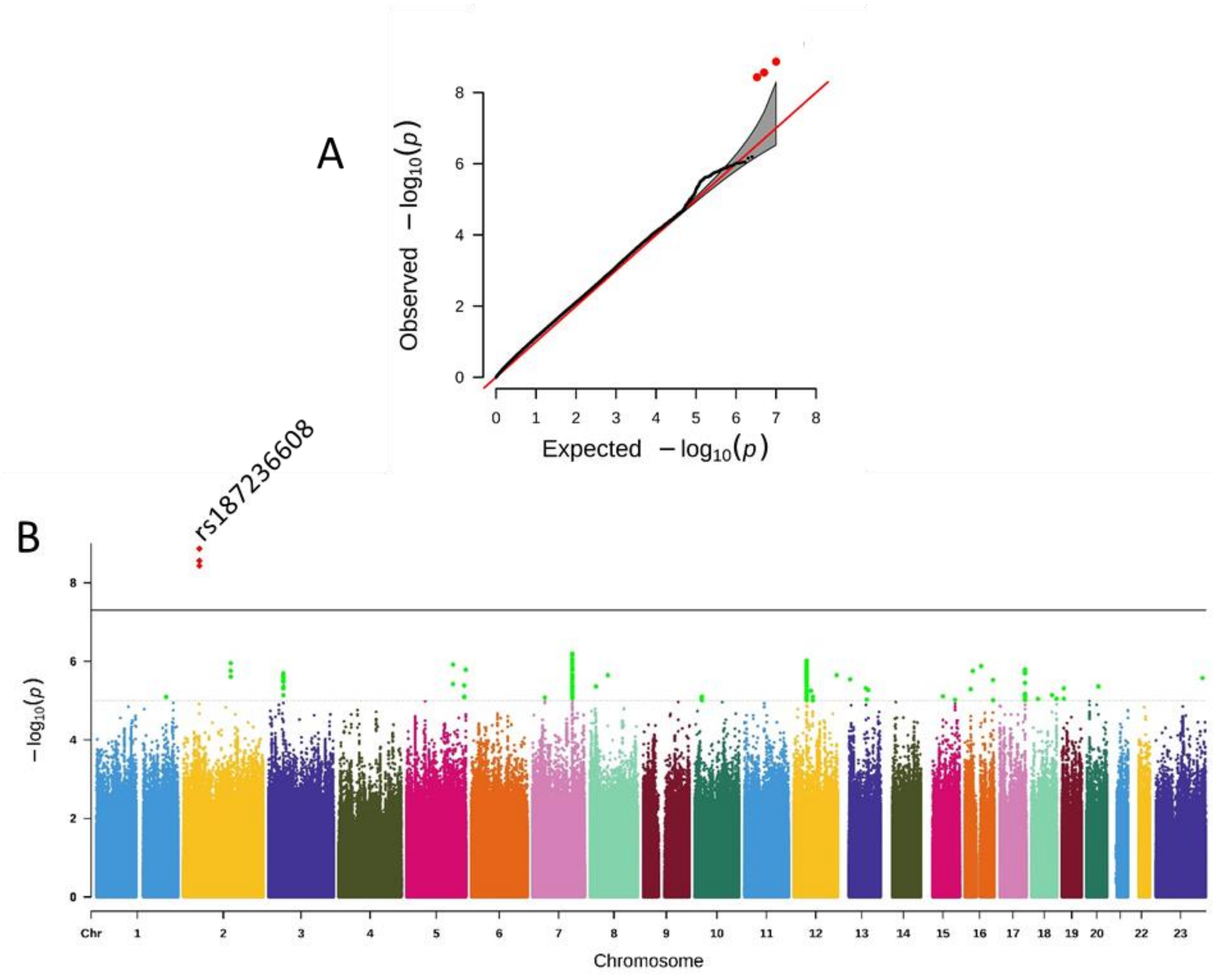


Figure 4.4. Quantile-Quantile and Manhattan Plot for the Square Face Shape Analysis. A) QQ plot. The distance of the point from the red line, y = x, measures the deviation of the distribution from the -log$_{10}$(p) that was expected under a gaussian distribution vs what was observed. The grey shaded area is the 95% confidence interval, and points in red indicate significant p-values (p < 5 x 10^{-8}). B) Manhattan plot. Near genome-wide significant variants (p < 1 x 10^{-5}) are illustrated in green and genome-wide significant variants (p < 5 x 10^{-8}) are shown in red

Table 4.2. Categorically Analyzed Face Shape META Results

GWAS Pheno*	***RSID***	***Chr***	***Position****	***META P***	***IUPUI OR***	***IUPUI P***	***PSU OR***	***PSU P***
Square	**rs187236608**	**2**	**48590930**	**1.35E-09**	**-0.28**	**6.80E-07**	**-0.32**	**3.80E-04**
Square	**rs111431304**	**2**	**48697075**	**2.75E-09**	**-0.30**	**2.83E-06**	**-0.29**	**2.21E-04**
Square	**rs113531385**	**2**	**48629079**	**3.72E-09**	**0.32**	**6.76E-06**	**0.34**	**1.34E-04**
Diamond	**rs59143906**	**18**	**9435944**	**2.92E-08**	**Not Found**	**Not Found**	**0.24**	**2.92E-08**
Diamond	rs16940196	18	13185735	1.52E-07	-1.87	1.55E-05	-1.44	0.00222403

*Note: Bold Denotes Genome Wide Significance (p < 5 x 10^{-8})

**Note: RSID position is based on hg19

Table 4.3. Categorically Analyzed Face Shape META Datamining Results

GWAS Pheno*	***RSID***	***EUR MAF****	***Type***	***Assoc Gene***	***Assoc PubMedID*****	***Assoc Phenotype***
Square	**rs187236608**	**1% (a)**	**Intronic**	**STON1; LHCGR**	**Yes; 28441456**	**Face Shape**
Square	**rs111431304**	**1% (a)**	**Intronic**	**STON1; LHCGR**	**Yes; 28441457**	**Face Shape**
Square	**rs113531385**	**2% (g)**	**Intonic**	**STON1; GTF2A1L**	**Yes; 28441458**	**Face Shape**
Diamond	**rs59143906**	**4% (aaaa)**	**Regulatory**	**No**	**No**	**NA**
Diamond	rs16940196	23% (t)	Upstream Gene Variant	LDLRAD4	Yes; 30048462; 30172743	Hip/Heel Bone Density

*Note: Bold Denotes Genome Wide Significance (p < 5 x 10^{-8})

**Note: Minor allele frequencies found in 1000 Genome Phase 3 European individuals

***Note: References listed as Pubmed IDs

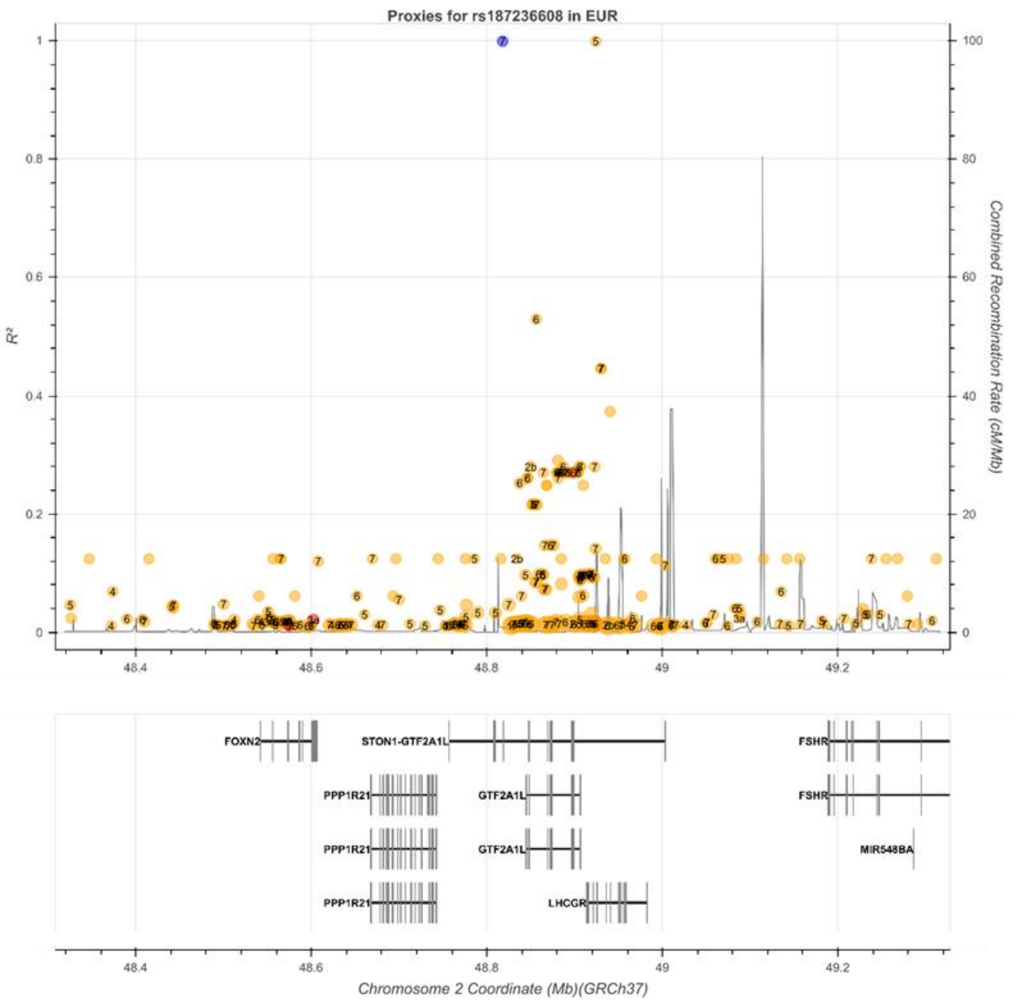


Figure 4.5. LD Plot of Rs187236608. Rs187236608 is shown in blue

The final square face shape hit of the trio is rs113531385, which was found between rs111431304 and rs187236608 and located in an intron of the *STON1-GTF2A1L* gene, which, like *LHCGR*, has been implicated in testes biology[153]. Importantly, all three variants on the square face META have low allele frequencies in Europeans (~1% in 1000 Genome Euro Reference populations and ~1.6% in gnomAD on average). Due to this lack of variance, a certain level of caution should be applied to the validity of these hits. Still, the presence of LD between all 3 hits ($R^2 = 1$) as well as the potential biological evidence may warrant a closer look via functional studies. In addition to its association signal, the variant was found in both the IUPUI and PSU cohorts and had similar directional effects (see Table 4.2 IUPUI and PSU OR).

The peak on chromosome 18 within the diamond META may also warrant further exploration (See Figure 4.6). Rs142553210, which was only found in the PSU cohort, mostly likely indicates that this is a false positive as it is monoallelic in the EUR population of the 1000 Genomes Project[30]. While not genome-wide significant, the closest near-genome wide significant ‘hit’ was rs16940196 ($p = 1.52 \times 10^{-7}$). Rs16940196, was found in both cohorts and was not fixed in European populations (MAF = 23%). Rs16940196 is situated 32kb upstream of *LDLRAD4* which functions as a negative regulator of TGF-beta signaling suggesting that it plays a role in cellular proliferation, differentiation, motility, apoptosis, immunosuppression, and extracellular matrix production[154]. *LDLRAD4* has also been associated with bone mineral density in the hip and heel[155,156]. Furthermore, rs16940196 is within the proximity of the cis-Regulatory Element EH37E1163681 (chr18:13185718-13186266), which has shown high H3K4me3 modification activity (+1.66 standard deviations higher than 210 other cell types measured) in mesenchymal stem cells and decreased levels of H3K27ac modification in osteoblasts (-0.95 standard deviations less than 136 cell types sampled)[130,131]. Therefore, as H3K27ac and H3K4me3 are epigenetic markers commonly associated with gene expression, this may suggest that rs16940196, based on its proximity to this regulatory element may affect bone formation and, by extension, may play a role in craniofacial development. Further functional analyses may add to this evidence.

As seen in Figure 4.3, there did not appear to be any additional significant or highly promising hits worth investigating in this study. However, by refining our analysis, we may be able to boost our power and bring several potential variants up past genome-wide significance.

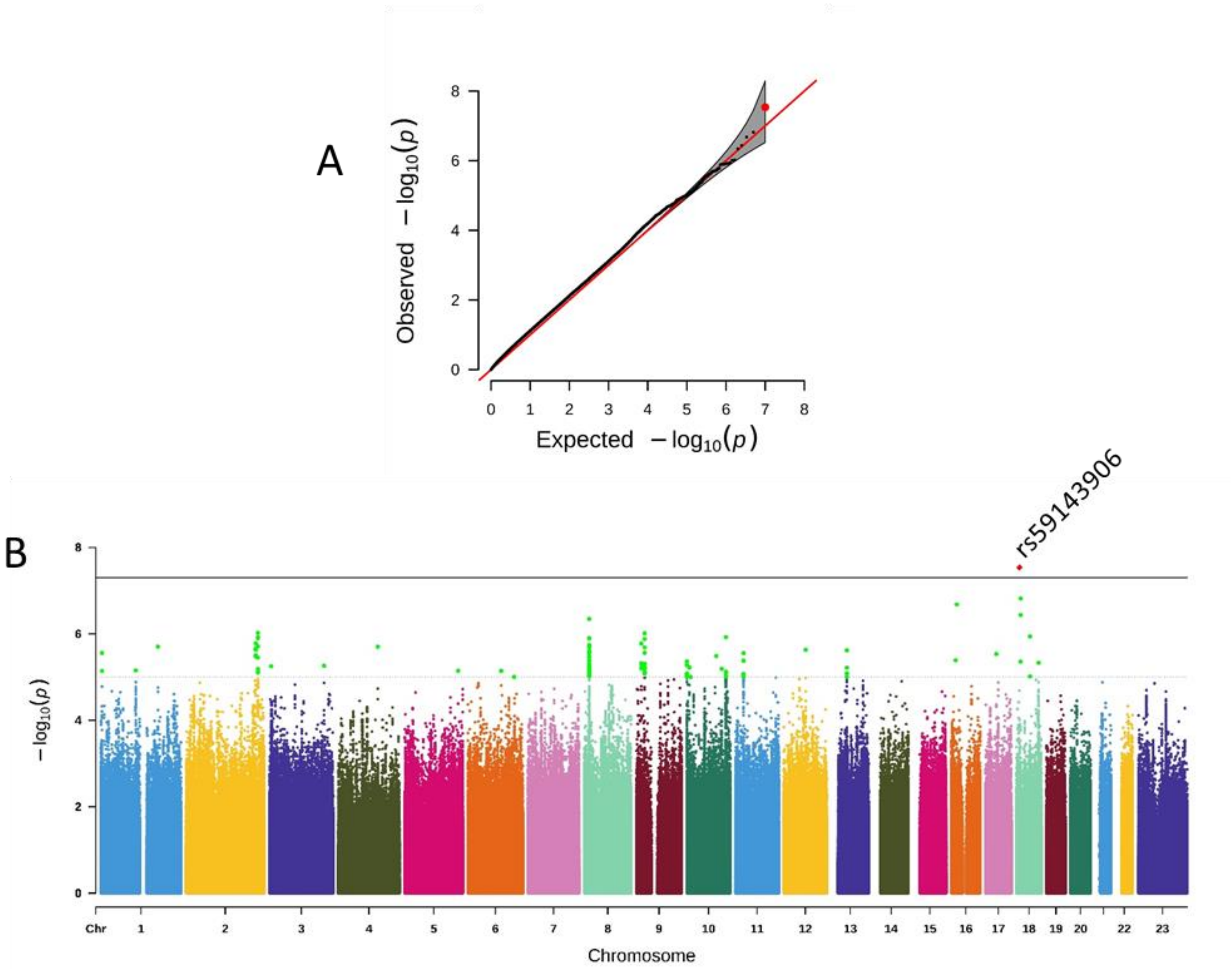


Figure 4.6. Quantile-Quantile and Manhattan Plot for the Diamond Face Shape Analysis. A) QQ plot. The distance of the point from the red line, y = x, measures the deviation of the distribution from the -$\log_{10}$(p) that was expected under a gaussian distribution vs what was observed. The grey shaded area is the 95% confidence interval, and points in red indicate significant p-values ($p < 5 \times 10^{-8}$). B) Manhattan plot. Near genome-wide significant variants ($p < 1 \times 10^{-5}$) are illustrated in green and genome-wide significant variants ($p < 5 \times 10^{-8}$) are shown in red

4.4 Conclusion

In conclusion, these analyses have shown for the first time that a basic phenotype defining the general shape of the face/jawline can be used to identify variants that may play a role in craniofacial development of the mandible and overall bone health. The aforementioned mice knockout models and enhancer regulatory studies add to the association evidence that indicates that the variants identified here, specifically the regions surrounding genes *STON1*, *LHCGR*, *GTF2A1L*, and *LDLRAD4*, warrant further investigation. However, as the META was performed with no validation dataset, it is important to validate these results on an additional, external dataset and/or utilize a functional approach for candidate validation. Still, these analyses have demonstrated that fast, non-technical, and categorical facial phenotyping is powerful enough to detect variants associated with our phenotype of interest and may also be expanded to other identifiable facial features (e.g. hook and button noses).

CHAPTER 5. ASSESSING VARIANT INTERACTION THROUGH STRUCTURAL EQUATION MODELING USING QUANTITATIVE FACIAL DEFINITIONS

Part of “Insights into the genetic architecture of the human face.” Julie D. White, Karlijne Indencleef, Sahin Naqvi, Ryan J. Eller, Jasmien Roosenboom, Myoung Keun Lee, Jiarui Li, Jaaved Mohammed, Stephen Richmond, Ellen E. Quillen, Heather L. Norton, Eleanor Feingold, Tomek Swigut, Mary L. Marazita, Hilde Peeters, Greet Hens, John R. Shaffer, Joanna Wysocka, Susan Walsh, Seth M. Weinberg, Mark D. Shriver, Peter Claes
[Submitted for Review]

5.1 Introduction

Due to the intrinsic nature of GWAS, the phenotype or trait of interest is typically a univariate variable. As mentioned previously in Chapter 4, facial GWAS have slowly evolved from performing GWAS on individual Euclidean distance measurements between facial landmarks, to performing tests on multivariate constructs that encapsulate multiple facial ‘dimensions’ into a ‘singular variable’. Essentially, these first studies have tried to convert a complex phenotype encompassing multiple measurements into a form that a traditional GWAS approach can analyze; a single input vector on a quantitative scale. Zhang et al., 2010 and Denny et al., 2010, have tried to adjust the GWAS analysis itself to accommodate tests on multiple correlated phenotypes[157,158]. Ultimately, these two solutions can be condensed down to a fundamental weakness inherit in most GWAS; that they are best suited to assess single, independent, and highly associative variants with an independent phenotype. GWAS can still be conducted in scenarios that do not fit these idealized parameters, however they often suffer in power as a phenotype that is multifaceted will ultimately cause the correlation signal to be spread out across analyses. Also, a test on a phenotype that is influenced by multiple correlated genotypes, will also cause the signal to be spread across genotypes within the same run. Since the threshold for significance is often set very conservatively ($p < 5 \times 10^{-8}$) to reduce the chance of false positives, these signals are often lost in the background, forcing researchers to perform META analyses in order to boost power and their signals.

Therefore, it is no surprise that a GWAS analysis of a multifaceted polygenic trait such as facial morphology, and the numerous interacting genotypes that regulate the concert of proteins that coordinate craniofacial development suffers from the inability to separate signal from noise. However, the weaknesses of performing GWAS on multifaceted phenotypes like facial morphology can be mitigated by developing novel methods of modeling that captures the face's complex phenotype without muting any of its intricate details. Ideally, if one can capture the multidimensional complexities of the face in combination with modeling the polygenic effects that a traditional GWAS is not capable of analyzing, this may allow the signal to noise ratio to be increased considerably. In this collaborative approach that utilizes a hierarchical spectral clustering method as seen previously in Claes et al.,2018 and White et al., 2019, we sought to understand facial morphology on the basis of measuring and accounting for localized variance within a facial region with a larger sample set than analyzed in Claes et al., 2018. In addition, although this approach could assess a multivariate phenotype, thus capturing a better representation of the human facial structure, it could still only explore the relative role multiple phenotypes have on a single variant at a time. Therefore, through our analysis we utilized structural equation modeling (SEM) on the output of the canonical correlation results in order to inform us of any structural relationship between the significantly associated variants within each mask. In addition, we explored the potential for epistatic interactions between the variants that are competing to express their biological contribution and thus their phenotypic effect within these data-driven facial segments.

5.2 Materials & Methods

5.2.1 Samples, Genotyping & Facial Imagery

The samples used for this study included a combination of three independently collected datasets from the United States (US) and one dataset from the United Kingdom (UK), for a total sample size of 8,246 (after processing and quality control). The US samples originated from Indiana University-Purdue University Indianapolis (IUPUI), the 3D Facial Norms cohort (3DFN), which were collected by the University of Pittsburgh, and from studies conducted by collaborators from the Pennsylvania State University (PSU). The UK dataset included samples from the Avon

Longitudinal Study of Parents and their Children (ALSPAC). Each cohort's collection is described below.

5.2.1.1 IUPUI Dataset

Refer to section 3.2.2.1. Of the individuals collected at IUPUI, 784 individuals who had 3D imagery and genotypic data were partitioned for use in this study.

5.2.1.2 Penn State University (PSU) Dataset

The number of individuals from the PSU dataset partitioned for this study included 1,990 individuals with 3D imagery and genotypic data. Additional details of participant recruitment can be found in Section 3.2.2.2. PSU participants were genotyped by 23andMe on the v3, (900K variants) and v4 arrays (600K variants) (Mountain View, CA). PSU sample images were obtained with either the 3dMDface (3dMD, Atlanta, GA) or Vectra H1 systems (Canfield, Parsippany, NJ).

5.2.1.3 University of Pittsburg Dataset

1,906 3D images and genotype data were obtained from the 3D Facial Norms repository[159]. The repository includes 3D facial surface images and self-reported demographic descriptors as well as basic anthropometric measurements from individuals recruited at four US sites: Pittsburgh, PA (PITT IRB PRO09060553 and RB0405013); Seattle, WA (Seattle Children's IRB 12107); Houston, TX (UT Health Committee for the Protection of Human Subjects HSC-DB-09-0508); and Iowa City, IA (University of Iowa Human Subjects Office IRB (200912764 and 200710721). Recruitment was limited to individuals aged 3 to 40 years old and of self-reported European ancestry. Individuals were excluded if they reported a personal or family history of any birth defect or syndrome affecting the head or face, a personal history of any significant facial trauma or facial surgery, or any medical condition that might alter the structure of the face. 3DFN sample genotyping was performed at the Center for Inherited Disease Research at Johns Hopkins University. Participants, including 70 duplicate samples and 72 HapMap control samples, were genotyped on the Illumina OmniExpress + Exome v1.2 array in addition to 4,322 investigator-chosen SNPs included to capture variation in specific regions of interest involved in the genetics of facial variation to yield a total variant count of 968K. Facial surface images were acquired using the 3dMDface camera system (3dMD, Atlanta, GA).

5.2.1.4 UK ALSPAC Dataset

The UK sample was derived from the ALSPAC dataset, a longitudinal birth cohort in which pregnant women residing in Avon with an expected delivery date between 1 April 1991 and 31 December 1992 were recruited[160]. At the time, 14,541 pregnant women were recruited, and DNA samples were collected for 11,343 children. Genome-wide data was available for 8,952 subjects of the B2261 study, titled "Exploring distinctive facial features and their association with known candidate variants." In addition to this, 4,731 3D images were available along with information on sex, age, weight, height, ancestry, and other body characteristics. The ALSPAC study website contains details of all the data that is available through a fully searchable data dictionary (http://www.bris.ac.uk/alspac/researchers/our-data/). Ethical approval for the study was obtained from the ALSPAC Ethics and Law Committee and the Local Research Ethics Committees. Informed consent for the use of data collected via questionnaires and clinics was obtained from participants following the recommendations of the ALSPAC Ethics and Law Committee at the time. Consent for biological samples has been collected in accordance with the Human Tissue Act (2004). ALSPAC samples were genotyped using the Illumina Human Hap550 quad genome-wide SNP genotyping platform by Sample Logistics and Genotyping Facilities at the Wellcome Trust Sanger Institute (Cambridge, UK) and the Laboratory Corporation of America (Burlington, NC), supported by 23andMe. Using the Hap550 quad genome-wide array yielded approximately 550K variants. The ALSPAC sample was imaged using a Konica Minolta Vivid 900 laser scanner (Konica Minolta Sensing Europe, Milton Keynes, UK). For this system, two high-resolution facial scans were taken and then processed, merged, and registered using a macro algorithm in Rapidform® software (INUS Technology Inc., Seoul, South Korea). Ultimately, the final subset of European individuals who also had imaging data, covariates, and genotypes was 3,566 individuals.

5.2.2 Facial Phenotyping

Phenotyping was performed by our collaborator at PSU. 3D surface images were imported in wavefront.obj format into Matlab 2017b to perform the spatially dense registration process using a series of in-house functions packaged together in the MeshMonk registration framework[29]. To study global and local effects on facial variation, a data-driven facial segmentation on the UK and US datasets combined was performed, as described previously[28]. Before segmentation, images in the two datasets were separately adjusted for sex, age, age-squared, height, weight, facial size, the first four genomic ancestry axes, and the camera system. After adjustment, facial segments were defined by grouping vertices that are strongly correlated using hierarchical spectral clustering. This resulted in the construction of 63 facial segments (See Figure 5.1), broken into five levels of the face. Following additional spatial configuration and Principal Components Analyses (PCA), in combination with parallel analysis, each facial segment's information was captured up using principal coordinates ranging from 8 to 70, depending on the segment.

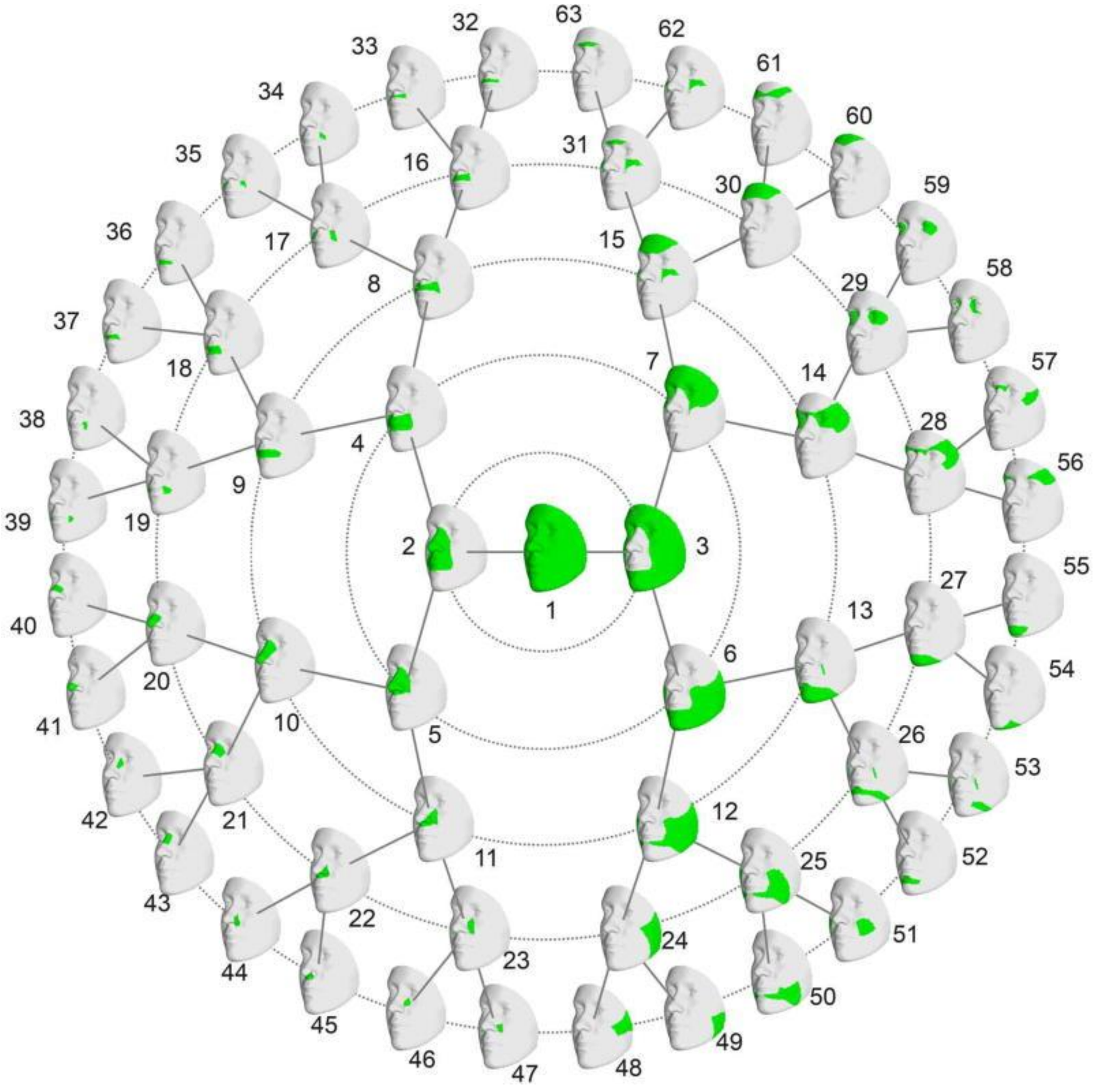


Figure 5.1. Hierarchical Clustering of Face Shape. Facial segmentation was performed on a dataset of 2329. First, a squared similarity matrix was constructed based on a landmark and all other neighboring 3D landmarks. Subsequently, a 5-level hierarchical spectral clustering was performed on this matrix that resulted in 63 total masks

5.2.3 Quality Control, Imputation & Facial GWAS Meta Analyses

For this part of the study, our PSU collaborator gathered all available raw genotypic data from the IUPUI, PSU, and University of Pittsburgh datasets, hereafter referred to as US. The US datasets were imputed separately, and then combined following Verma, et al., 2014[144]. For each dataset, standard data cleaning and quality assurance practices were performed based on the GRCh37 (hg19) genome assembly[145]. The genotypes were "harmonized" with 1000 Genomes Project (1000G) Phase 3[30] using Genotype Harmonizer[146] with a window size of 200 SNPs, a minimum of 10 variants, and alignment based on minor allele frequency (--mafAlign 0.1). This program was also used to filter out ambiguous SNPs, update the SNP id, and update the reference allele as needed, all in reference to the 1000G Phase 3 genotypes. After genotype harmonization, additional QC metrics, such as relatedness and ancestry analyses (selected individuals aligned with 1000 Genomes European populations only) were performed that led to a final merged dataset of 4,680 US participants with 7,417,619 SNPs for analysis. The raw genotype data from ALSPAC, hereafter referred to as UK, was not available and restrictions are in place against merging the ALSPAC genotypes with any other genotypes. For this reason, imputed UK genotypes were obtained directly from the ALSPAC database. After post-imputation quality control and ancestry analyses, the UK dataset contained 8,629,873 SNPs from 3,566 individuals for analysis.

The meta-analysis framework consisted of three steps: identification, verification, and meta-analysis (See Figure 5.2). For all analyses, the genotypes were coded as either 0, 1, or 2, based on the presence of the major allele. In the identification step (using both US, and UK data), within each of the 63 facial segments, each SNP was associated with phenotypic variation using canonical correlation analysis (CCA, canoncorr in Matlab 2017b). CCA is a multivariate analysis which extracts the linear combination of PCs that are maximally correlated with the SNP, which represent the direction of phenotypic effect in shape space (i.e. a phenotypic trait). In the verification step, the shape variables (PCs) of the non-identification dataset (i.e. the verification dataset) were projected onto the trait found in the identification stage, which returns a univariate variable (See UniVar in Figure 5.2). These univariate variables were then tested for genotype-phenotype associations in a standard linear regression with the SNP genotypes of the verification dataset as independent variables and the univariate trait projection score as the dependent variable. Next, the identification p-value (from the CCA) and the verification p-value (from the univariate regression) were combined in a meta-analysis using Stouffer's method[123,161]. This process was repeated, resulting in two meta-analysis p-values accompanied by two identified traits, per segment and per SNP: first using US in the identification stage and UK as verification ($\mathrm{META}_{\mathrm{US}}$), then using UK in the identification stage and US as verification ($\mathrm{META}_{\mathrm{UK}}$).

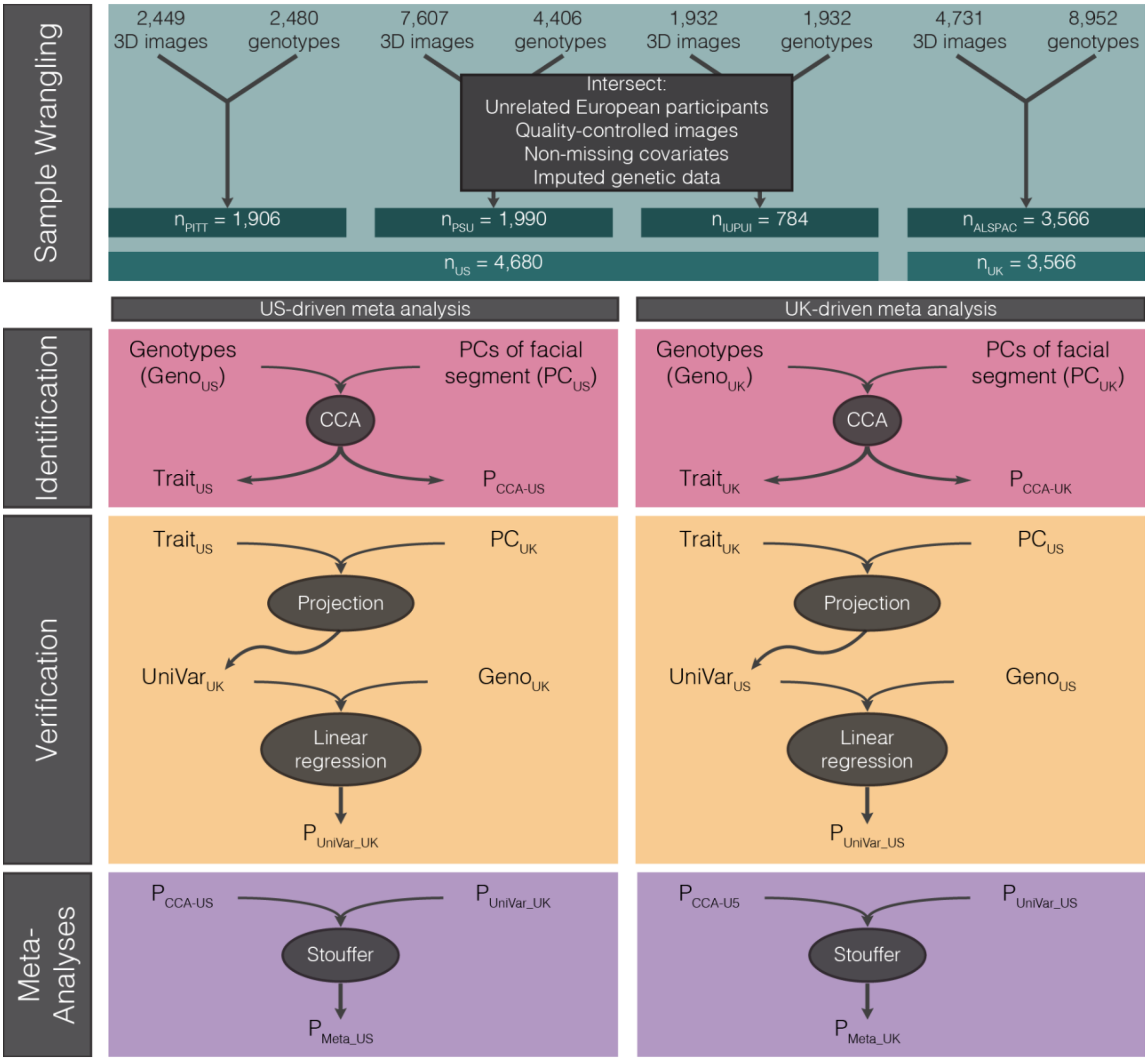


Figure 5.2. Study Design. *Sample Wrangling*: Images and genotypes from each study were intersected and unrelated participants of European ancestry, with quality-controlled images, covariates, and imputed genetic data were selected to obtain the analyzed data. *Identification*: Within each facial segment, canonical correlation analysis (CCA) was used to identify the facial principal components most correlated with the genotypes, which led to a p-value (P_{CCA-US} or P_{CCA-UK}) and facial trait most correlated with each SNP ($Trait_{US}$ and $Trait_{UK}$). *Verification*: The principal components of the other dataset were then projected onto this trait to obtain a univariate variable representing the distribution of participants from the verification dataset for the trait identified in the identification dataset ($UniVar_{UK}$ and $UniVar_{US}$). The genotypes of the verification dataset are then tested against this variable via linear regression, resulting in an additional p-value ($P_{UniVar-UK}$ and $P_{UniVar-US}$). *Meta-Analysis*: The p-values from identification and verification are meta-analyzed using Stouffer's method, resulting in the final set of p-values from each meta-analysis permutation ($P_{META-US}$ and $P_{META-UK}$)

5.2.4 Gene Annotation

Genes 500 kb up and downstream of the lead SNPs found during association analyses were identified using the Table Browser of the UCSC Genome Browser[162]. The most likely candidate gene per lead SNP was identified based on a three-step system. First, we investigated whether any gene in the window was previously associated with craniofacial development or morphology through normal-range facial association studies, genetic disorders with facial dysmorphology as a symptom, or animal models. If this was not the case, we checked whether the gene was a known contributor to facial development based on the paper of Hooper and colleagues, who used transcriptome data from critical periods of mouse face formation to assess gene activity across facial development[163]. If both methods did not deliver a suitable candidate gene, the most likely candidate gene was selected based on the FUMA gene prioritization algorithm[164]. To investigate the potential roles of the identified lead SNPs, enrichment analyses using FUMA and GREAT[165] were performed using default parameters (See Figure 5.3).

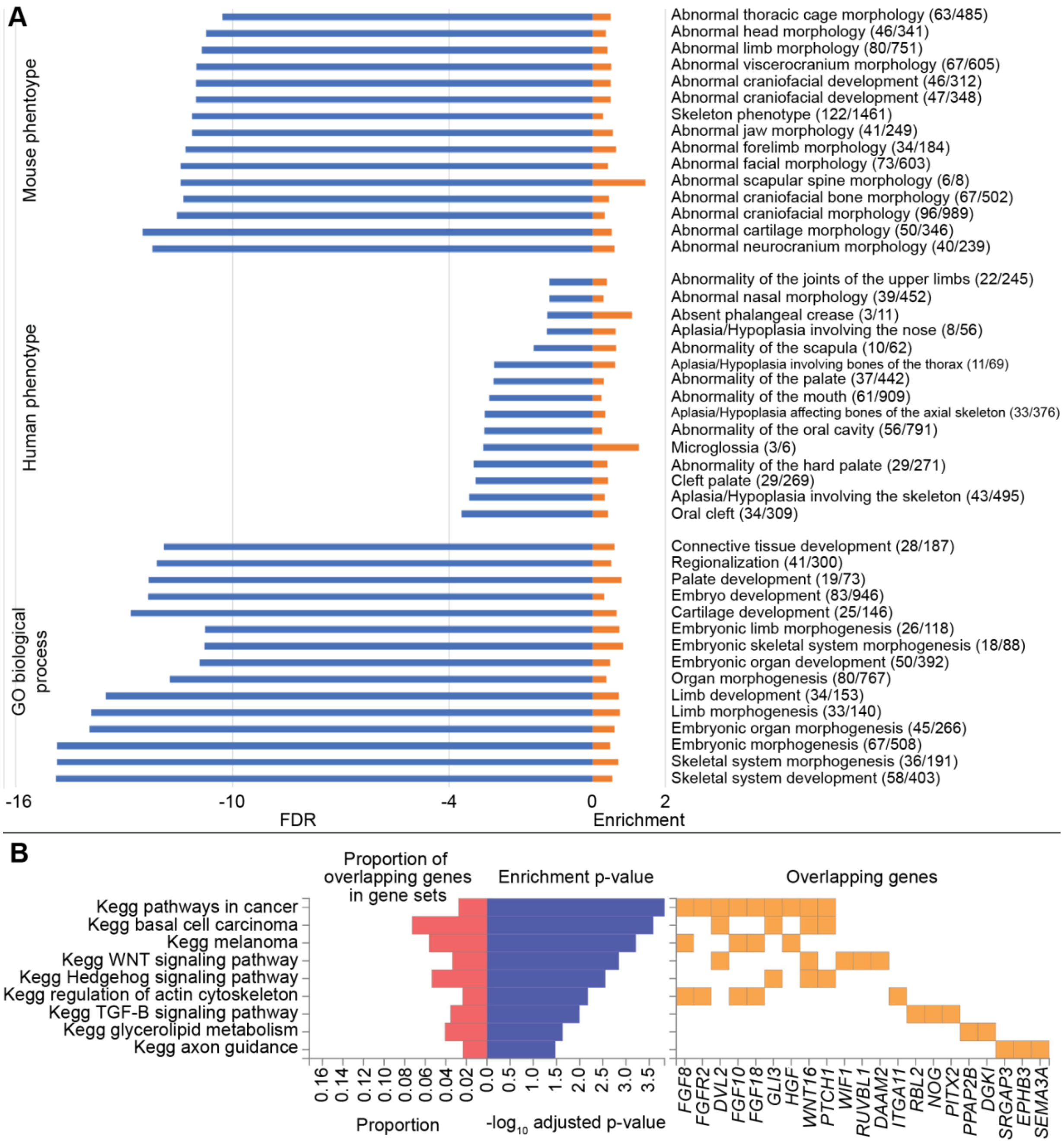
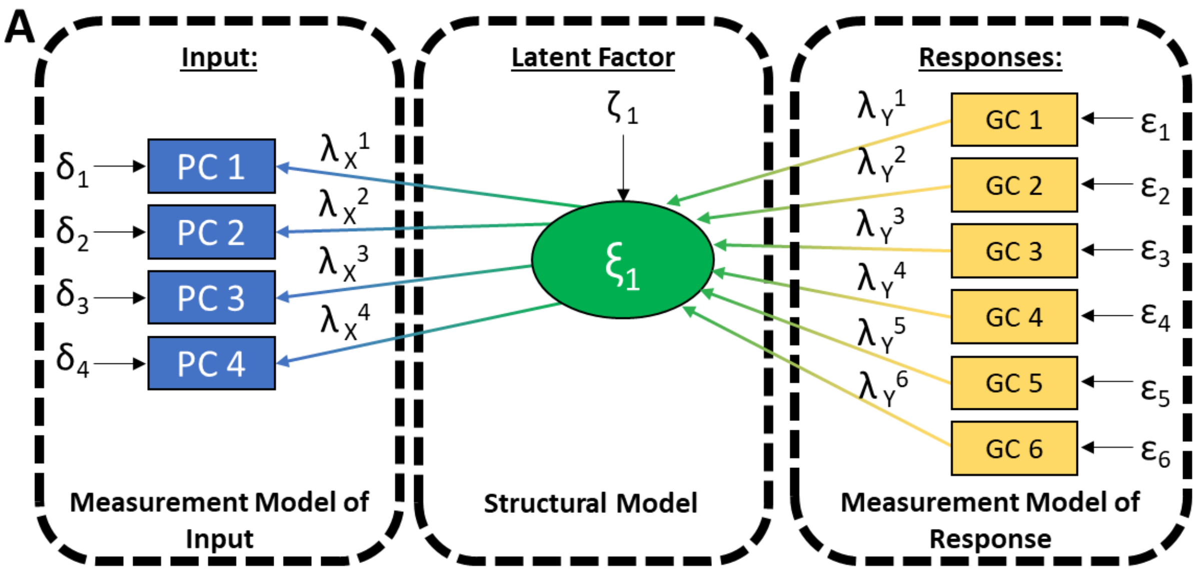


Figure 5.3. GREAT and FUMA Analyses Showing Enrichment for Craniofacial and Limb Development. A) GREAT analysis. Plotted is the binomial test FDR (blue) and binomial enrichment (orange). We indicate the actual number of genomic regions in the test set with the annotation compared to the observed region hits (expected number of genomic regions in the test set with the annotation) behind every term. B) FUMA analysis, indicating the KEGG pathways that were enriched in our results. Multiple pathways are relevant for craniofacial development. The right panel shows the genes that are involved in the pathways

5.2.5 Structural Equation Modeling on Multiple Univariate Facial Phenotypes

Structural equation modeling was performed on a collection of phenotypic and genotypic data collected from both US and UK-based cohorts (N = 8246). Phenotypic data was provided in the form of 8 to 70 principal components derived from quasi-landmarks collected from each of the 63 total facial masks (See Figure 5.1). Genotypic data included a list of 203 variants, scored additively (i.e. 0, 1, or 2), found to be significant during the canonical correlation GWAS mentioned previously. Missing genotypic data points were substituted with the mode genotype, or the most commonly seen genotype based on genotype frequencies. Covariates of age, sex, height, weight, and face size were also included in the model to control for effects such as BMI and sex-based facial dimorphism. Distributions of the continuous covariates of age, height, weight, and face shape were plotted and displayed near normal distributions. As genotypes were trichotomous, normality was not accessed, however the principal components that comprise the latent variable, were by nature, normally distributed. All SEM analyses were conducted in R using the Lavaan Package[82].

Since analyzing the entire dataset via a single SEM would require the modeling of thousands of interactions and would also require extensive computational resources, separate SEM's were conducted. Thus, several models were run iteratively, first to filter the original list of 203 variants down to the most influential by assessing model fit parameters and genotype regression weights. Significant regression weights of genotypes (evaluated at $p < 0.2$) were combined and included in the second SEM, which was used to output a ranked list of variants that were deemed most influential to the segment. Variants that significantly explained the segment (assessed at $p < 0.05$) were retained for the refined model used in follow-up epistasis and projection analyses. In the specific application of this multivariate technique, measured principal components from the full dataset (N = 8,246) of 3D facial images for each segment were used to generate the latent variable. The latent variable was regressed against these observed variables (PC's and covariates) subsequently explaining their relationship (See Figure 5.4).



B

Model Equations:

$$\xi_{(n\,x\,1)} = \frac{PC_{(q\,x\,1)} - \delta_{(q\,x\,1)}}{\lambda_{PC_{(q\,x\,1)}}}$$

$$\xi_{(p\,x\,1)} = GC_{(p\,x\,1)} - \lambda_{GC_{(p\,x\,n)}}\ x\ \zeta_{(n\,x\,1)}$$

Key:

ξ = Matrix of exogenous latent variables

n = The number of exogenous latent variables

PC = Matrix of indicators of the exogenous variables (e.g. A matrix of principal components of the facial segment)

q = The number of input indicators

δ = Measurement error of each indicator of exogenous variable

λ = Regression weight of specified indicator on exogenous latent variable

GC = Matrix of indicators of the exogenous variable

p = The number of response indicators

ζ = Disturbance term for exogenous latent variables (i.e. The disturbance, or error, for the facial segment)

Figure 5.4. SEM Model Structure for Facial Segments. A) A visual representation of the SEM that is comprised of Principal Components (PC) that create half the measurement model, the genotypes and covariates (GC) that comprise the other half. Together the two halves of the measurement with the structural model comprised of a singular unobserved latent factor (ξ_1) makes up the entire SEM model. B) The model's general mathematic equation

After pruning the SEM for variants that best explain our latent variable construct, I validated the accuracy of our model using the Chi-square, RMSEA, CFI, and SRMR fit indices populated by the model. Groups of variants per facial segment were also functionally validated by examining *in-silico* their H3K27ac activity across cell types between facial segments and comparing that to variant H3K27ac activity within facial segments using Spearman's rho.

5.2.6 Epistasis Analysis

The refined SEM model generated a latent univariate variable that was used to assess whether interactions between genotypes increase or decrease the effect (distribution) of a phenotype. Variants deemed significant by the refined SEM model ($p < 0.05$) were input into the Plink[37] analysis program where all diplotype combination effects on the dimensionally reduced latent segment constructs were assessed via a chi-square analysis[38]. Four diplotype combinations that were deemed significant at $p < 0.05$, after correcting for multiple testing, were reported and followed up with various data mining analyses.

To assess the genotypic contribution to epistatic masking (i.e. the combination of two variants reduce the output phenotype) and boosting (i.e. the combination of two variants elevate the output phenotype), all nine diplotypes and their phenotypic distributions were plotted along with marginal distributions using R and various analysis packages including Agricolae, Cowplot, ggplot2, ggpubr, gridExtra, gtable, grid, Hmisc, psych, and data.table[166–175]. The marginal phenotypic medians of the singular genotypes were averaged in order to visualize the predicted phenotypic distribution that would occur if the two genotypes were acting independently. The average median of the singular genotypes using the univariate latent phenotype variable (dashed blue line in Figure 5.6) was compared to the medians of the varying combined diplotypes (solid black line in figure 5.6). Significance was performed using a Mood's Median test to discern the epistatic significance of these diplotype combinations (reported P-value in Figure 5.6)[176]. Follow-up data mining on the results of the epistatic analyses was performed using various online research tools and databases including VarElect,[177] StringDB[178], and Encode[130,131].

5.3 Results and Discussion

The identification and verification analysis strategy yielded 126 p-values and 126 traits for every SNP, representing the 63 segments by two permutations. Per SNP, the lowest p-value was selected, and was noted in which version of the meta-analysis ($META_{US}$ or $META_{UK}$; "Best Permutation") and segment ("Best Segment") this p-value occurred (See Table 5.1). Peak selection on both genomic position and phenotypic effect resulted in 218 lead SNPs. Of these 218 lead SNPs, 203 showed consistent phenotypic effects in the US and UK in the Best Segment and were noted as genome-wide significant hits within the overall META GWAS. Of the 203 variants, 86 of the peaks these variants were found in overlap with GWAS conducted on facial phenotypes, 64 were observed at loci that contain genes potentially influencing craniofacial morphology (which were determined either from human malformations or animal studies), and 53 were novel with no connection with previously known craniofacial morphology (See Annotation Category column within Table 5.1).

Table 5.1. 203 Lead Facial SNPs

RSID*	Lowest P	Best Segment	Best Dataset	Candidate gene	Annotation category
rs76244841	1.5E-08	30	UK	PRDM16	Region previously implicated in normal-range facial morphology
rs1572037	2.6E-22	2	US	PRDM16	Region previously implicated in normal-range facial morphology
rs742071	1.1E-15	11	UK	PAX7	Region previously implicated in normal-range facial morphology using other analyses of these data
rs4912082	3.2E-08	10	UK	CAPZB	Candidate gene implicated in craniofacial morphology through animal model
rs16834081	2.4E-09	11	US	MATN1	Region or candidate gene implicated in craniofacial morphology through human dysmorphology
rs199971562	1.2E-11	30	UK	PLPP3	No previous association
rs4916071	1.9E-51	11	US	c1orf87	Region previously implicated in normal-range facial morphology using other analyses of these data
rs79297754	9.7E-10	61	US	SLC44A5	No previous association
rs12070922	3.7E-09	11	US	LPHN2	Region or candidate gene implicated in craniofacial morphology through human dysmorphology
rs7513680	7.0E-13	51	UK	TBX15	Region previously implicated in normal-range facial morphology
rs3936018	8.0E-58	14	US	WARS2	Region previously implicated in normal-range facial morphology
rs17023457	3.3E-15	48	UK	WARS2	Region previously implicated in normal-range facial morphology
rs11589479	3.9E-09	1	US	ADAM15	Region previously implicated in normal-range facial morphology
rs577676	1.1E-08	51	US	PRRX1	Region or candidate gene implicated in craniofacial morphology through human dysmorphology
rs10919462	2.7E-11	9	US	PRRX1	Region or candidate gene implicated in craniofacial morphology through human dysmorphology
rs2759656	1.2E-79	53	UK	CRB1	Region previously implicated in normal-range facial morphology
rs12039502	1.9E-10	13	US	MARK1	No previous association
rs7558413	1.3E-10	1	US	NT5C1B	No previous association
rs6715010	1.3E-14	22	US	OSR1	Region previously implicated in normal-range facial morphology
rs79037251	9.5E-11	24	UK	OSR1	Region previously implicated in normal-range facial morphology
rs1427539	2.5E-09	7	US	OSR1	Region previously implicated in normal-range facial morphology
rs6740960	3.4E-36	1	US	PKDCC	Region previously implicated in normal-range facial morphology

rs10189338	2.5E-08	11	US	PKDCC	Region previously implicated in normal-range facial morphology
rs7590268	1.1E-08	29	US	THADA	Region or candidate gene implicated in craniofacial morphology through human dysmorphology
rs921119	1.4E-46	36	US	SIX3	Region previously implicated in normal-range facial morphology using other analyses of these data
rs35395759	3.6E-10	44	US	SIX2	Region or candidate gene implicated in craniofacial morphology through human dysmorphology
rs1367228	6.4E-09	30	UK	EFEMP1	No previous association
rs13035645	2.4E-08	27	UK	BCL11A	Region or candidate gene implicated in craniofacial morphology through human dysmorphology
rs6546175	4.8E-12	58	US	SPRED2	No previous association
rs3891585	3.8E-10	15	UK	MEIS1	Candidate gene implicated in craniofacial morphology through animal model
rs11675008	2.8E-08	49	UK	PNO1	No previous association
rs17655927	2.6E-09	61	US	DYSF	No previous association
rs772154	1.3E-08	21	UK	NCAPH	Region or candidate gene implicated in craniofacial morphology through human dysmorphology
rs7597495	6.5E-10	25	US	FBLN7	Candidate gene implicated in craniofacial morphology through animal model
rs11692600	1.9E-12	14	UK	INSIG2	Candidate gene implicated in craniofacial morphology through animal model
rs332108	5.7E-12	14	UK	EN1	Candidate gene implicated in craniofacial morphology through animal model
rs148912137	3.4E-12	7	US	ZEB2	Region previously implicated in normal-range facial morphology
rs970797	5.4E-68	1	US	HOXD1	Region previously implicated in normal-range facial morphology
rs10178696	2.6E-11	24	UK	MTX2	Region previously implicated in normal-range facial morphology
rs8176501	4.1E-09	32	UK	CALCRL	No previous association
rs13035389	2.7E-13	53	US	SATB2	Region previously implicated in normal-range facial morphology
rs4675617	2.5E-13	18	UK	SATB2	Region or candidate gene implicated in craniofacial morphology through human dysmorphology
rs1370926	6.5E-62	11	US	PAX3	Region previously implicated in normal-range facial morphology
rs7579011	7.1E-14	2	US	FARSB	Region previously implicated in normal-range facial morphology
rs4686337	2.1E-08	11	UK	SRGAP3-SETD5	Region or candidate gene implicated in craniofacial morphology through human dysmorphology
rs73048344	3.0E-08	15	UK	THRB	No previous association

rs17054293	2.8E-24	2	UK	CACNA2D3	Region previously implicated in normal-range facial morphology
rs9310211	4.6E-22	23	US	FOXP1	Region or candidate gene implicated in craniofacial morphology through human dysmorphology
rs3072056	3.8E-11	2	US	VGLL3	Region or candidate gene implicated in craniofacial morphology through human dysmorphology
rs34199564	2.2E-08	30	US	ARL13B	No previous association
rs793487	2.0E-08	1	US	COL8A1	Region or candidate gene implicated in craniofacial morphology through human dysmorphology
rs1391361	3.9E-08	4	UK	CD96	Region or candidate gene implicated in craniofacial morphology through human dysmorphology
rs7373685	9.0E-49	18	UK	GATA2	No previous association
rs6795164	1.4E-09	1	UK	SLCO2A1	Region or candidate gene implicated in craniofacial morphology through human dysmorphology
rs792736	1.2E-16	3	US	SHOX2	No previous association
rs58022575	2.8E-26	5	US	EPHB3	Region previously implicated in normal-range facial morphology
rs56081252	4.7E-11	6	UK	EPHB3	Region previously implicated in normal-range facial morphology
rs74921869	3.5E-11	5	US	FGFRL1	Region or candidate gene implicated in craniofacial morphology through human dysmorphology
rs3910659	4.5E-09	25	UK	STX18	No previous association
rs13117653	4.2E-18	34	US	MSX1	Region previously implicated in normal-range facial morphology using other analyses of these data
rs7674010	4.4E-08	20	UK	LEF1	Candidate gene implicated in craniofacial morphology through animal model
rs514892	2.2E-09	15	UK	PITX2	Region or candidate gene implicated in craniofacial morphology through human dysmorphology
rs7655723	8.4E-09	31	US	MYOZ2	No previous association
rs7694450	1.3E-08	9	UK	PRDM5	Candidate gene implicated in craniofacial morphology through animal model
rs6852838	3.5E-11	24	US	FAT4	Region previously implicated in normal-range facial morphology
rs62324070	2.8E-29	2	US	INTU	Region or candidate gene implicated in craniofacial morphology through human dysmorphology
rs9995821	5.7E-65	2	US	DCHS2	Region previously implicated in normal-range facial morphology
rs4342159	3.7E-10	2	UK	PALLD	No previous association
rs4695846	1.7E-19	53	UK	HAND2	Region previously implicated in normal-range facial morphology

rs3054104	3.1E-21	53	UK	HPGD	Region previously implicated in normal-range facial morphology
rs4866909	2.4E-10	62	UK	FGF10	Region previously implicated in normal-range facial morphology using other analyses of these data
rs10078545	1.7E-09	2	US	XRCC4	Region previously implicated in normal-range facial morphology
rs76770688	3.7E-09	9	US	MCC	Region previously implicated in normal-range facial morphology using other analyses of these data
rs13182341	3.5E-15	6	US	FBN	No previous association
rs4582322	9.5E-09	5	US	FABP6	No previous association
rs6555969	1.0E-13	56	US	FGF18	Region previously implicated in normal-range facial morphology
rs17073930	4.0E-08	2	US	FGF18	Region previously implicated in normal-range facial morphology
rs4959652	2.2E-12	7	UK	MYLK4	No previous association
rs7755467	1.1E-09	2	UK	NRSN1	No previous association
rs9274522	2.2E-08	62	UK	COL11A2	Region previously implicated in normal-range facial morphology using other analyses of these data
rs1520	2.3E-22	21	UK	DAAM2	Candidate gene implicated in craniofacial morphology through animal model
rs227832	3.9E-37	23	US	SUPT3H	Region previously implicated in normal-range facial morphology
rs73457129	4.9E-08	21	UK	RUNX2	Region previously implicated in normal-range facial morphology
rs12055796	1.5E-08	22	UK	RCAN2	No previous association
rs9381923	7.3E-19	22	US	TFAP2B	Region previously implicated in normal-range facial morphology
rs6923760	1.0E-10	11	US	PKHD1	No previous association
rs4715567	1.7E-08	1	US	BMP5	Candidate gene implicated in craniofacial morphology through animal model
rs112411726	1.1E-09	15	US	TBX18	No previous association
rs6568401	4.6E-08	19	US	PRDM1	No previous association
rs9388518	3.7E-21	53	US	RSPO3	Region previously implicated in normal-range facial morphology using other analyses of these data
rs519332	6.5E-45	28	US	EYA4	Region previously implicated in normal-range facial morphology using other analyses of these data
rs34729823	9.2E-21	59	US	MTFR2	No previous association
rs2108791	2.6E-08	28	UK	COL28A1	No previous association
rs212672	1.3E-20	1	US	TWIST1	Region previously implicated in normal-range facial morphology using other analyses of these data

rs2465274	4.2E-09	9	UK	HOXA2	Region previously implicated in normal-range facial morphology using other analyses of these data
rs62443772	5.3E-16	22	UK	GLI3	Region previously implicated in normal-range facial morphology
rs12535551	4.6E-11	11	UK	IGFBP3	No previous association
rs7806852	1.9E-08	5	UK	HGF	No previous association
rs7807002	3.3E-10	61	US	SEMA3A	No previous association
rs4296976	8.0E-74	54	US	DLX6	Region previously implicated in normal-range facial morphology
rs884373	7.2E-17	26	US	WNT16	Candidate gene implicated in craniofacial morphology through animal model
rs798682	8.7E-24	45	US	FEZF1	No previous association
rs4732070	8.9E-09	3	US	CALD1	No previous association
rs6948022	6.1E-09	60	UK	DGK1	No previous association
rs2976940	8.6E-09	3	US	USP17L8	No previous association
rs12114954	1.1E-11	2	US	PPP1R3B	Region previously implicated in normal-range facial morphology
rs657913	4.0E-10	18	UK	MSRA	Region previously implicated in normal-range facial morphology
rs5029306	5.3E-10	1	UK	DPYSL2	Region previously implicated in normal-range facial morphology using other analyses of these data
rs149814396	2.7E-08	45	US	SFRP1	Candidate gene implicated in craniofacial morphology through animal model
rs56252175	1.7E-09	63	US	SNAI2	Region or candidate gene implicated in craniofacial morphology through human dysmorphology
rs116867275	1.9E-08	1	UK	PXDNL	No previous association
rs1588405	1.3E-08	60	UK	PKIA	No previous association
rs139016242	2.9E-10	63	UK	POP1	Region or candidate gene implicated in craniofacial morphology through human dysmorphology
rs60373317	2.0E-09	1	US	VPS13B	Region previously implicated in normal-range facial morphology
rs10089785	2.6E-08	10	UK	TRHR	Region previously implicated in normal-range facial morphology using other analyses of these data
rs2245221	1.9E-08	48	UK	TRPS1	Region or candidate gene implicated in craniofacial morphology through human dysmorphology
rs11337200	7.0E-13	2	US	TRPS1	Region or candidate gene implicated in craniofacial morphology through human dysmorphology
rs7843236	9.5E-10	2	UK	SNTB1	No previous association
rs2581548	8.5E-11	2	UK	HAS2	Candidate gene implicated in craniofacial morphology through animal model

rs871502	3.3E-08	4	UK	CCDC26	Region or candidate gene implicated in craniofacial morphology through human dysmorphology
rs7859005	2.7E-10	57	US	DMRT2	Region previously implicated in normal-range facial morphology
rs10758593	6.0E-14	3	UK	GLIS3	Region previously implicated in normal-range facial morphology
rs303751	2.8E-12	38	US	FREM1	Region previously implicated in normal-range facial morphology using other analyses of these data
rs10962767	5.1E-14	2	UK	BNC2	Region previously implicated in normal-range facial morphology
rs2230578	5.8E-09	2	UK	ROR2	Region previously implicated in normal-range facial morphology
rs12553508	5.9E-09	3	US	BARX1	Region previously implicated in normal-range facial morphology
rs145965565	1.2E-14	44	US	PTCH1	Region or candidate gene implicated in craniofacial morphology through human dysmorphology
rs1999464	1.7E-34	7	US	ELP1	No previous association
rs10448285	1.4E-10	57	UK	LMX1B	Region previously implicated in normal-range facial morphology
rs1902713	1.8E-08	20	US	GDF10	No previous association
rs9633535	4.3E-11	5	US	ARID5B	No previous association
rs138458666	7.5E-09	2	UK	KAT6B	Region or candidate gene implicated in craniofacial morphology through human dysmorphology
rs1907342	6.3E-19	9	US	C10orf11	Region previously implicated in normal-range facial morphology
rs1536446	4.0E-08	45	US	TCTN3	Region or candidate gene implicated in craniofacial morphology through human dysmorphology
rs11190970	7.5E-09	45	UK	FGF8	Candidate gene implicated in craniofacial morphology through animal model
rs242983	1.8E-20	1	US	EMX2	Region previously implicated in normal-range facial morphology
rs34988394	5.5E-14	62	UK	EMX2	Candidate gene implicated in craniofacial morphology through animal model
rs1696840	2.8E-08	34	UK	FGFR2	Region or candidate gene implicated in craniofacial morphology through human dysmorphology
rs6578283	2.5E-10	9	US	KCNQ1	Region previously implicated in normal-range facial morphology using other analyses of these data
rs34121257	1.1E-08	7	US	SOX6	Region or candidate gene implicated in craniofacial morphology through human dysmorphology
rs7121535	4.7E-09	1	US	ALX4	Region or candidate gene implicated in craniofacial morphology through human dysmorphology

rs10838269	8.9E-15	30	UK	ALX4	Region or candidate gene implicated in craniofacial morphology through human dysmorphology
rs10743452	4.7E-08	7	US	A2ML1	Region or candidate gene implicated in craniofacial morphology through human dysmorphology
rs11049359	4.2E-19	1	US	PTHLH	Candidate gene implicated in craniofacial morphology through animal model
rs7970054	6.6E-11	7	US	LRIG3	No previous association
rs60493340	4.0E-19	3	US	WIF1	No previous association
rs11175967	2.1E-15	28	UK	HMGA2	Region previously implicated in normal-range facial morphology
rs2695152	1.3E-08	35	US	NAV3	Region previously implicated in normal-range facial morphology using other analyses of these data
rs10160931	8.8E-14	13	US	NAV3	No previous association
rs58687115	5.3E-11	12	UK	gene desert	No previous association
rs4385969	2.5E-11	2	US	SLC6A15	No previous association
rs11609649	6.6E-29	1	US	ALX1	Region or candidate gene implicated in craniofacial morphology through human dysmorphology
rs10779162	4.1E-17	9	US	ALX1	Region previously implicated in normal-range facial morphology
rs2098990	6.7E-25	5	US	TBX3	Region previously implicated in normal-range facial morphology
rs66516258	1.3E-09	11	US	SPRY2	Region or candidate gene implicated in craniofacial morphology through human dysmorphology
rs11842203	7.2E-11	5	UK	TGDS-SOX21	Region or candidate gene implicated in craniofacial morphology through human dysmorphology
rs9524742	3.1E-15	1	US	TGDS-SOX21	Region or candidate gene implicated in craniofacial morphology through human dysmorphology
rs2390383	2.0E-08	14	UK	PCCA	Region previously implicated in normal-range facial morphology using other analyses of these data
rs9558696	9.6E-10	2	US	EFNB2	Region previously implicated in normal-range facial morphology using other analyses of these data
rs1411551	3.8E-12	20	US	IRS2	Region previously implicated in normal-range facial morphology
rs59081965	7.5E-09	1	US	PAX9	Region previously implicated in normal-range facial morphology using other analyses of these data
rs12878658	4.6E-09	11	US	BMP4	Region previously implicated in normal-range facial morphology
rs10047930	1.6E-08	53	UK	DACT1	No previous association

rs12881623	1.0E-09	18	UK	SIX1	Candidate gene implicated in craniofacial morphology through animal model
rs12890110	2.2E-09	17	UK	RAD51B	Region previously implicated in normal-range facial morphology
rs1542448	1.5E-13	7	US	BCL11B	Region or candidate gene implicated in craniofacial morphology through human dysmorphology
rs76810699	2.5E-09	1	US	BCL11B	Region or candidate gene implicated in craniofacial morphology through human dysmorphology
rs1257585	7.4E-11	56	UK	BCL11B	Region or candidate gene implicated in craniofacial morphology through human dysmorphology
rs148375239	4.7E-10	53	US	GREM1	Region previously implicated in normal-range facial morphology
rs34961041	9.4E-14	31	US	MEIS2	Region or candidate gene implicated in craniofacial morphology through human dysmorphology
rs2306022	5.9E-09	15	US	ITGA11	No previous association
rs34430707	6.9E-12	2	US	THSD4	Region previously implicated in normal-range facial morphology
rs112087864	7.1E-11	24	UK	THSD4	Region previously implicated in normal-range facial morphology
rs2401176	9.1E-11	11	UK	ADAMTSL3	No previous association
rs9923447	5.9E-27	9	UK	RPGRIP1L	Region or candidate gene implicated in craniofacial morphology through human dysmorphology
rs62051935	2.0E-13	19	US	FOXC2	Region previously implicated in normal-range facial morphology
rs2760734	2.1E-09	51	UK	DPH1	Region previously implicated in normal-range facial morphology
rs9899183	8.1E-13	26	UK	SHBG	Region previously implicated in normal-range facial morphology
rs7218433	1.1E-21	7	UK	NOG	Region or candidate gene implicated in craniofacial morphology through human dysmorphology
rs227727	1.0E-16	38	US	NOG	Region previously implicated in normal-range facial morphology using other analyses of these data
rs9893705	8.3E-12	49	US	TBX4	No previous association
rs34439270	8.1E-09	3	US	KCNJ2	Region previously implicated in normal-range facial morphology using other analyses of these data
rs9908442	3.3E-46	5	US	SOX9	Region previously implicated in normal-range facial morphology
rs9302943	9.8E-47	44	US	SOX9	Region previously implicated in normal-range facial morphology

rs2359442	3.3E-09	22	US	SETBP1	Region or candidate gene implicated in craniofacial morphology through human dysmorphology
rs11665450	1.3E-10	46	US	TCF4	Region previously implicated in normal-range facial morphology
rs634687	2.5E-11	11	US	TSHZ1	Region or candidate gene implicated in craniofacial morphology through human dysmorphology
rs287104	3.8E-37	5	UK	KCTD15	Region previously implicated in normal-range facial morphology using other analyses of these data
rs6139982	3.9E-09	60	US	BMP2	Region or candidate gene implicated in craniofacial morphology through human dysmorphology
rs116792	5.0E-11	11	US	MKKS	No previous association
rs6047635	2.3E-12	11	UK	PAX1	Region previously implicated in normal-range facial morphology using other analyses of these data
rs6047637	4.5E-20	31	UK	PAX1	Region previously implicated in normal-range facial morphology using other analyses of these data
rs6113624	7.1E-17	21	US	FOXA2	Region previously implicated in normal-range facial morphology
rs1474738	2.9E-10	1	UK	DNMT3B	Region or candidate gene implicated in craniofacial morphology through human dysmorphology
rs877313	1.1E-29	2	US	DHX35	Region previously implicated in normal-range facial morphology
rs6022641	3.0E-10	36	US	CYP24A1	No previous association
rs4811827	2.9E-09	23	UK	BMP7	Region previously implicated in normal-range facial morphology
rs736702	1.9E-08	30	UK	FBLN1	Candidate gene implicated in craniofacial morphology through animal model
rs6529847	4.2E-08	28	US	NLGN4X	No previous association
rs2520378	3.6E-10	1	US	EDA	Region or candidate gene implicated in craniofacial morphology through human dysmorphology
rs1045686	4.8E-08	60	UK	FAM133A	No previous association
rs11167418	3.9E-08	2	US	PLS3	Candidate gene implicated in craniofacial morphology through animal model
rs4829906	9.5E-14	63	UK	ZIC3	No previous association

*Variants organized by chromosome and position

5.3.1 Using SEM to Identify Groups of Variants that Explain Within Facial Segment Variation

As mentioned previously, GWAS themselves are unable to model complex multivariate relationships between multiple variants and several phenotypes. Therefore, to more fully explain the complex genetic interactions occurring within facial segments on the 203 variants identified as being significant both the US and UK datasets were input into a structural equation model. In general, the number of model parameters generated by the final refined SEM model for each segment ranged between 92 and 217, depending on the number of shape PCs and SNPs included in each model (See Table 5.2). As 8,246 participants were used, this led to a range of 38-90 participants per parameter, which is well above recommendations[179]. Additional statistical power was lent to our models by having a large number of samples and input variables per latent factor[180]. Of the 63 segments, the SEM models for 13 segments were discarded because they did not converge on a solution, which normally occurs when variants are non-informative for that segment or the variance of the segment is too low. For each of the 50 SEM models where the refinement process was successful, we evaluated the fit of each model by instituting cutoffs on the following indices: Chi-square (p-value < 0.05), comparative fit index (CFI > 0.90), root mean square error of approximation (RMSEA < 0.08), and standardized root mean square residual (SRMR < 0.08)[179,180], which generally indicate the strength of how well the SEM models the data. 18 models passed all recommended model fit parameters and 32 models passed all but one of the fit indices, leading to the conclusion that the refined SEM models fit our data well. Final model fit indices and model parameter estimates are provided in Table 5.3.

Table 5.2. Facial SEM Metadata

Segment Name	Quad	Number of PCs representing shape variation	Number of SNPs surviving model refinement	P-value cutoff for refinement and epistasis	Number of SNPs kept for epistasis analysis	Epistasis test performed?
Mask1	Full face	Model removed	Model removed	Model removed	Model removed	Model removed
Mask2	Nose	44	12	0.05	10	Yes
Mask3	Non-nose	60	20	0.05	19	Yes
Mask4	I	Model removed	Model removed	Model removed	Model removed	Model removed
Mask5	II	32	44	0.05	43	Yes
Mask6	III	41	26	0.05	20	Yes
Mask7	IV	51	52	0.1	50	Yes
Mask8	I	15	37	0.05	32	Yes
Mask9	I	19	34	0.05	32	Yes
Mask10	II	18	57	0.05	52	Yes
Mask11	II	29	42	0.05	39	Yes
Mask12	III	26	26	0.05	25	Yes
Mask13	III	31	49	0.05	46	Yes
Mask14	IV	31	38	0.05	36	Yes
Mask15	IV	45	11	0.05	5	Yes
Mask16	I	8	34	0.1	1	Not enough variables
Mask17	I	11	49	0.05	47	Yes
Mask18	I	Model removed	Model removed	Model removed	Model removed	Model removed
Mask19	I	16	37	0.05	33	Yes
Mask20	II	10	61	0.05	60	Yes
Mask21	II	14	32	0.05	29	Yes
Mask22	II	25	44	0.05	41	Yes
Mask23	II	13	36	0.05	23	Yes
Mask24	III	23	2	0.05	1	Not enough variables
Mask25	III	20	13	0.1	3	Yes
Mask26	III	23	35	0.05	34	Yes
Mask27	III	21	49	0.05	46	Yes
Mask28	IV	21	48	0.05	47	Yes
Mask29	IV	23	29	0.05	28	Yes
Mask30	IV	33	44	0.05	43	Yes
Mask31	IV	26	38	0.05	33	Yes
Mask32	I	8	36	0.05	29	Yes

Mask33	I	7	42	0.05	37	Yes
Mask34	I	11	31	0.05	25	Yes
Mask35	I	8	40	0.05	22	Yes
Mask36	I	Model removed	Model removed	Model removed	Model removed	Model removed
Mask37	I	Model removed	Model removed	Model removed	Model removed	Model removed
Mask38	I	Model removed	Model removed	Model removed	Model removed	Model removed
Mask39	I	11	37	0.05	32	Yes
Mask40	II	7	13	0.05	13	Yes
Mask41	II	7	43	0.05	42	Yes
Mask42	II	10	39	0.05	35	Yes
Mask43	II	Model removed	Model removed	Model removed	Model removed	Model removed
Mask44	II	Model removed	Model removed	Model removed	Model removed	Model removed
Mask45	II	Model removed	Model removed	Model removed	Model removed	Model removed
Mask46	II	7	58	0.05	56	Yes
Mask47	II	9	55	0.05	51	Yes
Mask48	III	17	39	0.05	30	Yes
Mask49	III	Model removed	Model removed	Model removed	Model removed	Model removed
Mask50	III	18	18	0.05	12	Yes
Mask51	III	Model removed	Model removed	Model removed	Model removed	Model removed
Mask52	III	13	44	0.05	37	Yes
Mask53	III	Model removed	Model removed	Model removed	Model removed	Model removed
Mask54	III	Model removed	Model removed	Model removed	Model removed	Model removed
Mask55	III	14	30	0.05	26	Yes
Mask56	IV	14	54	0.05	52	Yes
Mask57	IV	19	21	0.05	0	Not enough variables
Mask58	IV	13	32	0.05	24	Yes
Mask59	IV	13	35	0.05	32	Yes
Mask60	IV	27	50	0.05	49	Yes
Mask61	IV	21	51	0.05	48	Yes
Mask62	IV	22	25	0.05	22	Yes
Mask63	IV	25	37	0.05	35	Yes

Table 5.3. Facial SEM Fit Indices

Name	Quadrant	Model fit parameters*							
		Chi-Sq	Chi-Sq DF	Chi-Sq	CFI	RMSE	SRMR	TLI	GFI
				p < 0.05	> 0.90	< 0.08	< 0.08	> 0.95	> 0.95
Mask1	Full face	NA	NA	NA	NA	NA	NA	NA	NA
Mask2	Nose	1534	1848	1.000	1.000	0.000	0.009	1.073	0.992
Mask3	Non-nose	2357	3480	1.000	1.000	0.000	0.008	1.322	0.991
Mask4	I	NA	NA	NA	NA	NA	NA	NA	NA
Mask5	II	4401	2138	0.000	0.704	0.011	0.012	0.693	0.968
Mask6	III	2284	2219	0.166	0.987	0.002	0.009	0.986	0.987
Mask7	IV	5339	4324	0.000	0.840	0.005	0.010	0.836	0.976
Mask8	I	1539	748	0.000	0.862	0.011	0.010	0.851	0.976
Mask9	I	1638	944	0.000	0.871	0.009	0.010	0.862	0.980
Mask10	II	3049	1274	0.000	0.742	0.013	0.010	0.724	0.961
Mask11	II	4240	1833	0.000	0.692	0.013	0.012	0.679	0.966
Mask12	III	1419	1199	0.000	0.956	0.005	0.009	0.954	0.987
Mask13	III	3442	2204	0.000	0.812	0.008	0.010	0.805	0.974
Mask14	IV	2743	1874	0.000	0.852	0.008	0.010	0.846	0.979
Mask15	IV	866	1869	1.000	1.000	0.000	0.007	1.298	0.995
Mask16	I	499	328	0.000	0.965	0.008	0.007	0.959	0.985
Mask17	I	1343	634	0.000	0.877	0.012	0.008	0.864	0.972
Mask18	I	NA	NA	NA	NA	NA	NA	NA	NA
Mask19	I	1251	809	0.000	0.917	0.008	0.009	0.910	0.982
Mask20	II	1503	674	0.000	0.862	0.012	0.007	0.846	0.965
Mask21	II	1369	623	0.000	0.867	0.012	0.010	0.855	0.977
Mask22	II	3781	1571	0.000	0.708	0.013	0.012	0.693	0.965
Mask23	II	1633	617	0.000	0.834	0.014	0.010	0.818	0.971
Mask24	III	76	494	1.000	1.000	0.000	0.004	1.116	0.999
Mask25	III	407	607	1.000	1.000	0.000	0.007	1.050	0.995
Mask26	III	2158	1220	0.000	0.847	0.010	0.010	0.839	0.978
Mask27	III	2206	1369	0.000	0.860	0.009	0.009	0.852	0.975
Mask28	IV	2106	1349	0.000	0.872	0.008	0.009	0.864	0.977
Mask29	IV	1394	1088	0.000	0.941	0.006	0.009	0.938	0.986
Mask30	IV	2532	2223	0.000	0.940	0.004	0.009	0.937	0.982
Mask31	IV	2095	1499	0.000	0.890	0.007	0.010	0.884	0.981
Mask32	I	547	342	0.000	0.958	0.009	0.007	0.952	0.984
Mask33	I	636	326	0.000	0.938	0.011	0.007	0.927	0.979
Mask34	I	1089	454	0.000	0.886	0.013	0.010	0.873	0.977
Mask35	I	744	370	0.000	0.929	0.011	0.007	0.918	0.978
Mask36	I	NA	NA	NA	NA	NA	NA	NA	NA
Mask37	I	NA	NA	NA	NA	NA	NA	NA	NA

Mask38	I	NA	NA	NA	NA	NA	NA	NA	NA
Mask39	I	760	514	0.000	0.951	0.008	0.007	0.945	0.984
Mask40	II	305	152	0.000	0.967	0.011	0.009	0.961	0.990
Mask41	II	714	332	0.000	0.928	0.012	0.007	0.915	0.976
Mask42	II	874	476	0.000	0.923	0.010	0.008	0.913	0.979
Mask43	II	NA	NA	NA	NA	NA	NA	NA	NA
Mask44	II	NA	NA	NA	NA	NA	NA	NA	NA
Mask45	II	NA	NA	NA	NA	NA	NA	NA	NA
Mask46	II	1025	422	0.000	0.893	0.013	0.007	0.875	0.966
Mask47	II	1468	547	0.000	0.847	0.014	0.008	0.826	0.962
Mask48	III	1279	903	0.000	0.927	0.007	0.008	0.921	0.982
Mask49	III	NA	NA	NA	NA	NA	NA	NA	NA
Mask50	III	566	611	0.902	1.000	0.000	0.008	1.011	0.992
Mask51	III	NA	NA	NA	NA	NA	NA	NA	NA
Mask52	III	771	713	0.065	0.988	0.003	0.006	0.987	0.986
Mask53	III	NA	NA	NA	NA	NA	NA	NA	NA
Mask54	III	NA	NA	NA	NA	NA	NA	NA	NA
Mask55	III	696	597	0.003	0.979	0.004	0.007	0.977	0.988
Mask56	IV	1683	909	0.000	0.869	0.010	0.008	0.857	0.972
Mask57	IV	756	710	0.111	0.990	0.003	0.008	0.989	0.991
Mask58	IV	763	569	0.000	0.960	0.006	0.008	0.956	0.986
Mask59	IV	952	605	0.000	0.933	0.008	0.008	0.927	0.983
Mask60	IV	2464	1884	0.000	0.898	0.006	0.009	0.894	0.979
Mask61	IV	2012	1409	0.000	0.888	0.007	0.008	0.882	0.977
Mask62	IV	1044	944	0.012	0.979	0.004	0.009	0.978	0.989
Mask63	IV	1857	1403	0.000	0.915	0.006	0.009	0.910	0.982

*Note: Green highlights indicate a good model fit metric

Reassuringly, for segments that are closely related in the segmentation hierarchy (i.e. segments 5, 11, 23, and 47) there is an average overlap of 46% of the variants meeting the $p < 0.05$ cutoff for significance, compared to 13.6% average overlap for non-hierarchically related segments (i.e. segments 5 and 6). As further validation of the SEM's, the H3K27ac activity across all cell types was compared by a collaborator for significant variants both within and between segments using Spearman's rho (See Figure 5.5). In the analysis it was found that pairs of SNPs are significantly (p = 0.0062) more correlated if they originate from the same masks as opposed to originating from different masks, which should be the trend if the SEM models are accurately grouping SNP's based on the facial segments they influence most.

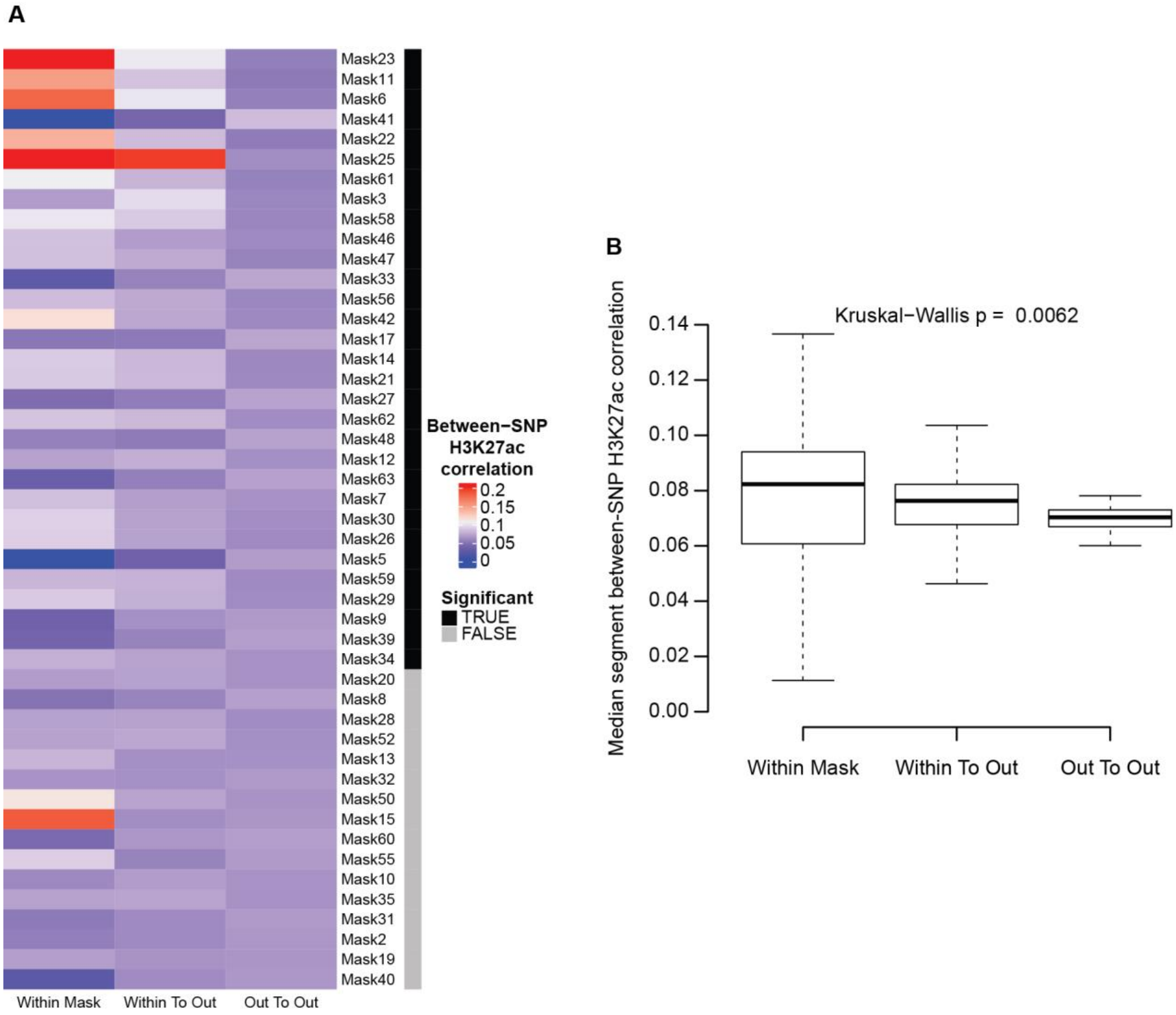
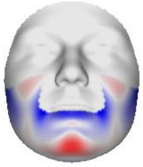
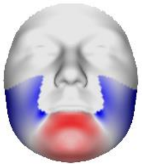
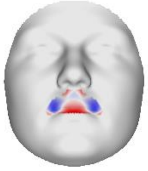
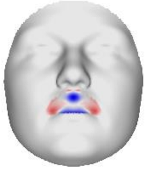
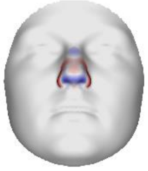
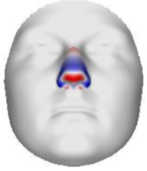
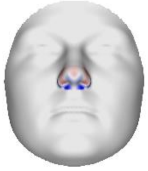
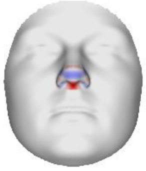


Figure 5.5. Correlation of H3K27ac Activity Among SEM Models. A) For all segments (aka "masks"), we compared the H3K27ac activity for significant SNPs from the refined SEM model for variation at that facial segment. Plotted is the Spearman's rho correlation between pairs of SNPs significant in the same SEM model ("Within Mask"); pairs of SNPs where one is from the SEM model and the other is not ("Within To Out"); and where both SNPs in the pair are from a different SEM model ("Out To Out"). Segments where the distribution of correlation across all cell types was significantly different ($p < 0.05$) are indicated in black. B) For all cell types, the median correlation across all segments is plotted for each of the three SNP groupings. Significance between the means was determined using Kruskal-Wallis test

5.3.2 Epistatic Results

50 of the 63 facial segments were analyzed for epistatic interactions based on the subset variant list output from the refined facial segment's SEM. 13 segment models were either unable to converge or did not have enough surviving variants to perform a test on epistasis. Within the 50 successful segments, four variant pairs (see Table 5.4) showed a significant pairwise epistatic interaction (multiple testing corrected $p < 0.05$) showing some form of a biological interaction on the face shape, which is illustrated by the 'Facial Morph' column in Table 5.4. Furthermore, the phenotypic and marginal distributions for the epistatic pairs can be seen in Figure 5.6 A-E. By looking at Figure 5.6 B, it is possible to see that there is a highly significant difference between the diplotype phenotype TTCC (i.e. the TTCC with interaction phenotype) shown as a solid black line, and the expected phenotype determined by observing the CC and TT genotypes separately (i.e. the expected phenotype of TTCC if TT and CC had no interaction effects) shown as a dotted blue line. Thus, the interaction of the two homozygous major alleles results in the depression of the phenotype. Conversely, scanning across the top row of Figure 5.6 B from right to left, it is evident that altering rs10838269 to include one copy of the minor allele, C, has the opposite effect boosting the phenotype higher than what is expected with the diplotype of CTCC. An addition of yet another minor allele to rs10838269 yielding a diplotype of CCCC, further increases the facial segments phenotype.

Table 5.4. Summary of the Epistatic Interactions Found in Four Pairs of Facial Segments. The segment in which the epistatic interaction was found is listed followed by the RSID's, the facial morphs, the associated genes, and the p-value associated with the linear regression-based test of epistasis. The facial morph is shown to better depict the potential phenotypic effects with red colorations indicating an outward direction while blue colorations indicate an inward direction

	SNP1			SNP2			
Segment	*RSID*	*Facial Morph*	*Assoc. Gene*	*RSID*	*Facial Morph*	*Assoc Gene*	*Pvalue*
6	rs10838269 Chr 11		TSPAN18	rs11175967 Chr12		IRAK3	9.94E-07
9	rs76244841 Chr1		GNB1	rs62443772 Chr7		GLI3	4.68E-06
11	rs6740960 Chr2		PKDCC	rs6795164 Chr3		SLCO2A1	5.21E-05
22	rs7373685 Chr3		GATA2	rs7843236 Chr8		SNTB1	7.10E-05

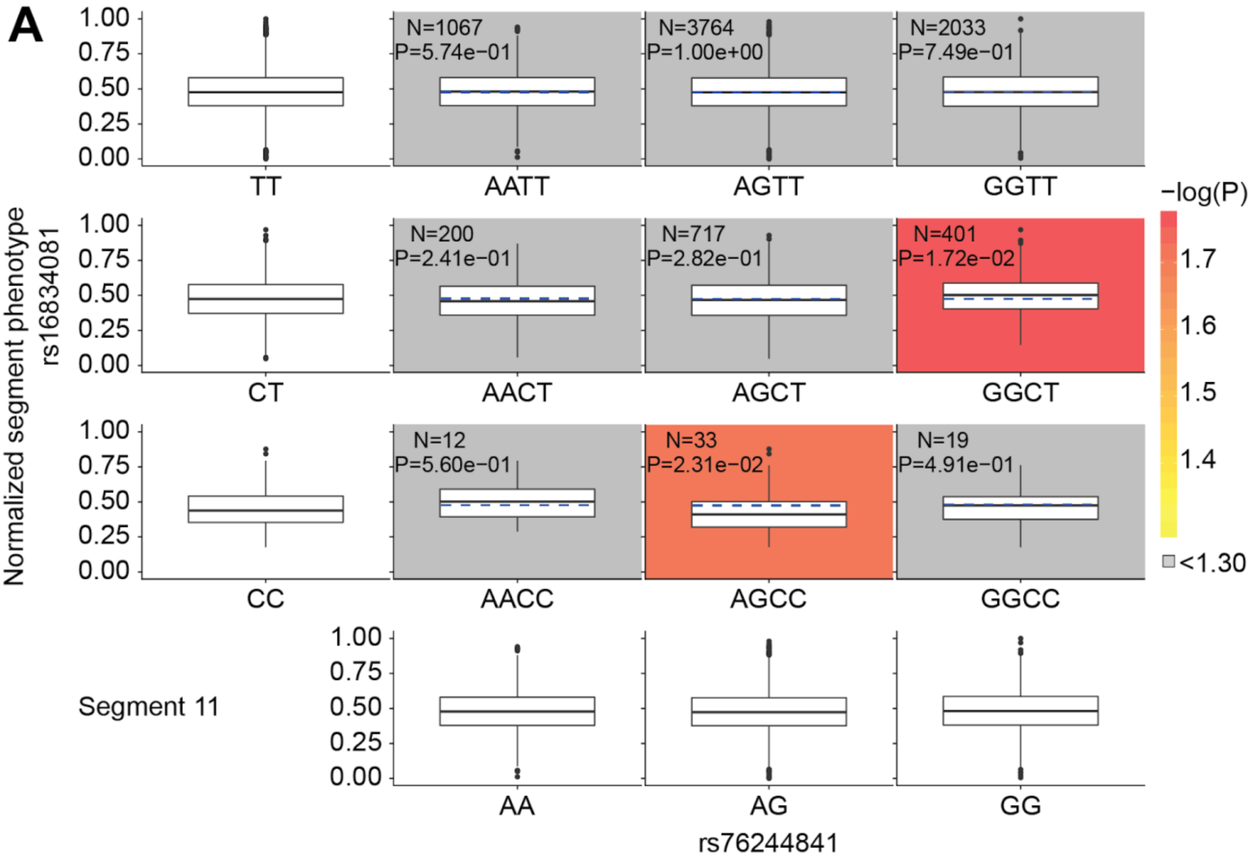


Figure 5.6. Phenotypic Effects of the Epistatic Interactions Found in Four Pairs of Facial Segments. Phenotypic and marginal distributions for diplotype combinations for A) a random SNP pairing and B-E) each significant epistasis pair. Boxplots are plotted to visualize the epistatic effect on the phenotype. The marginal phenotypic medians of the singular genotypes (first column and last row of boxplots) were used to calculate and visualize the predicted diplotype phenotypic distribution that would occur if the two genotypes were acting independently (dashed blue lines in colored boxplots). This median was compared to the observed medians of the diplotypes (solid black lines; colored boxplots) via Mood’s Median test with one degree of freedom. Log transformed p-values were used to color the boxplots if there was a significant ($p < 0.05$; $\log(p) > 1.30$) difference between the expected phenotype of the combined genotype and observed diplotype

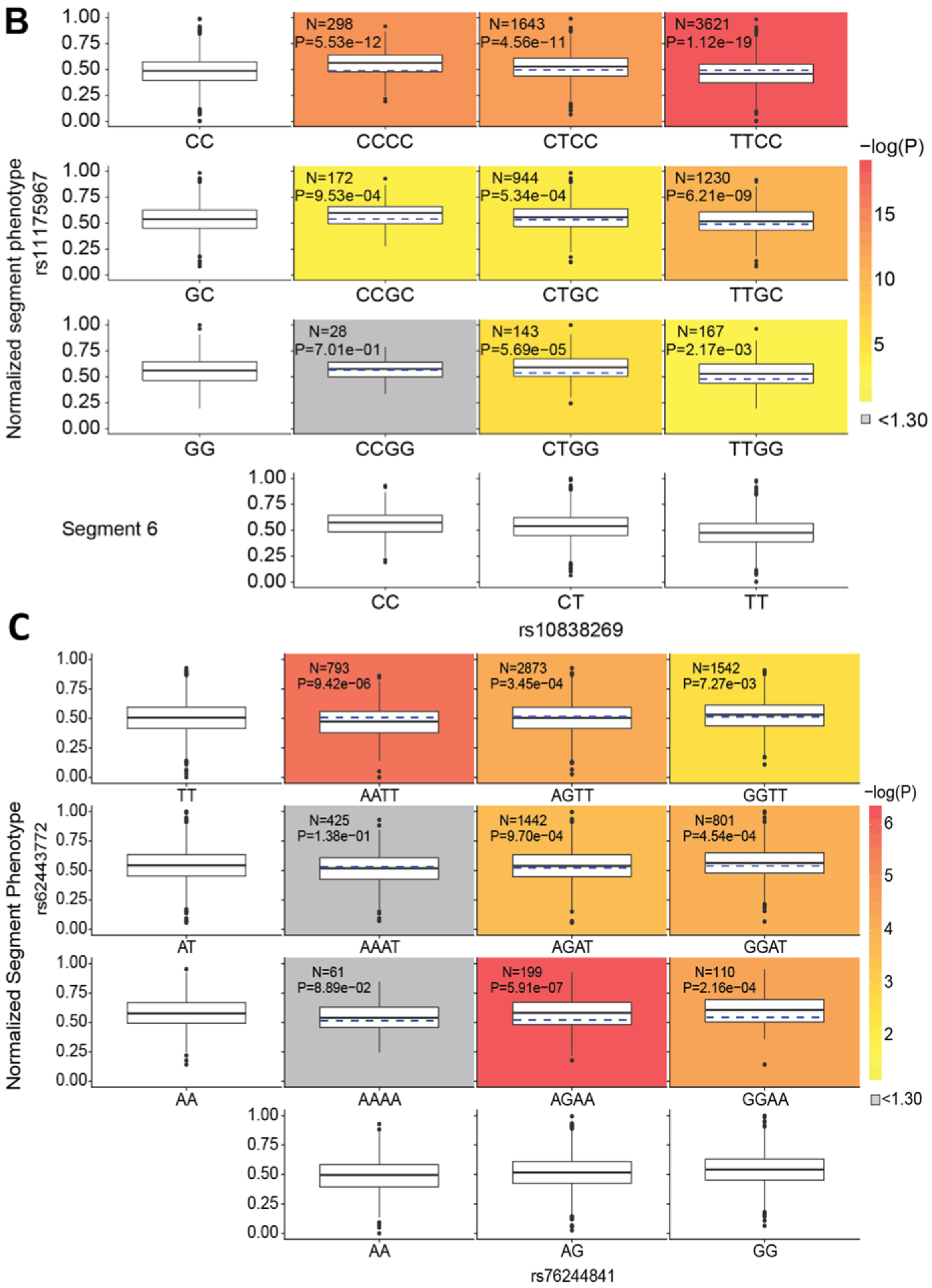


Figure 5.6. Continued

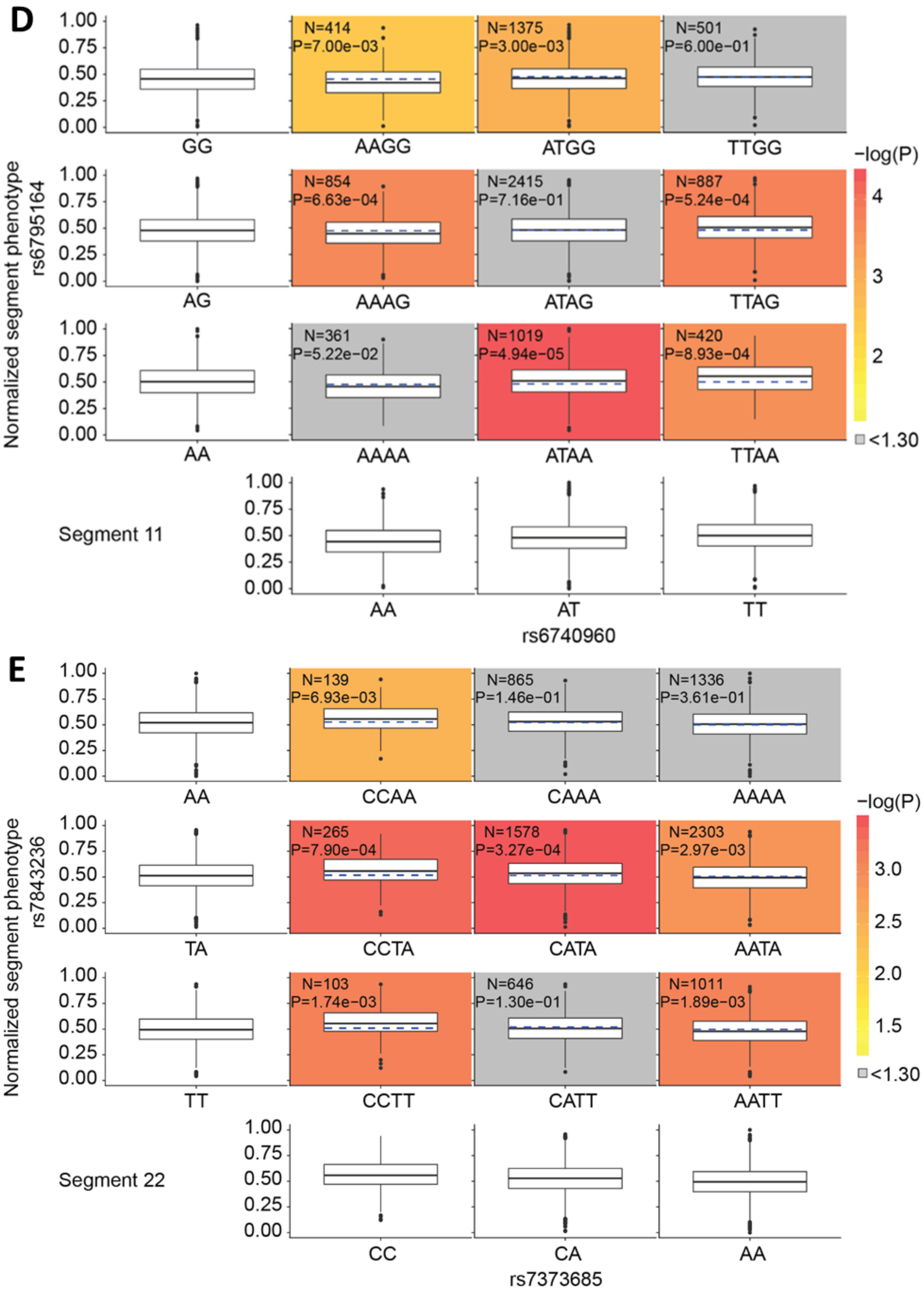


Figure 5.6. Continued

The strongest pairwise variant epistatic interaction ($p = 9.94 \times 10^{-7}$) between rs10838269 (*ALX4* associated) and rs11175967 (*HMGA2* associated) within segment 6 (See Figure 5.6 B), which covers the area of the face from the zygoma to the mandible, was explored further via a literature search. Rs11175967 is an intronic variant mapped to the *HMGA2* gene that encodes a high mobility group AT-hook protein, which forms components of enhancesomes and may also function as a transcriptional regulator[181]. HMGA2 has been associated with Silver-Russell Syndrome, symptoms of which include a triangular face shape and broad foreheads[182,183]. *HGMA2* has also been linked to H3K27ac chromatin regulation in osteoblasts [GEO:GSM733739, Bradley Bernstein, Broad]. Rs10838269, its epistatic partner, is an intergenic variant whose nearest protein coding gene is *ALX4*. *ALX4* encodes a paired-like homeodomain transcription factor that is known to be expressed in the mesenchyme of developing bones and has been shown to play a vital role in craniofacial development[184]. Based on experimental evidence from Encode, the variant is located in a region that is a known to be regulated by H3K27ac [GEO:GSM733656, Bradley Bernstein, Broad] in K562 (i.e. bone marrow) cells. *ALX4* has also been shown to be involved in Potocki-Shaffer Syndrome, which is characterized by abnormal bone development including deformations in the parietal bones that comprise the top and sides of the skull[185].

ALX4 knockouts in mice have also demonstrated abnormal fusion of nasal cartilage[186,187]. The connection between *HMGA2* and *ALX4* has already been documented in genome-wide analyses of adaptive loci in sheep, where the authors found both *HMGA2* and *ALX4* contributing to sheep phenotype with the former being tied to ear morphology and the latter being associated with "stature and morphology."[188]

In addition, genomic analyses on finches have shown that alterations of *ALX1* and *HMGA2* have been associated with beak shape[189] and size[190]. It is important to note that while *ALX1* and *ALX4* are separate proteins, the work by Qu et al., 1999 have found the two paired-like homeodomain transcription factors functionally redundant[187]. Therefore, beak morphology is affected by multiple interconnected genes working in tandem, which may be very similar to vertebrate craniofacial development, which is also controlled by complex multi-gene pathways[191,192]. Thus, this epistatic interaction between *ALX4* and *HMGA2* found *in silico* to have an effect on human craniofacial morphology also has support in the literature from non-human studies that illustrates its effect on face and beak shape. Future functional examination of this proposed epistatic connection may provide additional evidence of this gene masking effect.

Rs76244841 (*PRDM16* associated) and rs62443772 (*GLI3* associated) were also found to have a significant interaction ($p = 9.94 \times 10^{-7}$) in facial segment 9 (See Figure 5.6 C). *PRDM16* encodes a zinc finger transcription factor[193,194] and has been shown to repress TGFβ signaling[195]. While previous studies have demonstrated a link between TGFβ signaling and craniofacial development[196–199], Bjork et al., 2010 specifically suggests that *PRDM16's* modulation of TGFβ effects craniofacial development including palate shelf elevation[200]. *GLI3* encodes a transcriptional activator and a repressor of the sonic hedgehog (Shh) pathway, which has been shown to play a role in limb development[201–203]. In addition, there is evidence that mouse null *Gli3* mutants result in a broad nose phenotype[204] as well as genome-wide scans that identified *GLI3* as affecting nose morphology[27]. The connection between *PRDM16* and *GLI3* can be best seen through their interaction through the *SUFU* intermediary, which is a negative regulator of the hedgehog/smoothened signaling pathway[157,205–209]. Multiple studies conducted on Drosophila melanogaster have identified evidence for a tetrameric Hedgehog signaling complex comprising Fu, Ci (an ortholog of PRDM16), Cos2, and Su(fu) (an ortholog of SUFU), including evidence that Su(fu) binds directly to Ci[210–212]. Subsequently, SUFU has been shown to mediate the phosphorylation of GLI3 via GSK3[213] and has also been shown to interact with the GLI1-3 zinc-finger, DNA-binding proteins[206,214]. Thus, the literature suggests a link between *PRDM16* and *GLI3* via the *SUFU* intermediary, which contributes to palatal shelf elevation and nose morphology (specifically, nose width). Interestingly, the facial segment we observed this epistatic interaction within was segment 9, which covers the premaxillary soft tissue from the base of the columella to the oral commissure.

Rs6740960 (*PKDCC* associated) and rs6795164 (*SLCO2A1* associated) ($p = 5.21 \times 10^{-5}$), and rs7373685 (*GATA2* associated) and rs7843236 (*SNTB1* associated) ($p = 7.10 \times 10^{-5}$) were found in facial segments 11 and 22 (See Figures 5.6 D and E) respectively, which are hierarchical masks that include areas surrounding the base of the nose. Due to the overlapping nature of the masks, these variants were analyzed as a collective group. The nature of the relationship between these four variants is less clear, however some trends are evident. The first is that there appears to be a connection between *GATA2* and *SLCO2A1* through *AKT1*. AKT1 is one of 3 related serine/threonine-protein kinases first found in mouse models[215], which regulate multiple processes, such as metabolism, cell survival, growth, proliferation, and angiogenesis. It has also been indicated in Proteus syndrome whose symptoms include bone development abnormalities[216,217]. SLCO2A1 is a solute carrier involved in the release and transport of prostaglandin[218,219] and has also been shown to be involved in hypertrophic osteoarthropathy[220–222]. *SLCO2A1* regulates *AKT1* and the Akt pathway through prostaglandin[223]. Furthermore, the PI3K/Akt signal pathway has been shown to negatively regulate the transcriptional activator *GATA2*[224]. There were not any connections found with *PKDCC* and *SNTB1*, however, there was an interesting connection between *SNTB1* and *GATA2* via Dystrophin (DMD). DMD serves as a key component of the dystrophin-associated glycoprotein complex, which helps stabilize the sarcolemma[225]. SNTB1 is an adapter protein that has been suggested to link receptors to the dystrophin glycoprotein complex[226,227]. GATA2 has also been shown to be a transcriptional factor of DMD[215]. Finally, there is evidence in mouse models that supports a connection between the Akt signaling pathway and DMD[216], which serves as another underlying link between three of the four epistatic hits (*SLCO2A1*, *GATA2*, and *SNTB1*). While there were no evident links between *PKDCC* and the other epistatic hits, it may be worth noting that this tyrosine-protein kinase has been previously shown to be involved in bone growth[217–219].

Therefore, SEM conducted on a particular facial segment, consisting of a PC comprised latent variable, and utilizing a pool of 203 significantly associated facial variants was found to be capable of identifying relationships between variants that can then be used to assess for epistatic interactions. However, it is important to note that bioinformatic analyses such as GWAS and SEM only give us a starting point, or in this case a list of variants, that we can then use to create functional experiments in order to determine whether these variants found *in silico* are valid *in vivo*.

5.3.3 Using SEM for Prediction

The constructed refined SEM models can also be used to predict the latent factor phenotype given a genotypic profile. From these models, the predicted latent factor phenotype can then be translated back into principal components and used to construct a predicted 3D model of the particular facial segment. Using a refined model, several artificial genotype profiles, which contained the NCBI definition of a 100% ancestral, a 100% derived profile (based on NCBI), and 20 additional randomized genetic profiles were combined with the covariate mean (as determined from the US-UK combined dataset). These profiles were fed into the SEM model, which resulted in the prediction of the principal components for the corresponding segment. The principal components were subsequently projected into the initial PCA space and output visually. While these projections were done for all models that had a refined model, only the masks showing the most noticeable results are displayed in Figure 5.7. These projections are for display purposes only since there has been no attempt to assess the accuracy of these predictions, which are based on the genotypes of simulated individuals. In order to generate these projections, principal components are reinserted in the shape space of each segment (generated from the original set of individuals) and the image is exaggerated along the linear axis (Figure 5.7 “Exaggerated”) to better show the difference between the ancestral and derived projections.

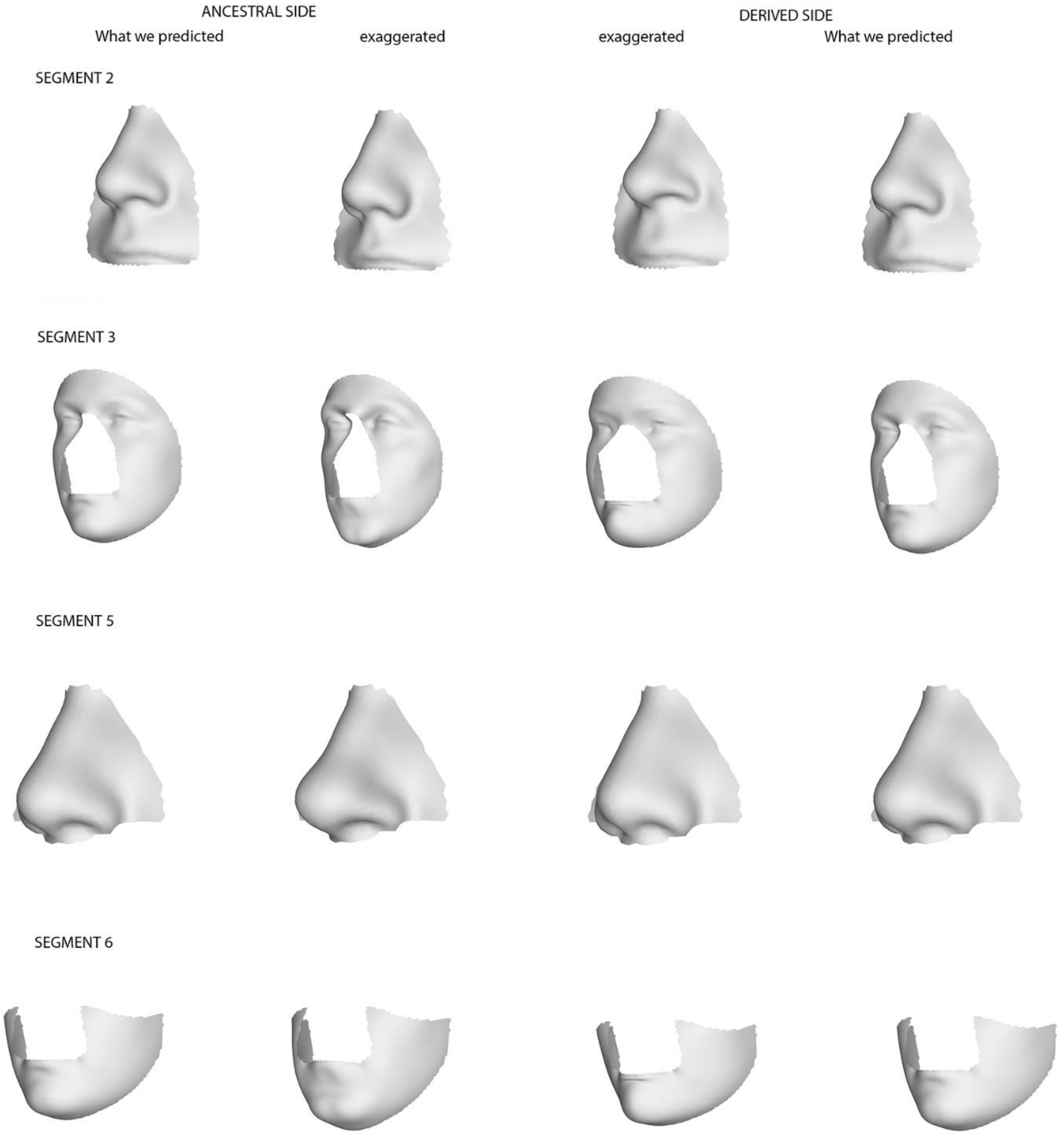


Figure 5.7. SEM Projections Based on Theoretical Ancestral and Derived Individuals Using 203 Variants. Segments 2, 3, 5, and 6 are shown with the ancestral profile being shown in the leftmost two columns and the derived profile being shown in the rightmost two columns. Since the ancestral and derived profiles did not encapsulate the extreme phenotypes like we had assumed, we mathematically exaggerated the phenotypes (columns 1 and 3; “exaggerated”) to better visually demonstrate phenotypic differences

5.4 Conclusion

In conclusion, by employing a multivariate technique that is capable of modeling interactions between correlated genotypes and a multifaceted phenotype, I have shown that it is possible to delve deeper into traditional GWAS results in order to explore the cross talk that occurs between variants that may influence facial morphology. This analysis shows that it is possible to assign groups of variants to certain segments of the face and measure their contribution towards understanding the variation that exists within that segment. It is also possible to identify pairs of variants that are directly masking/enhancing phenotypic distributions due to their diplotype combinations, which warrant further functional investigation. Through these advanced methods (e.g. CCA and SEM) we are bridging the gap between statistically correlated variants, their potential biological interactions, and measuring their phenotypic variance.

CHAPTER 6. CONCLUSION

Over the course of this research I found novel results by innovating and implementing new phenotyping and analysis methods. For example, many previous iris pigmentation studies have phenotyped iris color by assigning eyes into three simple categories: blue, intermediate, and brown. Even quantitative methods more recently applied could not capture true iris color as they either extracted a small segment of color or averaged color across every image pixel, which is not a true reflection of the phenotype. However, by digitally training and quantifying irises, we were better able to capture the mixture of color seen within irises. Due to this approach, several variants were identified that may warrant further exploration through replication and functional studies. Rs3820285, located within *CELA3A*, was found to be associated with blue iris phenotypes. While no biological link was found between *CELA3A* and pigment, we did find the variant nearby various regulatory features that were within 17kb of a gene, *CDC42*, known to affect melanin deposition[135]. Rs77373930, an intronic variant located within *AC007389.1*, was also identified in both green iris and the PCA-FA reduced phenotype GWAS, as well as rs6420484, an intronic SNP located within *NPLOC4* and *TSPAN10*. The biological evidence for a pigment connection was low for rs77373930, but was high for rs6420484, which was found in a previous iris color GWAS[108] and is also in LD with several variants located in *PDE6G*, a gene known to cause Retinitis Pigmentosa[139,140]. In order to gain power in our GWAS to bring the near-genome-wide significant variants past the significance threshold, we will continue to refine our phenotyping training and classification. One trend that we saw over the course of the analysis was that some individuals who may be deemed as having blue eyes were quantified as brown due to some of the thresholds that were set within the training parameters. By improving the training and quantitation procedures, we may further increase analyses power and remove background noise.

Currently the standard output for iris prediction is a categorical output of blue, intermediate, or brown. Here I illustrated that by implementing a neural network, we are capable of outputting a model that can predict a quantitative amount of four iris color classes with a mean absolute error (MAE) of 13.57% +/- 0.36% across all eye colors. In addition, when the model was compared with a currently published model, the results were very similar, although we were only testing five individuals so more extensive testing is required. Nevertheless, after nearly a decade since the first categorical iris prediction tool was released[228], it is long past due for a quantitative prediction of iris color. However, the application of 'black-box' neural networks and deep neural networks must be treated with caution in terms of feature selection and model structure. With that caveat, we will continue to improve our quantitative prediction tool by further increasing its accuracy and usability. Although the prediction model had trouble predicting individuals with high green iris content, it was still able to capture some portion of perceived green within the iris. The aim during the association analyses was to try to find new variants associated with that phenotype, however there were no clear hits. There are candidates that required further study, and perhaps examining an increased number of green-eyed individuals for these markers, diplotype combinations, and/or epistatic interactions may allow a better prediction of this difficult phenotype. As green was the most underrepresented iris category in the dataset, we may be able to improve the model by either enriching our testing set with additional green irises, or by reducing the number of non-green irises in the training set. We also plan on consulting with computer scientists in the creation of an 'artificially' created iris output for prediction visualization purposes instead of relying on dataset retrieved irises, which due to their 'identifying' biometric capabilities would not be ideal to use as output for the prediction of an 'unknown' person. One of the avenues we plan to investigate in particular is the implementation of a generative adversarial network (GAN), which has showed promise in recreating human irises from a dataset of iris images[229].

To date many facial morphology studies have focused on examining distance measurements between facial landmarks. However, none had analyzed simple face or jaw shape. Here it was found that despite having a simpler phenotype, we were still able to detect variants that may affect face shape. Rs187236608, and intronic variant within *STON1*, rs11431304, an intronic variant found within *LHCGR*, and rs113531385, an intronic variant located within *STON1-GTF2A1L*, were all identified in this manner. While *STON1* did not return any connections with facial morphology, *LHCGR* was associated with delayed bone maturation and reduced bone formation in a mouse study. *STON1-GTF2A1L* has links to *LHCGR* through testes biology and may result in similar developmental disorders. We will need follow-up replication studies in order to validate our results and, if validated, subsequent functional studies to explore the biological role of these variants. It may also be advantageous to find a more objective method of classifying categorical face shape in order to boost power, reduce classifier bias, and increase the speed of phenotyping even further. One avenue we may explore is deep learning. Specifically, we may try to implement a convolutional neural network that is capable of classifying image inputs to better assess jaw shape and even entire face shape.

In addition to looking at a simpler method of facial phenotyping, we also implemented a network-based analysis method to analyze the result output of a more complex facial phenotyping analysis. By using the facial phenotyping methods as discussed in Claes et al., 2018 and White et al., 2019, I was able to implement a structural equation model on 203 variants output from their canonical correlation GWAS. By giving structure to their results through modeling the variants, the PC phenotypes, and the latent factor comprised of PC's, we were able to assess how each variant affected this complex network and how they worked together via epistatic interactions to either amplify or suppress a facial phenotype. It should also be highlighted that while SEM's have long been used to model structure within simple questionnaire data[230–232], far fewer have attempted a SEM of this magnitude on variables as complex as genomic variants[233,234]. Still, to follow up on our work, we may attempt to delve into some of the canonical correlation intermediates that are created prior to the creation of the single latent variable that was used in the analysis that was submitted for review. We may also explore other types of multi-trait GWAS analyses[235], in order to better analyze this multi-faceted phenotype.

As our analysis and phenotyping methods improve, so too must our bioinformatic tools that make these research projects possible. One of the major hurdles in performing bioinformatics, especially for first-time users, is the lack of inter-operability between various analysis programs. By creating the *Odyssey* pipeline, I attempted to rectify this problem for several types of commonly used GWAS analysis. This pipeline will drastically decrease the analysis time of phasing and imputation as well as offer an offline imputation solution for those who are unable to use other online options such as the Michigan Imputation Server or the Sanger Sequencing Service. I also developed several 'add-on' modules that should benefit researchers wanting to QC their data, perform phenotype normalization, and wish to correct for ancestry prior to performing a GWAS. However, much can be improved with *Odyssey* from making the code cleaner, improving the user-interface, and expanding the amount of analysis options.

While there is still much to explore, our understanding of human facial morphology and iris pigmentation continues to expand. As we continue to innovate new phenotyping methods and perform new types of analyses, we hope to incorporate the variant associations we find in a variety of applications such as prediction modeling for forensic and anthropological purposes. These variants may also lay the groundwork in order to understand various pigment and morphological diseases, which may someday lead to treatment. Hopefully, the advances made through the course of this research can be used as yet another stepping-stone to further advance these two fields and help those who will reap the benefits of their potential applications.

APPENDIX A. ODYSSEY USER GUIDE

Odyssey

Version: 2.1.0

Odyssey is semi-autonomous workflow that aids in the preparation, phasing, and imputation of genomic data. Odyssey also provides support for custom reference panel merging, population structure (admixture) analysis, phenotype normalization, genetic data quality control, genome-wide-association studies (GWAS), and visualization of analysis results.

If you use Odyssey in any published work, please cite the following paper:

Eller, Ryan J., Sarath C. Janga, and Susan Walsh. "Odyssey: a semi-automated pipeline for phasing, imputation, and analysis of genome-wide genetic data." BMC bioinformatics 20.1 (2019): 364. https://doi.org/10.1186/s12859-019-2964-5

General Overview

Odyssey relies primarily on the idea that the creation of imputed genomic data must go through 4 basic steps: pre-imputation QC, phasing, imputation, and post imputation cleanup and QC. Once Odyssey is setup, users can automate these 4 steps on a High Performance System (HPS) running some form of Linux or a desktop (if you are not concerned about time requirements as phasing/imputation is normally resource intensive). Odyssey contains 6 main folders. The '0_DataPrepModule' directory is where the user places his or her target data that needs to be remapped and/or 'fixed' to a reference genome (to assure proper genomic positions and allele 'flippige'). The '1_Target' directory is where the cleaned data from '0_DataPrepModule' is automatically put following the data being cleaned. The 'Reference' directory is where the reference data that will be used by Shapeit2/Eagle2 and Impute4/Minimac4. The '2_Phasing' directory is where phasing scripts, logs, and phased data will be housed while the '3_Impute' directory is where an imputed dosage VCF will be stored. The '4_GWAS' directory is where GWAS analyses are conducted and also contains the Phenotype sub-directory where phenotype normalization may be performed. A 'SubModule' directory contains the 'PopStratModule,' which performs admixture analysis on a given dataset, 'Remapping_Made_Easy_Submodule, which allows for semi-autonomous genome remapping on a given dataset, and the 'HGDP_Starter-SubModule', which formats a version of the HGDP dataset for Odyssey use. Lastly the '5_QuickResultsis' directory populates metadata collected over the range of analysis for a given imputation or GWAS project. Each imputation run and GWAS analysis are stored in discrete project folders within the main 4 directories (e.g. 1_Target, 2_Phasing, 3_Impute, and 4_GWAS).

Odyssey is setup based on a pre-made downloadable Singularity container, which contains all but 2 of Odyssey's dependencies. Impute4 must be downloaded separtely due to licensing restrictions, and GNU-Parallel (which is an optional dependency that speeds up runtime) must be configured separtely. All of Odyssey's dependencies are housed within the 'Configuration' directory.

Users should also make note of the Settings.conf file, which guides users in setting Odyssey's variables, which controls the workflow. This method of user control was used instead of the more traditional command line arguments in order to simplify the workflow while allowing the users to visualize all the options Odyssey has to offer.

Once Odyssey is setup and the workflow settings are configured in Settings.conf, the user can execute the various scripts in the main Odyssey folder which cleans, phases, imputes, and performs a GWAS on the data. Odyssey is optimized for human data, but it is theoretically possible to analyze data from other organisms using Odyssey as well.

Odyssey Tutorial

As an added reference, an Odyssey Tutorial has been provided which contains a 100 sample HGDP dataset (provided in the 'Tutorial_Data' directory) as well as an 'Odyssey Tutorial [v#]' document, which walks you through the data prep, phasing, imputation, post imputation QC that is needed to create a dosage VCF, as well as some of the sub modules. The tutorial as well as configuration instructions can be found within the 'Odyssey_Literature' directory. I would highly recommend new users to utilize the detailed explanation that can be found in the tutorial in order to become more familiar with Odyssey. Running through the tutorial covers all the essential steps, offers some tips, and provides a look at more advanced settings that will hopefully make data prep, QC, phasing, imputation, post Imputation QC, GWAS and admixture analysis, and phenotype normalization very simple and easy. It will only take around 3-6 hour to complete the tutorial given adequate computational resources. One can even truncate the analysis to a few chromosomes to decrease the amount of time to completion even further.

Quick Setup

Note: It is best to use Odyssey on a system that has Singularity already installed. There are non-Singularity setup methods of setting up Odyssey, but doing so is not recommended. A detailed explanation of Odyssey setup can be found in 'Odyssey Installation Instructions' found within the Readme folder

Download Odyssey via Github:

1. On the Odyssey Github page navigate to the Github Release tab

   The release should have 3 files for download: OdysseyContainer.sif.gz + Source code (zip) + Source code (tar.gz)

2. Download and extract the Source code to your working directory

```
3.  cd /Directory/You/Want/To/Work/From
4.  unzip Odyssey-[#.#.#]
```

5. Download and extract the OdysseyContainer.sif.gz to the Singularity directory:

```
6.  cd ./Odyssey-[#.#.#]/Configuration/Singularity
7.  unzip OdysseyContainer.sif.gz
```

8. Due to licensing restrictions, you will need to download and extract the Impute4 (if you choose to use that imputation software over Minimac4) executable to the Impute4 directory within Configuration

   a) Request access to Impute4 here: https://jmarchini.org/impute-4/

   b) After downloading Impute4 to the Impute4 directory you will need to alter the following line in ./Configuration/Setup/Programs-Singularity.conf:

   ```
   Impute_Exec4="${WorkingDir}Configuration/Impute4/impute4.1.1_r294.2";
   ```

   d) Replace impute4.1.1_r294.2 with the appropriate version you just downloaded

9. You’re done. Start using Odyssey

Extra Note: While the pre-made releases are an easy way to download everything you need to run Odyssey, I am constantly updating Odyssey's Github page so you may use

Git to pull changes from Github between releases, if you want the latest and greatest improvements.

Setup Reference Data:

Reference data to be used with phasing/imputation is installed by the workflow by default, but can be changed in Settings.conf so that you can use your own reference data. To do so, simply populate the Reference Data Folder with your preferred reference data. By default reference data is downloaded from the IMPUTE2/Minimac3 site under their reference data for "1,000 Genomes haplotypes -- Phase 3 integrated variant set release in NCBI build 37 (hg19) coordinates" (Updated 3 Aug 2015)

If you choose to use a custom reference dataset, then several adjustments may need to be made to the naming of the reference files (.legend, .map, and .hap) and also to the Settings.conf file. These adjustments are explained in greater detail in the Settings.conf file. Users should pay particular attention to make sure that custom reference data is sync'ed to the target data (i.e. don't try and use a Target dataset mapped to GRCH 37 to a reference dataset mapped to GRCH 38).

Setup Target Data:

As an optional first step to cleanup your genomic data (hereafter referred to as "Target" data) you may put your Plink formatted data (.bed/.bim/.fam) in ./0_DataPrepModule/PLACE_DATA_2B_FIXED_HERE/. Running the '0_DataPrepCleanup.sh' script from Odyssey's main directory will give you the option of fixing your target data to a GRCH 37 reference build and will remove positional duplicates if they exist. This step is not required, but highly recommended as positional duplicates will cause the pipeline to error out. Fixing the data to a reference genome also helps reduce the chances of encountering a workflow error caused by 'dirty' data. This optional step will deposit your data in the ./1_Target/PLACE_NEW_PROJECT_TARGET_DATA_HERE/.

Additional Note: As data that will be imputed should match the genomic build of the reference dataset, it may be helpful to utilize the Remapping_Made_Easy_SubModule found within the 'SubModules' Directory. It utilizes NCBI's Remap Service to semi-automate the remapping of your data to a build of your choice (that is supported by NCBI of course). More information on running this sub-module can be found within the Odyssey Tutorial.

The first required step of Odyssey is placing data in the ./Target/PLACE_NEW_PROJECT_TARGET_DATA_HERE/ directory. Data placed here undergo basic quality control prior to phasing (e.g. filtering individuals and variants based on missingness, minor allele frequency, and Hardy-Weinbergy Equilibrium). The cutoffs are set by default to industry standards, however users may choose to adjust the thresholds in Settings.conf. In addition, at this step users may opt to interactively visualize their dataset exploring the quality control measures mentioned above, as well as evaluate sample relatedness via IBD. Dataset QC plots (e.g. IBD, HWE, and Missingness) are saved automatically to a results folder within ./Target/PLACE_NEW_PROJECT_TARGET_DATA_HERE/.

Fill out Settings.conf:

Settings.conf file is responsible for setting the variables that will be used to execute the scripts in the home directory. Essentially, all the main scripts on the home directory "phone home" back to the Settings file to lookup their variables. Because of this, unless additional customization is needed, you should never have to modify any of the main scripts in the home directory (however, each script within Odyssey Modules and Submodules are heavily commented to allow for easy navigation when attempting more advanced customization not supported by the Settings file). Most of these variables are relating to toggling steps of Odyssey (to allow for more user control and to help with troubleshooting), specifying the home directory (i.e. './Odyssey/'), etc. Step-by-step instructions on how to setup Settings.conf variables can be found in the Settings file itself and a more detailed explanation can be found in the Odyssey Tutorial.

A Note on Odyssey File Organization:

Odyssey has an organization scheme to keep all imputation results separate from each other so the user does not have to "reset" the Odyssey folder after each imputation run. Odyssey does this by organizing files into 'Imputatation Projects'. Each project will create a folder that is identified by a BaseName, or a name that is specified in Settings.conf at the beginning of the analysis to identify the imputation run. This will allow for the creation of identifiable folders within the Target, Phase, Impute, GWAS, and QuickResults folders. For example, if I have a dataset of Homo sapien target DNA that I want imputed, I will setup an Imputation Project named "Human_Impute1" (the name must not contain whitespaces). Odyssey will then create a target folder (within the Target directory) specific to my imputation run and move my data into it. Odyssey will then deposit phased and imputed scripts, outputs, and results within the Phase and Impute folders respectively. If

I then want to impute a different set of data, I simply create a new Imputation project, which will separately house the target, phase, and impute data from my second imputation run.

Running Odyssey

Step 1: Pre-Imputation QC and Setup

1. Once the target data has been cleaned and is deposited in the 'PLACE_NEW_PROJECT_TARGET_DATA_HERE' folder within the Target folder directory, the first script, "1_ImputeProjectSetup-QC-Split.sh" can be run from the home directory. Simply use a command prompt to navigate to the home directory (e.g. $ cd /path/to/Odyssey-v[#.#.#]/) and execute the script (e.g. $ bash 1_ImputeProjectSetup-QC-Split.sh) which will setup an Imputation Project Folder, move your Target Data into this Project Folder, and will provide a small amount of pre-imputation QC which includes (by default):

   a) Filtering for individual missingness (removes individuals missing more than 5% of genomic content)

   b) Filtering for genetic variant missingness (removes variants missing in more than 5% of individuals)

   c) Filtering for minor allele frequencies (removes variants that contain a minor allele frequency of 2.5% or less)

   d) Filtering for Hardy-Weinberg Equilibrium (removes variants that have a HWE p-value of 1e-5 or below). This test is very lenient to allow for diverse target data.

2. SHAPEIT/EAGLE requires data to be split by chromosome so the last step is splitting the dataset into their respective chromosomes By default the script looks for chromosomes 1-26 (the default for human samples) into their respective chromosomes.

Step 2: Phasing

1. Odyssey organizes phased data into an Imputation Project Folder created within the Phase folder. The name of this folder is specified by the Imputation Project Name variable and will contain subdirectories that house the phasing scripts, outputs, and results.

2. No additional files outside of those created in Step 1 need to be created to run the Phasing step. Each step builds on the next and contains all the files necessary to run the next step.

3. Phasing is carried out using SHAPEIT/EAGLE recommended settings (shown below) and a reference data map provided by IMPUTE2 (by default) or the user.

   a) The SHAPEIT command has the following general form: shapeit --thread [# threads] --input-bed [PlinkTargetBedFile] --input-map [ReferenceMapFile] --output-max [OutputPhasedName] --output-log [OutputPhasedLogName]

   b) The EAGLE command has the following general form: eagle --numThreads [# threads] --bfile [PlinkTargetBedFile] --geneticMapFile=[ReferenceMapFile] --outPrefix=[OutputPhasedName] --chrom [Chrom #]

4. Phasing customization can be set via altering settings found in Settings.conf

5. Phased output, logs, and scripts are deposited within the Imputation Project directory placed within the Phase directory

Step 3a: Imputation

1. Odyssey organizes imputed data into an Imputation Project Folder created within the Impute folder. The name of this folder is specified by the Imputation Project Name variable and will contain subdirectories that house the imputation scripts, outputs, and results. Note that either Impute or Minimac can be used with a pre-phased dataset run through either Shapeit or Eagle.

2. Imputation is carried out using IMPUTE/Minimac recommended settings using reference data (genetic, hap, and map files) provided by either IMPUTE/Minimac (by default) or the user. The General IMPUTE/Minimac commands are listed below. Note that the reason why Impute4 is used is because it has superior speed in comparison to Impute2. Minimac is another imputation solution and is setup

by default for users who do not need to utilize Impute's abilities to more accurately impute admixed datasets.

a) impute4 -g [PhasedHapsFile] -s [PhasedSampleFile] -m [ReferenceGeneticMapFile] -h [ReferenceHapsFile] -l [ReferenceLegendFile] -int [StartChromosomeChunkSegment EndChromosomeChunkSegment] -maf_align -Ne [20000] -o [OutputName]

b) minimac4 --cpus [# cpus] --allTypedSites --minRatio 0.00001 --refHaps [ReferenceHapsFile] --haps [PhasedHapsFileFromSHAPEIT] --prefix [OutputName]

3. Imputation customization can be set via altering settings found in Settings.conf.
4. Imputed output, logs, and scripts are deposited within the Imputation Project directory placed within the Raw_Imputation folder within the Impute directory

Step 3b: Post Imputation Cleaning and Concatenation

Post imputation cleaning is performed differently depending on whether Impute or Minimac was used.

For Impute Workflows:

1. Since imputted files are divided by chromosome and by segment, these files must be concatenated. Odyssey does this through 3b_ConcatConvert.sh which does a simple concat command with all the imputed chromsomal segments housed within the Raw Imputation folder and re-assigns their chromosome number (which isn't explicitly assigned during imputation)
2. SNPTEST then creates a SNP Report which calculates the INFO imputation QC metric. This will later be used to filter the VCF file (the INFO cutoff is set to 0.3 by default, but may be adjusted in Settings.conf)
3. Concatenated chromsomal .GEN files are converted to a dosage VCF file (.VCF) using Plink 2.0 and filtered by INFO score
4. The dosage VCF files are concatenated via BCFTools
5. The final output within the 'ConcatImputation' Folder contains the following:

a) .snpstat (SNPTEST snp report that contains several metrics on the imputed chromosome including the INFO score)

b) .snpstatOut (is a log file for SNPTEST which contains the run results from SNPTEST AND a count of the total number of variants imputed for the particular chromosome and how many are left after INFO filtering)

c) Most importantly is the 1Imputed_[BaseName].vcf.gz file which is the final imputation product. This dosage VCF file can be inputted into analysis programs such as Plink 2.0, SNPTEST, GenAble, etc. for further analysis

For Impute Workflows:

1. By default Minimac concatenates imputed chromosomal segments into imputed chromosomal.vcf.gz files which already contains the Minimac R2 (very similar to Impute's INFO scores) imputation QC metric.

2. These chromosomal.vcf.gz files are then filtered via Plink

3. Filtered vcf.gz imputation files are then concatenated via BCFTools

4. The final output within the 'ConcatImputation' Folder contains the following:

   a) .log and .out files detailing the actions performed on individual dosage chromosomal vcf.gz files

   b) Most importantly is the 1Imputed_[BaseName].vcf.gz file which is the final imputation product. This dosage VCF file can be inputted into analysis programs such as Plink 2.0, SNPTEST, GenAble, etc. for further analysis. Unlike the Impute vcf.gz product, the Minimac vcf.gz has the R2 INFO metrics baked into the vcf.gz file itself for all imputed and typed variants

Step 4: Setup GWAS Project and Run GWAS Analysis

The last step in Odyssey will be to setup a GWAS Analysis Project where a dosage VCF can either be manually imported into Plink or an Imputation Project can be specified, which will allow Odyssey to automatically lookup the dosage VCF file and corresponding sex sample file (the .fam file for the dataset which contains sex information) and perform an analysis. How Plink imports the data can be setup via Settings.conf. The GWAS analysis is designed to perform and visualize a genotype:phenotype analysis (i.e. a Genome-Wide-

Association Study). Specifically, the genotypic data and inputted phenotypic data is fit on a general linear model or a logistic model, and R is used to visualize the output.

1. Users will need to setup a GWAS Project for the GWAS analysis by completing the GWAS Project Variables Section of Settings.conf. More specific details on how to fill out the variables are included within the Settings file itself and the tutorial, but briefly:

   a) Specify the GWAS Project Name

   b) List the Imputation Project the user wishes to analyze or manually list the path of a dosage VCF

   c) Specify the name of the covariate/phenotype file that correspond to the imputation files (which is placed in ./4_GWAS/Phenotype).

   Note: It is important that the user read the instructions in the Plink manual regarding the formatting of the phenotype and covariate files.

   d) List any additional Plink options to be run during the analysis withing the "Plink_Options" variable

2. Run the '5_AutomatePlink.sh' script from the Odyssey directory

3. A new GWAS Project directory should be created which upon analyis completion should contain the following:

   a) An analysis log file

   b) A QQPlot

   c) A Plotly interactive Manhattan Plot (with its Plotly dependency folder)

   d) A summary table that contains the top 10000 variants with the lowest unadjusted p-values (as well as multiple comparison corrections, effect sizes, etc.)

   e) A gzipped file that contains the raw results from Plink

Additional Odyssey Tasks

Population Stratification Analysis

Odyssey is capable of performing an Eigensoft-like analysis to assess a datasets ancestry structure using reference datasets such as the 1000 Genomes or HGDP datasets. Follow the instructions found within 00_PopStrat_Helper.xlsx, which is an Excel file that already has 1000 Genomes and HGDP datasets pre-populated and organized. Briefly, a Population Stratification workflow follows the following steps:

1. 2 datasets, a reference dataset that contains known ancestry and a target dataset that contains ancestries that will be determined, are placed in the ./SubModules/PopStratModule/PLACE_Target-Ref_Datasets_HERE/ directory

2. Common genotypes are found between the 2 datasets in order to perform an inner-merge (i.e. the 2 datasets are merged so that only the common genotypes are retained)

3. The resulting dataset undergoes some basic quality control measures (i.e. missingness and HWE) and then pruned based on linkage disequilibrium

4. A PCA analysis is performed on the combined, QC'ed, and prunned dataset

5. Files that contain the PCA eigenvectors and eigenvalues are used in tandem with an ancestry file (which is a Plink formatted ID list of reference samples that contain the ancestry the user wishes to keep) that the user uploads

6. An R script is then used to construct an X-dimensional centroid based on the eigenvectors of the samples that are of the ancestry the user wishes to retain. Outliers that fall outside of the X-dimensional centroid are determined based on a specified standard deviation or inter quartile range cutoff.

7. The output of the analysis is an interactive 3D Plotly scatterplot that color coordinates individuals who are deemed outliers, reference samples, and samples that should be included in the analysis. A text document recommending which individuals should be removed based on ancestry is also provided.

GWAS Phenotype Normalization

Non-normal phenotypes occasionally negatively affect GWAS analyses. In response Odyssey contains an Analyze/Transform Phenotypes script housed within the ./4_GWAS/Phenotype directory. Running this script will prompt for the path to a Plink formatted phenotype file to analyze and will then proceed to show the distribution of the phenotype by column (with each column, m>2, being a different phenotype). See https://www.cog-genomics.org/plink/1.9/input#pheno for more information on phenotype file formatting. This interactive script will then ask the user whether to perform a Yeo-Johnson, Rank-Order Inverse Normalization, or no normalization on the selected phenotype. If normalizing the script will visualize the before and after, ask whether the transformation is to be accepted, and if accepted will append the transformed phenotype onto the end of the phenotype file.

Custom Imputation Ref Panel Creation (BETA)

As more sequence data is being collected it is now possible for users to create highly customizable imputation reference panels that better suits their needs. If a user has several reference datasets (e.g. custom collected data in addition to 1000 Genomes data) and wishes to merge them into a single reference dataset, then this can be accomplished using the MergeRefSets Submodule found within ./Reference/MergeRefSets/. In general, the workflow is as follows:

1. Put Reference dataset 1 in ./Reference/MergeRefSets/Reference1
2. Put Reference dataset 2 in ./Reference/MergeRefSets/Reference2
3. Add the proper genetic map files in ./Reference/MergeRefSets/GeneticMaps
4. Configure settings in Merge.conf
5. From ./Reference/MergeRefSets run

```
6.  $ bash ./MergeRef.sh
```

7. The two reference datasets will be merged using Impute2's merging function (explained here: https://mathgen.stats.ox.ac.uk/impute/merging_reference_panels.html) and deposited in ./Reference/MergeRefSets/MergedRefPanels

Note: Depending on the size of the reference panels, this operation may take an incredibly long time since the operation is essentially imputing the reference sets to each other, which is essentially a dual-imputation procedure.

APPENDIX B. ODYSSEY PERFORMANCE METRICS

Table B1 and B2. Benchmarks for *Odyssey*. Benchmarks were conducted on an admixed HGDP sample set of 940 individuals (542K variants after quality control) with a 1000 Genome Phase 3 Reference dataset (80M variants) on a High-Performance Cluster using a SHAPEIT-IMPUTE and a Eagle-Minimac workflow. While Odyssey can be parallelized almost infinitely, benchmarks were conducted on 3 CPUs throughout the pipeline unless programs were unable to utilize the additional cores or the use of additional cores would be unlikely to increase performance. Benchmarks performed for each step are listed and illustrate the maximum amount of RAM and time required to complete each step. Since phasing and imputation can be massively parallelized their benchmarking times are divided into “Time per Job,” the time it took to complete a single job for a benchmarking sample for the given step and “Time per Step,” the total amount of time it took to complete the entire step for the entire HGDP dataset running on up to 150 concurrent jobs on the HPS. Two tables are shown for two specific workflows, the SHAPEIT-IMPUTE workflow (left) and Eagle-Minimac (right)

Step	Max Memory Required	CPU's Used	Benchmarking Sample	Time per Job on HPS	Time per Step on HPS
Step 0 - Clean-up	0.9 GB	1	HGDP Dataset	0:18:15	0:18:15
Step 1 - Pre-QC	0.1 GB	1	HGDP Dataset	0:01:07	0:01:07
Step 2 – Phase*	1.5 GB	3	Chr 1	1:27:53	1:27:53
Step 3a – Impute*	1.7 GB	1	Chr 1 Segment 1	0:05:38	0:18:31
Step 3b - Convert	2.2 GB	3	HGDP Dataset	5:14:40	5:14:40
Step 4 - Analyze	4.8 GB	3	HGDP Dataset	0:11:25	0:11:25
Step 4 - Visualize	17.7 GB	1	HGDP Dataset	0:12:45	0:12:45
Pop Strat Add-in	17.2 GB	1	HGDP Target – 1K Genomes Reference	0:18:42	0:18:42
Total Estimation	17.7 GB	-	-	-	8:03:18

****SHAPEIT2-IMPUTE4 Workflow**

Step	Max Memory Required	CPU's Used	Benchmarking Sample	Time per Job on HPS	Time per Step on HPS
Step 0 - Clean-up	0.9 GB	1	HGDP Dataset	0:18:15	0:18:15
Step 1 - Pre-QC	0.1 GB	1	HGDP Dataset	0:01:07	0:01:07
Step 2 – Phase*	1.9 GB	3	Chr 1	0:24:22	0:24:22
Step 3a – Impute*	3.2 GB	3	Chr 1	0:44:52	0:44:52
Step 3b - Convert	3.7 GB	3	HGDP Dataset	1:29:32	1:29:32
Step 4 - Analyze	4.8 GB	3	HGDP Dataset	0:11:25	0:11:25
Step 4 - Visualize	17.7 GB	1	HGDP Dataset	0:12:45	0:12:45
Pop Strat Add-in	17.2 GB	1	HGDP Target – 1K Genomes Reference	0:18:42	0:18:42
Total Estimation	17.7 GB	-	-	-	3:00:41

****Eagle2-Minimac4 Workflow**

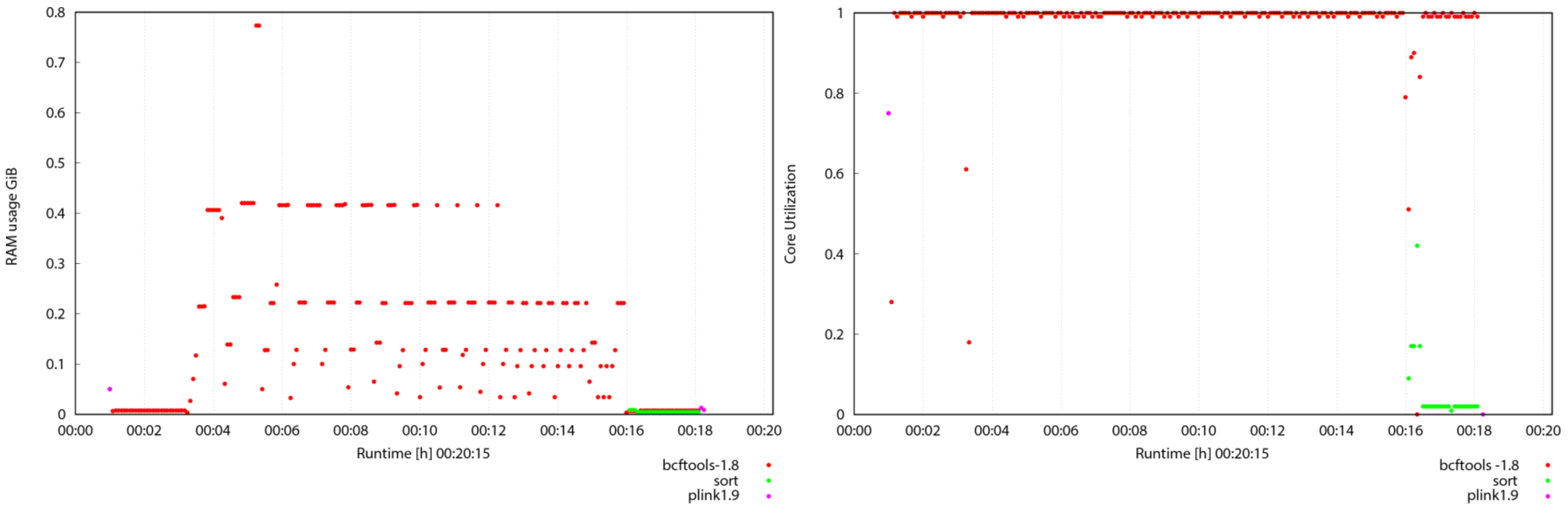


Figure B1. CPU and Ram Utilization for Step 0 – Data Cleanup. Odyssey Step 0, which fixes strand orientation of the user's dataset to a reference genome, was performed on the entire HGDP dataset containing 940 admixed individuals and approximately 542K markers. Collectl was used to monitor the CPU (left) and RAM (right) usage for each process (i.e. program) in the step as well as the total time for completion. A buffer of 120 seconds was added to the end of the step for technical reasons. One hyperthreaded CPU core and 32 GB RAM were allotted for Step 0. Using the entire specified dataset, the data cleanup step would require approximately 18 minutes and 1 GB RAM

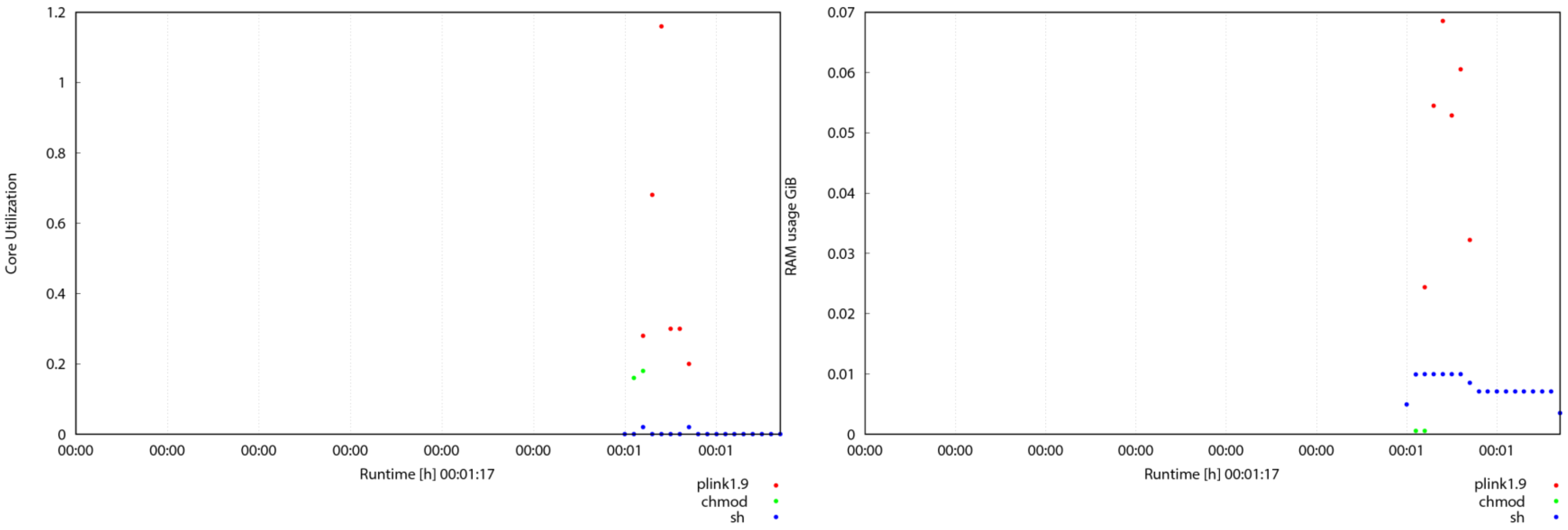


Figure B2. CPU and Ram Utilization for Step 1 – Data Quality Control. Odyssey Step 1, which performs quality control metrics of missingness, minor allele frequency, and Hardy-Weinberg equilibrium filters as well as divide the dataset into individual chromosomes, was performed on the entire HGDP dataset containing 940 admixed individuals and approximately 542K markers. Collectl was used to monitor the CPU (left) and RAM (right) usage for each process (i.e. program) in the step as well as the total time for completion. A buffer of 10 seconds was added to the end of the step for technical reasons. One hyperthreaded CPU core and 32 GB RAM were allotted for Step 1. Using the entire specified dataset, the data quality control step would require approximately 1 minute and <1 GB RAM

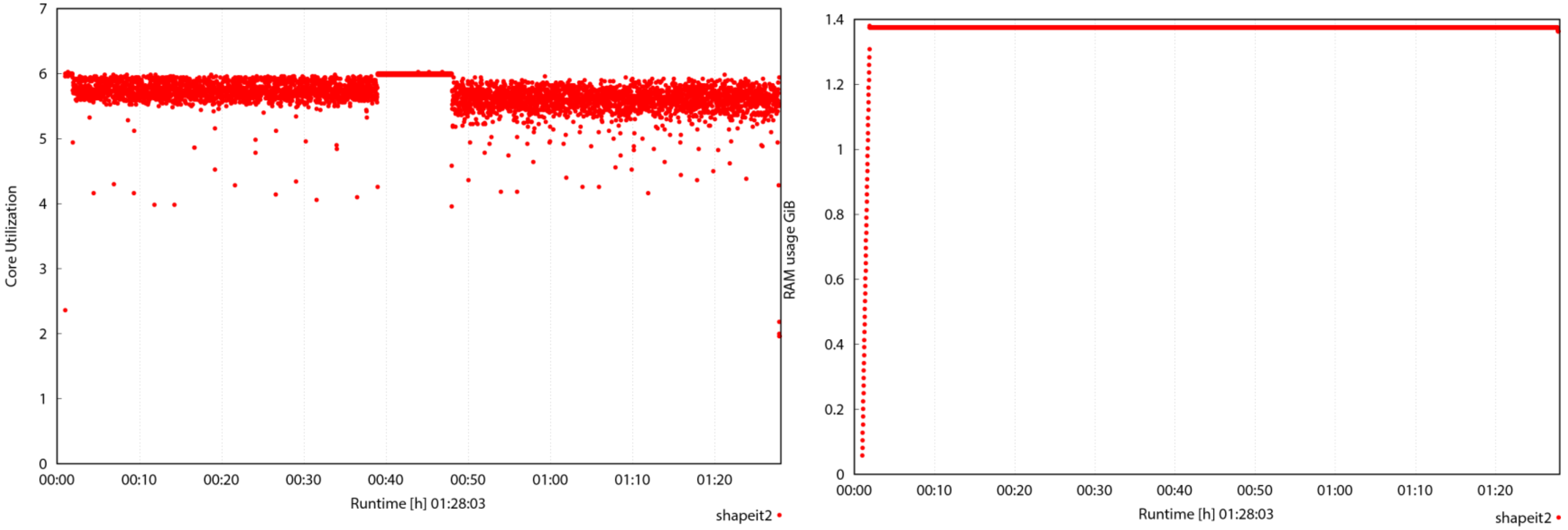


Figure B3. CPU and Ram Utilization for Step 2 – Phasing [SHAPEIT-IMPUTE Workflow]. Odyssey Step 2, which performs phasing via SHAPEIT2, was performed on chromosome 1 of the HGDP dataset containing 940 admixed individuals and approximately 542K markers (30893 markers were phased on Chromosome 1). Collectl was used to monitor the CPU (left) and RAM (right) usage for each process (i.e. program) in the step as well as the total time for completion. A buffer of 10 seconds was added to the end of the step for technical reasons. Three hyperthreaded CPU cores were allotted to SHAPEIT2 (which was programmed to run on six threads) as well as 32 GB RAM. Phasing 1 chromosome from the specified dataset would require approximately 1.5 hours and 1.5 GB RAM

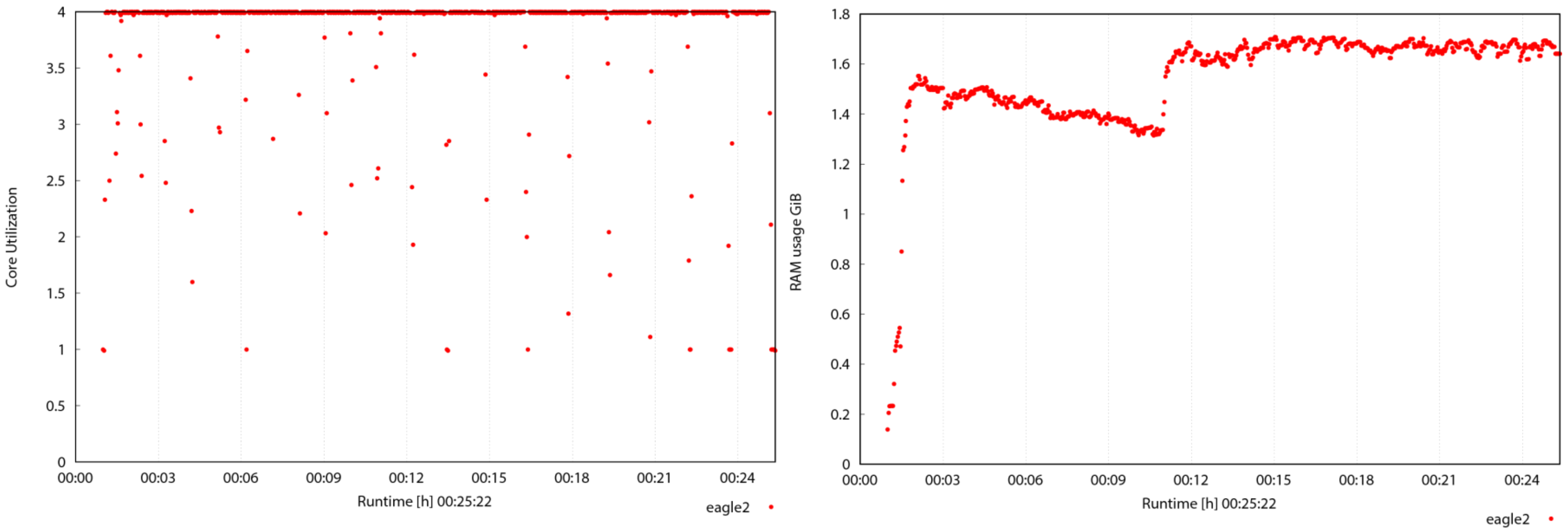


Figure B4. CPU and Ram Utilization for Step 2 – Phasing [Eagle-Minimac Workflow]. Odyssey Step 2, which performs phasing via Eagle2, was performed on chromosome 1 of the HGDP dataset containing 940 admixed individuals and approximately 542K markers (36543 markers were phased on Chromosome 1). Collectl was used to monitor the CPU (left) and RAM (right) usage for each process (i.e. program) in the step as well as the total time for completion. Three hyperthreaded CPU cores were allotted to Eagle2 (which was programmed to run on six threads) as well as 32 GB RAM. Phasing 1 chromosome from the specified dataset would require approximately 24 minutes and 1.9 GB RAM

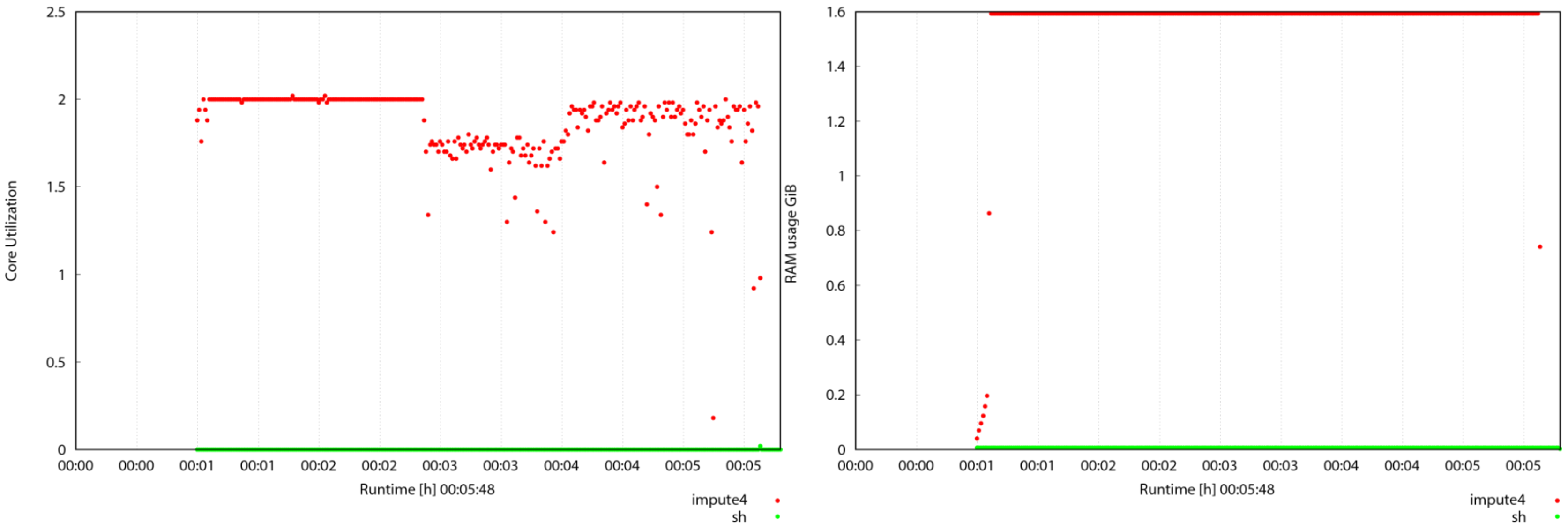


Figure B5. CPU and Ram Utilization for Step 3a – Imputation [SHAPEIT-IMPUTE Workflow]. Odyssey Step 3a, which performs imputation via IMPUTE4, was performed on the first 5 megabase segment of chromosome 1 of the HGDP dataset containing 940 admixed individuals and approximately 542K markers (4316 phased markers were used for imputation on the first 5MB segment on Chromosome 1 resulting in 161268 total imputed and genotyped markers). Collectl was used to monitor the CPU (left) and RAM (right) usage for each process (i.e. program) in the step as well as the total time for completion. A buffer of 10 seconds was added to the end of the step for technical reasons. One hyperthreaded CPU core and 32 GB RAM was allotted to IMPUTE4 for Step 3a. Imputing the first 5 megabase segment of 1 chromosome from the specified dataset would require approximately 6 minutes and 1.7 GB RAM

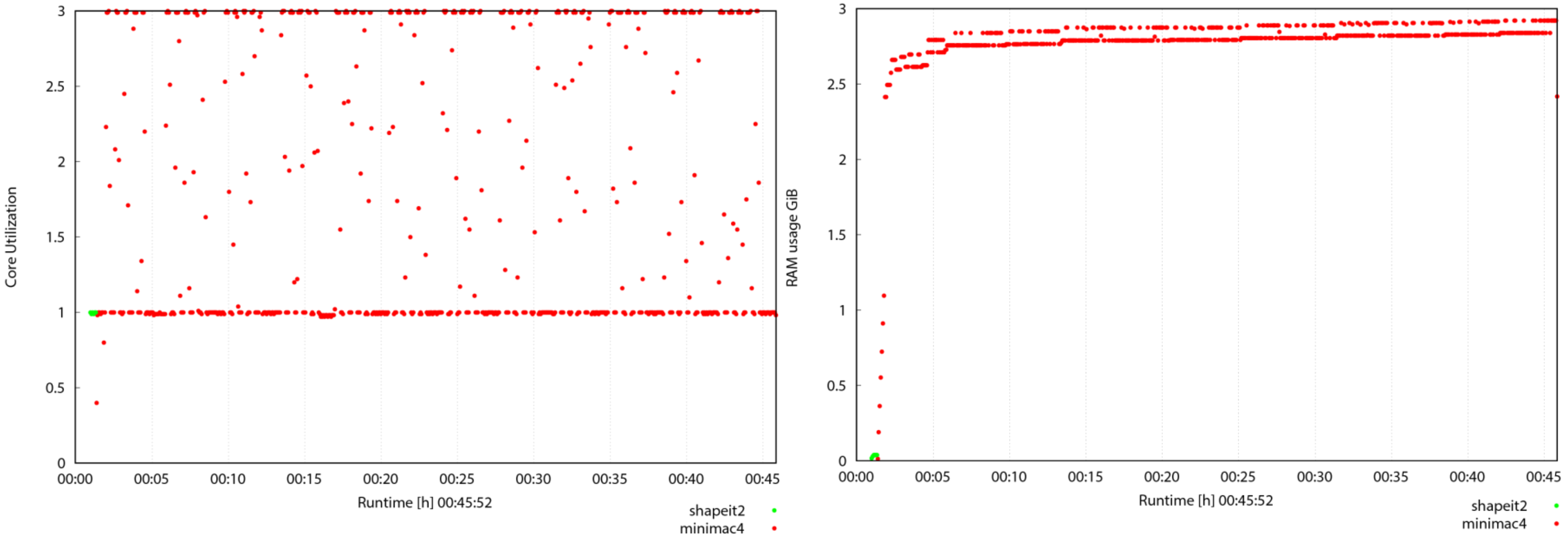


Figure B6. CPU and Ram Utilization for Step 3a – Imputation [Eagle-Minimac Workflow]. Odyssey Step 3a, which performs imputation via Minimac4, was performed on chromosome 1 of the HGDP dataset containing 940 admixed individuals and approximately 542K markers (36543 Eagle2 phased markers were used for imputation on Chromosome 1 resulting in 3745840 total imputed and genotyped markers). Collectl was used to monitor the CPU (left) and RAM (right) usage for each process (i.e. program) in the step as well as the total time for completion. Three hyperthreaded CPU cores and 32 GB RAM was allotted to Minimac4 (which was programmed to run on six threads) for Step 3a. Imputing Chromosome 1 from the specified dataset would require approximately 44 minutes and 3.2 GB RAM

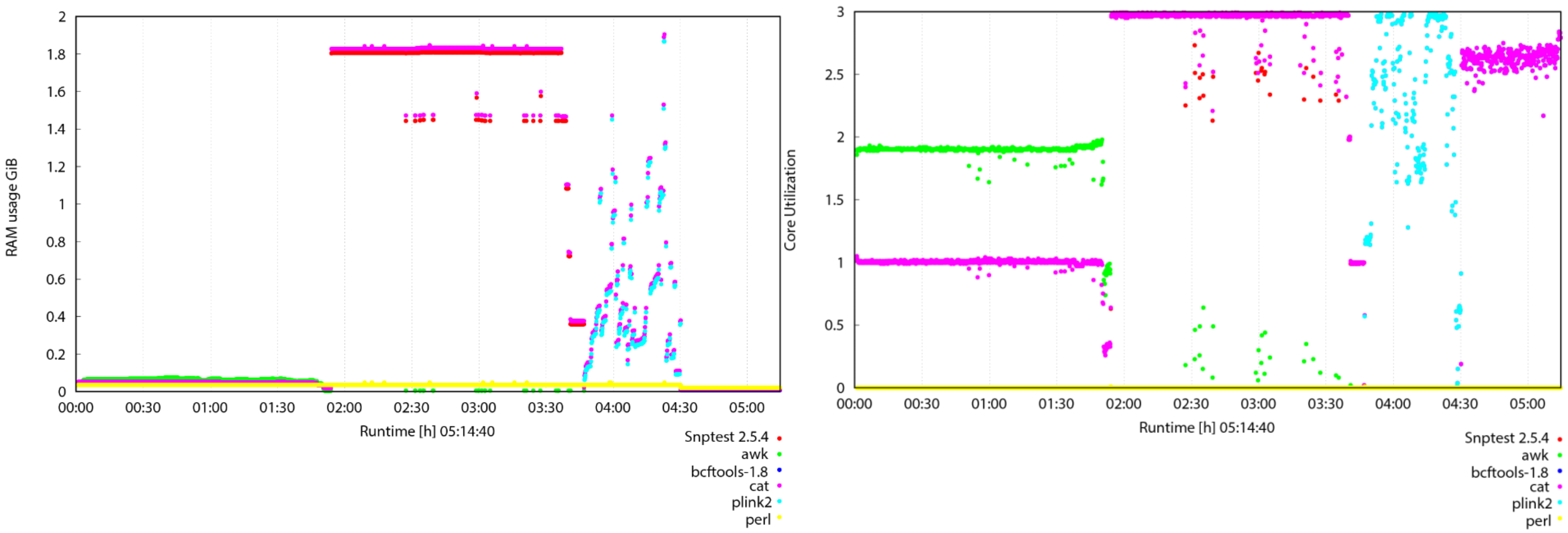


Figure B7. CPU and Ram Utilization for Step 3b – Concatenation, Post-Imputation Quality Control, and Conversion [SHAPEIT-IMPUTE Workflow]. Odyssey Step 3b calculates post-imputation quality control metrics via SNPTEST, converts the segmented chromosomal .gen files to VCF files while filtering based on a specified QC metric (i.e. the INFO metric since IMPUTE4 is being used), and concatenates the chromosomal VCF files file for analysis. This step was performed on the entire HGDP dataset containing 940 admixed individuals and approximately 39 million imputed and genotyped markers after post-imputation quality-control measures. Collectl was used to monitor the CPU (left) and RAM (right) usage for each process (i.e. program) in the step as well as the total time for completion. Three hyperthreaded CPU cores and 32 GB RAM were allotted for Step 3b. Using the entire specified dataset, Step 3b would require approximately 5.2 hours and 2 GB RAM. Note that BCFTools and cat are overlapped on the core utilization figure between time points 4:30 and 5:14

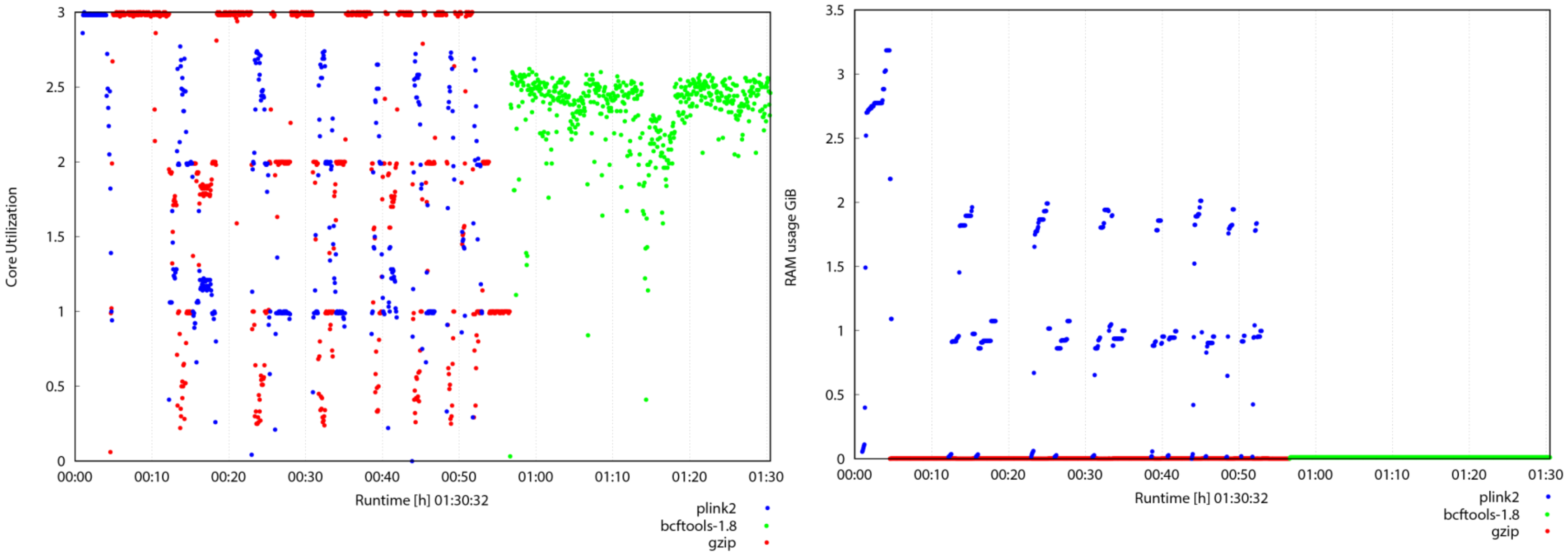


Figure B8. CPU and Ram Utilization for Step 3b – Concatenation, Post-Imputation Quality Control, and Conversion [Eagle-Minimac Workflow]. Odyssey Step 3b filters the dosage.vcf.gz based on a specified QC metric (i.e. the R^2 metric since Minimac4 is being used), converts, and then merges all chromosomal dosage .vcf.gz files into a single dosage vcf.gz file for analysis. This step was performed on the entire HGDP dataset containing 940 admixed individuals and approximately 25.4 million imputed and genotyped markers after post-imputation quality-control measures. Collectl was used to monitor the CPU (left) and RAM (right) usage for each process (i.e. program) in the step as well as the total time for completion. Three hyperthreaded CPU cores and 32 GB RAM were allotted for Step 3b. Using the entire specified dataset, Step 3b would require approximately 1.5 hours and 3.7 GB RAM

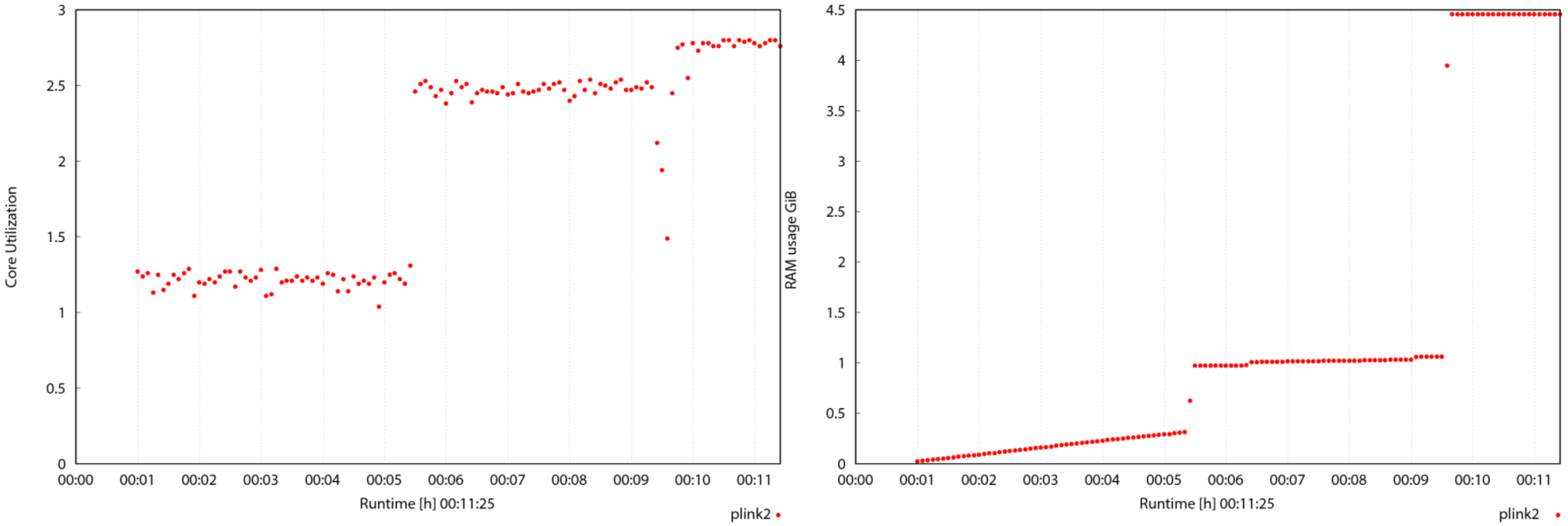


Figure B9. CPU and Ram Utilization for Step 4 – Analysis. Odyssey Step 4 performs a simple Genome-Wide-Association test by using PLINK2 to create a general linear regression model on the 39 million imputed and genotyped variants (generated from the SHAPEIT-IMPUTE workflow) in 940 admixed individuals using sex as a covariate and using randomly calculated phenotype data that contains values from zero to one. Three hyperthreaded CPU cores and 32 GB RAM were allotted for Step 4. Collectl was used to monitor the CPU (left) and RAM (right) usage for each process (i.e. program) in the step as well as the total time for completion. Using the entire specified dataset, analysis would require approximately 11.5 minutes and 4.5 GB RAM

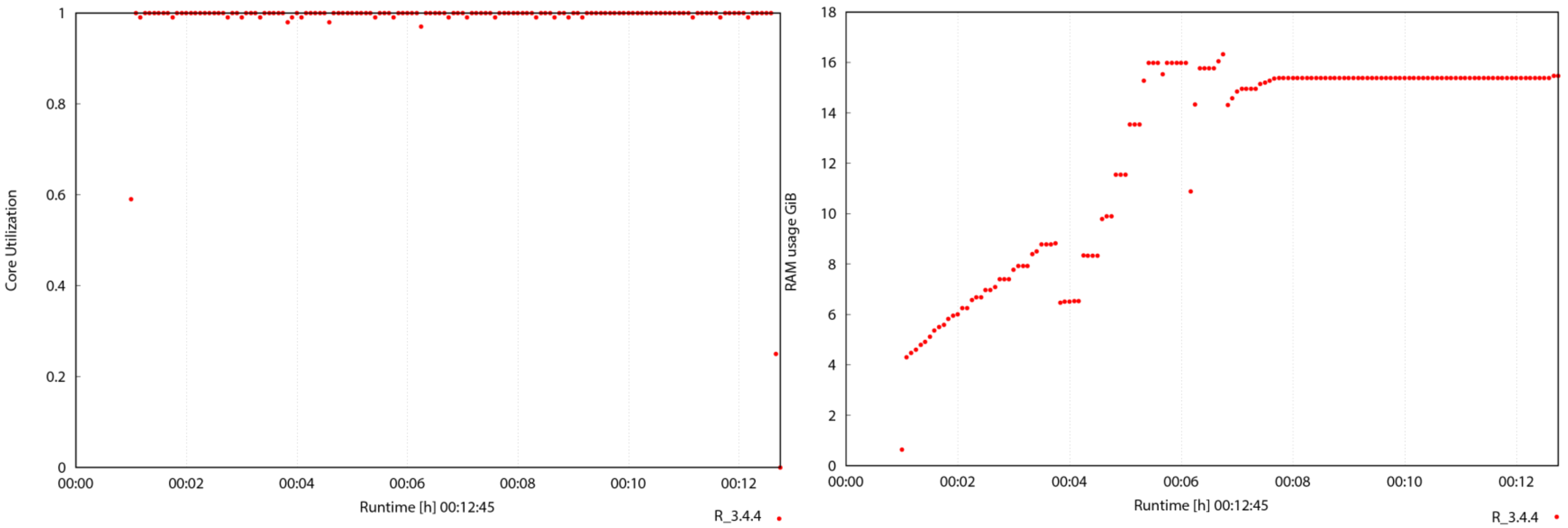


Figure B10. CPU and Ram Utilization for Step 4 – Visualization. Odyssey Step 4 uses R-3.4.4 to read and clean the results file output from the analysis portion of Step 4. R performs a Bonferroni and Benjamini-Hochberg procedure to account for multiple comparisons, sorts the data, and outputs 1) a sorted table of the top 10000 variants with the smallest unadjusted p-values, 2) a qqPlot, and 3) an interactive Manhattan plot. The raw data file is then gunzipped. One hyperthreaded CPU and 32 GB RAM were allotted for Step 4. Collectl was used to monitor the CPU (left) and RAM (right) usage for each process (i.e. program) in the step as well as the total time for completion. Step 3b would require approximately 13 minutes and 18 GB RAM

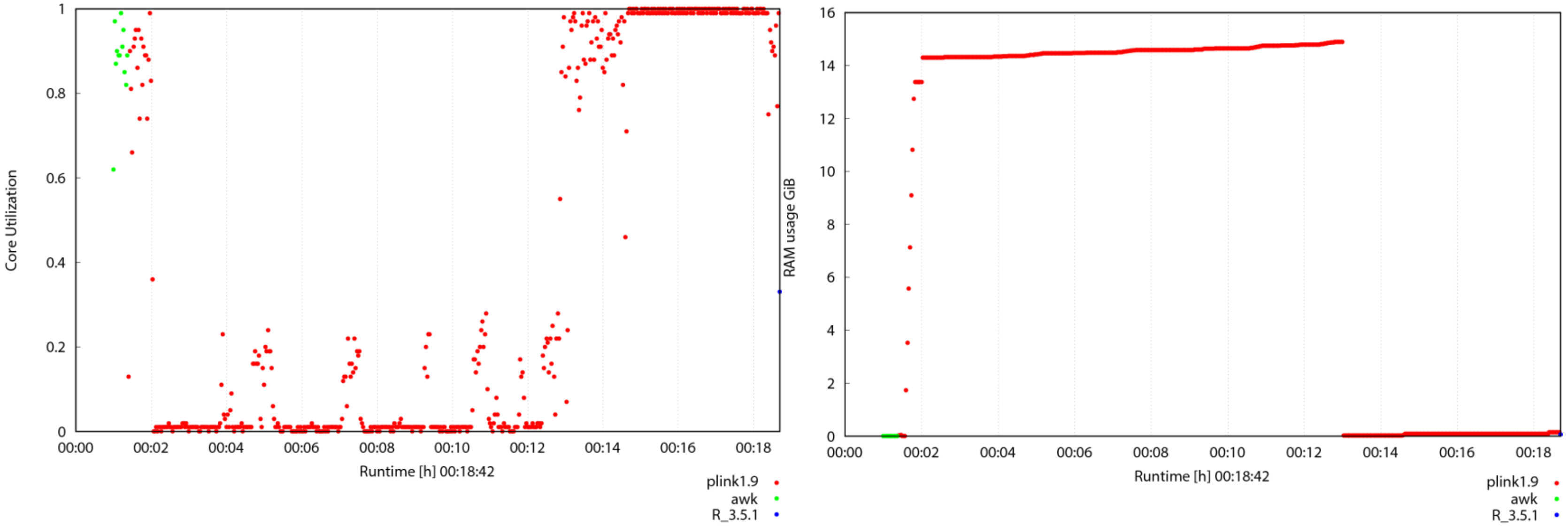


Figure B11. CPU and Ram Utilization for Population Stratification Module. The optional Population Stratification Module uses an admixed reference set of known ancestries to predict the ancestries of a target dataset via a Principal Component analysis (PCA), uses a user-defined subset of reference data corresponding to an ancestry of interest to calculate an "ancestral centroid", which is then used to remove individuals falling outside of the desired ancestral group. R-3.5.1 is used to visualize the PCA including individuals kept for GWAS, removed due to being ancestral outliers, and those that were used as a reference. The processed PCA results in addition to a list of individuals who should be dropped from the GWAS is also output by R. In this benchmark the 1000 Genomes Phase 3 dataset was used as a reference, the 940 admixed individual HGDP dataset was used as the target dataset, the ancestral group of interest that was selected was European, principal components that contributed 1% or more were selected for centroid creation, and individuals that fell outside of 3 standard deviations of the centroid's dimensions were considered ancestral outliers and removed. 1 hyperthreaded CPU and 32 GB RAM were allotted to run the Population Stratification add-in. Collectl was used to monitor the CPU (left) and RAM (right) usage for each process (i.e. program) as well as the total time for completion. The add-in would require approximately 19 minutes and 18 GB RAM

APPENDIX C. IUPUI DATASET DEMOGRAPHICS

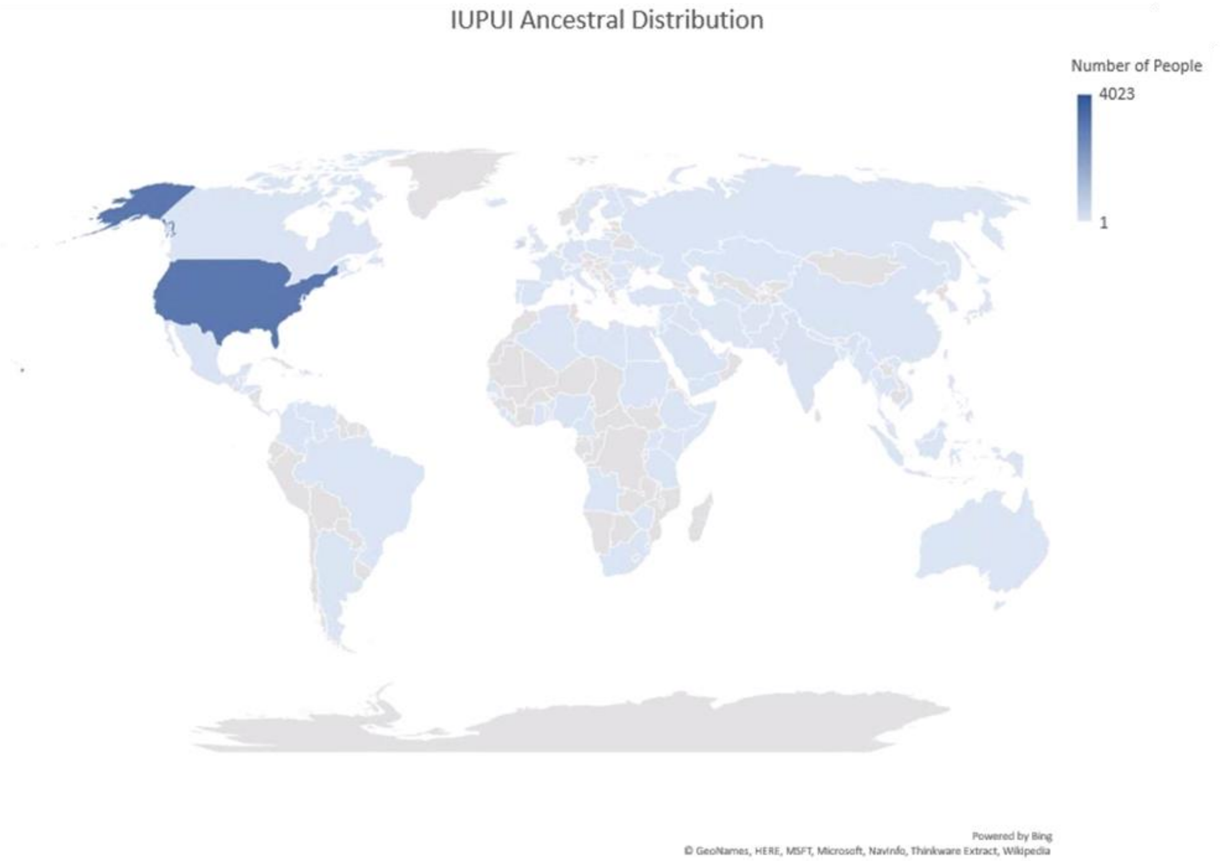

IUPUI Ancestral Distribution
Number of People
4023
1
Powered by Bing
© GeoNames, HERE, MSFT, Microsoft, NavInfo, Thinkware Extract, Wikipedia

Figure C1. Ancestry Breakdown of Individuals Collected in the IUPUI Study

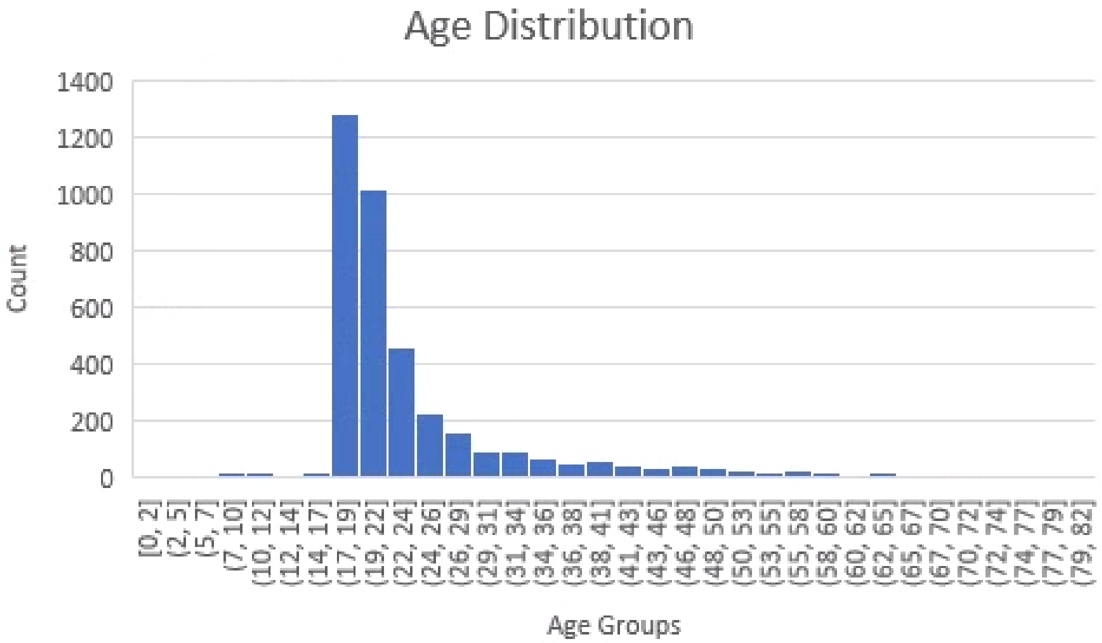


Figure C2. Age Distribution of Individuals Collected in the IUPUI Study

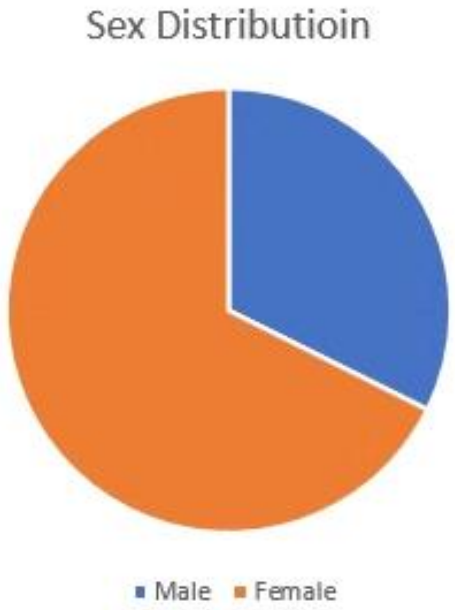


Figure C3. Sex Distribution of Individuals Collected in the IUPUI Study

APPENDIX D. IRISQUANTER TRAINING IMAGES

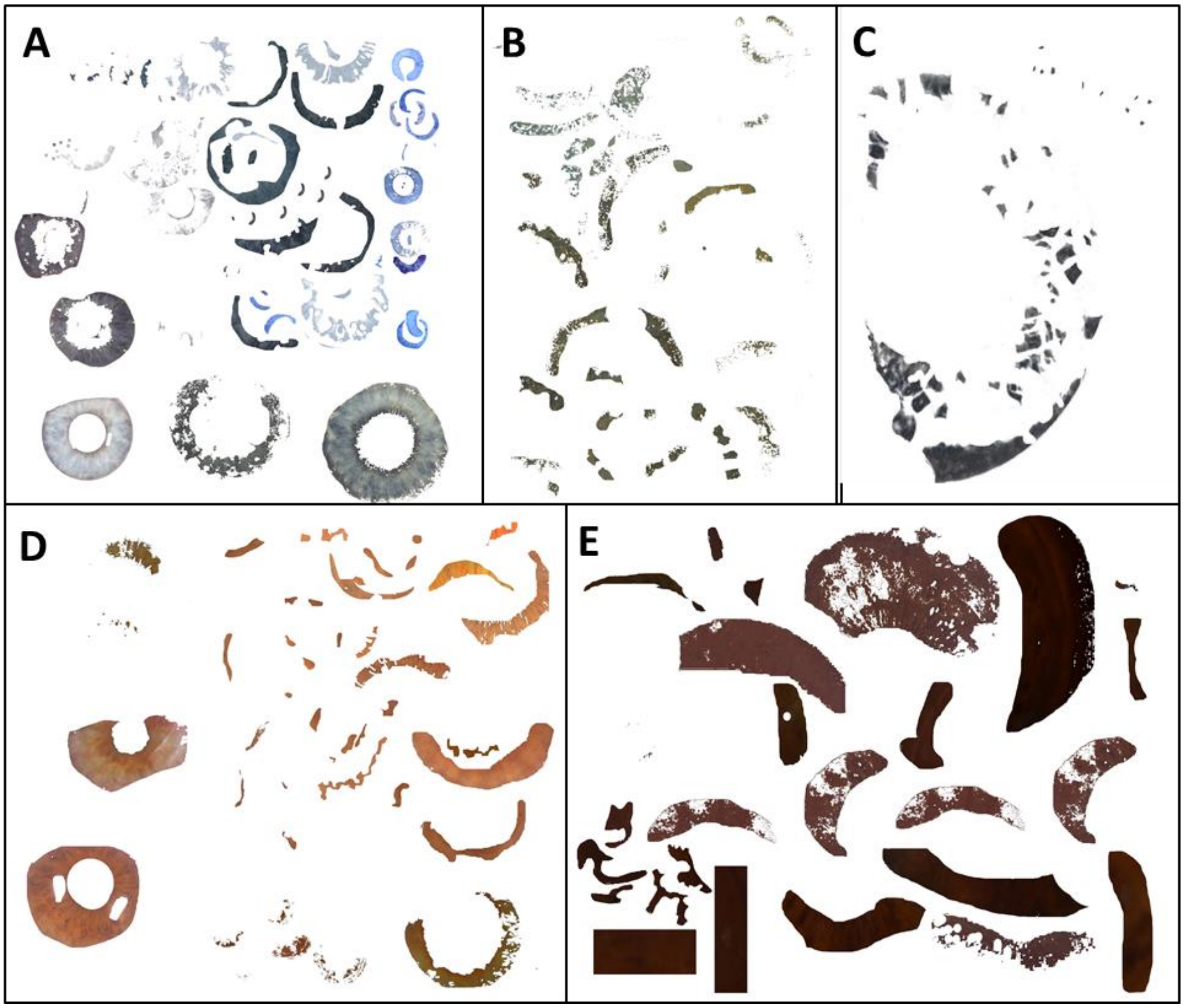
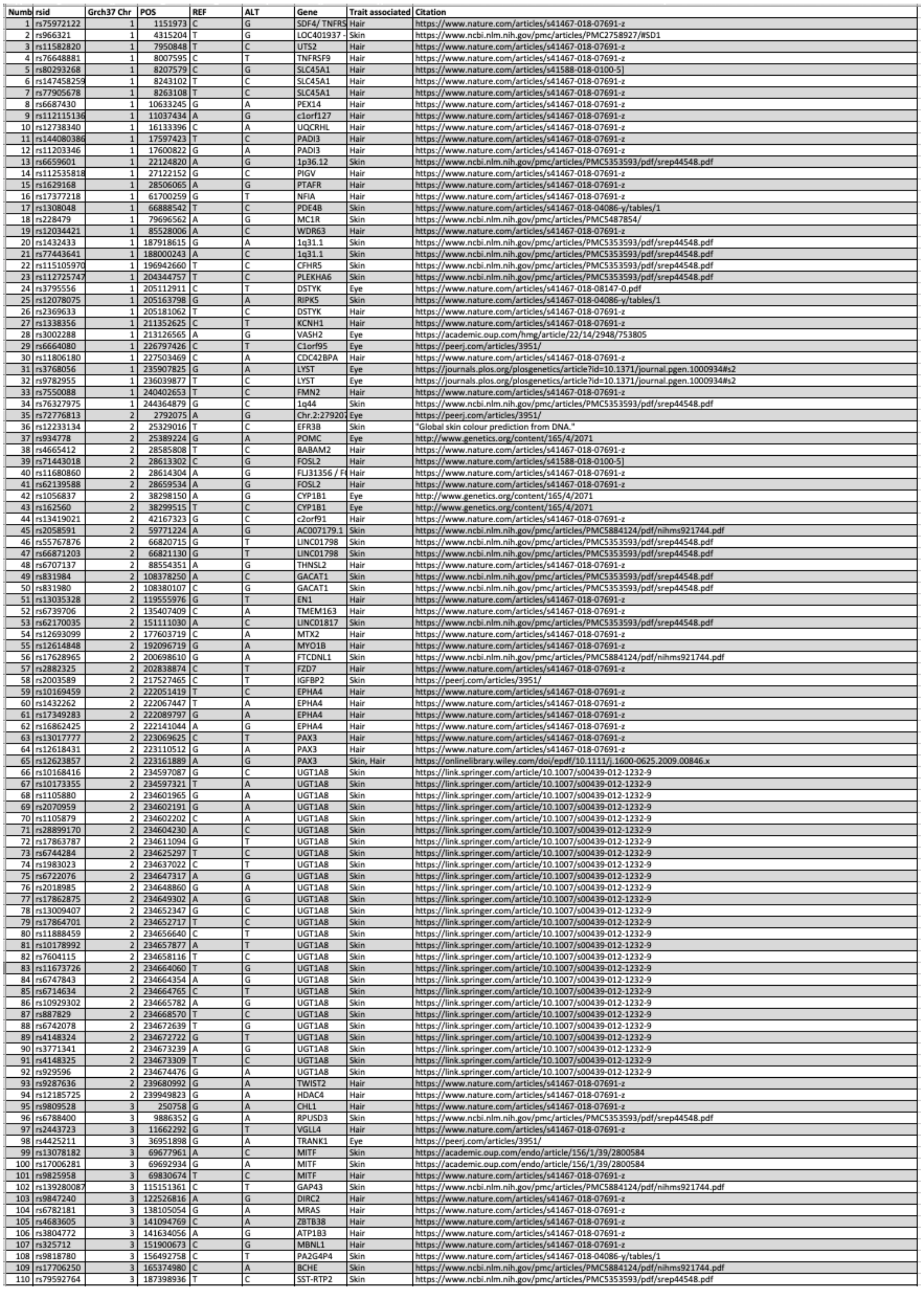


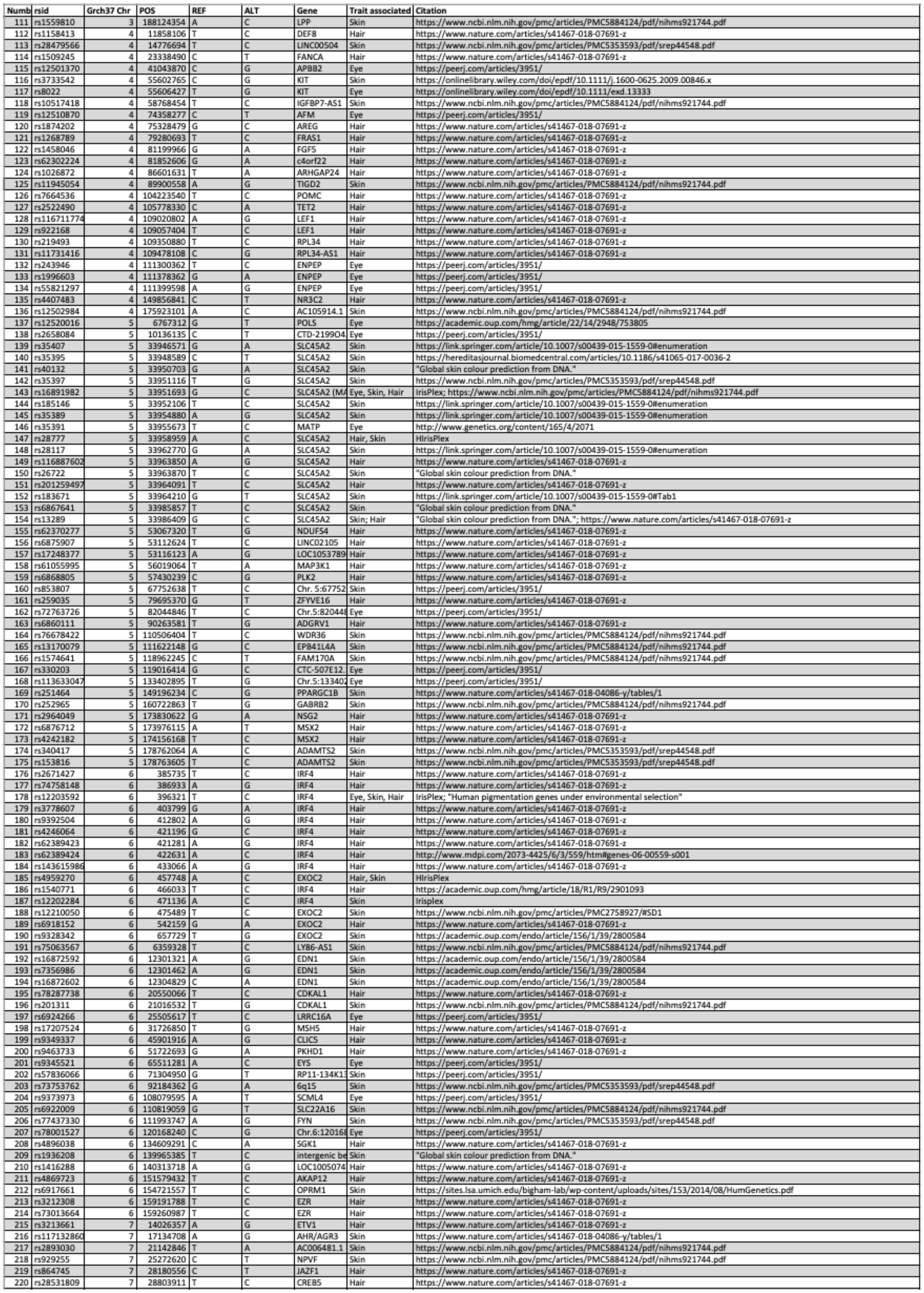
Figure D1. IrisQuanter Training Images. A) Blue, B) green, C) crypt, D) light brown, and E) dark brown images were collected, processed, and used in the iris training of the IrisQuanter program

APPENDIX E. PIGMENTATION VARIANT LITERATURE REVIEW

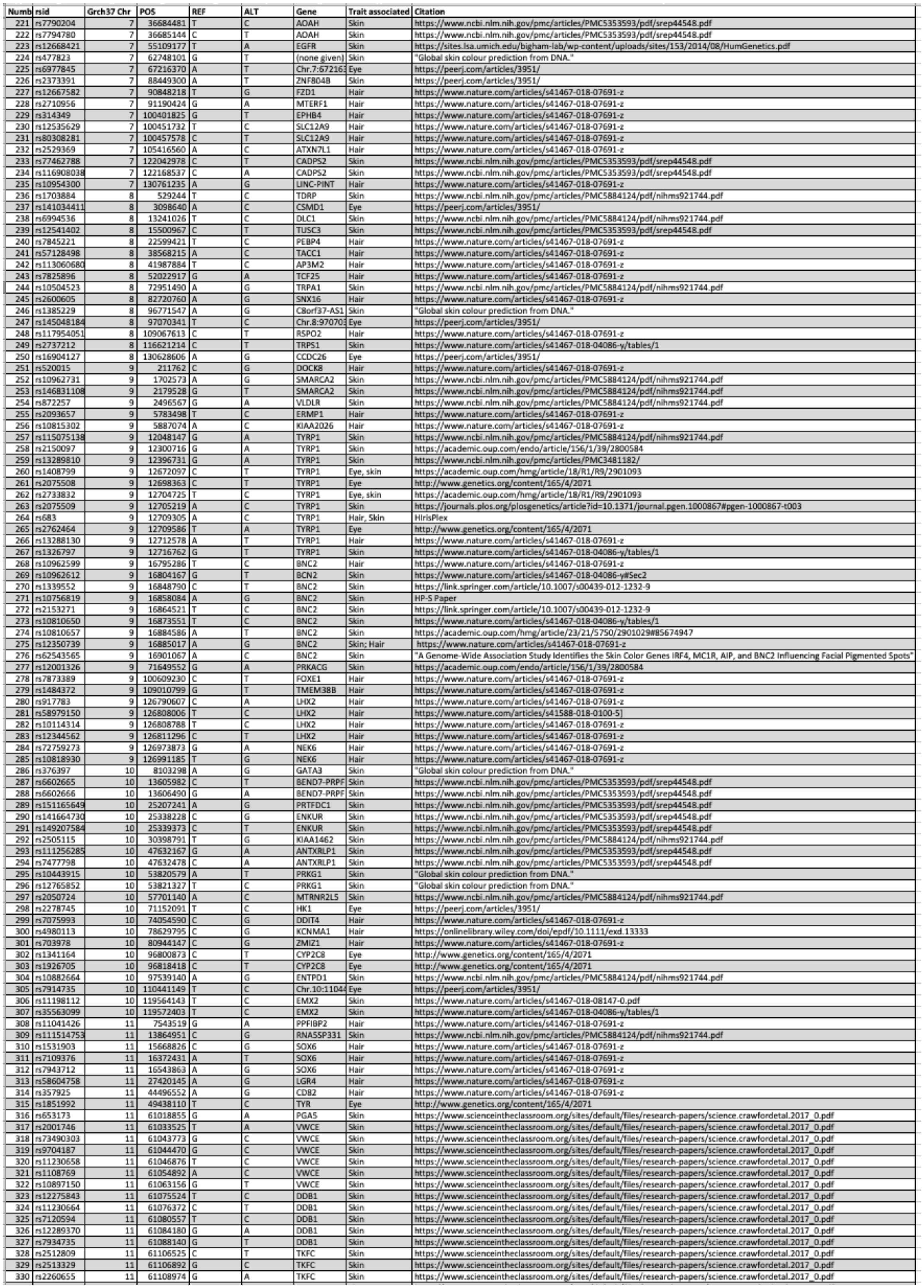
Table E1. 981 Pigmentation-Related Variants Based on a Literature Review



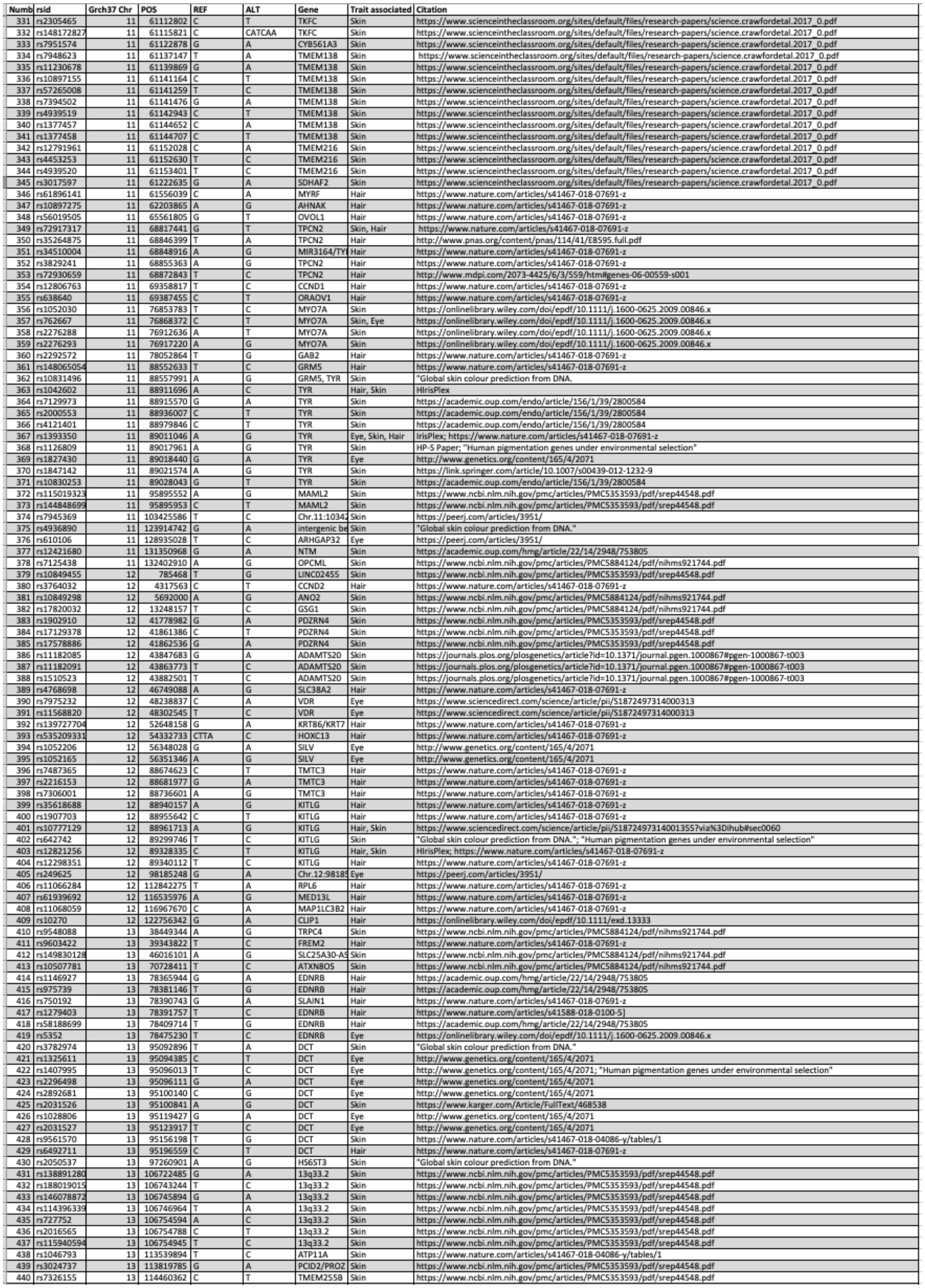
Numb	rsid	Grch37 Chr	POS	REF	ALT	Gene	Trait associated	Citation
1	rs75972122	1	1151973	C	G	SDF4/ TNFRS	Hair	https://www.nature.com/articles/s41467-018-07691-z
2	rs966321	1	4315204	T	G	LOC401937 -	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2758927/#SD1
3	rs11582820	1	7950848	T	C	UTS2	Hair	https://www.nature.com/articles/s41467-018-07691-z
4	rs76648881	1	8007595	C	T	TNFRSF9	Hair	https://www.nature.com/articles/s41467-018-07691-z
5	rs80293268	1	8207579	C	G	SLC45A1	Hair	https://www.nature.com/articles/s41588-018-0100-5]
6	rs147458259	1	8243102	T	C	SLC45A1	Hair	https://www.nature.com/articles/s41467-018-07691-z
7	rs77905678	1	8263108	T	C	SLC45A1	Hair	https://www.nature.com/articles/s41467-018-07691-z
8	rs6687430	1	10633245	G	A	PEX14	Hair	https://www.nature.com/articles/s41467-018-07691-z
9	rs112115136	1	11037434	A	G	c1orf127	Hair	https://www.nature.com/articles/s41467-018-07691-z
10	rs12738340	1	16133396	C	A	UQCRHL	Hair	https://www.nature.com/articles/s41467-018-07691-z
11	rs144080386	1	17597423	T	C	PADI3	Hair	https://www.nature.com/articles/s41467-018-07691-z
12	rs11203346	1	17600822	G	A	PADI3	Hair	https://www.nature.com/articles/s41467-018-07691-z
13	rs6659601	1	22124820	A	G	1p36.12	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
14	rs112535818	1	27122152	G	C	PIGV	Hair	https://www.nature.com/articles/s41467-018-07691-z
15	rs1629168	1	28506065	A	G	PTAFR	Hair	https://www.nature.com/articles/s41467-018-07691-z
16	rs17377218	1	61700259	G	T	NFIA	Hair	https://www.nature.com/articles/s41467-018-07691-z
17	rs1308048	1	66888542	T	C	PDE4B	Skin	https://www.nature.com/articles/s41467-018-04086-y/tables/1
18	rs228479	1	79696562	A	G	MC1R	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5487854/
19	rs12034421	1	85528006	A	C	WDR63	Hair	https://www.nature.com/articles/s41467-018-07691-z
20	rs1432433	1	187918615	G	A	1q31.1	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
21	rs77443641	1	188000243	A	C	1q31.1	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
22	rs115105970	1	196942660	T	C	CFHR5	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
23	rs112725747	1	204344757	T	C	PLEKHA6	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
24	rs3795556	1	205112911	C	T	DSTYK	Eye	https://www.nature.com/articles/s41467-018-08147-0.pdf
25	rs12078075	1	205163798	G	A	RIPK5	Skin	https://www.nature.com/articles/s41467-018-04086-y/tables/1
26	rs2369633	1	205181062	T	C	DSTYK	Hair	https://www.nature.com/articles/s41467-018-07691-z
27	rs1338356	1	211352625	C	T	KCNH1	Hair	https://www.nature.com/articles/s41467-018-07691-z
28	rs3002288	1	213126565	A	G	VASH2	Eye	https://academic.oup.com/hmg/article/22/14/2948/753805
29	rs6664080	1	226797426	C	T	C1orf95	Eye	https://peerj.com/articles/3951/
30	rs11806180	1	227503469	C	A	CDC42BPA	Hair	https://www.nature.com/articles/s41467-018-07691-z
31	rs3768056	1	235907825	G	A	LYST	Eye	https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000934#s2
32	rs9782955	1	236039877	T	C	LYST	Eye	https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000934#s2
33	rs7550088	1	240402653	T	C	FMN2	Hair	https://www.nature.com/articles/s41467-018-07691-z
34	rs76327975	1	244364879	G	C	1q44	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
35	rs72776813	2	2792075	A	G	Chr.2:279207	Eye	https://peerj.com/articles/3951/
36	rs12233134	2	25329016	T	C	EFR3B	Skin	"Global skin colour prediction from DNA."
37	rs934778	2	25389224	G	A	POMC	Eye	http://www.genetics.org/content/165/4/2071
38	rs4665412	2	28585808	T	C	BABAM2	Hair	https://www.nature.com/articles/s41467-018-07691-z
39	rs71443018	2	28613302	C	G	FOSL2	Hair	https://www.nature.com/articles/s41588-018-0100-5]
40	rs11680860	2	28614304	A	G	FLJ31356 / F	Hair	https://www.nature.com/articles/s41467-018-07691-z
41	rs62139588	2	28659534	A	G	FOSL2	Hair	https://www.nature.com/articles/s41467-018-07691-z
42	rs1056837	2	38298150	A	G	CYP1B1	Eye	http://www.genetics.org/content/165/4/2071
43	rs162560	2	38299515	T	C	CYP1B1	Eye	http://www.genetics.org/content/165/4/2071
44	rs13419021	2	42167323	G	C	c2orf91	Hair	https://www.nature.com/articles/s41467-018-07691-z
45	rs2058591	2	59771224	A	G	AC007179.1	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5884124/pdf/nihms921744.pdf
46	rs55767876	2	66820715	G	T	LINC01798	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
47	rs66871203	2	66821130	G	T	LINC01798	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
48	rs6707137	2	88554351	A	G	THNSL2	Hair	https://www.nature.com/articles/s41467-018-07691-z
49	rs831984	2	108378250	A	C	GACAT1	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
50	rs831980	2	108380107	C	G	GACAT1	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
51	rs13035328	2	119555976	G	T	EN1	Hair	https://www.nature.com/articles/s41467-018-07691-z
52	rs6739706	2	135407409	C	A	TMEM163	Hair	https://www.nature.com/articles/s41467-018-07691-z
53	rs62170035	2	151111030	A	C	LINC01817	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
54	rs12693099	2	177603719	C	A	MTX2	Hair	https://www.nature.com/articles/s41467-018-07691-z
55	rs12614848	2	192096719	G	A	MYO1B	Hair	https://www.nature.com/articles/s41467-018-07691-z
56	rs17628965	2	200698610	G	A	FTCDNL1	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5884124/pdf/nihms921744.pdf
57	rs2882325	2	202838874	C	T	FZD7	Hair	https://www.nature.com/articles/s41467-018-07691-z
58	rs2003589	2	217527465	C	T	IGFBP2	Skin	https://peerj.com/articles/3951/
59	rs10169459	2	222051419	T	C	EPHA4	Hair	https://www.nature.com/articles/s41467-018-07691-z
60	rs1432262	2	222067447	T	A	EPHA4	Hair	https://www.nature.com/articles/s41467-018-07691-z
61	rs17349283	2	222089797	G	A	EPHA4	Hair	https://www.nature.com/articles/s41467-018-07691-z
62	rs16862425	2	222141044	A	G	EPHA4	Hair	https://www.nature.com/articles/s41467-018-07691-z
63	rs13017777	2	223069625	C	T	PAX3	Hair	https://www.nature.com/articles/s41467-018-07691-z
64	rs12618431	2	223110512	G	A	PAX3	Hair	https://www.nature.com/articles/s41467-018-07691-z
65	rs12623857	2	223161889	A	G	PAX3	Skin, Hair	https://onlinelibrary.wiley.com/doi/epdf/10.1111/j.1600-0625.2009.00846.x
66	rs10168416	2	234597087	G	C	UGT1A8	Skin	https://link.springer.com/article/10.1007/s00439-012-1232-9
67	rs10173355	2	234597321	T	A	UGT1A8	Skin	https://link.springer.com/article/10.1007/s00439-012-1232-9
68	rs1105880	2	234601965	G	A	UGT1A8	Skin	https://link.springer.com/article/10.1007/s00439-012-1232-9
69	rs2070959	2	234602191	G	A	UGT1A8	Skin	https://link.springer.com/article/10.1007/s00439-012-1232-9
70	rs1105879	2	234602202	C	A	UGT1A8	Skin	https://link.springer.com/article/10.1007/s00439-012-1232-9
71	rs28899170	2	234604230	A	C	UGT1A8	Skin	https://link.springer.com/article/10.1007/s00439-012-1232-9
72	rs17863787	2	234611094	G	T	UGT1A8	Skin	https://link.springer.com/article/10.1007/s00439-012-1232-9
73	rs6744284	2	234625297	T	C	UGT1A8	Skin	https://link.springer.com/article/10.1007/s00439-012-1232-9
74	rs1983023	2	234637022	C	T	UGT1A8	Skin	https://link.springer.com/article/10.1007/s00439-012-1232-9
75	rs6722076	2	234647317	A	G	UGT1A8	Skin	https://link.springer.com/article/10.1007/s00439-012-1232-9
76	rs2018985	2	234648860	G	A	UGT1A8	Skin	https://link.springer.com/article/10.1007/s00439-012-1232-9
77	rs17862875	2	234649302	A	G	UGT1A8	Skin	https://link.springer.com/article/10.1007/s00439-012-1232-9
78	rs13009407	2	234652347	G	C	UGT1A8	Skin	https://link.springer.com/article/10.1007/s00439-012-1232-9
79	rs17864701	2	234652717	T	C	UGT1A8	Skin	https://link.springer.com/article/10.1007/s00439-012-1232-9
80	rs11888459	2	234656640	C	T	UGT1A8	Skin	https://link.springer.com/article/10.1007/s00439-012-1232-9
81	rs10178992	2	234657877	A	T	UGT1A8	Skin	https://link.springer.com/article/10.1007/s00439-012-1232-9
82	rs7604115	2	234658116	T	C	UGT1A8	Skin	https://link.springer.com/article/10.1007/s00439-012-1232-9
83	rs11673726	2	234664060	T	G	UGT1A8	Skin	https://link.springer.com/article/10.1007/s00439-012-1232-9
84	rs6747843	2	234664354	A	G	UGT1A8	Skin	https://link.springer.com/article/10.1007/s00439-012-1232-9
85	rs6714634	2	234664765	C	T	UGT1A8	Skin	https://link.springer.com/article/10.1007/s00439-012-1232-9
86	rs10929302	2	234665782	A	G	UGT1A8	Skin	https://link.springer.com/article/10.1007/s00439-012-1232-9
87	rs887829	2	234668570	T	C	UGT1A8	Skin	https://link.springer.com/article/10.1007/s00439-012-1232-9
88	rs6742078	2	234672639	T	G	UGT1A8	Skin	https://link.springer.com/article/10.1007/s00439-012-1232-9
89	rs4148324	2	234672722	G	T	UGT1A8	Skin	https://link.springer.com/article/10.1007/s00439-012-1232-9
90	rs3771341	2	234673239	A	G	UGT1A8	Skin	https://link.springer.com/article/10.1007/s00439-012-1232-9
91	rs4148325	2	234673309	T	C	UGT1A8	Skin	https://link.springer.com/article/10.1007/s00439-012-1232-9
92	rs929596	2	234674476	G	A	UGT1A8	Skin	https://link.springer.com/article/10.1007/s00439-012-1232-9
93	rs9287636	2	239680992	G	A	TWIST2	Hair	https://www.nature.com/articles/s41467-018-07691-z
94	rs12185725	2	239949823	G	A	HDAC4	Hair	https://www.nature.com/articles/s41467-018-07691-z
95	rs9809528	3	250758	G	A	CHL1	Hair	https://www.nature.com/articles/s41467-018-07691-z
96	rs6788400	3	9886352	G	A	RPUSD3	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
97	rs2443723	3	11662292	G	T	VGLL4	Hair	https://www.nature.com/articles/s41467-018-07691-z
98	rs4425211	3	36951898	G	A	TRANK1	Eye	https://peerj.com/articles/3951/
99	rs13078182	3	69677961	A	C	MITF	Skin	https://academic.oup.com/endo/article/156/1/39/2800584
100	rs17006281	3	69692934	G	A	MITF	Skin	https://academic.oup.com/endo/article/156/1/39/2800584
101	rs9825958	3	69830674	T	C	MITF	Hair	https://www.nature.com/articles/s41467-018-07691-z
102	rs139280087	3	115151361	C	T	GAP43	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5884124/pdf/nihms921744.pdf
103	rs9847240	3	122526816	A	G	DIRC2	Hair	https://www.nature.com/articles/s41467-018-07691-z
104	rs6782181	3	138105054	G	A	MRAS	Hair	https://www.nature.com/articles/s41467-018-07691-z
105	rs4683605	3	141094769	C	A	ZBTB38	Hair	https://www.nature.com/articles/s41467-018-07691-z
106	rs3804772	3	141634056	A	G	ATP1B3	Hair	https://www.nature.com/articles/s41467-018-07691-z
107	rs325712	3	151900673	C	G	MBNL1	Hair	https://www.nature.com/articles/s41467-018-07691-z
108	rs9818780	3	156492758	C	T	PA2G4P4	Skin	https://www.nature.com/articles/s41467-018-04086-y/tables/1
109	rs17706250	3	165374980	C	A	BCHE	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5884124/pdf/nihms921744.pdf
110	rs79592764	3	187398936	T	C	SST-RTP2	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf



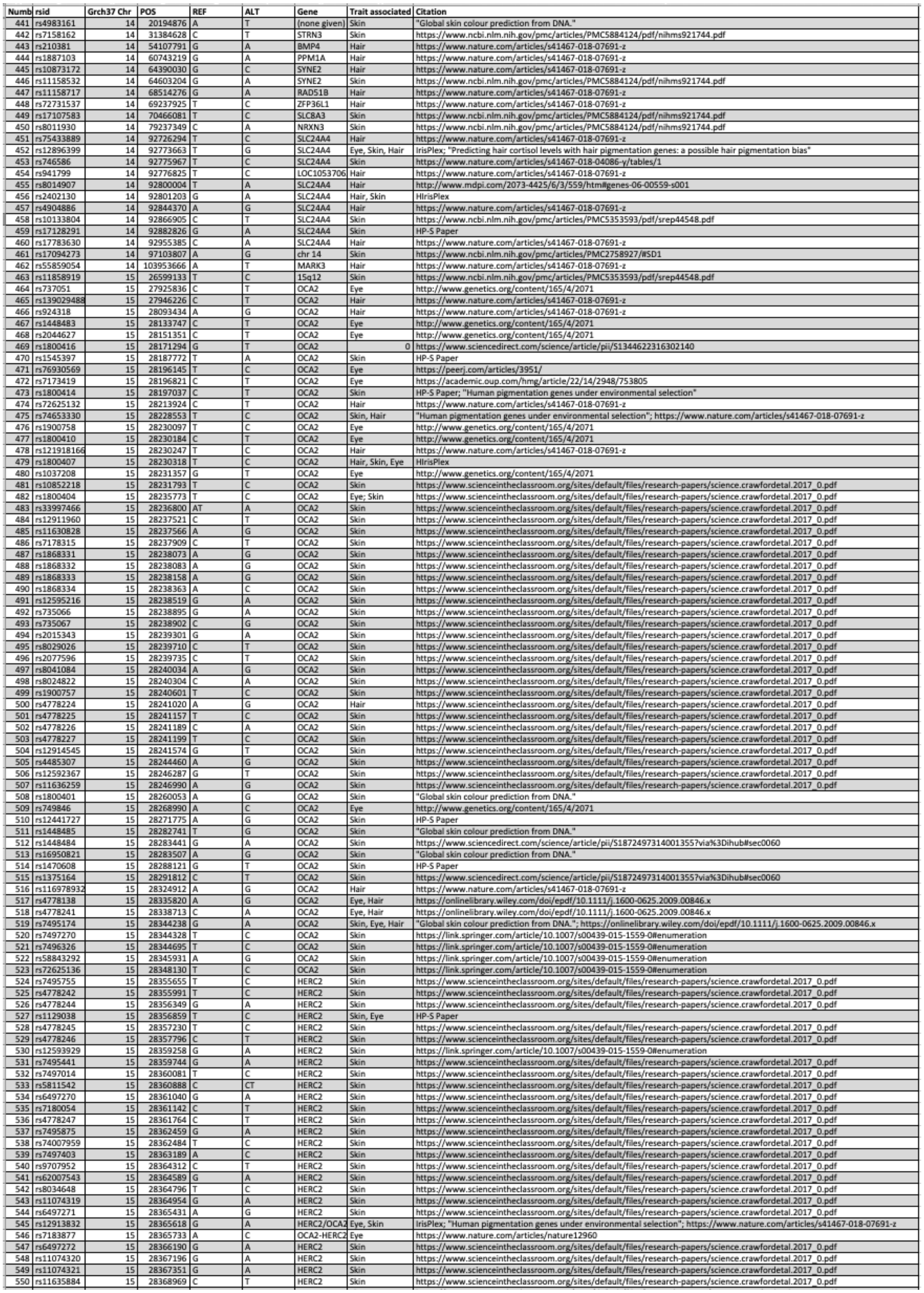
Numb	rsid	Grch37 Chr	POS	REF	ALT	Gene	Trait associated	Citation
111	rs1559810	3	188124354	A	C	LPP	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5884124/pdf/nihms921744.pdf
112	rs1158413	4	11858106	T	C	DEF8	Hair	https://www.nature.com/articles/s41467-018-07691-z
113	rs28479566	4	14776694	T	C	LINC00504	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
114	rs1509245	4	23338490	C	T	FANCA	Hair	https://www.nature.com/articles/s41467-018-07691-z
115	rs12501370	4	41043870	C	G	APBB2	Eye	https://peerj.com/articles/3951/
116	rs3733542	4	55602765	C	G	KIT	Skin	https://onlinelibrary.wiley.com/doi/epdf/10.1111/j.1600-0625.2009.00846.x
117	rs8022	4	55606427	T	G	KIT	Eye	https://onlinelibrary.wiley.com/doi/epdf/10.1111/exd.13333
118	rs10517418	4	58768454	T	C	IGFBP7-AS1	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5884124/pdf/nihms921744.pdf
119	rs12510870	4	74358277	C	T	AFM	Eye	https://peerj.com/articles/3951/
120	rs1874202	4	75328479	G	C	AREG	Hair	https://www.nature.com/articles/s41467-018-07691-z
121	rs1268789	4	79280693	T	C	FRAS1	Hair	https://www.nature.com/articles/s41467-018-07691-z
122	rs1458046	4	81199966	G	A	FGF5	Hair	https://www.nature.com/articles/s41467-018-07691-z
123	rs62302224	4	81852606	G	A	c4orf22	Hair	https://www.nature.com/articles/s41467-018-07691-z
124	rs1026872	4	86601631	T	A	ARHGAP24	Hair	https://www.nature.com/articles/s41467-018-07691-z
125	rs11945054	4	89900558	A	G	TIGD2	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5884124/pdf/nihms921744.pdf
126	rs7664536	4	104223540	T	C	POMC	Hair	https://www.nature.com/articles/s41467-018-07691-z
127	rs2522490	4	105778330	C	A	TET2	Hair	https://www.nature.com/articles/s41467-018-07691-z
128	rs116711774	4	109020802	A	G	LEF1	Hair	https://www.nature.com/articles/s41467-018-07691-z
129	rs922168	4	109057404	T	C	LEF1	Hair	https://www.nature.com/articles/s41467-018-07691-z
130	rs219493	4	109350880	T	C	RPL34	Hair	https://www.nature.com/articles/s41467-018-07691-z
131	rs11731416	4	109478108	C	G	RPL34-AS1	Hair	https://www.nature.com/articles/s41467-018-07691-z
132	rs243946	4	111300362	T	C	ENPEP	Eye	https://peerj.com/articles/3951/
133	rs1996603	4	111378362	G	A	ENPEP	Eye	https://peerj.com/articles/3951/
134	rs55821297	4	111399598	A	G	ENPEP	Eye	https://peerj.com/articles/3951/
135	rs4407483	4	149856841	C	T	NR3C2	Hair	https://www.nature.com/articles/s41467-018-07691-z
136	rs12502984	4	175923101	A	C	AC105914.1	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5884124/pdf/nihms921744.pdf
137	rs12520016	5	6767312	G	T	POLS	Eye	https://academic.oup.com/hmg/article/22/14/2948/753805
138	rs2658084	5	10136135	C	T	CTD-2199O4	Eye	https://peerj.com/articles/3951/
139	rs35407	5	33946571	G	A	SLC45A2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
140	rs35395	5	33948589	C	T	SLC45A2	Skin	https://hereditasjournal.biomedcentral.com/articles/10.1186/s41065-017-0036-2
141	rs40132	5	33950703	G	A	SLC45A2	Skin	"Global skin colour prediction from DNA."
142	rs35397	5	33951116	T	G	SLC45A2	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
143	rs16891982	5	33951693	G	C	SLC45A2 (MA	Eye, Skin, Hair	IrisPlex; https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5884124/pdf/nihms921744.pdf
144	rs185146	5	33952106	T	C	SLC45A2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
145	rs35389	5	33954880	A	G	SLC45A2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
146	rs35391	5	33955673	T	C	MATP	Eye	http://www.genetics.org/content/165/4/2071
147	rs28777	5	33958959	A	C	SLC45A2	Hair, Skin	HIrisPlex
148	rs28117	5	33962770	G	A	SLC45A2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
149	rs116887602	5	33963850	A	G	SLC45A2	Hair	https://www.nature.com/articles/s41467-018-07691-z
150	rs26722	5	33963870	T	C	SLC45A2	Skin	"Global skin colour prediction from DNA."
151	rs201259497	5	33964091	T	C	SLC45A2	Hair	https://www.nature.com/articles/s41467-018-07691-z
152	rs183671	5	33964210	G	T	SLC45A2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#Tab1
153	rs6867641	5	33985857	T	C	SLC45A2	Skin	"Global skin colour prediction from DNA."
154	rs13289	5	33986409	G	C	SLC45A2	Skin; Hair	"Global skin colour prediction from DNA."; https://www.nature.com/articles/s41467-018-07691-z
155	rs62370277	5	53067320	T	G	NDUFS4	Hair	https://www.nature.com/articles/s41467-018-07691-z
156	rs6875907	5	53112624	T	C	LINC02105	Hair	https://www.nature.com/articles/s41467-018-07691-z
157	rs17248377	5	53116123	A	G	LOC1053789	Hair	https://www.nature.com/articles/s41467-018-07691-z
158	rs61055995	5	56019064	T	A	MAP3K1	Hair	https://www.nature.com/articles/s41467-018-07691-z
159	rs6868805	5	57430239	C	G	PLK2	Hair	https://www.nature.com/articles/s41467-018-07691-z
160	rs853807	5	67752638	T	C	Chr. 5:67752	Skin	https://peerj.com/articles/3951/
161	rs259035	5	79695370	G	T	ZFYVE16	Hair	https://www.nature.com/articles/s41467-018-07691-z
162	rs72763726	5	82044846	T	C	Chr.5:820448	Eye	https://peerj.com/articles/3951/
163	rs6860111	5	90263581	T	G	ADGRV1	Hair	https://www.nature.com/articles/s41467-018-07691-z
164	rs76678422	5	110506404	T	C	WDR36	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5884124/pdf/nihms921744.pdf
165	rs13170079	5	111622148	G	C	EPB41L4A	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5884124/pdf/nihms921744.pdf
166	rs1574641	5	118962245	C	T	FAM170A	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5884124/pdf/nihms921744.pdf
167	rs330203	5	119016414	G	C	CTC-507E12.	Eye	https://peerj.com/articles/3951/
168	rs113633047	5	133402895	T	G	Chr.5:133402	Eye	https://peerj.com/articles/3951/
169	rs251464	5	149196234	C	G	PPARGC1B	Skin	https://www.nature.com/articles/s41467-018-04086-y/tables/1
170	rs252965	5	160722863	T	G	GABRB2	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5884124/pdf/nihms921744.pdf
171	rs2964049	5	173830622	G	A	NSG2	Hair	https://www.nature.com/articles/s41467-018-07691-z
172	rs6876712	5	173976115	A	T	MSX2	Hair	https://www.nature.com/articles/s41467-018-07691-z
173	rs4242182	5	174156168	T	C	MSX2	Hair	https://www.nature.com/articles/s41467-018-07691-z
174	rs340417	5	178762064	A	C	ADAMTS2	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
175	rs153816	5	178763605	T	C	ADAMTS2	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
176	rs2671427	6	385735	T	C	IRF4	Hair	https://www.nature.com/articles/s41467-018-07691-z
177	rs74758148	6	386933	A	G	IRF4	Hair	https://www.nature.com/articles/s41467-018-07691-z
178	rs12203592	6	396321	T	C	IRF4	Eye, Skin, Hair	IrisPlex; "Human pigmentation genes under environmental selection"
179	rs3778607	6	403799	G	A	IRF4	Hair	https://www.nature.com/articles/s41467-018-07691-z
180	rs9392504	6	412802	A	G	IRF4	Hair	https://www.nature.com/articles/s41467-018-07691-z
181	rs4246064	6	421196	G	C	IRF4	Hair	https://www.nature.com/articles/s41467-018-07691-z
182	rs62389423	6	421281	A	G	IRF4	Hair	https://www.nature.com/articles/s41467-018-07691-z
183	rs62389424	6	422631	A	C	IRF4	Hair	http://www.mdpi.com/2073-4425/6/3/559/htm#genes-06-00559-s001
184	rs143615986	6	433066	A	G	IRF4	Hair	https://www.nature.com/articles/s41467-018-07691-z
185	rs4959270	6	457748	A	C	EXOC2	Hair, Skin	HIrisPlex
186	rs1540771	6	466033	T	C	IRF4	Hair	https://academic.oup.com/hmg/article/18/R1/R9/2901093
187	rs12202284	6	471136	A	C	IRF4	Skin	Irisplex
188	rs12210050	6	475489	T	C	EXOC2	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2758927/#SD1
189	rs6918152	6	542159	G	A	EXOC2	Hair	https://www.nature.com/articles/s41467-018-07691-z
190	rs9328342	6	657729	T	G	EXOC2	Skin	https://academic.oup.com/endo/article/156/1/39/2800584
191	rs75063567	6	6359328	T	C	LY86-AS1	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5884124/pdf/nihms921744.pdf
192	rs16872592	6	12301321	A	G	EDN1	Skin	https://academic.oup.com/endo/article/156/1/39/2800584
193	rs7356986	6	12301462	A	G	EDN1	Skin	https://academic.oup.com/endo/article/156/1/39/2800584
194	rs16872602	6	12304829	C	A	EDN1	Skin	https://academic.oup.com/endo/article/156/1/39/2800584
195	rs78287738	6	20550066	T	C	CDKAL1	Hair	https://www.nature.com/articles/s41467-018-07691-z
196	rs201311	6	21016532	T	G	CDKAL1	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5884124/pdf/nihms921744.pdf
197	rs6924266	6	25505617	T	C	LRRC16A	Eye	https://peerj.com/articles/3951/
198	rs17207524	6	31726850	T	G	MSH5	Hair	https://www.nature.com/articles/s41467-018-07691-z
199	rs9349337	6	45901916	A	G	CLIC5	Hair	https://www.nature.com/articles/s41467-018-07691-z
200	rs9463733	6	51722693	G	A	PKHD1	Hair	https://www.nature.com/articles/s41467-018-07691-z
201	rs9345521	6	65511281	A	C	EYS	Eye	https://peerj.com/articles/3951/
202	rs57836066	6	71304950	G	T	RP11-134K13	Skin	https://peerj.com/articles/3951/
203	rs73753762	6	92184362	G	A	6q15	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
204	rs9373973	6	108079595	A	T	SCML4	Eye	https://peerj.com/articles/3951/
205	rs6922009	6	110819059	G	T	SLC22A16	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5884124/pdf/nihms921744.pdf
206	rs77437330	6	111993747	A	G	FYN	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
207	rs78001527	6	120168240	C	G	Chr.6:120168	Eye	https://peerj.com/articles/3951/
208	rs4896038	6	134609291	C	A	SGK1	Hair	https://www.nature.com/articles/s41467-018-07691-z
209	rs1936208	6	139965385	T	C	intergenic be	Skin	"Global skin colour prediction from DNA."
210	rs1416288	6	140313718	A	G	LOC1005074	Hair	https://www.nature.com/articles/s41467-018-07691-z
211	rs4869723	6	151579432	T	C	AKAP12	Hair	https://www.nature.com/articles/s41467-018-07691-z
212	rs6917661	6	154721557	T	C	OPRM1	Skin	https://sites.lsa.umich.edu/bigham-lab/wp-content/uploads/sites/153/2014/08/HumGenetics.pdf
213	rs3212308	6	159191788	T	C	EZR	Hair	https://www.nature.com/articles/s41467-018-07691-z
214	rs73013664	6	159260987	T	C	EZR	Hair	https://www.nature.com/articles/s41467-018-07691-z
215	rs3213661	7	14026357	A	G	ETV1	Hair	https://www.nature.com/articles/s41467-018-07691-z
216	rs117132860	7	17134708	A	G	AHR/AGR3	Skin	https://www.nature.com/articles/s41467-018-04086-y/tables/1
217	rs2893030	7	21142846	T	A	AC006481.1	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5884124/pdf/nihms921744.pdf
218	rs929255	7	25272620	C	T	NPVF	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5884124/pdf/nihms921744.pdf
219	rs864745	7	28180556	C	T	JAZF1	Hair	https://www.nature.com/articles/s41467-018-07691-z
220	rs28531809	7	28803911	T	C	CREB5	Hair	https://www.nature.com/articles/s41467-018-07691-z



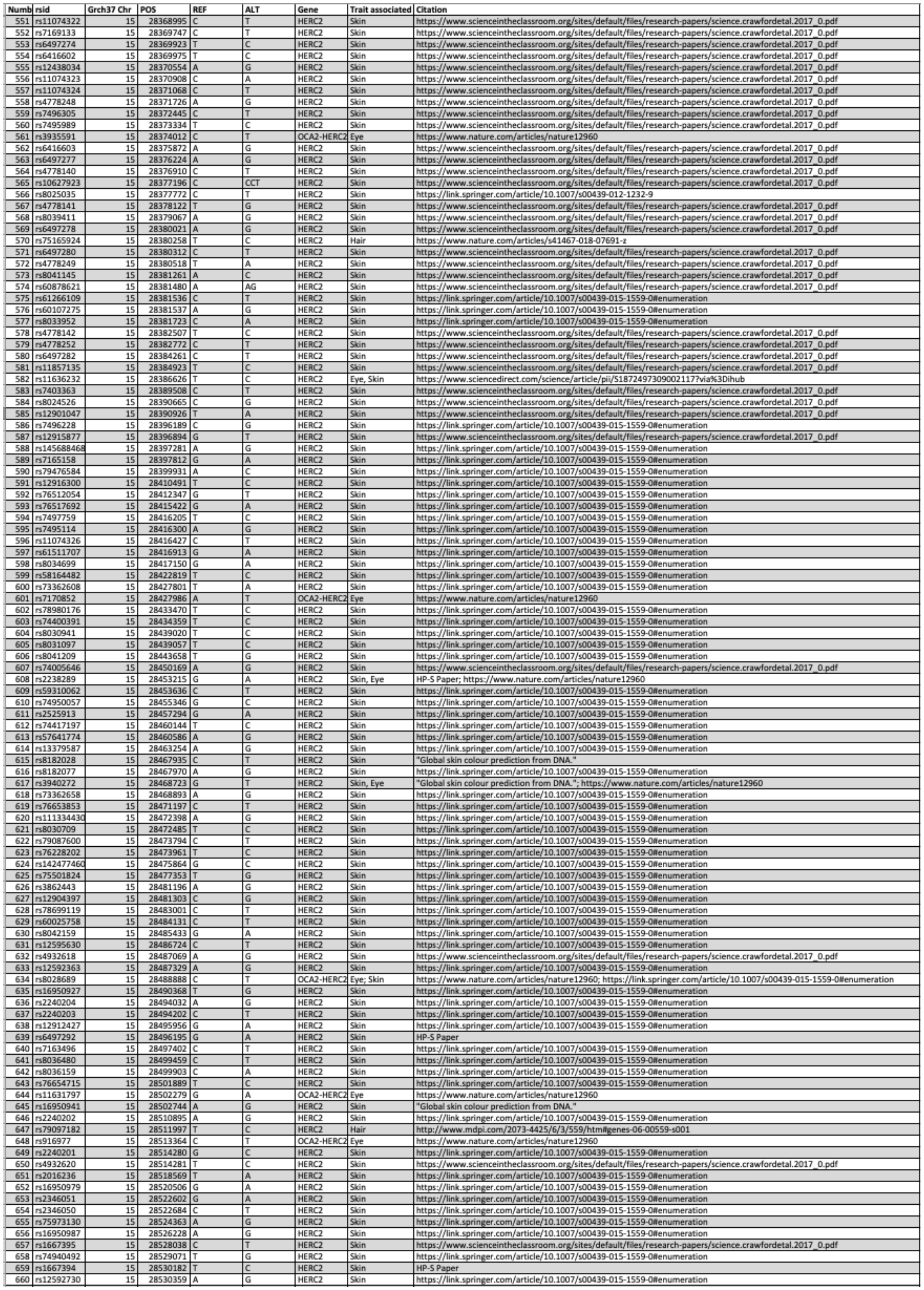
Numb	rsid	Grch37 Chr	POS	REF	ALT	Gene	Trait associated	Citation
221	rs7790204	7	36684481	T	C	AOAH	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
222	rs7794780	7	36685144	C	T	AOAH	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
223	rs12668421	7	55109177	T	A	EGFR	Skin	https://sites.lsa.umich.edu/bigham-lab/wp-content/uploads/sites/153/2014/08/HumGenetics.pdf
224	rs477823	7	62748101	G	T	(none given)	Skin	"Global skin colour prediction from DNA."
225	rs6977845	7	67216370	A	T	Chr.7:67216	Eye	https://peerj.com/articles/3951/
226	rs2373391	7	88449300	A	T	ZNF804B	Skin	https://peerj.com/articles/3951/
227	rs12667582	7	90848218	T	G	FZD1	Hair	https://www.nature.com/articles/s41467-018-07691-z
228	rs2710956	7	91190424	G	A	MTERF1	Hair	https://www.nature.com/articles/s41467-018-07691-z
229	rs314349	7	100401825	G	T	EPHB4	Hair	https://www.nature.com/articles/s41467-018-07691-z
230	rs12535629	7	100451732	T	C	SLC12A9	Hair	https://www.nature.com/articles/s41467-018-07691-z
231	rs80308281	7	100457578	C	T	SLC12A9	Hair	https://www.nature.com/articles/s41467-018-07691-z
232	rs2529369	7	105416560	A	C	ATXN7L1	Hair	https://www.nature.com/articles/s41467-018-07691-z
233	rs77462788	7	122042978	C	T	CADPS2	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
234	rs116908038	7	122168537	C	A	CADPS2	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
235	rs10954300	7	130761235	A	G	LINC-PINT	Hair	https://www.nature.com/articles/s41467-018-07691-z
236	rs1703884	8	529244	T	C	TDRP	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5884124/pdf/nihms921744.pdf
237	rs141034411	8	3098640	A	C	CSMD1	Eye	https://peerj.com/articles/3951/
238	rs6994536	8	13241026	T	C	DLC1	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5884124/pdf/nihms921744.pdf
239	rs12541402	8	15500967	C	T	TUSC3	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
240	rs7845221	8	22599421	T	C	PEBP4	Hair	https://www.nature.com/articles/s41467-018-07691-z
241	rs57128498	8	38568215	A	C	TACC1	Hair	https://www.nature.com/articles/s41467-018-07691-z
242	rs113060680	8	41987884	T	C	AP3M2	Hair	https://www.nature.com/articles/s41467-018-07691-z
243	rs7825896	8	52022917	G	A	TCF25	Hair	https://www.nature.com/articles/s41467-018-07691-z
244	rs10504523	8	72951490	A	G	TRPA1	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5884124/pdf/nihms921744.pdf
245	rs2600605	8	82720760	A	G	SNX16	Hair	https://www.nature.com/articles/s41467-018-07691-z
246	rs1385229	8	96771547	A	G	C8orf37-AS1	Skin	"Global skin colour prediction from DNA."
247	rs145048184	8	97070341	T	C	Chr.8:97070	Eye	https://peerj.com/articles/3951/
248	rs117954051	8	109067613	C	T	RSPO2	Hair	https://www.nature.com/articles/s41467-018-07691-z
249	rs2737212	8	116621214	C	T	TRPS1	Skin	https://www.nature.com/articles/s41467-018-04086-y/tables/1
250	rs16904127	8	130628606	A	G	CCDC26	Eye	https://peerj.com/articles/3951/
251	rs520015	9	211762	C	G	DOCK8	Hair	https://www.nature.com/articles/s41467-018-07691-z
252	rs10962731	9	1702573	A	G	SMARCA2	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5884124/pdf/nihms921744.pdf
253	rs146831108	9	2179528	G	T	SMARCA2	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5884124/pdf/nihms921744.pdf
254	rs872257	9	2496567	G	A	VLDLR	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5884124/pdf/nihms921744.pdf
255	rs2093657	9	5783498	T	C	ERMP1	Hair	https://www.nature.com/articles/s41467-018-07691-z
256	rs10815302	9	5887074	A	C	KIAA2026	Hair	https://www.nature.com/articles/s41467-018-07691-z
257	rs115075138	9	12048147	G	A	TYRP1	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5884124/pdf/nihms921744.pdf
258	rs2150097	9	12300716	G	A	TYRP1	Skin	https://academic.oup.com/endo/article/156/1/39/2800584
259	rs13289810	9	12396731	G	A	TYRP1	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3481182/
260	rs1408799	9	12672097	C	T	TYRP1	Eye, skin	https://academic.oup.com/hmg/article/18/R1/R9/2901093
261	rs2075508	9	12698363	C	T	TYRP1	Eye	http://www.genetics.org/content/165/4/2071
262	rs2733832	9	12704725	T	C	TYRP1	Eye, skin	https://academic.oup.com/hmg/article/18/R1/R9/2901093
263	rs2075509	9	12705219	A	C	TYRP1	Skin	https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000867#pgen-1000867-t003
264	rs683	9	12709305	A	C	TYRP1	Hair, Skin	HIrisPlex
265	rs2762464	9	12709586	T	A	TYRP1	Eye	http://www.genetics.org/content/165/4/2071
266	rs13288130	9	12712578	A	T	TYRP1	Hair	https://www.nature.com/articles/s41467-018-07691-z
267	rs1326797	9	12716762	G	T	TYRP1	Skin	https://www.nature.com/articles/s41467-018-04086-y/tables/1
268	rs10962599	9	16795286	T	C	BNC2	Hair	https://www.nature.com/articles/s41467-018-07691-z
269	rs10962612	9	16804167	G	T	BCN2	Skin	https://www.nature.com/articles/s41467-018-04086-y#Sec2
270	rs1339552	9	16848790	C	T	BNC2	Skin	https://link.springer.com/article/10.1007/s00439-012-1232-9
271	rs10756819	9	16858084	A	G	BNC2	Skin	HP-S Paper
272	rs2153271	9	16864521	T	C	BNC2	Skin	https://link.springer.com/article/10.1007/s00439-012-1232-9
273	rs10810650	9	16873551	T	C	BNC2	Skin	https://www.nature.com/articles/s41467-018-04086-y/tables/1
274	rs10810657	9	16884586	A	T	BNC2	Skin	https://academic.oup.com/hmg/article/23/21/5750/2901029#85674947
275	rs12350739	9	16885017	A	G	BNC2	Skin; Hair	https://www.nature.com/articles/s41467-018-07691-z
276	rs62543565	9	16901067	A	C	BNC2	Skin	"A Genome-Wide Association Study Identifies the Skin Color Genes IRF4, MC1R, AIP, and BNC2 Influencing Facial Pigmented Spots"
277	rs12001326	9	71649552	G	A	PRKACG	Skin	https://academic.oup.com/endo/article/156/1/39/2800584
278	rs7873389	9	100609230	C	T	FOXE1	Hair	https://www.nature.com/articles/s41467-018-07691-z
279	rs1484372	9	109010799	G	T	TMEM38B	Hair	https://www.nature.com/articles/s41467-018-07691-z
280	rs917783	9	126790607	C	A	LHX2	Hair	https://www.nature.com/articles/s41467-018-07691-z
281	rs58979150	9	126808006	T	C	LHX2	Hair	https://www.nature.com/articles/s41588-018-0100-5]
282	rs10114314	9	126808788	T	C	LHX2	Hair	https://www.nature.com/articles/s41467-018-07691-z
283	rs12344562	9	126811296	C	T	LHX2	Hair	https://www.nature.com/articles/s41467-018-07691-z
284	rs72759273	9	126973873	G	A	NEK6	Hair	https://www.nature.com/articles/s41467-018-07691-z
285	rs10818930	9	126991185	T	G	NEK6	Hair	https://www.nature.com/articles/s41467-018-07691-z
286	rs376397	10	8103298	A	G	GATA3	Skin	"Global skin colour prediction from DNA."
287	rs6602665	10	13605982	C	T	BEND7-PRPF	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
288	rs6602666	10	13606490	G	A	BEND7-PRPF	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
289	rs151165649	10	25207241	A	G	PRTFDC1	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
290	rs141664730	10	25338228	C	G	ENKUR	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
291	rs149207584	10	25339373	C	T	ENKUR	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
292	rs2505115	10	30398791	T	G	KIAA1462	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5884124/pdf/nihms921744.pdf
293	rs111256285	10	47632167	G	A	ANTXRLP1	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
294	rs7477798	10	47632478	C	A	ANTXRLP1	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
295	rs10443915	10	53820579	A	T	PRKG1	Skin	"Global skin colour prediction from DNA."
296	rs12765852	10	53821327	T	C	PRKG1	Skin	"Global skin colour prediction from DNA."
297	rs2050724	10	57701140	A	C	MTRNR2L5	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5884124/pdf/nihms921744.pdf
298	rs2278745	10	71152091	T	C	HK1	Eye	https://peerj.com/articles/3951/
299	rs7075993	10	74054590	C	G	DDIT4	Hair	https://www.nature.com/articles/s41467-018-07691-z
300	rs4980113	10	78629795	C	G	KCNMA1	Hair	https://onlinelibrary.wiley.com/doi/epdf/10.1111/exd.13333
301	rs703978	10	80944147	C	G	ZMIZ1	Hair	https://www.nature.com/articles/s41467-018-07691-z
302	rs1341164	10	96800873	C	T	CYP2C8	Eye	http://www.genetics.org/content/165/4/2071
303	rs1926705	10	96818418	C	T	CYP2C8	Eye	http://www.genetics.org/content/165/4/2071
304	rs10882664	10	97539140	A	G	ENTPD1	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5884124/pdf/nihms921744.pdf
305	rs7914735	10	110441149	T	C	Chr.10:11044	Eye	https://peerj.com/articles/3951/
306	rs11198112	10	119564143	T	C	EMX2	Skin	https://www.nature.com/articles/s41467-018-08147-0.pdf
307	rs35563099	10	119572403	T	C	EMX2	Skin	https://www.nature.com/articles/s41467-018-04086-y/tables/1
308	rs11041426	11	7543519	G	A	PPFIBP2	Hair	https://www.nature.com/articles/s41467-018-07691-z
309	rs111514753	11	13864951	C	G	RNA5SP331	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5884124/pdf/nihms921744.pdf
310	rs1531903	11	15668826	C	G	SOX6	Hair	https://www.nature.com/articles/s41467-018-07691-z
311	rs7109376	11	16372431	A	T	SOX6	Hair	https://www.nature.com/articles/s41467-018-07691-z
312	rs7943712	11	16543863	A	G	SOX6	Hair	https://www.nature.com/articles/s41467-018-07691-z
313	rs58604758	11	27420145	A	G	LGR4	Hair	https://www.nature.com/articles/s41467-018-07691-z
314	rs357925	11	44496552	A	G	CD82	Hair	https://www.nature.com/articles/s41467-018-07691-z
315	rs1851992	11	49438110	T	C	TYR	Eye	http://www.genetics.org/content/165/4/2071
316	rs653173	11	61018855	G	A	PGA5	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
317	rs2001746	11	61033525	T	A	VWCE	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
318	rs73490303	11	61043773	G	C	VWCE	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
319	rs9704187	11	61044470	G	C	VWCE	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
320	rs11230658	11	61046876	T	C	VWCE	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
321	rs1108769	11	61054892	A	C	VWCE	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
322	rs10897150	11	61063156	G	T	VWCE	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
323	rs12275843	11	61075524	T	C	DDB1	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
324	rs11230664	11	61076372	C	T	DDB1	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
325	rs7120594	11	61080557	T	C	DDB1	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
326	rs12289370	11	61084180	G	A	DDB1	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
327	rs7934735	11	61088140	G	T	DDB1	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
328	rs2512809	11	61106525	C	T	TKFC	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
329	rs2513329	11	61106892	G	C	TKFC	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
330	rs2260655	11	61108974	G	A	TKFC	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf



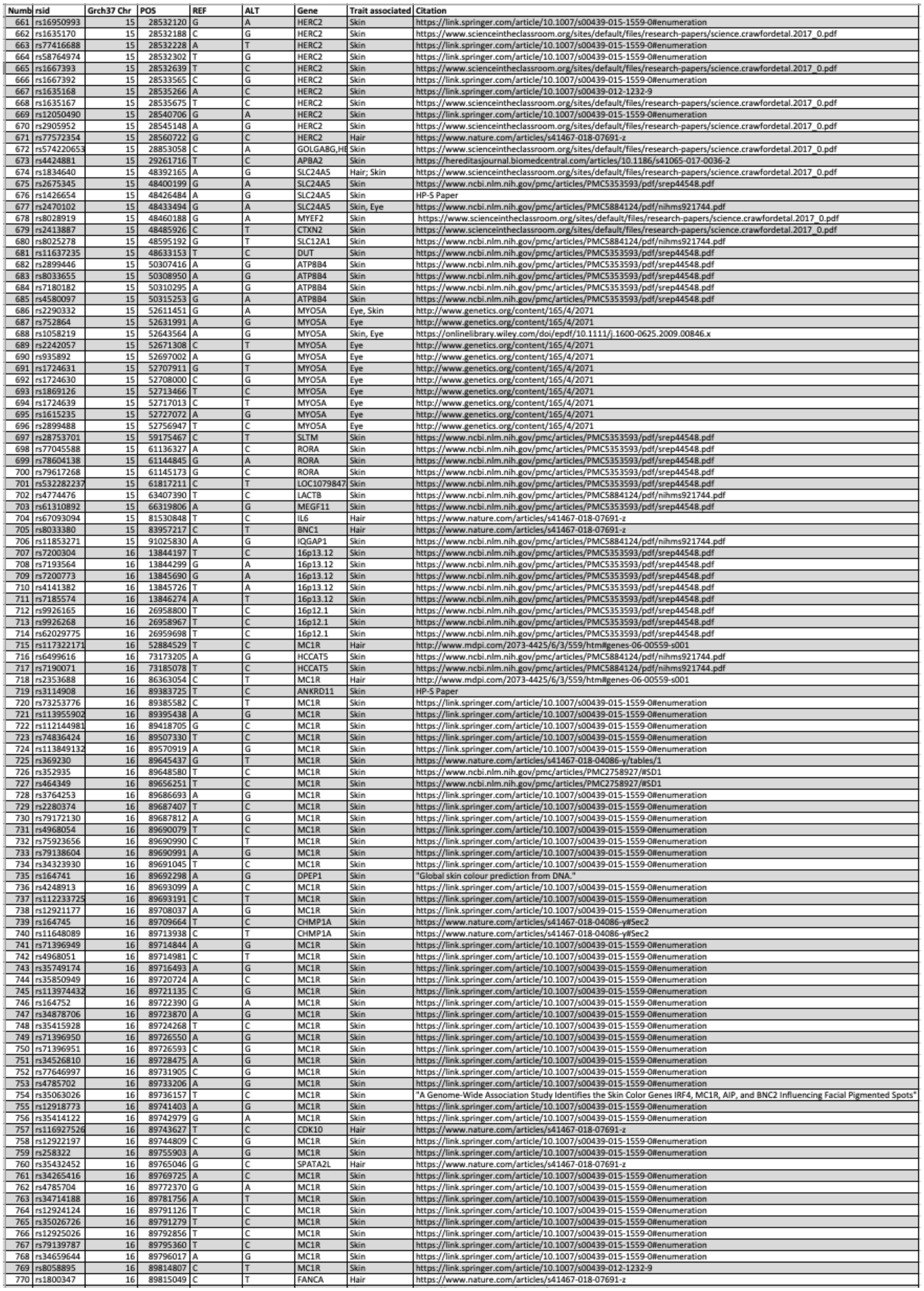
Numb	rsid	Grch37 Chr	POS	REF	ALT	Gene	Trait associated	Citation
331	rs2305465	11	61112802	C	T	TKFC	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
332	rs148172827	11	61115821	C	CATCAA	TKFC	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
333	rs7951574	11	61122878	G	A	CYB561A3	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
334	rs7948623	11	61137147	T	A	TMEM138	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
335	rs11230678	11	61139869	G	A	TMEM138	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
336	rs10897155	11	61141164	C	T	TMEM138	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
337	rs57265008	11	61141259	T	C	TMEM138	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
338	rs7394502	11	61141476	G	A	TMEM138	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
339	rs4939519	11	61142943	C	T	TMEM138	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
340	rs1377457	11	61144652	C	A	TMEM138	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
341	rs1377458	11	61144707	C	T	TMEM138	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
342	rs12791961	11	61152028	C	A	TMEM216	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
343	rs4453253	11	61152630	T	C	TMEM216	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
344	rs4939520	11	61153401	T	C	TMEM216	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
345	rs3017597	11	61222635	G	A	SDHAF2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
346	rs61896141	11	61556039	C	A	MYRF	Hair	https://www.nature.com/articles/s41467-018-07691-z
347	rs10897275	11	62203865	A	G	AHNAK	Hair	https://www.nature.com/articles/s41467-018-07691-z
348	rs56019505	11	65561805	G	T	OVOL1	Hair	https://www.nature.com/articles/s41467-018-07691-z
349	rs72917317	11	68817441	G	T	TPCN2	Skin, Hair	https://www.nature.com/articles/s41467-018-07691-z
350	rs35264875	11	68846399	T	A	TPCN2	Hair	http://www.pnas.org/content/pnas/114/41/E8595.full.pdf
351	rs34510004	11	68848916	A	G	MIR3164/TYI	Hair	https://www.nature.com/articles/s41467-018-07691-z
352	rs3829241	11	68855363	A	G	TPCN2	Hair	https://www.nature.com/articles/s41467-018-07691-z
353	rs72930659	11	68872843	T	C	TPCN2	Hair	http://www.mdpi.com/2073-4425/6/3/559/htm#genes-06-00559-s001
354	rs12806763	11	69358817	T	C	CCND1	Hair	https://www.nature.com/articles/s41467-018-07691-z
355	rs638640	11	69387455	C	T	ORAOV1	Hair	https://www.nature.com/articles/s41467-018-07691-z
356	rs1052030	11	76853783	T	C	MYO7A	Skin	https://onlinelibrary.wiley.com/doi/epdf/10.1111/j.1600-0625.2009.00846.x
357	rs762667	11	76868372	C	T	MYO7A	Skin, Eye	https://onlinelibrary.wiley.com/doi/epdf/10.1111/j.1600-0625.2009.00846.x
358	rs2276288	11	76912636	A	T	MYO7A	Skin	https://onlinelibrary.wiley.com/doi/epdf/10.1111/j.1600-0625.2009.00846.x
359	rs2276293	11	76917220	A	G	MYO7A	Skin	https://onlinelibrary.wiley.com/doi/epdf/10.1111/j.1600-0625.2009.00846.x
360	rs2292572	11	78052864	T	G	GAB2	Hair	https://www.nature.com/articles/s41467-018-07691-z
361	rs148065054	11	88552633	T	C	GRM5	Hair	https://www.nature.com/articles/s41467-018-07691-z
362	rs10831496	11	88557991	A	G	GRM5, TYR	Skin	"Global skin colour prediction from DNA.
363	rs1042602	11	88911696	A	C	TYR	Hair, Skin	HIrisPlex
364	rs7129973	11	88915570	G	A	TYR	Skin	https://academic.oup.com/endo/article/156/1/39/2800584
365	rs2000553	11	88936007	C	T	TYR	Skin	https://academic.oup.com/endo/article/156/1/39/2800584
366	rs4121401	11	88979846	C	T	TYR	Skin	https://academic.oup.com/endo/article/156/1/39/2800584
367	rs1393350	11	89011046	A	G	TYR	Eye, Skin, Hair	IrisPlex; https://www.nature.com/articles/s41467-018-07691-z
368	rs1126809	11	89017961	A	G	TYR	Skin	HP-S Paper; "Human pigmentation genes under environmental selection"
369	rs1827430	11	89018440	G	A	TYR	Eye	http://www.genetics.org/content/165/4/2071
370	rs1847142	11	89021574	A	G	TYR	Skin	https://link.springer.com/article/10.1007/s00439-012-1232-9
371	rs10830253	11	89028043	G	T	TYR	Skin	https://academic.oup.com/endo/article/156/1/39/2800584
372	rs115019323	11	95895552	A	G	MAML2	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
373	rs144848699	11	95895953	C	T	MAML2	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
374	rs7945369	11	103425586	T	C	Chr.11:10342	Skin	https://peerj.com/articles/3951/
375	rs4936890	11	123914742	G	A	intergenic be	Skin	"Global skin colour prediction from DNA."
376	rs610106	11	128935028	T	C	ARHGAP32	Eye	https://peerj.com/articles/3951/
377	rs12421680	11	131350968	G	A	NTM	Skin	https://academic.oup.com/hmg/article/22/14/2948/753805
378	rs7125438	11	132402910	A	G	OPCML	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5884124/pdf/nihms921744.pdf
379	rs10849455	12	785468	T	G	LINC02455	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
380	rs3764032	12	4317563	C	T	CCND2	Hair	https://www.nature.com/articles/s41467-018-07691-z
381	rs10849298	12	5692000	A	G	ANO2	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5884124/pdf/nihms921744.pdf
382	rs17820032	12	13248157	T	C	GSG1	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5884124/pdf/nihms921744.pdf
383	rs1902910	12	41778982	G	A	PDZRN4	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
384	rs17129378	12	41861386	C	T	PDZRN4	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
385	rs17578886	12	41862536	G	A	PDZRN4	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
386	rs11182085	12	43847683	G	A	ADAMTS20	Skin	https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000867#pgen-1000867-t003
387	rs11182091	12	43863773	T	C	ADAMTS20	Skin	https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000867#pgen-1000867-t003
388	rs1510523	12	43882501	T	C	ADAMTS20	Skin	https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000867#pgen-1000867-t003
389	rs4768698	12	46749088	A	G	SLC38A2	Hair	https://www.nature.com/articles/s41467-018-07691-z
390	rs7975232	12	48238837	C	A	VDR	Eye	https://www.sciencedirect.com/science/article/pii/S1872497314000313
391	rs11568820	12	48302545	T	C	VDR	Eye	https://www.sciencedirect.com/science/article/pii/S1872497314000313
392	rs139727704	12	52648158	G	A	KRT86/KRT7	Hair	https://www.nature.com/articles/s41467-018-07691-z
393	rs535209331	12	54332733	CTTA	C	HOXC13	Hair	https://www.nature.com/articles/s41467-018-07691-z
394	rs1052206	12	56348028	G	A	SILV	Eye	http://www.genetics.org/content/165/4/2071
395	rs1052165	12	56351346	A	G	SILV	Eye	http://www.genetics.org/content/165/4/2071
396	rs7487365	12	88674623	C	T	TMTC3	Hair	https://www.nature.com/articles/s41467-018-07691-z
397	rs2216153	12	88681977	G	A	TMTC3	Hair	https://www.nature.com/articles/s41467-018-07691-z
398	rs7306001	12	88736601	A	G	TMTC3	Hair	https://www.nature.com/articles/s41467-018-07691-z
399	rs35618688	12	88940157	A	G	KITLG	Hair	https://www.nature.com/articles/s41467-018-07691-z
400	rs1907703	12	88955642	C	T	KITLG	Hair	https://www.nature.com/articles/s41467-018-07691-z
401	rs10777129	12	88961713	A	G	KITLG	Hair, Skin	https://www.sciencedirect.com/science/article/pii/S1872497314001355?via%3Dihub#sec0060
402	rs642742	12	89299746	T	C	KITLG	Skin	"Global skin colour prediction from DNA."; "Human pigmentation genes under environmental selection"
403	rs12821256	12	89328335	C	T	KITLG	Hair, Skin	HIrisPlex; https://www.nature.com/articles/s41467-018-07691-z
404	rs12298351	12	89340112	T	C	KITLG	Hair	https://www.nature.com/articles/s41467-018-07691-z
405	rs249625	12	98185248	G	A	Chr.12:98185	Eye	https://peerj.com/articles/3951/
406	rs11066284	12	112842275	T	A	RPL6	Hair	https://www.nature.com/articles/s41467-018-07691-z
407	rs61939692	12	116535976	A	G	MED13L	Hair	https://www.nature.com/articles/s41467-018-07691-z
408	rs11068059	12	116967670	C	A	MAP1LC3B2	Hair	https://www.nature.com/articles/s41467-018-07691-z
409	rs10270	12	122756342	G	A	CLIP1	Hair	https://onlinelibrary.wiley.com/doi/epdf/10.1111/exd.13333
410	rs9548088	13	38449344	A	G	TRPC4	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5884124/pdf/nihms921744.pdf
411	rs9603422	13	39343822	T	C	FREM2	Hair	https://www.nature.com/articles/s41467-018-07691-z
412	rs149830128	13	46016101	A	G	SLC25A30-AS	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5884124/pdf/nihms921744.pdf
413	rs10507781	13	70728411	T	C	ATXN8OS	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5884124/pdf/nihms921744.pdf
414	rs1146927	13	78365944	G	A	EDNRB	Hair	https://academic.oup.com/hmg/article/22/14/2948/753805
415	rs975739	13	78381146	T	G	EDNRB	Hair	https://academic.oup.com/hmg/article/22/14/2948/753805
416	rs750192	13	78390743	G	A	SLAIN1	Hair	https://www.nature.com/articles/s41467-018-07691-z
417	rs1279403	13	78391757	T	C	EDNRB	Hair	https://www.nature.com/articles/s41588-018-0100-5]
418	rs58188699	13	78409714	T	G	EDNRB	Hair	https://academic.oup.com/hmg/article/22/14/2948/753805
419	rs5352	13	78475230	T	C	EDNRB	Eye	https://onlinelibrary.wiley.com/doi/epdf/10.1111/j.1600-0625.2009.00846.x
420	rs3782974	13	95092896	T	A	DCT	Skin	"Global skin colour prediction from DNA."
421	rs1325611	13	95094385	C	T	DCT	Eye	http://www.genetics.org/content/165/4/2071
422	rs1407995	13	95096013	T	C	DCT	Eye	http://www.genetics.org/content/165/4/2071; "Human pigmentation genes under environmental selection"
423	rs2296498	13	95096111	G	A	DCT	Eye	http://www.genetics.org/content/165/4/2071
424	rs2892681	13	95100140	C	G	DCT	Eye	http://www.genetics.org/content/165/4/2071
425	rs2031526	13	95100841	A	G	DCT	Skin	https://www.karger.com/Article/FullText/468538
426	rs1028806	13	95119427	G	A	DCT	Eye	http://www.genetics.org/content/165/4/2071
427	rs2031527	13	95123917	T	C	DCT	Eye	http://www.genetics.org/content/165/4/2071
428	rs9561570	13	95156198	T	G	DCT	Skin	https://www.nature.com/articles/s41467-018-04086-y/tables/1
429	rs6492711	13	95196559	C	T	DCT	Hair	https://www.nature.com/articles/s41467-018-07691-z
430	rs2050537	13	97260901	A	G	HS6ST3	Skin	"Global skin colour prediction from DNA."
431	rs138891280	13	106722485	G	A	13q33.2	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
432	rs188019015	13	106743244	T	C	13q33.2	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
433	rs146078872	13	106745894	G	A	13q33.2	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
434	rs114396339	13	106746964	T	A	13q33.2	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
435	rs727752	13	106754594	A	C	13q33.2	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
436	rs2016565	13	106754788	C	T	13q33.2	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
437	rs115940594	13	106754945	T	C	13q33.2	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
438	rs1046793	13	113539894	T	C	ATP11A	Skin	https://www.nature.com/articles/s41467-018-04086-y/tables/1
439	rs3024737	13	113819785	G	A	PCID2/PROZ	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
440	rs7326155	13	114460362	C	T	TMEM255B	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf



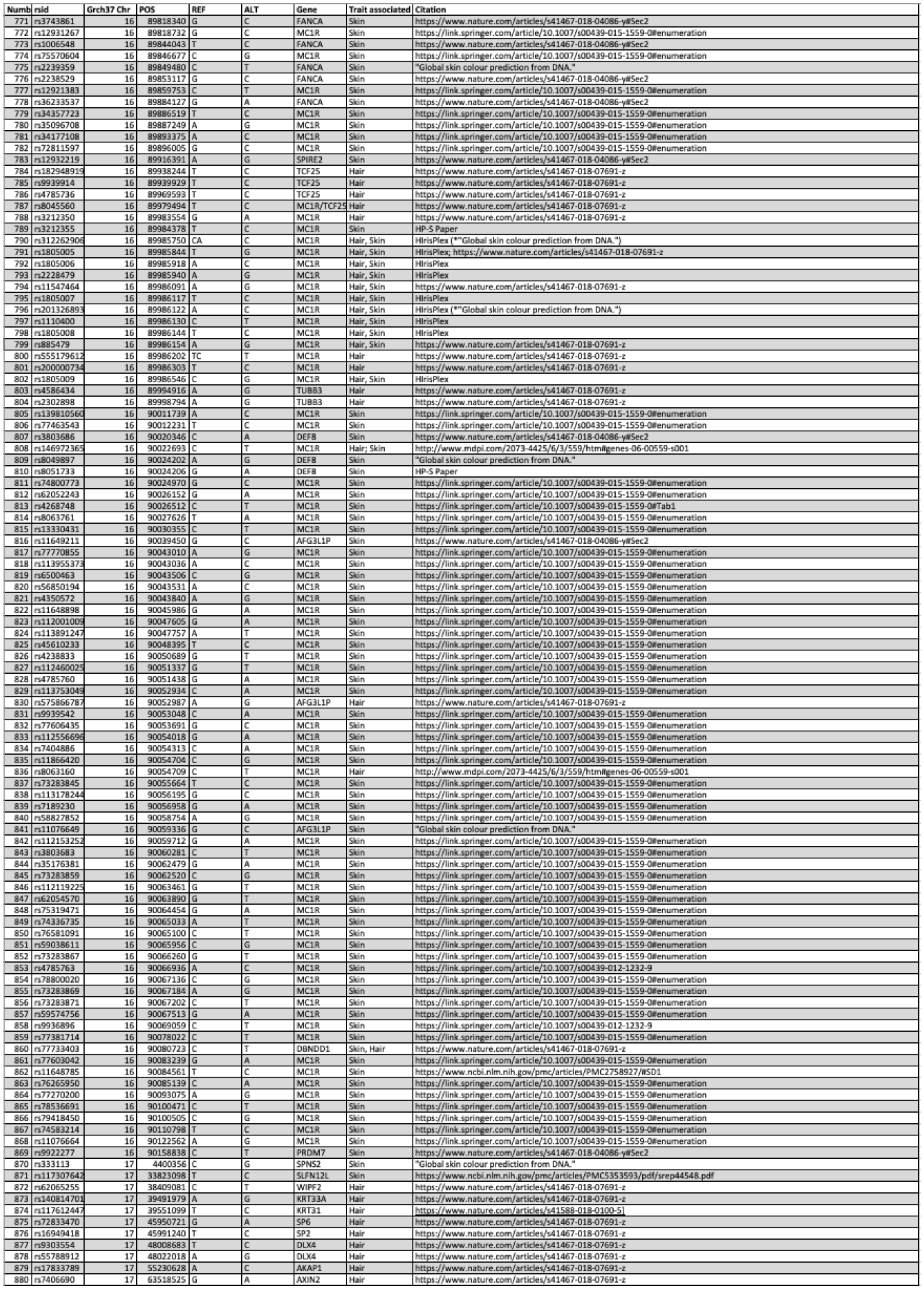
Numb	rsid	Grch37 Chr	POS	REF	ALT	Gene	Trait associated	Citation
441	rs4983161	14	20194876	A	T	(none given)	Skin	"Global skin colour prediction from DNA."
442	rs7158162	14	31384628	C	T	STRN3	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5884124/pdf/nihms921744.pdf
443	rs210381	14	54107791	G	A	BMP4	Hair	https://www.nature.com/articles/s41467-018-07691-z
444	rs1887103	14	60743219	G	A	PPM1A	Hair	https://www.nature.com/articles/s41467-018-07691-z
445	rs10873172	14	64390030	G	C	SYNE2	Hair	https://www.nature.com/articles/s41467-018-07691-z
446	rs11158532	14	64603204	G	A	SYNE2	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5884124/pdf/nihms921744.pdf
447	rs11158717	14	68514276	G	A	RAD51B	Hair	https://www.nature.com/articles/s41467-018-07691-z
448	rs72731537	14	69237925	T	C	ZFP36L1	Hair	https://www.nature.com/articles/s41467-018-07691-z
449	rs17107583	14	70466081	T	C	SLC8A3	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5884124/pdf/nihms921744.pdf
450	rs8011930	14	79237349	C	A	NRXN3	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5884124/pdf/nihms921744.pdf
451	rs75433889	14	92726294	T	C	SLC24A4	Hair	https://www.nature.com/articles/s41467-018-07691-z
452	rs12896399	14	92773663	T	G	SLC24A4	Eye, Skin, Hair	IrisPlex; "Predicting hair cortisol levels with hair pigmentation genes: a possible hair pigmentation bias"
453	rs746586	14	92775967	T	C	SLC24A4	Skin	https://www.nature.com/articles/s41467-018-04086-y/tables/1
454	rs941799	14	92776825	T	C	LOC1053706	Hair	https://www.nature.com/articles/s41467-018-07691-z
455	rs8014907	14	92800004	T	A	SLC24A4	Hair	http://www.mdpi.com/2073-4425/6/3/559/htm#genes-06-00559-s001
456	rs2402130	14	92801203	G	A	SLC24A4	Hair, Skin	HIrisPlex
457	rs4904886	14	92844370	A	G	SLC24A4	Hair	https://www.nature.com/articles/s41467-018-07691-z
458	rs10133804	14	92866905	C	T	SLC24A4	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
459	rs17128291	14	92882826	G	A	SLC24A4	Skin	HP-S Paper
460	rs17783630	14	92955385	C	A	SLC24A4	Hair	https://www.nature.com/articles/s41467-018-07691-z
461	rs17094273	14	97103807	A	G	chr 14	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2758927/#SD1
462	rs55859054	14	103953666	A	T	MARK3	Hair	https://www.nature.com/articles/s41467-018-07691-z
463	rs11858919	15	26599133	T	C	15q12	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
464	rs737051	15	27925836	C	T	OCA2	Eye	http://www.genetics.org/content/165/4/2071
465	rs139029488	15	27946226	C	T	OCA2	Hair	https://www.nature.com/articles/s41467-018-07691-z
466	rs924318	15	28093434	A	G	OCA2	Hair	https://www.nature.com/articles/s41467-018-07691-z
467	rs1448483	15	28133747	C	T	OCA2	Eye	http://www.genetics.org/content/165/4/2071
468	rs2044627	15	28151351	C	T	OCA2	Eye	http://www.genetics.org/content/165/4/2071
469	rs1800416	15	28171294	G	T	OCA2	0	https://www.sciencedirect.com/science/article/pii/S1344622316302140
470	rs1545397	15	28187772	T	A	OCA2	Skin	HP-S Paper
471	rs76930569	15	28196145	T	C	OCA2	Eye	https://peerj.com/articles/3951/
472	rs7173419	15	28196821	C	T	OCA2	Eye	https://academic.oup.com/hmg/article/22/14/2948/753805
473	rs1800414	15	28197037	C	T	OCA2	Skin	HP-S Paper; "Human pigmentation genes under environmental selection"
474	rs72625132	15	28213924	C	T	OCA2	Hair	https://www.nature.com/articles/s41467-018-07691-z
475	rs74653330	15	28228553	T	C	OCA2	Skin, Hair	"Human pigmentation genes under environmental selection"; https://www.nature.com/articles/s41467-018-07691-z
476	rs1900758	15	28230097	T	C	OCA2	Eye	http://www.genetics.org/content/165/4/2071
477	rs1800410	15	28230184	C	T	OCA2	Eye	http://www.genetics.org/content/165/4/2071
478	rs121918166	15	28230247	T	C	OCA2	Hair	https://www.nature.com/articles/s41467-018-07691-z
479	rs1800407	15	28230318	T	C	OCA2	Hair, Skin, Eye	HIrisPlex
480	rs1037208	15	28231357	G	T	OCA2	Eye	http://www.genetics.org/content/165/4/2071
481	rs10852218	15	28231793	T	C	OCA2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
482	rs1800404	15	28235773	T	C	OCA2	Eye; Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
483	rs33997466	15	28236800	AT	A	OCA2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
484	rs12911960	15	28237521	C	T	OCA2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
485	rs11630828	15	28237566	A	G	OCA2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
486	rs7178315	15	28237909	C	T	OCA2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
487	rs1868331	15	28238073	A	G	OCA2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
488	rs1868332	15	28238083	A	G	OCA2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
489	rs1868333	15	28238158	A	G	OCA2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
490	rs1868334	15	28238363	A	C	OCA2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
491	rs12595216	15	28238519	G	A	OCA2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
492	rs735066	15	28238895	G	A	OCA2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
493	rs735067	15	28238902	C	G	OCA2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
494	rs2015343	15	28239301	G	A	OCA2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
495	rs8029026	15	28239710	C	T	OCA2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
496	rs2077596	15	28239735	C	T	OCA2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
497	rs8041084	15	28240034	A	G	OCA2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
498	rs8024822	15	28240304	C	A	OCA2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
499	rs1900757	15	28240601	T	C	OCA2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
500	rs4778224	15	28241020	A	G	OCA2	Hair	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
501	rs4778225	15	28241157	T	C	OCA2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
502	rs4778226	15	28241189	C	A	OCA2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
503	rs4778227	15	28241199	T	C	OCA2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
504	rs12914545	15	28241574	G	T	OCA2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
505	rs4485307	15	28244460	A	G	OCA2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
506	rs12592367	15	28246287	G	T	OCA2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
507	rs11636259	15	28246990	A	G	OCA2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
508	rs1800401	15	28260053	A	G	OCA2	Skin	"Global skin colour prediction from DNA."
509	rs749846	15	28268990	A	C	OCA2	Eye	http://www.genetics.org/content/165/4/2071
510	rs12441727	15	28271775	A	G	OCA2	Skin	HP-S Paper
511	rs1448485	15	28282741	T	G	OCA2	Skin	"Global skin colour prediction from DNA."
512	rs1448484	15	28283441	G	A	OCA2	Skin	https://www.sciencedirect.com/science/article/pii/S1872497314001355?via%3Dihub#sec0060
513	rs16950821	15	28283507	A	G	OCA2	Skin	"Global skin colour prediction from DNA."
514	rs1470608	15	28288121	G	T	OCA2	Skin	HP-S Paper
515	rs1375164	15	28291812	C	T	OCA2	Skin	https://www.sciencedirect.com/science/article/pii/S1872497314001355?via%3Dihub#sec0060
516	rs116978932	15	28324912	A	G	OCA2	Hair	https://www.nature.com/articles/s41467-018-07691-z
517	rs4778138	15	28335820	A	G	OCA2	Eye, Hair	https://onlinelibrary.wiley.com/doi/epdf/10.1111/j.1600-0625.2009.00846.x
518	rs4778241	15	28338713	C	A	OCA2	Eye, Hair	https://onlinelibrary.wiley.com/doi/epdf/10.1111/j.1600-0625.2009.00846.x
519	rs7495174	15	28344238	G	A	OCA2	Skin, Eye, Hair	"Global skin colour prediction from DNA."; https://onlinelibrary.wiley.com/doi/epdf/10.1111/j.1600-0625.2009.00846.x
520	rs7497270	15	28344328	T	C	OCA2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
521	rs7496326	15	28344695	T	C	OCA2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
522	rs58843292	15	28345931	A	G	OCA2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
523	rs72625136	15	28348130	T	C	OCA2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
524	rs7495755	15	28355655	T	C	HERC2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
525	rs4778242	15	28355991	T	C	HERC2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
526	rs4778244	15	28356349	G	A	HERC2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
527	rs1129038	15	28356859	T	C	HERC2	Skin, Eye	HP-S Paper
528	rs4778245	15	28357230	T	C	HERC2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
529	rs4778246	15	28357796	C	T	HERC2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
530	rs12593929	15	28359258	G	A	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
531	rs7495441	15	28359744	G	A	HERC2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
532	rs7497014	15	28360081	T	C	HERC2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
533	rs5811542	15	28360888	C	CT	HERC2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
534	rs6497270	15	28361040	G	A	HERC2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
535	rs7180054	15	28361142	C	T	HERC2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
536	rs4778247	15	28361764	C	T	HERC2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
537	rs7495875	15	28362459	G	A	HERC2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
538	rs74007959	15	28362484	T	C	HERC2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
539	rs7497403	15	28363189	A	C	HERC2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
540	rs9707952	15	28364312	C	T	HERC2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
541	rs62007543	15	28364589	G	A	HERC2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
542	rs8034648	15	28364796	T	C	HERC2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
543	rs11074319	15	28364954	G	A	HERC2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
544	rs6497271	15	28365431	A	G	HERC2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
545	rs12913832	15	28365618	G	A	HERC2/OCA2	Eye, Skin	IrisPlex; "Human pigmentation genes under environmental selection"; https://www.nature.com/articles/s41467-018-07691-z
546	rs7183877	15	28365733	A	C	OCA2-HERC2	Eye	https://www.nature.com/articles/nature12960
547	rs6497272	15	28366190	G	A	HERC2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
548	rs11074320	15	28367196	G	A	HERC2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
549	rs11074321	15	28367351	G	A	HERC2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
550	rs11635884	15	28368969	C	T	HERC2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf



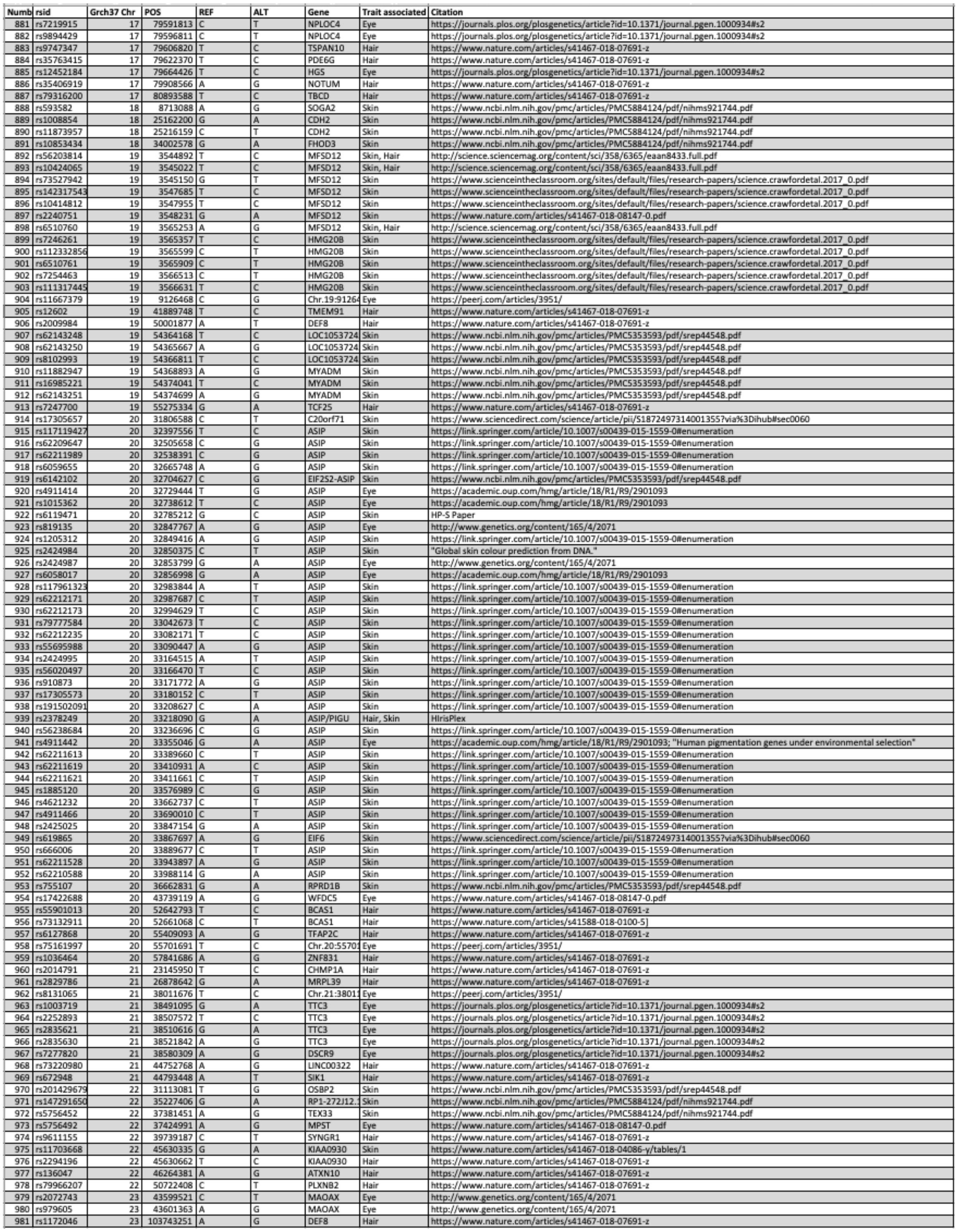
Numb	rsid	Grch37 Chr	POS	REF	ALT	Gene	Trait associated	Citation
551	rs11074322	15	28368995	C	T	HERC2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
552	rs7169133	15	28369747	C	T	HERC2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
553	rs6497274	15	28369923	T	C	HERC2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
554	rs6416602	15	28369975	T	C	HERC2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
555	rs12438034	15	28370554	A	G	HERC2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
556	rs11074323	15	28370908	C	A	HERC2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
557	rs11074324	15	28371068	C	T	HERC2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
558	rs4778248	15	28371726	A	G	HERC2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
559	rs7496305	15	28372445	C	T	HERC2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
560	rs7495989	15	28373334	T	C	HERC2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
561	rs3935591	15	28374012	C	T	OCA2-HERC2	Eye	https://www.nature.com/articles/nature12960
562	rs6416603	15	28375872	A	G	HERC2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
563	rs6497277	15	28376224	A	G	HERC2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
564	rs4778140	15	28376910	C	T	HERC2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
565	rs10627923	15	28377196	C	CCT	HERC2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
566	rs8025035	15	28377772	C	T	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-012-1232-9
567	rs4778141	15	28378122	T	G	HERC2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
568	rs8039411	15	28379067	A	G	HERC2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
569	rs6497278	15	28380021	A	G	HERC2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
570	rs75165924	15	28380258	T	C	HERC2	Hair	https://www.nature.com/articles/s41467-018-07691-z
571	rs6497280	15	28380312	C	T	HERC2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
572	rs4778249	15	28380518	T	A	HERC2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
573	rs8041145	15	28381261	A	C	HERC2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
574	rs60878621	15	28381480	A	AG	HERC2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
575	rs61266109	15	28381536	C	T	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
576	rs60107275	15	28381537	A	G	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
577	rs8033952	15	28381723	C	A	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
578	rs4778142	15	28382507	T	C	HERC2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
579	rs4778252	15	28382772	C	T	HERC2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
580	rs6497282	15	28384261	C	T	HERC2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
581	rs11857135	15	28384923	T	C	HERC2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
582	rs11636232	15	28386626	T	C	HERC2	Eye, Skin	https://www.sciencedirect.com/science/article/pii/S1872497309002117?via%3Dihub
583	rs7403363	15	28389508	C	T	HERC2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
584	rs8024526	15	28390665	C	G	HERC2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
585	rs12901047	15	28390926	T	A	HERC2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
586	rs7496228	15	28396189	C	G	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
587	rs12915877	15	28396894	G	T	HERC2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
588	rs145688468	15	28397281	A	G	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
589	rs7165158	15	28397812	G	A	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
590	rs79476584	15	28399931	A	C	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
591	rs12916300	15	28410491	T	C	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
592	rs76512054	15	28412347	G	T	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
593	rs76517692	15	28415422	G	A	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
594	rs7497759	15	28416205	T	C	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
595	rs7495114	15	28416300	A	G	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
596	rs11074326	15	28416427	C	T	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
597	rs61511707	15	28416913	G	A	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
598	rs8034699	15	28417150	G	A	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
599	rs58164482	15	28422819	T	C	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
600	rs73362608	15	28427801	T	A	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
601	rs7170852	15	28427986	A	T	OCA2-HERC2	Eye	https://www.nature.com/articles/nature12960
602	rs78980176	15	28433470	T	C	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
603	rs74400391	15	28434359	T	C	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
604	rs8030941	15	28439020	T	C	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
605	rs8031097	15	28439057	T	C	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
606	rs8041209	15	28443658	T	G	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
607	rs74005646	15	28450169	A	G	HERC2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
608	rs2238289	15	28453215	G	A	HERC2	Skin, Eye	HP-S Paper; https://www.nature.com/articles/nature12960
609	rs59310062	15	28453636	C	T	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
610	rs74950057	15	28455346	G	C	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
611	rs2525913	15	28457294	G	A	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
612	rs74417197	15	28460144	T	C	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
613	rs57641774	15	28460586	A	G	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
614	rs13379587	15	28463254	A	G	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
615	rs8182028	15	28467935	C	T	HERC2	Skin	"Global skin colour prediction from DNA."
616	rs8182077	15	28467970	A	G	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
617	rs3940272	15	28468723	G	T	HERC2	Skin, Eye	"Global skin colour prediction from DNA."; https://www.nature.com/articles/nature12960
618	rs73362658	15	28468893	A	G	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
619	rs76653853	15	28471197	C	T	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
620	rs111334430	15	28472398	A	G	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
621	rs8030709	15	28472485	T	C	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
622	rs79087600	15	28473794	C	T	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
623	rs76228202	15	28473961	T	C	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
624	rs142477460	15	28475864	G	C	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
625	rs75501824	15	28477353	T	G	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
626	rs3862443	15	28481196	A	G	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
627	rs12904397	15	28481303	C	G	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
628	rs78699119	15	28483001	C	T	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
629	rs60025758	15	28484131	C	T	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
630	rs8042159	15	28485433	G	A	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
631	rs12595630	15	28486724	C	T	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
632	rs4932618	15	28487069	A	G	HERC2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
633	rs12592363	15	28487329	A	G	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
634	rs8028689	15	28488888	C	T	OCA2-HERC2	Eye; Skin	https://www.nature.com/articles/nature12960; https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
635	rs16950927	15	28490368	T	G	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
636	rs2240204	15	28494032	A	G	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
637	rs2240203	15	28494202	C	T	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
638	rs12912427	15	28495956	G	A	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
639	rs6497292	15	28496195	G	A	HERC2	Skin	HP-S Paper
640	rs7163496	15	28497402	C	T	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
641	rs8036480	15	28499459	C	T	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
642	rs8036159	15	28499903	C	A	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
643	rs76654715	15	28501889	T	C	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
644	rs11631797	15	28502279	G	A	OCA2-HERC2	Eye	https://www.nature.com/articles/nature12960
645	rs16950941	15	28502744	A	G	HERC2	Skin	"Global skin colour prediction from DNA."
646	rs2240202	15	28510895	A	G	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
647	rs79097182	15	28511997	T	C	HERC2	Hair	http://www.mdpi.com/2073-4425/6/3/559/htm#genes-06-00559-s001
648	rs916977	15	28513364	C	T	OCA2-HERC2	Eye	https://www.nature.com/articles/nature12960
649	rs2240201	15	28514280	G	C	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
650	rs4932620	15	28514281	T	C	HERC2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
651	rs2016236	15	28518569	T	A	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
652	rs16950979	15	28520506	G	A	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
653	rs2346051	15	28522602	G	A	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
654	rs2346050	15	28522684	C	T	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
655	rs75973130	15	28524363	A	G	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
656	rs16950987	15	28526228	A	G	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
657	rs1667395	15	28528038	C	T	HERC2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
658	rs74940492	15	28529071	T	G	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
659	rs1667394	15	28530182	T	C	HERC2	Skin	HP-S Paper
660	rs12592730	15	28530359	A	G	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration



Numb	rsid	Grch37 Chr	POS	REF	ALT	Gene	Trait associated	Citation
661	rs16950993	15	28532120	G	A	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
662	rs1635170	15	28532188	C	G	HERC2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
663	rs77416688	15	28532228	A	T	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
664	rs58764974	15	28532302	T	G	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
665	rs1667393	15	28532639	T	C	HERC2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
666	rs1667392	15	28533565	C	G	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
667	rs1635168	15	28535266	A	C	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-012-1232-9
668	rs1635167	15	28535675	T	C	HERC2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
669	rs12050490	15	28540706	G	A	HERC2	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
670	rs2905952	15	28545148	A	G	HERC2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
671	rs77572354	15	28560722	G	C	HERC2	Hair	https://www.nature.com/articles/s41467-018-07691-z
672	rs574220653	15	28853058	C	A	GOLGA8G,HE	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
673	rs4424881	15	29261716	T	C	APBA2	Skin	https://hereditasjournal.biomedcentral.com/articles/10.1186/s41065-017-0036-2
674	rs1834640	15	48392165	A	G	SLC24A5	Hair; Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
675	rs2675345	15	48400199	G	A	SLC24A5	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
676	rs1426654	15	48426484	A	G	SLC24A5	Skin	HP-S Paper
677	rs2470102	15	48433494	G	A	SLC24A5	Skin, Eye	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5884124/pdf/nihms921744.pdf
678	rs8028919	15	48460188	G	A	MYEF2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
679	rs2413887	15	48485926	C	T	CTXN2	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
680	rs8025278	15	48595192	G	T	SLC12A1	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5884124/pdf/nihms921744.pdf
681	rs11637235	15	48633153	T	C	DUT	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
682	rs2899446	15	50307416	A	G	ATP8B4	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
683	rs8033655	15	50308950	A	G	ATP8B4	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
684	rs7180182	15	50310295	A	G	ATP8B4	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
685	rs4580097	15	50315253	G	A	ATP8B4	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
686	rs2290332	15	52611451	G	A	MYO5A	Eye, Skin	http://www.genetics.org/content/165/4/2071
687	rs752864	15	52631991	A	G	MYO5A	Eye	http://www.genetics.org/content/165/4/2071
688	rs1058219	15	52643564	A	G	MYO5A	Skin, Eye	https://onlinelibrary.wiley.com/doi/epdf/10.1111/j.1600-0625.2009.00846.x
689	rs2242057	15	52671308	C	T	MYO5A	Eye	http://www.genetics.org/content/165/4/2071
690	rs935892	15	52697002	A	G	MYO5A	Eye	http://www.genetics.org/content/165/4/2071
691	rs1724631	15	52707911	G	T	MYO5A	Eye	http://www.genetics.org/content/165/4/2071
692	rs1724630	15	52708000	C	G	MYO5A	Eye	http://www.genetics.org/content/165/4/2071
693	rs1869126	15	52713466	T	C	MYO5A	Eye	http://www.genetics.org/content/165/4/2071
694	rs1724639	15	52717013	C	T	MYO5A	Eye	http://www.genetics.org/content/165/4/2071
695	rs1615235	15	52727072	A	G	MYO5A	Eye	http://www.genetics.org/content/165/4/2071
696	rs2899488	15	52756947	T	C	MYO5A	Eye	http://www.genetics.org/content/165/4/2071
697	rs28753701	15	59175467	C	T	SLTM	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
698	rs77045588	15	61136327	A	C	RORA	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
699	rs78604138	15	61144845	G	A	RORA	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
700	rs79617268	15	61145173	G	C	RORA	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
701	rs532282237	15	61817211	C	T	LOC1079847	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
702	rs4774476	15	63407390	T	C	LACTB	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5884124/pdf/nihms921744.pdf
703	rs61310892	15	66319806	A	G	MEGF11	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
704	rs67093094	15	81530848	T	C	IL6	Hair	https://www.nature.com/articles/s41467-018-07691-z
705	rs8033380	15	83957217	C	T	BNC1	Hair	https://www.nature.com/articles/s41467-018-07691-z
706	rs11853271	15	91025830	A	G	IQGAP1	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5884124/pdf/nihms921744.pdf
707	rs7200304	16	13844197	T	C	16p13.12	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
708	rs7193564	16	13844299	G	A	16p13.12	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
709	rs7200773	16	13845690	G	A	16p13.12	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
710	rs4141382	16	13845726	T	A	16p13.12	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
711	rs7185574	16	13846274	A	T	16p13.12	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
712	rs9926165	16	26958800	T	C	16p12.1	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
713	rs9926268	16	26958967	T	C	16p12.1	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
714	rs62029775	16	26959698	T	C	16p12.1	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
715	rs117322171	16	52884529	T	C	MC1R	Hair	http://www.mdpi.com/2073-4425/6/3/559/htm#genes-06-00559-s001
716	rs6499616	16	73173205	A	G	HCCAT5	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5884124/pdf/nihms921744.pdf
717	rs7190071	16	73185078	T	C	HCCAT5	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5884124/pdf/nihms921744.pdf
718	rs2353688	16	86363054	C	T	MC1R	Hair	http://www.mdpi.com/2073-4425/6/3/559/htm#genes-06-00559-s001
719	rs3114908	16	89383725	T	C	ANKRD11	Skin	HP-S Paper
720	rs73253776	16	89385582	C	T	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
721	rs113955902	16	89395438	A	G	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
722	rs112144981	16	89418705	G	C	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
723	rs74836424	16	89507330	T	C	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
724	rs113849132	16	89570919	A	G	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
725	rs369230	16	89645437	G	T	MC1R	Skin	https://www.nature.com/articles/s41467-018-04086-y/tables/1
726	rs352935	16	89648580	T	C	MC1R	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2758927/#SD1
727	rs464349	16	89656251	T	C	MC1R	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2758927/#SD1
728	rs3764253	16	89686693	A	G	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
729	rs2280374	16	89687407	T	C	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
730	rs79172130	16	89687812	A	G	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
731	rs4968054	16	89690079	T	C	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
732	rs75923656	16	89690990	C	T	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
733	rs79138604	16	89690991	A	G	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
734	rs34323930	16	89691045	T	C	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
735	rs164741	16	89692298	A	G	DPEP1	Skin	"Global skin colour prediction from DNA."
736	rs4248913	16	89693099	A	C	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
737	rs112233725	16	89693191	C	T	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
738	rs12921177	16	89708037	A	G	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
739	rs164745	16	89709664	T	C	CHMP1A	Skin	https://www.nature.com/articles/s41467-018-04086-y#Sec2
740	rs11648089	16	89713938	C	T	CHMP1A	Skin	https://www.nature.com/articles/s41467-018-04086-y#Sec2
741	rs71396949	16	89714844	A	G	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
742	rs4968051	16	89714981	C	T	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
743	rs35749174	16	89716493	A	G	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
744	rs35850949	16	89720724	A	C	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
745	rs113974432	16	89721135	C	G	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
746	rs164752	16	89722390	G	A	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
747	rs34878706	16	89723870	A	G	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
748	rs35415928	16	89724268	T	C	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
749	rs71396950	16	89726550	A	G	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
750	rs71396951	16	89726593	C	G	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
751	rs34526810	16	89728475	A	G	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
752	rs77646997	16	89731905	C	G	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
753	rs4785702	16	89733206	A	G	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
754	rs35063026	16	89736157	T	C	MC1R	Skin	"A Genome-Wide Association Study Identifies the Skin Color Genes IRF4, MC1R, AIP, and BNC2 Influencing Facial Pigmented Spots"
755	rs12918773	16	89741403	A	G	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
756	rs35414122	16	89742979	G	A	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
757	rs116927526	16	89743627	T	C	CDK10	Hair	https://www.nature.com/articles/s41467-018-07691-z
758	rs12922197	16	89744809	C	G	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
759	rs258322	16	89755903	A	G	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
760	rs35432452	16	89765046	G	C	SPATA2L	Hair	https://www.nature.com/articles/s41467-018-07691-z
761	rs34265416	16	89769725	A	C	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
762	rs4785704	16	89772370	G	A	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
763	rs34714188	16	89781756	A	T	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
764	rs12924124	16	89791126	T	C	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
765	rs35026726	16	89791279	T	C	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
766	rs12925026	16	89792856	T	C	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
767	rs79139787	16	89795360	T	C	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
768	rs34659644	16	89796017	A	G	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
769	rs8058895	16	89814807	C	T	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-012-1232-9
770	rs1800347	16	89815049	C	T	FANCA	Hair	https://www.nature.com/articles/s41467-018-07691-z



Numb	rsid	Grch37 Chr	POS	REF	ALT	Gene	Trait associated	Citation
771	rs3743861	16	89818340	G	C	FANCA	Skin	https://www.nature.com/articles/s41467-018-04086-y#Sec2
772	rs12931267	16	89818732	G	C	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
773	rs1006548	16	89844043	T	C	FANCA	Skin	https://www.nature.com/articles/s41467-018-04086-y#Sec2
774	rs75570604	16	89846677	C	G	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
775	rs2239359	16	89849480	C	T	FANCA	Skin	"Global skin colour prediction from DNA."
776	rs2238529	16	89853117	G	C	FANCA	Skin	https://www.nature.com/articles/s41467-018-04086-y#Sec2
777	rs12921383	16	89859753	C	T	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
778	rs36233537	16	89884127	G	A	FANCA	Skin	https://www.nature.com/articles/s41467-018-04086-y#Sec2
779	rs34357723	16	89886519	T	C	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
780	rs35096708	16	89887249	A	G	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
781	rs34177108	16	89893375	A	C	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
782	rs72811597	16	89896005	G	C	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
783	rs12932219	16	89916391	A	G	SPIRE2	Skin	https://www.nature.com/articles/s41467-018-04086-y#Sec2
784	rs182948919	16	89938244	T	C	TCF25	Hair	https://www.nature.com/articles/s41467-018-07691-z
785	rs9939914	16	89939929	T	C	TCF25	Hair	https://www.nature.com/articles/s41467-018-07691-z
786	rs4785736	16	89969593	T	C	TCF25	Hair	https://www.nature.com/articles/s41467-018-07691-z
787	rs8045560	16	89979494	T	C	MC1R/TCF25	Hair	https://www.nature.com/articles/s41467-018-07691-z
788	rs3212350	16	89983554	G	A	MC1R	Hair	https://www.nature.com/articles/s41467-018-07691-z
789	rs3212355	16	89984378	T	C	MC1R	Skin	HP-S Paper
790	rs312262906	16	89985750	CA	C	MC1R	Hair, Skin	HIrisPlex (*"Global skin colour prediction from DNA.")
791	rs1805005	16	89985844	T	G	MC1R	Hair, Skin	HIrisPlex; https://www.nature.com/articles/s41467-018-07691-z
792	rs1805006	16	89985918	A	C	MC1R	Hair, Skin	HIrisPlex
793	rs2228479	16	89985940	A	G	MC1R	Hair, Skin	HIrisPlex
794	rs11547464	16	89986091	A	G	MC1R	Hair, Skin	https://www.nature.com/articles/s41467-018-07691-z
795	rs1805007	16	89986117	T	C	MC1R	Hair, Skin	HIrisPlex
796	rs201326893	16	89986122	A	C	MC1R	Hair, Skin	HIrisPlex (*"Global skin colour prediction from DNA.")
797	rs1110400	16	89986130	C	T	MC1R	Hair, Skin	HIrisPlex
798	rs1805008	16	89986144	T	C	MC1R	Hair, Skin	HIrisPlex
799	rs885479	16	89986154	A	G	MC1R	Hair, Skin	https://www.nature.com/articles/s41467-018-07691-z
800	rs555179612	16	89986202	TC	T	MC1R	Hair	https://www.nature.com/articles/s41467-018-07691-z
801	rs200000734	16	89986303	T	C	MC1R	Hair	https://www.nature.com/articles/s41467-018-07691-z
802	rs1805009	16	89986546	C	G	MC1R	Hair, Skin	HIrisPlex
803	rs4586434	16	89994916	A	G	TUBB3	Hair	https://www.nature.com/articles/s41467-018-07691-z
804	rs2302898	16	89998794	A	G	TUBB3	Hair	https://www.nature.com/articles/s41467-018-07691-z
805	rs139810560	16	90011739	A	C	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
806	rs77463543	16	90012231	T	C	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
807	rs3803686	16	90020346	C	A	DEF8	Skin	https://www.nature.com/articles/s41467-018-04086-y#Sec2
808	rs146972365	16	90022693	C	T	MC1R	Hair; Skin	http://www.mdpi.com/2073-4425/6/3/559/htm#genes-06-00559-s001
809	rs8049897	16	90024202	A	G	DEF8	Skin	"Global skin colour prediction from DNA."
810	rs8051733	16	90024206	G	A	DEF8	Skin	HP-S Paper
811	rs74800773	16	90024970	G	C	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
812	rs62052243	16	90026152	G	A	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
813	rs4268748	16	90026512	C	T	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#Tab1
814	rs8063761	16	90027626	T	A	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
815	rs13330431	16	90030355	C	T	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
816	rs11649211	16	90039450	G	C	AFG3L1P	Skin	https://www.nature.com/articles/s41467-018-04086-y#Sec2
817	rs77770855	16	90043010	A	G	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
818	rs113955373	16	90043036	A	C	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
819	rs6500463	16	90043506	C	G	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
820	rs56850194	16	90043531	A	C	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
821	rs4350572	16	90043840	A	G	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
822	rs11648898	16	90045986	G	A	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
823	rs112001009	16	90047605	G	A	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
824	rs113891247	16	90047757	A	T	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
825	rs45610233	16	90048395	T	C	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
826	rs4238833	16	90050689	G	T	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
827	rs112460025	16	90051337	G	T	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
828	rs4785760	16	90051438	G	A	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
829	rs113753049	16	90052934	C	A	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
830	rs575866787	16	90052987	A	G	AFG3L1P	Hair	https://www.nature.com/articles/s41467-018-07691-z
831	rs9939542	16	90053048	C	A	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
832	rs77606435	16	90053691	G	C	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
833	rs112556696	16	90054018	G	A	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
834	rs7404886	16	90054313	C	A	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
835	rs11866420	16	90054704	C	G	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
836	rs8063160	16	90054709	C	T	MC1R	Hair	http://www.mdpi.com/2073-4425/6/3/559/htm#genes-06-00559-s001
837	rs73283845	16	90055664	T	C	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
838	rs113178244	16	90056195	G	C	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
839	rs7189230	16	90056958	G	A	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
840	rs58827852	16	90058754	A	G	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
841	rs11076649	16	90059336	G	C	AFG3L1P	Skin	"Global skin colour prediction from DNA."
842	rs112153252	16	90059712	G	A	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
843	rs3803683	16	90060281	C	T	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
844	rs35176381	16	90062479	G	A	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
845	rs73283859	16	90062520	C	G	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
846	rs112119225	16	90063461	G	T	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
847	rs62054570	16	90063890	G	T	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
848	rs75319471	16	90064454	G	A	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
849	rs74336735	16	90065033	A	T	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
850	rs76581091	16	90065100	C	T	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
851	rs59038611	16	90065956	C	G	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
852	rs73283867	16	90066260	G	T	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
853	rs4785763	16	90066936	A	C	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-012-1232-9
854	rs78800020	16	90067136	C	G	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
855	rs73283869	16	90067184	A	G	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
856	rs73283871	16	90067202	C	T	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
857	rs59574756	16	90067513	G	A	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
858	rs9936896	16	90069059	C	T	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-012-1232-9
859	rs77381714	16	90078022	C	T	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
860	rs77733403	16	90080723	C	T	DBNDD1	Skin, Hair	https://www.nature.com/articles/s41467-018-07691-z
861	rs77603042	16	90083239	G	A	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
862	rs11648785	16	90084561	T	C	MC1R	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2758927/#SD1
863	rs76265950	16	90085139	C	A	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
864	rs77270200	16	90093075	A	G	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
865	rs78536691	16	90100471	C	T	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
866	rs79418450	16	90100505	C	G	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
867	rs74583214	16	90110798	T	C	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
868	rs11076664	16	90122562	A	G	MC1R	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
869	rs9922277	16	90158838	C	T	PRDM7	Skin	https://www.nature.com/articles/s41467-018-04086-y#Sec2
870	rs333113	17	4400356	C	G	SPNS2	Skin	"Global skin colour prediction from DNA."
871	rs117307642	17	33823098	T	C	SLFN12L	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
872	rs62065255	17	38409081	C	T	WIPF2	Hair	https://www.nature.com/articles/s41467-018-07691-z
873	rs140814701	17	39491979	A	G	KRT33A	Hair	https://www.nature.com/articles/s41467-018-07691-z
874	rs117612447	17	39551099	T	C	KRT31	Hair	https://www.nature.com/articles/s41588-018-0100-5
875	rs72833470	17	45950721	G	A	SP6	Hair	https://www.nature.com/articles/s41467-018-07691-z
876	rs16949418	17	45991240	T	C	SP2	Hair	https://www.nature.com/articles/s41467-018-07691-z
877	rs9303554	17	48008683	T	C	DLX4	Hair	https://www.nature.com/articles/s41467-018-07691-z
878	rs55788912	17	48022018	A	G	DLX4	Hair	https://www.nature.com/articles/s41467-018-07691-z
879	rs17833789	17	55230628	A	C	AKAP1	Hair	https://www.nature.com/articles/s41467-018-07691-z
880	rs7406690	17	63518525	G	A	AXIN2	Hair	https://www.nature.com/articles/s41467-018-07691-z



Numb	rsid	Grch37 Chr	POS	REF	ALT	Gene	Trait associated	Citation
881	rs7219915	17	79591813	C	T	NPLOC4	Eye	https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000934#s2
882	rs9894429	17	79596811	C	T	NPLOC4	Eye	https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000934#s2
883	rs9747347	17	79606820	T	C	TSPAN10	Hair	https://www.nature.com/articles/s41467-018-07691-z
884	rs35763415	17	79622370	T	C	PDE6G	Hair	https://www.nature.com/articles/s41467-018-07691-z
885	rs12452184	17	79664426	T	C	HGS	Eye	https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000934#s2
886	rs35406919	17	79908566	A	G	NOTUM	Hair	https://www.nature.com/articles/s41467-018-07691-z
887	rs79316200	17	80893588	T	C	TBCD	Hair	https://www.nature.com/articles/s41467-018-07691-z
888	rs593582	18	8713088	A	G	SOGA2	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5884124/pdf/nihms921744.pdf
889	rs1008854	18	25162200	G	A	CDH2	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5884124/pdf/nihms921744.pdf
890	rs11873957	18	25216159	C	T	CDH2	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5884124/pdf/nihms921744.pdf
891	rs10853434	18	34002578	G	A	FHOD3	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5884124/pdf/nihms921744.pdf
892	rs56203814	19	3544892	T	C	MFSD12	Skin, Hair	http://science.sciencemag.org/content/sci/358/6365/eaan8433.full.pdf
893	rs10424065	19	3545022	T	C	MFSD12	Skin, Hair	http://science.sciencemag.org/content/sci/358/6365/eaan8433.full.pdf
894	rs73527942	19	3545150	G	T	MFSD12	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
895	rs142317543	19	3547685	T	C	MFSD12	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
896	rs10414812	19	3547955	T	C	MFSD12	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
897	rs2240751	19	3548231	G	A	MFSD12	Skin	https://www.nature.com/articles/s41467-018-08147-0.pdf
898	rs6510760	19	3565253	A	G	MFSD12	Skin, Hair	http://science.sciencemag.org/content/sci/358/6365/eaan8433.full.pdf
899	rs7246261	19	3565357	T	C	HMG20B	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
900	rs112332856	19	3565599	C	T	HMG20B	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
901	rs6510761	19	3565909	C	T	HMG20B	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
902	rs7254463	19	3566513	C	T	HMG20B	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
903	rs111317445	19	3566631	T	C	HMG20B	Skin	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
904	rs11667379	19	9126468	C	G	Chr.19:91264	Eye	https://peerj.com/articles/3951/
905	rs12602	19	41889748	T	C	TMEM91	Hair	https://www.nature.com/articles/s41467-018-07691-z
906	rs2009984	19	50001877	A	T	DEF8	Hair	https://www.nature.com/articles/s41467-018-07691-z
907	rs62143248	19	54364168	T	C	LOC1053724	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
908	rs62143250	19	54365667	A	G	LOC1053724	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
909	rs8102993	19	54366811	T	C	LOC1053724	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
910	rs11882947	19	54368893	A	G	MYADM	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
911	rs16985221	19	54374041	T	C	MYADM	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
912	rs62143251	19	54374699	A	G	MYADM	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
913	rs7247700	19	55275334	G	A	TCF25	Hair	https://www.nature.com/articles/s41467-018-07691-z
914	rs17305657	20	31806588	C	T	C20orf71	Skin	https://www.sciencedirect.com/science/article/pii/S1872497314001355?via%3Dihub#sec0060
915	rs117119427	20	32397556	T	C	ASIP	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
916	rs62209647	20	32505658	C	G	ASIP	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
917	rs62211989	20	32538391	C	G	ASIP	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
918	rs6059655	20	32665748	A	G	ASIP	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
919	rs6142102	20	32704627	C	G	EIF2S2-ASIP	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
920	rs4911414	20	32729444	T	G	ASIP	Eye	https://academic.oup.com/hmg/article/18/R1/R9/2901093
921	rs1015362	20	32738612	T	C	ASIP	Eye	https://academic.oup.com/hmg/article/18/R1/R9/2901093
922	rs6119471	20	32785212	G	C	ASIP	Skin	HP-S Paper
923	rs819135	20	32847767	A	G	ASIP	Eye	http://www.genetics.org/content/165/4/2071
924	rs1205312	20	32849416	A	G	ASIP	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
925	rs2424984	20	32850375	C	T	ASIP	Skin	"Global skin colour prediction from DNA."
926	rs2424987	20	32853799	G	A	ASIP	Eye	http://www.genetics.org/content/165/4/2071
927	rs6058017	20	32856998	G	A	ASIP	Eye	https://academic.oup.com/hmg/article/18/R1/R9/2901093
928	rs117961323	20	32983844	A	T	ASIP	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
929	rs62212171	20	32987687	C	T	ASIP	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
930	rs62212173	20	32994629	T	C	ASIP	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
931	rs79777584	20	33042673	T	C	ASIP	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
932	rs62212235	20	33082171	T	C	ASIP	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
933	rs55695988	20	33090447	A	G	ASIP	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
934	rs2424995	20	33164515	A	T	ASIP	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
935	rs56020497	20	33166470	T	C	ASIP	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
936	rs910873	20	33171772	A	G	ASIP	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
937	rs17305573	20	33180152	C	T	ASIP	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
938	rs191502091	20	33208627	C	A	ASIP	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
939	rs2378249	20	33218090	G	A	ASIP/PIGU	Hair, Skin	HIrisPlex
940	rs56238684	20	33236696	C	G	ASIP	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
941	rs4911442	20	33355046	G	A	ASIP	Eye	https://academic.oup.com/hmg/article/18/R1/R9/2901093; "Human pigmentation genes under environmental selection"
942	rs62211613	20	33389660	C	T	ASIP	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
943	rs62211619	20	33410931	A	C	ASIP	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
944	rs62211621	20	33411661	C	T	ASIP	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
945	rs1885120	20	33576989	C	G	ASIP	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
946	rs4621232	20	33662737	C	T	ASIP	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
947	rs4911466	20	33690010	C	T	ASIP	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
948	rs2425025	20	33847154	G	A	ASIP	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
949	rs619865	20	33867697	A	G	EIF6	Skin	https://www.sciencedirect.com/science/article/pii/S1872497314001355?via%3Dihub#sec0060
950	rs666006	20	33889677	C	T	ASIP	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
951	rs62211528	20	33943897	A	G	ASIP	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
952	rs62210588	20	33988114	G	A	ASIP	Skin	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
953	rs755107	20	36662831	G	A	RPRD1B	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
954	rs17422688	20	43739119	A	G	WFDC5	Eye	https://www.nature.com/articles/s41467-018-08147-0.pdf
955	rs55901013	20	52642793	T	C	BCAS1	Hair	https://www.nature.com/articles/s41467-018-07691-z
956	rs73132911	20	52661068	C	T	BCAS1	Hair	https://www.nature.com/articles/s41588-018-0100-5]
957	rs6127868	20	55409093	A	G	TFAP2C	Hair	https://www.nature.com/articles/s41467-018-07691-z
958	rs75161997	20	55701691	T	C	Chr.20:55701	Eye	https://peerj.com/articles/3951/
959	rs1036464	20	57841686	A	G	ZNF831	Hair	https://www.nature.com/articles/s41467-018-07691-z
960	rs2014791	21	23145950	T	C	CHMP1A	Hair	https://www.nature.com/articles/s41467-018-07691-z
961	rs2829786	21	26878642	G	A	MRPL39	Hair	https://www.nature.com/articles/s41467-018-07691-z
962	rs8131065	21	38011676	T	C	Chr.21:3801	Eye	https://peerj.com/articles/3951/
963	rs1003719	21	38491095	G	A	TTC3	Eye	https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000934#s2
964	rs2252893	21	38507572	T	C	TTC3	Eye	https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000934#s2
965	rs2835621	21	38510616	G	A	TTC3	Eye	https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000934#s2
966	rs2835630	21	38521842	A	G	TTC3	Eye	https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000934#s2
967	rs7277820	21	38580309	A	G	DSCR9	Eye	https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000934#s2
968	rs73220980	21	44752768	A	G	LINC00322	Hair	https://www.nature.com/articles/s41467-018-07691-z
969	rs672948	21	44793448	A	T	SIK1	Hair	https://www.nature.com/articles/s41467-018-07691-z
970	rs201429679	22	31113081	T	G	OSBP2	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
971	rs147291650	22	35227406	G	A	RP1-272J12.1	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5884124/pdf/nihms921744.pdf
972	rs5756452	22	37381451	A	G	TEX33	Skin	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5884124/pdf/nihms921744.pdf
973	rs5756492	22	37424991	A	G	MPST	Eye	https://www.nature.com/articles/s41467-018-08147-0.pdf
974	rs9611155	22	39739187	C	T	SYNGR1	Hair	https://www.nature.com/articles/s41467-018-07691-z
975	rs11703668	22	45630335	G	A	KIAA0930	Skin	https://www.nature.com/articles/s41467-018-04086-y/tables/1
976	rs2294196	22	45630662	T	C	KIAA0930	Hair	https://www.nature.com/articles/s41467-018-07691-z
977	rs136047	22	46264381	A	G	ATXN10	Hair	https://www.nature.com/articles/s41467-018-07691-z
978	rs79966207	22	50722408	C	T	PLXNB2	Hair	https://www.nature.com/articles/s41467-018-07691-z
979	rs2072743	23	43599521	C	T	MAOAX	Eye	http://www.genetics.org/content/165/4/2071
980	rs979605	23	43601363	A	G	MAOAX	Eye	http://www.genetics.org/content/165/4/2071
981	rs1172046	23	103743251	A	G	DEF8	Hair	https://www.nature.com/articles/s41467-018-07691-z

APPENDIX F. NEURAL NETWORK PIGMENTATION VARIANTS

Table F1. Feature-Selected Pigmentation Variants Input into the Iris Prediction Neural Network

Variant	Gene	Associated Trait	Z	P	Citation
rs12203592	IRF4	Eye, Skin, Hair	10.59	0	https://www.ncbi.nlm.nih.gov/pubmed/20457092; https://genomebiology.biomedcentral.com/articles/10.1186/gb-2012-13-9-248
rs1129038	HERC2	Skin, Eye	-25.78	0	https://www.sciencedirect.com/science/article/pii/S1872497318302205
rs1800407	OCA2	Hair, Skin, Eye	7.71	1.22E-14	https://www.sciencedirect.com/science/article/pii/S1872497312001810; "Predicting hair cortisol levels with hair pigmentation genes: a possible hair pigmentation bias"; https://www.nature.com/articles/s41467-018-07691-z
rs35763415	PDE6G	Hair	-4.77	1.89E-06	https://www.nature.com/articles/s41467-018-07691-z
rs28117	SLC45A2	Skin	-4.60	4.17E-06	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
rs12202284	IRF4	Skin	4.37	1.25E-05	https://reader.elsevier.com/reader/sd/87B2B08122BFE636349CAE57D03D85A7EFABEE31D332CF5F397F45C23A85CAA4556DD7771188E0C40F1E3EA8E83ABEEB
rs121918166	OCA2	Hair	4.20	2.71E-05	https://www.nature.com/articles/s41467-018-07691-z
rs1426654	SLC24A5	Skin	4.10	4.16E-05	https://www.sciencedirect.com/science/article/pii/S1872497318302205; https://genomebiology.biomedcentral.com/articles/10.1186/gb-2012-13-9-248; https://www.nature.com/articles/nature12960; https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf; https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
rs74653330	OCA2	Skin, Hair	3.78	0.000156	https://genomebiology.biomedcentral.com/articles/10.1186/gb-2012-13-9-248; https://www.nature.com/articles/s41467-018-07691-z
rs1393350	TYR	Eye, Skin, Hair	3.76	0.000173	https://www.ncbi.nlm.nih.gov/pubmed/20457092; https://www.nature.com/articles/s41467-018-07691-z
rs1408799	TYRP1	Eye, skin	-3.72	0.000199	https://academic.oup.com/hmg/article/18/R1/R9/2901093; https://genomebiology.biomedcentral.com/articles/10.1186/gb-2012-13-9-248

rs3212355	MC1R	Skin	3.44	0.000577	https://www.sciencedirect.com/science/article/pii/S1872497318302205
rs6917661	OPRM1	Skin	-3.41	0.000644	https://reader.elsevier.com/reader/sd/87B2B08122BFE636349CAE57D03D85A7EFABEE31D332CF5F397F45C23A85CAA4556DD7771188E0C40F1E3EA8E83ABEEB
rs941799	LOC105370627	Hair	3.37	0.000746	https://www.nature.com/articles/s41467-018-07691-z
rs7495875	HERC2	Skin	3.27	0.001058	https://www.scienceintheclassroom.org/sites/default/files/research-papers/science.crawfordetal.2017_0.pdf
rs1448484	OCA2	Skin	3.27	0.001073	https://www.sciencedirect.com/science/article/pii/S1872497314001355?via%3Dihub#sec0060
rs11731416	RPL34-AS1	Hair	3.27	0.001085	https://www.nature.com/articles/s41467-018-07691-z
rs11066284	RPL6	Hair	-3.23	0.001236	https://www.nature.com/articles/s41467-018-07691-z
rs7183877	OCA2-HERC2	Eye	3.19	0.001403	https://www.nature.com/articles/nature12960
rs16950821	OCA2	Skin	-2.95	0.003171	https://link.springer.com/article/10.1007/s00439-017-1808-5
rs79316200	TBCD	Hair	2.86	0.004232	https://www.nature.com/articles/s41467-018-07691-z
rs251464	PPARGC1B	Skin	-2.80	0.005136	https://www.nature.com/articles/s41467-018-04086-y/tables/1
rs6739706	TMEM163	Hair	2.80	0.005226	https://www.nature.com/articles/s41467-018-07691-z
rs11806180	CDC42BPA	Hair	2.79	0.005297	https://www.nature.com/articles/s41467-018-07691-z
rs2075508	TYRP1	Eye	-2.72	0.006469	http://www.genetics.org/content/165/4/2071
rs1448483	OCA2	Eye	2.72	0.006534	http://www.genetics.org/content/165/4/2071
rs8049897	DEF8	Skin	2.60	0.009446	https://link.springer.com/article/10.1007/s00439-017-1808-5
rs4248913	MC1R	Skin	-2.51	0.012012	https://link.springer.com/article/10.1007/s00439-015-1559-0#enumeration
rs6924266	LRRC16A	Eye	2.45	0.0142	https://peerj.com/articles/3951/
rs28753701	SLTM	Skin	-2.45	0.014222	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353593/pdf/srep44548.pdf
rs57836066	RP11-134K13.4	Skin	-2.44	0.014595	https://peerj.com/articles/3951/
rs17820032	GSG1	Skin	2.39	0.016712	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5884124/pdf/nihms921744.pdf
rs4424881	APBA2	Skin	2.35	0.018707	https://reader.elsevier.com/reader/sd/87B2B08122BFE636349CAE57D03D85A7EFABEE31D332CF5F397F45C23A85CAA4556DD7771188E0C40F1E3EA8E83ABEEB

rs2050537	HS6ST3	Skin	-2.35	0.018709	https://link.springer.com/article/10.1007/s00439-017-1808-5
rs4980113	KCNMA1	Hair	-2.31	0.020861	https://onlinelibrary.wiley.com/doi/epdf/10.1111/exd.13333
rs76648881	TNFRSF9	Hair	2.25	0.024138	https://www.nature.com/articles/s41467-018-07691-z
rs1052030	MYO7A	Skin	-2.08	0.037436	https://onlinelibrary.wiley.com/doi/epdf/10.1111/j.1600-0625.2009.00846.x
rs1458046	FGF5	Hair	-2.08	0.037907	https://www.nature.com/articles/s41467-018-07691-z
rs1667394	HERC2	Skin	-2.02	0.043176	https://www.sciencedirect.com/science/article/pii/S1872497318302205

REFERENCES

1. Kayser, M. Forensic DNA Phenotyping: Predicting human appearance from crime scene material for investigative purposes. *Forensic Science International: Genetics* **18**, 33–48 (2015).
2. King, T. E. *et al.* Identification of the remains of King Richard III. *Nat Commun* **5**, 1–8 (2014).
3. Sulem, P. *et al.* Two newly identified genetic determinants of pigmentation in Europeans. *Nat Genet* **40**, 835–837 (2008).
4. Sulem, P. *et al.* Genetic determinants of hair, eye and skin pigmentation in Europeans. *Nat Genet* **39**, 1443–52 (2007).
5. Sturm, R. A. *et al.* A single SNP in an evolutionary conserved region within intron 86 of the HERC2 gene determines human blue-brown eye color. *Am J Hum Genet* **82**, 424–31 (2008).
6. Duffy, D. L. *et al.* A three-single-nucleotide polymorphism haplotype in intron 1 of OCA2 explains most human eye-color variation. *Am J Hum Genet* **80**, 241–52 (2007).
7. Chaitanya, L. *et al.* The HIrisPlex-S system for eye, hair and skin colour prediction from DNA: Introduction and forensic developmental validation. *Forensic Science International: Genetics* **35**, 123–135 (2018).
8. Walsh, S. *et al.* Global skin colour prediction from DNA. *Human genetics* **136**, 847–863 (2017).
9. Walsh, S. *et al.* Developmental validation of the IrisPlex system: determination of blue and brown iris colour for forensic intelligence. *Forensic Sci Int Genet* **5**, 464–71 (2011).
10. Ruiz, Y. *et al.* Further development of forensic eye color predictive tests. *Forensic Science International: Genetics* **7**, 28–40 (2013).
11. Liu, F. *et al.* Eye color and the prediction of complex phenotypes from genotypes. *Curr Biol* **19**, R192-3 (2009).
12. Kastelic, V., Pośpiech, E., Draus-Barini, J., Branicki, W. & Drobnič, K. Prediction of eye color in the Slovenian population using the IrisPlex SNPs. *Croatian medical journal* **54**, 381–386 (2013).

13. Pośpiech, E., Draus-Barini, J., Kupiec, T., Wojas-Pelc, A. & Branicki, W. Prediction of Eye Color from Genetic Data Using Bayesian Approach*. *Journal of Forensic Sciences* **57**, 880–886 (2012).

14. Beaty, T. H. *et al.* A genome-wide association study of cleft lip with and without cleft palate identifies risk variants near MAFB and ABCA4. *Nat Genet* **42**, 525–9 (2010).

15. Boehringer, S. *et al.* Genetic determination of human facial morphology: links between cleft-lips and normal variation. *Eur J Hum Genet* **19**, 1192–7 (2011).

16. Jugessur, A. *et al.* Genetic variants in IRF6 and the risk of facial clefts: single-marker and haplotype-based analyses in a population-based case-control study of facial clefts in Norway. *Genet Epidemiol* **32**, 413–24 (2008).

17. Park, J. W. *et al.* Association between IRF6 and nonsyndromic cleft lip with or without cleft palate in four populations. *Genet Med* **9**, 219–27 (2007).

18. Zucchero, T. M. *et al.* Interferon regulatory factor 6 (IRF6) gene variants and the risk of isolated cleft lip or palate. *N Engl J Med* **351**, 769–80 (2004).

19. Guihard-Costa, A.-M., Khung, S., Delbecque, K., Ménez, F. & Delezoide, A.-L. Biometry of face and brain in fetuses with trisomy 21. *Pediatr. Res.* **59**, 33–38 (2006).

20. Jesuino, F. A. S. & Valladares-Neto, J. Craniofacial morphological differences between Down syndrome and maxillary deficiency children. *Eur J Orthod* **35**, 124–130 (2013).

21. Austin, J. H., Preger, L., Siris, E. & Taybi, H. Short hard palate in newborn: roentgen sign of mongolism. *Radiology* **92**, 775-776 passim (1969).

22. Jensen, G. M., Cleall, J. F. & Yip, A. S. Dentoalveolar morphology and developmental changes in Down's syndrome (trisomy 21). *Am J Orthod* **64**, 607–618 (1973).

23. Fink, G. B., Madaus, W. K. & Walker, G. F. A quantitative study of the face in Down's syndrome. *Am J Orthod* **67**, 540–553 (1975).

24. Liu, F. *et al.* A genome-wide association study identifies five loci influencing facial morphology in Europeans. *PLoS Genet* **8**, e1002932 (2012).

25. Paternoster, L. *et al.* Genome-wide association study of three-dimensional facial morphology identifies a variant in PAX3 associated with nasion position. *Am J Hum Genet* **90**, 478–85 (2012).

26. Cole, J. B. *et al.* Genomewide Association Study of African Children Identifies Association of SCHIP1 and PDE8A with Facial Size and Shape. *PLOS Genetics* **12**, e1006174 (2016).

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

28. Claes, P. *et al.* Genome-wide mapping of global-to-local genetic effects on human facial shape. *Nat Genet* **50**, 414–423 (2018).

29. White, J. D. *et al.* MeshMonk: Open-source large-scale intensive 3D phenotyping. *Sci Rep* **9**, 1–11 (2019).

30. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

31. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature genetics* **39**, 906 (2007).

32. Bycroft, C. *et al.* Genome-wide genetic data on~ 500,000 UK Biobank participants. *BioRxiv* 166298 (2017).

33. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nature genetics* **48**, 1284 (2016).

34. Galinsky, K. J. *et al.* Fast principal-component analysis reveals convergent evolution of ADH1B in Europe and East Asia. *The American Journal of Human Genetics* **98**, 456–472 (2016).

35. Bland, J. M. & Altman, D. G. Multiple significance tests: the Bonferroni method. *Bmj* **310**, 170 (1995).

36. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* **57**, 289–300 (1995).

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

38. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics* **81**, 559–575 (2007).

39. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).

40. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nature genetics* **44**, 821 (2012).

41. Brenner, M. & Hearing, V. J. The Protective Role of Melanin Against UV Damage in Human Skin†. *Photochemistry and Photobiology* **84**, 539–549 (2008).

42. Mackey, D. A., Wilkinson, C. H., Kearns, L. S. & Hewitt, A. W. Classification of iris colour: review and refinement of a classification schema. *Clinical & Experimental Ophthalmology* **39**, 462–471 (2011).

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

44. Scherer, D. & Kumar, R. Genetics of pigmentation in skin cancer--a review. *Mutat. Res.* **705**, 141–153 (2010).

45. Litwack, G. Chapter 13 - Metabolism of Amino Acids. in *Human Biochemistry* (ed. Litwack, G.) 359–394 (Academic Press, 2018). doi:10.1016/B978-0-12-383864-3.00013-2.

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

47. Horrell, E. M. W., Wilson, K. & D'Orazio, J. A. Melanoma — Epidemiology, Risk Factors, and the Role of Adaptive Pigmentation. *Melanoma - Current Clinical Management and Future Therapeutics* (2015) doi:10.5772/58994.

48. Costin, G.-E. & Hearing, V. J. Human skin pigmentation: melanocytes modulate skin color in response to stress. *FASEB J.* **21**, 976–994 (2007).

49. Levy, C., Khaled, M. & Fisher, D. E. MITF: master regulator of melanocyte development and melanoma oncogene. *Trends in Molecular Medicine* **12**, 406–414 (2006).

50. Yamaguchi, Y., Brenner, M. & Hearing, V. J. The Regulation of Skin Pigmentation. *J. Biol. Chem.* **282**, 27557–27561 (2007).

51. Wakamatsu, K., Hu, D.-N., McCormick, S. A. & Ito, S. Characterization of melanin in human iridal and choroidal melanocytes from eyes with various colored irides. *Pigment Cell & Melanoma Research* **21**, 97–105 (2008).

52. Eagle, R. C. Iris pigmentation and pigmented lesions: an ultrastructural study. *Trans Am Ophthalmol Soc* **86**, 581–687 (1988).

53. Wilkerson, C. L., Syed, N. A., Fisher, M. R., Robinson, N. L. & Albert, D. M. Melanocytes and iris color: light microscopic findings. *Archives of ophthalmology* **114**, 437–442 (1996).

54. Iris (anatomy). *Wikipedia* (2019).

55. Hogan, M. J. Histology of the human eye. *an Atlas and Textbook* (1971).

56. Remington, L. A. & Goodwin, D. *Clinical anatomy of the visual system E-Book*. (Elsevier Health Sciences, 2011).

57. 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).

58. Bustamante, J., Bredeston, L., Malanga, G. & Mordoh, J. Role of Melanin as a Scavenger of Active Oxygen Species. *Pigment Cell Research* **6**, 348–353 (1993).

59. Fertl, D. & Rosel, P. E. Albinism. in *Encyclopedia of Marine Mammals (Second Edition)* (eds. Perrin, W. F., Würsig, B. & Thewissen, J. G. M.) 24–26 (Academic Press, 2009). doi:10.1016/B978-0-12-373553-9.00006-7.

60. Wobmann, P. R. & Fine, B. S. The clump cells of Koganei: A light and electron microscopic study. *American journal of ophthalmology* **73**, 90–101 (1972).

61. Eiberg, H. *et al.* Blue eye color in humans may be caused by a perfectly associated founder mutation in a regulatory element located within the HERC2 gene inhibiting OCA2 expression. *Hum. Genet.* **123**, 177–187 (2008).

62. Frudakis, T. *et al.* Sequences associated with human iris pigmentation. *Genetics* **165**, 2071–2083 (2003).

63. Graf, J., Hodgson, R. & van Daal, A. Single nucleotide polymorphisms in the MATP gene are associated with normal human pigmentation variation. *Hum. Mutat.* **25**, 278–284 (2005).

64. 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**, 411–23 (2008).

65. Wollstein, A. *et al.* Novel quantitative pigmentation phenotyping enhances genetic association, epistasis, and prediction of human eye colour. *Scientific Reports* **7**, 43359 (2017).

66. Johannsdottir, B., Thorarinsson, F., Thordarson, A. & Magnusson, T. E. Heritability of craniofacial characteristics between parents and offspring estimated from lateral cephalograms. *Am J Orthod Dentofacial Orthop* **127**, 200–7; quiz 260–1 (2005).

67. Hubbe, M., Hanihara, T. & Harvati, K. Climate signatures in the morphological differentiation of worldwide modern human populations. *Anat Rec (Hoboken)* **292**, 1720–33 (2009).

68. Harvati, K. & Weaver, T. D. Human cranial anatomy and the differential preservation of population history and climate signatures. *Anat Rec A Discov Mol Cell Evol Biol* **288**, 1225–33 (2006).

69. Williams, S. E. & Slice, D. E. Regional shape change in adult facial bone curvature with age. *Am J Phys Anthropol* **143**, 437–47 (2010).

70. Tsukamoto, K. *et al.* Cloning and characterization of the inversion breakpoint at chromosome 2q35 in a patient with Waardenburg syndrome type I. *Hum Mol Genet* **1**, 315–7 (1992).

71. Rinne, T., Brunner, H. G. & van Bokhoven, H. p63-associated disorders. *Cell Cycle* **6**, 262–8 (2007).

72. Leoyklang, P., Siriwan, P. & Shotelersuk, V. A mutation of the p63 gene in non-syndromic cleft lip. *J Med Genet* **43**, e28 (2006).

73. Thomason, H. A., Dixon, M. J. & Dixon, J. Facial clefting in Tp63 deficient mice results from altered Bmp4, Fgf8 and Shh signaling. *Dev Biol* **321**, 273–82 (2008).

74. Peng, S. *et al.* Detecting genetic association of common human facial morphological variation using high density 3D image registration. *PLoS Comput Biol* **9**, e1003375 (2013).

75. Ermakov, S., Rosenbaum, M. G., Malkin, I. & Livshits, G. Family-based study of association between ENPP1 genetic variants and craniofacial morphology. *Ann Hum Biol* **37**, 754–66 (2010).

76. Yamaguchi, T., Maki, K. & Shibasaki, Y. Growth hormone receptor gene variant and mandibular height in the normal Japanese population. *Am J Orthod Dentofacial Orthop* **119**, 650–3 (2001).

77. Tomoyasu, Y. *et al.* Further evidence for an association between mandibular height and the growth hormone receptor gene in a Japanese population. *Am J Orthod Dentofacial Orthop* **136**, 536–41 (2009).

78. Zhou, J. *et al.* The growth hormone receptor gene is associated with mandibular height in a Chinese population. *J Dent Res* **84**, 1052–6 (2005).

79. Coussens, A. K. & van Daal, A. Linkage disequilibrium analysis identifies an FGFR1 haplotype-tag SNP associated with normal variation in craniofacial shape. *Genomics* **85**, 563–73 (2005).

80. Robert, P. & Escoufier, Y. A Unifying Tool for Linear Multivariate Statistical Methods: The RV- Coefficient. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **25**, 257–265 (1976).

81. Arbuckle, J. L. Amos (version 23.0)[computer program]. *Chicago: IBM SpSS* (2014).

82. Rosseel, Y. lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software* **48**, 1–36 (2012).

83. Muthén, L. & Muthén, B. Mplus user's guide sixth edition. *Los Angeles, CA: Muthén & Muthén* **2013**, (1998).

84. Howie, B. N., Donnelly, P. & Marchini, J. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLOS Genetics* **5**, e1000529 (2009).

85. Li, Y., Willer, C. J., Ding, J., Scheet, P. & Abecasis, G. R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic epidemiology* **34**, 816–834 (2010).

86. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nature Reviews Genetics* **11**, 499 (2010).

87. International HapMap Consortium. The international HapMap project. *Nature* **426**, 789 (2003).

88. the Haplotype Reference Consortium *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics* **48**, 1279–1283 (2016).

89. NIH. NIH announces national enrollment date for All of Us Research Program to advance precision medicine. (2018).

90. Van Rheenen, W. *et al.* Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis. *Nature genetics* **48**, 1043 (2016).

91. Lemieux Perreault, L.-P., Legault, M.-A., Asselin, G. & Dube, M.-P. genipe: an automated genome-wide imputation pipeline with automatic reporting and statistical tools. *Bioinformatics* **32**, 3661–3663 (2016).

92. Johnston, H. R. *et al.* Identifying tagging SNPs for African specific genetic variation from the African Diaspora Genome. *Scientific Reports* **7**, 46398 (2017).

93. Manolio, T. A., Rodriguez, L. L., Brooks, L., Abecasis, G. & GAIN Collaborative Research Group. New models of collaboration in genome-wide association studies: the Genetic Association Information Network. *Nature genetics* **39**, 1045 (2007).

94. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661 (2007).

95. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome research* (2009).

96. Loh, P.-R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature genetics* **47**, 284 (2015).

97. Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M. & Price, A. L. Advantages and pitfalls in the application of mixed-model association methods. *Nature genetics* **46**, 100 (2014).

98. Plotly. *Collaborative data science*. (Plotly Technologies Inc., 2015).

99. Price, A. L., Zaitlen, N. A., Reich, D. & Patterson, N. New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics* **11**, 459 (2010).

100. Kurtzer, G. M., Sochat, V. & Bauer, M. W. Singularity: Scientific containers for mobility of compute. *PloS one* **12**, e0177459 (2017).

101. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

102. Delaneau, O., Marchini, J. & Zagury, J.-F. A linear complexity phasing method for thousands of genomes. *Nature methods* **9**, 179 (2012).

103. Loh, P.-R. *et al.* Reference-based phasing using the Haplotype Reference Consortium panel. *Nature genetics* **48**, 1443 (2016).

104. Tange, O. Gnu parallel-the command-line power tool. *The USENIX Magazine* **36**, 42–47 (2011).

105. Cann, H. M. *et al.* A human genome diversity cell line panel. *science* **296**, 261–262 (2002).

106. Liu, Q. *et al.* Systematic assessment of imputation performance using the 1000 Genomes reference panels. *Briefings in bioinformatics* **16**, 549–562 (2014).

107. Köster, J. & Rahmann, S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* **28**, 2520–2522 (2012).

108. Liu, F. *et al.* Digital quantification of human eye color highlights genetic association of three new loci. *PLoS Genet* **6**, e1000934 (2010).

109. Beleza, S. *et al.* Genetic Architecture of Skin and Eye Color in an African-European Admixed Population. *PLOS Genetics* **9**, e1003372 (2013).

110. 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).

111. Andersen, J. D. *et al.* Genetic analyses of the human eye colours using a novel objective method for eye colour classification. *Forensic Sci Int Genet* **7**, 508–515 (2013).

112. 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 research* **29**, 141–162 (2016).

113. McKinney, W. Data Structures for Statistical Computing in Python. in *Proceedings of the 9th Python in Science Conference* (eds. Walt, S. van der & Millman, J.) 51–56 (2010).

114. Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).

115. Wiredfool *et al. Pillow: 3.1.0*. (Zenodo, 2016). doi:10.5281/ZENODO.44297.

116. van der Walt, S., Colbert, S. C. & Varoquaux, G. The NumPy Array: A Structure for Efficient Numerical Computation. *Comput. Sci. Eng.* **13**, 22–30 (2011).

117. Church, D. M. *et al.* Modernizing reference genome assemblies. *PLoS Biol.* **9**, e1001091 (2011).

118. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).

119. Delaneau, O. & Marchini, J. Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. *Nat Commun* **5**, 1–9 (2014).

120. Durbin, R. Efficient haplotype matching and storage using the positional Burrows-Wheeler transform (PBWT). *Bioinformatics* **30**, 1266–1272 (2014).

121. Zhou, X. & Stephens, M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature Methods* **11**, 407–409 (2014).

122. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).

123. Stouffer, S. A., Suchman, E. A., Devinney, L. C., Star, S. A. & Williams Jr., R. M. *The American soldier: Adjustment during army life. (Studies in social psychology in World War II), Vol. 1*. (Princeton Univ. Press, 1949).

124. Hysi, P. G. *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).

125. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *Journal of machine learning research* **12**, 2825–2830 (2011).

126. Abadi, M. *et al.* TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. *Software available from tensorflow. org* **1**, (2015).

127. Chollet, F. & others. *Keras*. (2015).

128. Nickolls, J., Buck, I., Garland, M. & Skadron, K. Scalable Parallel Programming with CUDA. *Queue* **6**, 40–53 (2008).

129. Tani, T., Ohsumi, J., Mita, K. & Takiguchi, Y. Identification of a novel class of elastase isozyme, human pancreatic elastase III, by cDNA and genomic gene cloning. *J. Biol. Chem.* **263**, 1231–1239 (1988).

130. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

131. Davis, C. A. *et al.* The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.* **46**, D794–D801 (2018).

132. Modzelewska, K., Newman, L. P., Desai, R. & Keely, P. J. Ack1 Mediates Cdc42-dependent Cell Migration and Signaling to p130Cas. *J. Biol. Chem.* **281**, 37527–37535 (2006).

133. Oceguera-Yanez, F. *et al.* Ect2 and MgcRacGAP regulate the activation and function of Cdc42 in mitosis. *J Cell Biol* **168**, 221–232 (2005).

134. Scott, G., Leopardi, S., Printup, S. & Madden, B. C. Filopodia are conduits for melanosome transfer to keratinocytes. *J. Cell. Sci.* **115**, 1441–1451 (2002).

135. Singh, S. K. *et al.* Melanin transfer in human skin cells is mediated by filopodia—a model for homotypic and heterotypic lysosome-related organelle transfer. *The FASEB Journal* **24**, 3756–3769 (2010).

136. Forrest, A. R. *et al.* A promoter-level mammalian expression atlas. *Nature* **507**, 462 (2014).

137. Ou, X., Andres, A., Pivik, R., Cleves, M. A. & Badger, T. M. Brain gray and white matter differences in healthy normal weight and obese children. *Journal of Magnetic Resonance Imaging* **42**, 1205–1213 (2015).

138. Hisa, T. *et al.* Hematopoietic, angiogenic and eye defects in Meis1 mutant animals. *EMBO J.* **23**, 450–459 (2004).

139. Dvir, L. *et al.* Autosomal-recessive early-onset retinitis pigmentosa caused by a mutation in PDE6G, the gene encoding the gamma subunit of rod cGMP phosphodiesterase. *Am. J. Hum. Genet.* **87**, 258–264 (2010).

140. Zhang, Z. & Artemyev, N. O. Determinants for phosphodiesterase-6 inhibition by its γ-subunit. *Biochemistry* **49**, 3862–3867 (2010).

141. Lee, M. K. *et al.* Genome-wide association study of facial morphology reveals novel associations with FREM1 and PARK2. *PLoS ONE* **12**, e0176566 (2017).

142. Do, T. Finding the Right Frames for Your Face. (2014).

143. San Francisco Plastic Surgery Shape | Mabrie Facial Cosmetic. https://www.yourfaceinourhands.com/your-face-perfected/the-science-of-beauty/shape/.

144. Verma, S. S. *et al.* Imputation and quality control steps for combining multiple genome-wide datasets. *Front. Genet.* **5**, (2014).

145. Stults, D. M., Killen, M. W., Pierce, H. H. & Pierce, A. J. Genomic architecture and inheritance of human ribosomal RNA gene clusters. *Genome Res.* **18**, 13–18 (2008).

146. Deelen, P. *et al.* Genotype harmonizer: automatic strand alignment and format conversion for genotype data integration. *BMC Res Notes* **7**, 901 (2014).

147. Maritzen, T., Podufall, J. & Haucke, V. Stonins--specialized adaptors for synaptic vesicle recycling and beyond? *Traffic* **11**, 8–15 (2010).

148. Costagliola, S. *et al.* Tyrosine sulfation is required for agonist recognition by glycoprotein hormone receptors. *EMBO J.* **21**, 504–513 (2002).

149. Liu, G. *et al.* Leydig-cell tumors caused by an activating mutation of the gene encoding the luteinizing hormone receptor. *N. Engl. J. Med.* **341**, 1731–1736 (1999).

150. Gromoll, J., Eiholzer, U., Nieschlag, E. & Simoni, M. Male hypogonadism caused by homozygous deletion of exon 10 of the luteinizing hormone (LH) receptor: differential action of human chorionic gonadotropin and LH. *J. Clin. Endocrinol. Metab.* **85**, 2281–2286 (2000).

151. Limer, K. L. *et al.* Genetic variation in sex hormone genes influences heel ultrasound parameters in middle-aged and elderly men: results from the European Male Aging Study (EMAS). *J. Bone Miner. Res.* **24**, 314–323 (2009).

152. Yarram, S. J. *et al.* Luteinizing hormone receptor knockout (LuRKO) mice and transgenic human chorionic gonadotropin (hCG)-overexpressing mice (hCG alphabeta+) have bone phenotypes. *Endocrinology* **144**, 3555–3564 (2003).

153. Han, S. Y. *et al.* TFIIAalpha/beta-like factor is encoded by a germ cell-specific gene whose expression is up-regulated with other general transcription factors during spermatogenesis in the mouse. *Biol. Reprod.* **64**, 507–517 (2001).

154. Nakano, N. *et al.* C18 ORF1, a Novel Negative Regulator of Transforming Growth Factor-β Signaling. *J Biol Chem* **289**, 12680–12692 (2014).

155. Kim, S. K. Identification of 613 new loci associated with heel bone mineral density and a polygenic risk score for bone mineral density, osteoporosis and fracture. *PLoS ONE* **13**, e0200785 (2018).

156. Liang, X. *et al.* Assessing the genetic correlations between early growth parameters and bone mineral density: A polygenic risk score analysis. *Bone* **116**, 301–306 (2018).

157. Zhang, H., Liu, C.-T. & Wang, X. An Association Test for Multiple Traits Based on the Generalized Kendall's Tau. *J Am Stat Assoc* **105**, 473–481 (2010).

158. Denny, J. C. *et al.* PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics* **26**, 1205–1210 (2010).

159. Weinberg, S. M. *et al.* The 3D Facial Norms Database: Part 1. A Web-Based Craniofacial Anthropometric and Image Repository for the Clinical and Research Community. *Cleft Palate Craniofac. J.* **53**, e185–e197 (2016).

160. Golding, J., Pembrey, M. & Jones, R. ALSPAC--the Avon Longitudinal Study of Parents and Children. I. Study methodology. *Paediatr Perinat Epidemiol* **15**, 74–87 (2001).

161. Hayton, J. C., Allen, D. G. & Scarpello, V. Factor Retention Decisions in Exploratory Factor Analysis: a Tutorial on Parallel Analysis. *Organizational Research Methods* **7**, 191–205 (2004).

162. Karolchik, D. *et al.* The UCSC Table Browser data retrieval tool. *Nucleic Acids Res* **32**, D493–D496 (2004).

163. Hooper, J. E. *et al.* Systems biology of facial development: contributions of ectoderm and mesenchyme. *Developmental Biology* **426**, 97–114 (2017).

164. Watanabe, K., Taskesen, E., Bochoven, A. van & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat Commun* **8**, 1–11 (2017).

165. McLean, C. Y. *et al.* GREAT improves functional interpretation of cis -regulatory regions. *Nat Biotechnol* **28**, 495–501 (2010).

166. De, M. Agricolae: statistical procedures for agricultural research. (2014).

167. Wilke, C. O. & cowplot), Rs. (Copyright for ggplot2 code copied to. *cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'*. (2019).

168. Wickham, H. & Chang, W. ggplot2: Create elegant data visualisations using the grammar of graphics. *R package version* **2**, (2016).

169. Kassambara, A. *ggpubr: 'ggplot2' Based Publication Ready Plots*. (2018).

170. Auguie, B. & Antonov, A. *gridExtra: Miscellaneous Functions for 'Grid' Graphics*. (2017).

171. Wickham, H., Pedersen, T. L. & RStudio. *gtable: Arrange 'Grobs' in Tables*. (2019).

172. Jr, F. E. H. & others, with contributions from C. D. and many. *Hmisc: Harrell Miscellaneous*. (2019).

173. Revelle, W. *psych: Procedures for Psychological, Psychometric, and Personality Research*. (2019).

174. Hlavac, M. *stargazer: Well-Formatted Regression and Summary Statistics Tables*. (2018).

175. Dowle, M. *et al. data.table: Extension of 'data.frame'*. (2019).

176. Brown, G. W. & Mood, A. M. On median tests for linear hypotheses. in (The Regents of the University of California, 1951).

177. Stelzer, G. *et al.* VarElect: the phenotype-based variation prioritizer of the GeneCards Suite. *BMC Genomics* **17**, 444 (2016).

178. STRING: functional protein association networks. https://string-db.org/.

179. Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A. & King, J. Reporting structural equation modeling and confirmatory factor analysis results: A review. *The Journal of educational research* **99**, 323–338 (2006).

180. Wolf, E. J., Harrington, K. M., Clark, S. L. & Miller, M. W. Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. *Educational and psychological measurement* **73**, 913–934 (2013).

181. Noro, B. *et al.* Molecular Dissection of the Architectural Transcription Factor HMGA2. *Biochemistry* **42**, 4569–4577 (2003).

182. Silver, H. K., Kiyasu, W., George, J. & Deamer, W. C. Syndrome of congenital hemihypertrophy, shortness of stature, and elevated urinary gonadotropins. *Pediatrics* **12**, 368–376 (1953).

183. Russell, A. A syndrome of intra-uterine dwarfism recognizable at birth with cranio-facial dysostosis, disproportionately short arms, and other anomalies (5 examples). *Proc. R. Soc. Med.* **47**, 1040–1044 (1954).

184. Kayserili, H. *et al.* ALX4 dysfunction disrupts craniofacial and epidermal development. *Hum Mol Genet* **18**, 4357–4366 (2009).

185. Hall, C. R., Wu, Y., Shaffer, L. G. & Hecht, J. T. Familial case of Potocki-Shaffer syndrome associated with microdeletion of EXT2 and ALX4. *Clin. Genet.* **60**, 356–359 (2001).

186. Beverdam, A., Brouwer, A., Reijnen, M., Korving, J. & Meijlink, F. Severe nasal clefting and abnormal embryonic apoptosis in Alx3/Alx4 double mutant mice. *Development* **128**, 3975–3986 (2001).

187. Qu, S., Tucker, S. C., Zhao, Q., deCrombrugghe, B. & Wisdom, R. Physical and genetic interactions between Alx4 and Cart1. *Development* **126**, 359–369 (1999).

188. Rochus, C. M. *et al.* Revealing the selection history of adaptive loci using genome-wide scans for selection: an example from domestic sheep. *BMC Genomics* **19**, 71 (2018).

189. Lamichhaney, S. *et al.* Evolution of Darwin’s finches and their beaks revealed by genome sequencing. *Nature* **518**, 371–375 (2015).

190. Lamichhaney, S. *et al.* A beak size locus in Darwin’s finches facilitated character displacement during a drought. *Science* **352**, 470–474 (2016).

191. Foppiano, S., Hu, D. & Marcucio, R. S. Signaling by bone morphogenetic proteins directs formation of an ectodermal signaling center that regulates craniofacial development. *Dev. Biol.* **312**, 103–114 (2007).

192. Minoux, M. & Rijli, F. M. Molecular mechanisms of cranial neural crest cell migration and patterning in craniofacial development. *Development* **137**, 2605–2621 (2010).

193. Nishikata, I. *et al.* A novel EVI1 gene family, MEL1, lacking a PR domain (MEL1S) is expressed mainly in t(1;3)(p36;q21)-positive AML and blocks G-CSF-induced myeloid differentiation. *Blood* **102**, 3323–3332 (2003).

194. Mochizuki, N. *et al.* A novel gene, MEL1, mapped to 1p36.3 is highly homologous to the MDS1/EVI1 gene and is transcriptionally activated in t(1;3)(p36;q21)-positive leukemia cells. *Blood* **96**, 3209–3214 (2000).

195. Takahata, M. *et al.* SKI and MEL1 cooperate to inhibit transforming growth factor-beta signal in gastric cancer cells. *J. Biol. Chem.* **284**, 3334–3344 (2009).

196. Ito, Y. *et al.* Conditional inactivation of Tgfbr2 in cranial neural crest causes cleft palate and calvaria defects. *Development* **130**, 5269–5280 (2003).

197. Kaartinen, V. *et al.* Abnormal lung development and cleft palate in mice lacking TGF-beta 3 indicates defects of epithelial-mesenchymal interaction. *Nat. Genet.* **11**, 415–421 (1995).

198. Xu, X. *et al.* Cell autonomous requirement for Tgfbr2 in the disappearance of medial edge epithelium during palatal fusion. *Dev. Biol.* **297**, 238–248 (2006).

199. Sanford, L. P. *et al.* TGFbeta2 knockout mice have multiple developmental defects that are non-overlapping with other TGFbeta knockout phenotypes. *Development* **124**, 2659–2670 (1997).

200. Bjork, B. C., Turbe-Doan, A., Prysak, M., Herron, B. J. & Beier, D. R. Prdm16 is required for normal palatogenesis in mice. *Hum Mol Genet* **19**, 774–789 (2010).

201. Stamataki, D., Ulloa, F., Tsoni, S. V., Mynett, A. & Briscoe, J. A gradient of Gli activity mediates graded Sonic Hedgehog signaling in the neural tube. *Genes Dev.* **19**, 626–641 (2005).

202. Koyabu, Y., Nakata, K., Mizugishi, K., Aruga, J. & Mikoshiba, K. Physical and Functional Interactions between Zic and Gli Proteins. *J. Biol. Chem.* **276**, 6889–6892 (2001).

203. Wang, B., Fallon, J. F. & Beachy, P. A. Hedgehog-Regulated Processing of Gli3 Produces an Anterior/Posterior Repressor Gradient in the Developing Vertebrate Limb. *Cell* **100**, 423–434 (2000).

204. Hui, C. & Joyner, A. L. A mouse model of Greig cephalo–polysyndactyly syndrome: the extra–toes J mutation contains an intragenic deletion of the Gli3 gene. *Nature Genetics* **3**, 241 (1993).

205. Taylor, M. D. *et al.* Mutations in SUFU predispose to medulloblastoma. *Nat. Genet.* **31**, 306–310 (2002).

206. Kogerman, P. *et al.* Mammalian Suppressor-of-Fused modulates nuclear–cytoplasmic shuttling of GLI-1. *Nature Cell Biology* **1**, 312 (1999).

207. Raducu, M. *et al.* SCF (Fbxl17) ubiquitylation of Sufu regulates Hedgehog signaling and medulloblastoma development. *EMBO J.* **35**, 1400–1416 (2016).

208. Merchant, M. *et al.* Suppressor of fused regulates Gli activity through a dual binding mechanism. *Mol. Cell. Biol.* **24**, 8627–8641 (2004).

209. Cherry, A. L. *et al.* Structural basis of SUFU-GLI interaction in human Hedgehog signalling regulation. *Acta Crystallogr. D Biol. Crystallogr.* **69**, 2563–2579 (2013).

210. Stegman, M. A. *et al.* Identification of a Tetrameric Hedgehog Signaling Complex. *J. Biol. Chem.* **275**, 21809–21812 (2000).

211. Methot, N. & Basler, K. Suppressor of fused opposes hedgehog signal transduction by impeding nuclear accumulation of the activator form of Cubitus interruptus. *Development* **127**, 4001–4010 (2000).

212. Monnier, V., Dussillol, F., Alves, G., Lamour-Isnard, C. & Plessis, A. Suppressor of fused links Fused and Cubitus interruptus on the Hedgehog signalling pathway. *Current Biology* **8**, 583-S2 (1998).

213. Kise, Y., Morinaka, A., Teglund, S. & Miki, H. Sufu recruits GSK3β for efficient processing of Gli3. *Biochemical and Biophysical Research Communications* **387**, 569–574 (2009).

214. Humke, E. W., Dorn, K. V., Milenkovic, L., Scott, M. P. & Rohatgi, R. The output of Hedgehog signaling is controlled by the dynamic association between Suppressor of Fused and the Gli proteins. *Genes Dev.* **24**, 670–682 (2010).

215. Tian, L. *et al.* Unveiling transcription factor regulation and differential co-expression genes in Duchenne muscular dystrophy. *Diagnostic Pathology* **9**, 210 (2014).

216. Peter, A. K. & Crosbie, R. H. Hypertrophic response of Duchenne and limb-girdle muscular dystrophies is associated with activation of Akt pathway. *Experimental Cell Research* **312**, 2580–2591 (2006).

217. Imuta, Y., Nishioka, N., Kiyonari, H. & Sasaki, H. Short limbs, cleft palate, and delayed formation of flat proliferative chondrocytes in mice with targeted disruption of a putative protein kinase gene, Pkdcc (AW548124). *Dev. Dyn.* **238**, 210–222 (2009).

218. Kinoshita, M., Era, T., Jakt, L. M. & Nishikawa, S.-I. The novel protein kinase Vlk is essential for stromal function of mesenchymal cells. *Development* **136**, 2069–2079 (2009).

219. Estrada, K. *et al.* Genome-wide meta-analysis identifies 56 bone mineral density loci and reveals 14 loci associated with risk of fracture. *Nat Genet* **44**, 491–501 (2012).

220. Kraus, P. & Lufkin, T. Dlx homeobox gene control of mammalian limb and craniofacial development. *American journal of medical genetics Part A* **140**, 1366–1374 (2006).

221. Alappat, S., Zhang, Z. Y. & Chen, Y. P. Msx homeobox gene family and craniofacial development. *Cell research* **13**, 429 (2003).

222. Kyrylkova, K., Iwaniec, U. T., Philbrick, K. A. & Leid, M. BCL11B regulates sutural patency in the mouse craniofacial skeleton. *Developmental biology* **415**, 251–260 (2016).

223. Lessel, D. *et al.* BCL11B mutations in patients affected by a neurodevelopmental disorder with reduced type 2 innate lymphoid cells. *Brain* **141**, 2299–2311 (2018).

224. Mammoto, A. *et al.* A mechanosensitive transcriptional mechanism that controls angiogenesis. *Nature* **457**, 1103–1108 (2009).

225. Haenggi, T. & Fritschy, J.-M. Role of dystrophin and utrophin for assembly and function of the dystrophin glycoprotein complex in non-muscle tissue. *Cell. Mol. Life Sci.* **63**, 1614–1631 (2006).

226. Ahn, A. H. *et al.* Cloning of human basic A1, a distinct 59-kDa dystrophin-associated protein encoded on chromosome 8q23-24. *Proc. Natl. Acad. Sci. U.S.A.* **91**, 4446–4450 (1994).

227. Erriquez, D., Perini, G. & Ferlini, A. Non-Coding RNAs in Muscle Dystrophies. *International Journal of Molecular Sciences* **14**, 19681–19704 (2013).

228. 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**, 170–80 (2011).

229. Minaee, S. & Abdolrashidi, A. Iris-GAN: Learning to Generate Realistic Iris Images Using Convolutional GAN. *arXiv:1812.04822 [cs]* (2018).

230. Igbaria, M., Zinatelli, N., Cragg, P. & Cavaye, A. L. Personal computing acceptance factors in small firms: A structural equation model. *MIS quarterly* **21**, (1997).

231. Ko, D.-W. & Stewart, W. P. A structural equation model of residents' attitudes for tourism development. *Tourism Management* **23**, 521–530 (2002).

232. Hellier, P. K., Geursen, G. M., Carr, R. A. & Rickard, J. A. Customer repurchase intention: A general structural equation model. *European journal of marketing* **37**, 1762–1800 (2003).

233. Kaakinen, M. *et al.* Life-Course Analysis of a Fat Mass and Obesity-Associated (FTO) Gene Variant and Body Mass Index in the Northern Finland Birth Cohort 1966 Using Structural Equation Modeling. *Am J Epidemiol* **172**, 653–665 (2010).

234. Kim, J.-Y., Namkung, J.-H., Lee, S.-M. & Park, T.-S. Application of Structural Equation Models to Genome-wide Association Analysis. *Genomics & Informatics* **8**, 150–158 (2010).

235. Turley, P. *et al.* Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nat Genet* **50**, 229–237 (2018).

PUBLICATIONS

Eller R.J., Janga S.C., & Walsh S. “Odyssey: a semi-automated pipeline for phasing, imputation, and analysis of genome-wide genetic data.” *BMC bioinformatics* 20, 364 (2019).

Peng F., Zhu G., Hysi P.G., Eller R.J., Chen Y., Li Y., Hamer M.A., Zeng C., Hopkins R.L., Jacobus C.L., and Wallace P.L. “Genome-wide association studies identify multiple genetic loci influencing eyebrow color variation in Europeans.” *Journal of Investigative Dermatology* (2019).

White J.D., Karlijne I., Naqvi S., Eller R.J, Roosenboom J., Lee M.K., Li J., Mohammed J., Richmond S., Quillen E.E., Norton H.L., Feingold E., Swigut T., Marazita M.L., Peeters H., Hens G., Shaffer J.R., Wysocka J., Walsh S., Weinberg S.M., Shriver M.D., and Claes P. “Insights into the genetic architecture of the human face.”
[Submitted for Review]