PREDICTIVE MODELING FOR COMPLEX TRAITS: NORMAL HUMAN PIGMENTATION VARIATION

By

Robert Keams Valenzuela

_____

A Dissertation Submitted to the Faculty of the

GRADUATE INTERDISCIPLINARY PROGRAM IN GENETICS

In Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

In the Graduate College

THE UNIVERSITY OF ARIZONA

2011

THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Dissertation Committee, we certify that we have read the dissertation prepared by Robert Keams Valenzuela

Entitled Predictive Modelling for Complex Traits: Normal Human Pigmentation Variation

And recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy

_____ Date: 4/15/2011
Dr. Murray H. Brilliant

_____ Date: 4/15/2011
Dr. J. Bruce Walsh

_____ Date: 4/15/2011
Dr. Leslie J. Baier

_____ Date: 4/15/2011
Dr. Tsu-Shuen Tsao

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

_____ Date: 4/15/2011
Dissertation Director: Dr. Murray H. Brilliant

STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of requirements for an advanced degree at the University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that accurate acknowledgment of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the head of the major department or the Dean of the Graduate College when in his or her judgment the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.

SIGNED: Robert Keams Valenzuela

Acknowledgments

I would like to acknowledge the people whose efforts have contributed either directly or indirectly into enabling me to develop on a professional and personal level to acquire the skills required to complete my dissertation. In a sense, the work contained in this dissertation represents a mosaic of their efforts.

First, with the utmost respect and gratitude, I would like to acknowledge Dr. Murray H. Brilliant for his guidance and instruction throughout my years as his graduate student. At a broader level, Murray's selection and arrangement of core classes of the Genetics Graduate Interdisciplinary Program, during the time he served as the director, has broadened my view of genetics and related fields. On a more personal level, Murray's patience, firmness, persistence, and professionalism have not only contributed to my academic knowledge, but to my personal growth and maturity. Murray's understanding of my situation of being a single parent (of a once infant and now toddler) coupled with his firmness and persistence that I continue to advance in my graduate progress has made completion of this work possible, I can never thank Murray enough for all his help. Murray has been an excellent mentor.

I would also like to thank my committee members Dr. Leslie Baier, Dr. Bruce Walsh, Dr. Tsu-Shuen Tsao, and Dr. Michael Hammer. In particular, I would like to thank Dr. Leslie Baier for allowing me the opportunity to work as a pre-doctoral fellow in her laboratory and allowing me to gain additional experience under the supervision of Dr. Yunhua-Li Muller, whom I also thank. I thank Dr. Bruce Walsh and Dr. John Meaney for their statistical guidance and feedback. I also thank Dr. Robert P. Erickson (Bob) and Dr. Michael Hammer for allowing me to rotate in their laboratories and gain additional experience. I thank Bob for reviewing my manuscripts and for his encouragement and support during critical periods of my graduate career.

I am thankful to Dr. Maria Teresa Velez and Donna Treloar for their efforts in attaining funding (Sloan Foundation), encouragement, direction, and guidance. I also thank the Sloan foundation for funding throughout my graduate career; SACNAS for the Genome Scholarship (2007-2008); the IGERT program (directed by Dr. Nachman) for funding (2008-2009) and educational support; and the Marshfield Clinic Research Foundation (2009-2011) for funding, educational, and professional development.

I thank my high school cross country and track coach, Kyle Sawyer, for instilling hard work, discipline, and positive motivation in me. I thank the late Dr. Joaquin Bustoz for allowing me the educational opportunity to be a part of the Math Science and Honors Program as an undergraduate; Dr. Phil Service for his mentorship while I was a master's student; Meff Thompson for her friendship, support, and instruction of cell culture techniques; Elaine Forbes for her instruction on cloning; and Adam Hoying for his sharing of knowledge and lab equipment.

I also thank Leanne Schwartz for her help during critical times; Athena Ganchorre for her friendship and support; Ted Kandell for his friendship, support, and "interesting insights." Finally, I thank Lori Taylor and Dr. Gio Bosco and my fellow Genetics GIDP colleagues.

Dedication

I dedicate this to my dad, Robert K. Chiago, for his support and to my son, Brent D. Valenzuela, for his endurance of this journey with me.

TABLE OF CONTENTS

LIST OF FIGURES

ABSTRACT

Melanin pigmentation is a complex trait governed by many genes. Variation in melanin pigmentation within, and between, populations makes it an important trait for assisting in physical identification of an individual in forensic investigations. Utilizing a training sample (n=789) comprised of various ethnicities and SNPs (75) in 24 genes previously implicated in human or animal pigmentation studies, I determined three-SNP multiple linear regression models that accounted for large proportions of pigmentation variation in skin (45.7%), eye color (76.4%), and hair [eumelanin-to-pheomelanin (43.2%) and total melanin (76.3%)], independent of ethnic origin. Rather than implementing stepwise regression, to ascertain the three-SNP predictive models, I devised an algorithm that is likely more robust than stepwise regression. The algorithm consisted of two steps: the first step reduced the pool of 75 SNPs to a pool of 40 by selection of SNPs that were significant ($p<0.05$) by one-way ANOVA; the second step enabled selection of SNPs for model incorporation based on their frequency in the best-fitted models of all possible combinations of three-SNP models (i.e., 40 choose 3).

Prediction models were validated utilizing an independent cohort (n=242, test sample) that was very similar in ethnic composition to the training sample. Relative shrinkage was moderate for skin reflectance (23.4%), eye color (19.4%), and eumelanin-to-pheomelanin (37.3%) of hair, and largest for total melanin (67%) of hair. Additionally, we refined our model-building algorithm, enabling visual comparison of the frequency and co-linearity due to linkage or co-inheritance of SNPs of the best-fitted

models. Application of our algorithm to the test sample yielded the same or similar models as the training sample. Two of the three SNPs composing the models were the same, with some variability in the third SNP of the model.

I. INTRODUCTION

Explanation of dissertation format

In compliance with the dissertation format manual of the University of Arizona, the research presented in this dissertation has been or will be submitted for publication, and is therefore in Appendix format. Following the literature review is the conclusion section that pertains to the research presented in the appendices.

1.1 Literature review

1.1.1 Melanin Pigmentation

Melanin pigmentation is responsible for much of the variation observed in skin, hair, and eye color within, and between, human populations. Other factors that have a lesser influence on skin color have been attributed to hemoglobin and beta-carotene (Edwards & Duntley 1939). Skin pigmentation in human populations has been shown to correlate with latitude (Figure 1.1) or average incidence of ultraviolet radiation (UVR) (Relethford 1997; Jablonski & Chaplin 2000), suggesting that pigmentation is an adaptive trait. At lower latitudes (i.e., close to the equator), where there is a relatively high average incidence of UVR, indigenous populations produce more pigment. Whereas, at higher latitudes (i.e., towards the poles), where there is a relatively low average incidence of UVR, indigenous populations produce less pigmentation.

The relationship of pigmentation and sun exposure is correlated with secondary biological effects of the skin that directly affect fitness. At lower latitudes, skin pigmentation protects against sunburn and the photolysis of folate resulting from UVR exposure (Branda & Eaton 1978). Folate is important in the development of the neural tube of fetuses, while the lack of folate results in neural tube defects of developing fetuses (reviewed in (Jablonski 1999). At higher latitudes, less skin pigmentation (lighter skin) enables UVR to penetrate the skin and catalyze the conversion of pro-vitamin D into its active form, vitamin D_3. Vitamin D_3 can also be obtained through the diet; some of the more important functions affected by vitamin D_3 include absorption of calcium into the bones, fertility, and immunity (reviewed in (Barrett & McElduff 2010)). Therefore, in regions with decreased sun exposure, lighter skin is important for vitamin D_3 production for the aforementioned reasons.

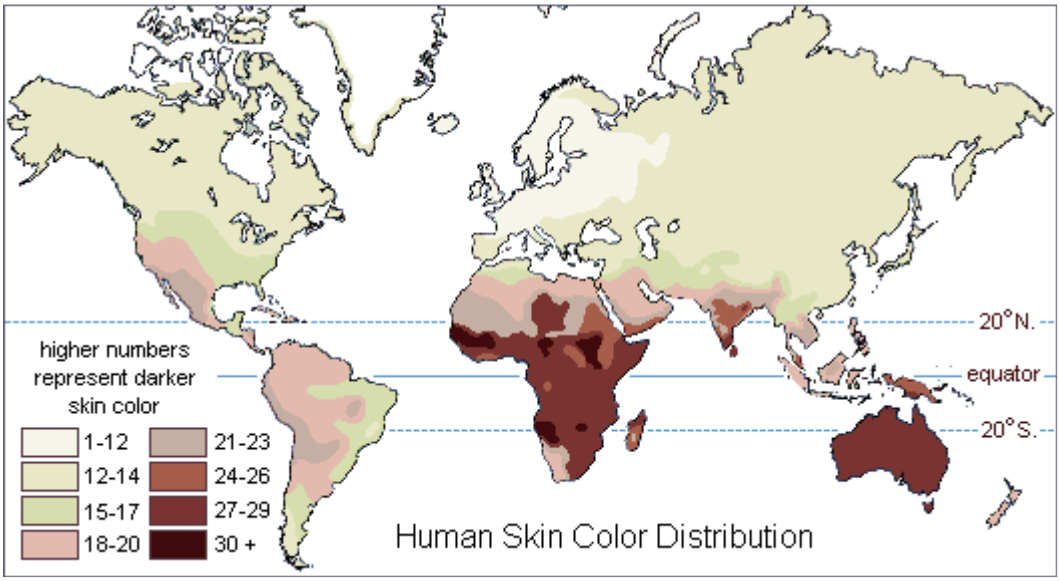

Figure 1.1 – Global map of indigenous skin coloration. Figure reproduced from O'Neil (2007).

Pair-wise comparisons of Whites and Blacks (Staricco & Pinkus 1957), various Asian groups (South Asian (Szabo 1959), East Asian (Srisopark 1976)) with Whites, and Australian Aborigines with Whites (Mitchell 1963) suggest that these populations have approximately the same number of the melanin producing cells, or melanocytes, per corresponding unit area of skin. More recent studies only partially support these earlier studies. Alaluf et al. (2003) measured melanocyte numbers across various ethnic groups (African, Indian, Mexican, Chinese, and Europeans) and found they were the same, with the exception of Europeans, where melanocyte number was higher per unit area of skin measured (Alaluf et al. 2003). This is not consistent with the approximately equal number of melanocytes that Staricco & Pinkus (1957) observed between Whites and Blacks, hence additional studies are needed to determine this relationship.

Early humans likely had darkly pigmented skin (Jablonski & Chaplin 2000). However, as ancestral groups migrated to less intense UV regions, lighter skin evolved. Both East Asian and White populations exhibit decreased amounts of pigmentation compared to sub-Saharan African populations. Specifically, the decrease in pigmentation production has been attributed to a decrease in function (Fuller, Spaulding, & Smith 2001) of the key sub-cellular enzyme, tyrosinase, that is responsible for the production of melanin. Genetic evidence suggests that skin pigmentation in East Asians and Whites are due to different genetic variants. In other words, a decrease in constitutive pigmentation of these two groups is due to convergent evolution (Yuasa et al. 2007; Norton et al. 2007; Edwards et al. 2010).

Pigmentation traits are continuously distributed within and between populations, as opposed to being discretely distributed as is often the case for simple Mendelian traits, suggesting that pigmentation is a polygenic/quantitative/complex trait.

1.1.2 Loss-of-function mutations

As with many traits, genes initially associated with pigmentation were identified by comparing and contrasting genetic variants of wildtype and mutant phenotypes in mice, humans, and zebrafish. Genetic mutations resulting in severe hypopigmentation (albinism) are classified into various forms. The four forms of albinism that result from genetic mutations whose products are involved in melanin synthesis result in oculocutaneous albinism (OCA) (reviewed in (Summers 2009)). OCA results in minimal to no melanin pigment in the hair, skin, and eyes. OCA1, or tyrosinanse negative albinism, results from mutations in the tyrosinase gene (*TYR*). OCA2 results from mutations in the *OCA2* gene, also called the P gene. OCA3 results from mutations in the tyrosinase-related protein 1 gene (*TYRP1*). OCA4 results from mutations in the gene that encodes the solute-carrier transport protein, *SLC45A2*, also called *MATP* (membrane-associated transport protein), or *AIM1* (antigen in melanoma 1), or *uw* (underwhite). The proteins associated with these forms of albinism are discussed below.

1.1.3 Melanin

Melanin is a biopolymer produced in specialized organelles, or lysome-related organelles (reviewed in (Raposo & Marks 2007)), called melanosomes within the

melanocyte according to the Raper-Mason-Prota scheme (reviewed in (Ito & Wakamatsu 2008)). There are two types of melanin that the body can produce: eumelanin (brown/black) and pheomelanin (yellow/red). The absolute quantity of melanin and the relative amounts of eumelanin and pheomelanin significantly contribute to the macroscopic variations of skin, hair, and eye color observed within, and between, populations.

1.1.3.1 Biosynthesis of melanin: eumelanin and pheomelanin

The starting molecule for the chemical reaction of either eumelanin or pheomelanin is tyrosine. The enzyme tyrosinase (TYR) converts tyrosine into dopaquinone (Figure 1.2). In the presence of cysteine, dopaquinone is converted into pheomelanin (reviewed in (Prota 1980)). Two other enzymes that display high homology to TYR and are involved in eumelanin synthesis are tyrosinase-related protein 1 (TYRP1) and tyrosinase-related protein 2 (TYRP2) (reviewed in (Olivares & Solano 2009)). Kinetic studies indicate that the synthesis of pheomelanin occurs much faster than the synthesis of eumelanin. Biochemical studies by Agrup et al. (1982) led them to propose the "casing" model of mixed melanogenesis (Agrup et al. 1982). The casing model posits that the chemical composition of a melanin granule is such that its core is composed of pheomelanin and its shell is composed of eumelanin. The casing model has recently been corroborated in a couple of studies (Bush et al. 2006; Peles et al. 2009). Using photoemission electron microscopy, Peles et al. (2009) showed that the outer core

of iris melanin granules is composed of eumelanin and the inner core is composed of pheomelanin (Figure 1.3).

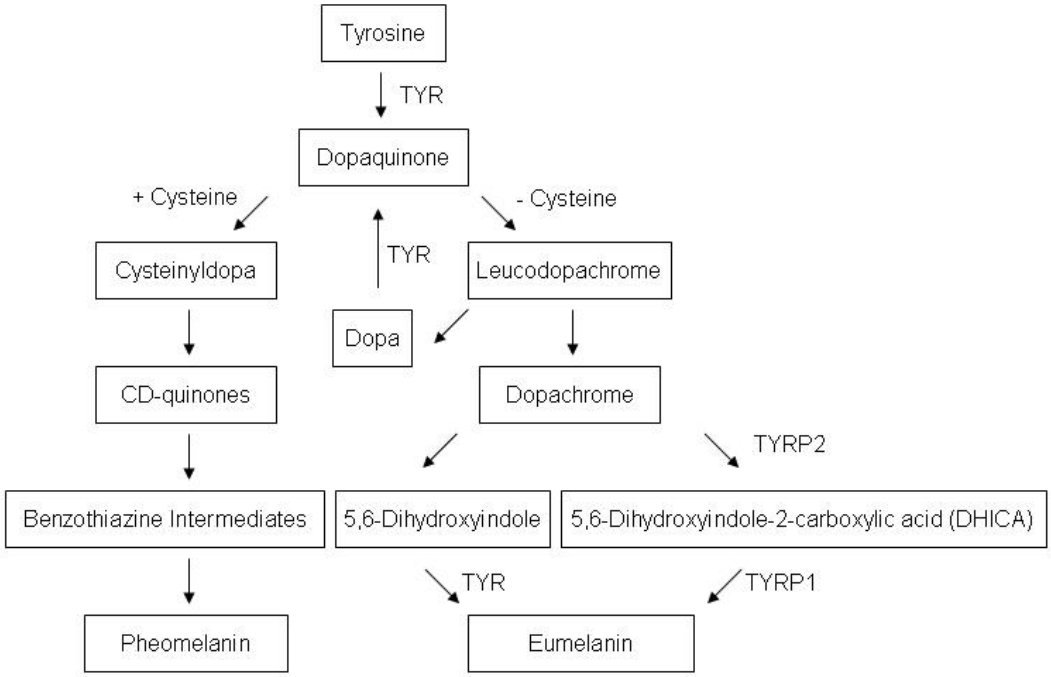

Figure 1.2 – Raper-Mason-Prota scheme (adapted from Wakamatsu and Ito (2002)).

Tyrosinase (TYR) is an endomembrane protein of the melanosome and is the key rate-limiting enzyme within the melanosome (Seiji et al. 1963) that is directly involved in the synthesis of melanin (melanogenesis) by converting the amino acid tyrosine ultimately into melanin. Studies suggest that TYR is trafficked to stage II melanosomes (Costin et al. 2003) from the trans-golgi-network. At least initially, TYR localizes to the endomembrane of the melanosome and has been reported to be associated with tyrosinase-related proteins 1 and 2 in a multimeric complex (Orlow et al. 1994). A more

recent study (Kobayashi & Hearing 2007), using murine cell lines, support that TYR is complexed with TYRP1, but not with TYRP2. Some studies have found that the amounts of TYR mRNA and protein do not differ between skin melanocytes of Blacks and Whites (Iozumi et al. 1993; Alaluf et al. 2003); another study found that the amounts of TYR did not vary between various ethnic groups (African, Indian, Mexican, Chinese, and European). In contrast, a more recent study (Cook et al. 2009) found that TYR abundance varied between the light- and dark-skin associated variants of SLC45A2. Although not explicitly stated, presumably, the subjects in the study by Alaluf et al. (2003) included both SLC45A2 variants. Therefore, additional studies that quantify and control additional genetic variants known to affect skin pigmentation are needed to clarify the relationship between TYR abundance and amount of pigmentation.

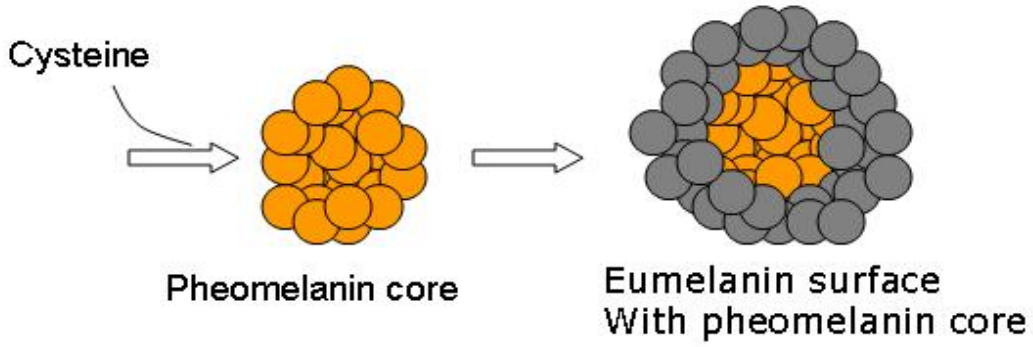

Figure 1.3 – Casing model of melanin (adapted from Simon et al. (2010)).

Tyrosinase-related protein 1 (TYRP1) is involved in melanin synthesis (Boissy et al. 1996), and has recently been shown to localize to late stage melanosomes by an

endosomal intermediate (Truschel et al. 2009). Interestingly, defects in TYRP1 (OCA3) result in a form of hypopigmentation that mostly affects eumelanin (Boissy et al. 1996). Another closely related gene, tyrosinase-related protein 2 (TYRP2), is involved in eumelanin synthesis (Tsukamoto et al. 1992), but has not been associated with any form of albinism.

1.1.4 Melanosome

Melanosomes are produced in the melanin producing cell, melanocytes (Seiji, Fitzpatrick, & Birbeck 1961). All melanosomes that contain melanin are composed of varying ratios of eumelanin and pheomelanin. Melanosomes that have a high quantity of eumelanin are termed eumelanosomes and tend to be ellipsoidal in structure, while melanosomes that contain mostly pheomelanin are termed pheomelanosomes and tend to be spherical in structure (Inazu & Mishima 1993). Both eumelanosomes and pheomelanosomes are derived from endosomes of the golgi apparatus. The development of eumelanosomes, but not pheomelanosomes, has been studied in detail and is characterized by four stages (stages I-IV) of development (Figure 1.4). Stage I melanosomes, or premelanosomes, lack any pigmentation. Stage II melanosomes are characterized by microfibrils (Mα subunits of PMEL17 protein) (reviewed in Raposo 2007)). Stage III melanosomes are characterized by deposition of melanin on the microfibrils. Stage IV melanosomes are fully pigmented and are transferred (in skin and hair) from the dendritic processes of the melanocyte into keratinocytes by a process described in the filopodia-phagocytosis model (Singh et al. 2010). In the keratinocytes,

the melanosomes, or melanin granules, cover the nucleus (supranuclear)(Gates & Zimmermann 1953), shielding it from the mutagenic effects of UVR exposure from the sun. Within the keratinocyte the melanosomes tends to be degraded as the keratinocyte migrates from the stratum basal to the stratum corneum (reviewed in (Borovanský & Elleder 2003)).

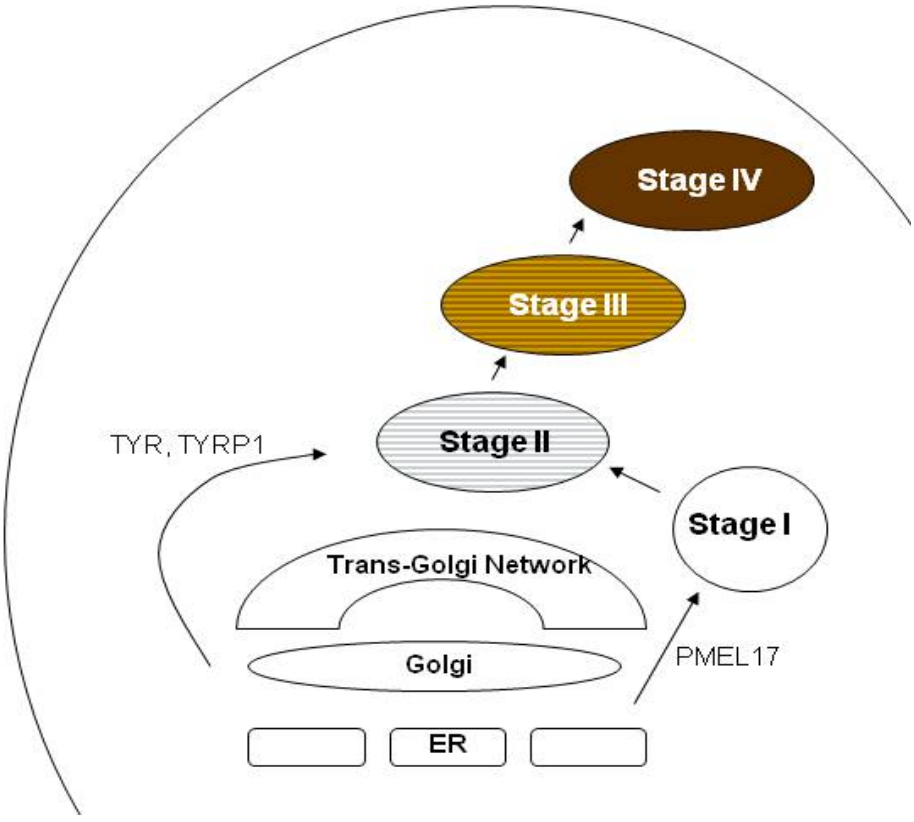

Figure 1.4 – Stages of melanosome development (adapted from Costin et al. (2003)).

Many studies have shown that the putative transmembrane proteins OCA2, SLC45A2 (OCA4), and SLC24A5 (solute carrier transport protein, family 24, member 5;

also called NCKX5) are essential for melanin production and are likely involved in regulating ion transport and pH within the melanosome. Variation within these genes has been associated with variation in pigmentation. SLC45A2 and OCA2 have been localized to the melanosome surface (Gardner et al. 1992; Newton et al. 2001), while studies have been conflicting on the locality of SLC24A5 [cf. (Chi et al. 2006; Ginger et al. 2008)].

The OCA2 gene codes for a putative 12 trans-membrane protein (Gardner et al., 1992). The OCA2 protein has homology with anion transporters and is thought to be involved in influencing the pH of the melanosome (Puri, Gardner, & Brilliant 2000; Chen, Manga, & Orlow 2002; Cheli et al. 2009), and either directly or indirectly involved in the trafficking of internal melanosomal proteins, tyrosinase (TYR) and tyrinosinase-related protein 1 (TYRP1) (Manga et al. 2001). The particular stage of melanosomal development affected by OCA2 is not known, but OCA2 mutant melanocytes do not produce stage III and IV melanosomes (Moyer 1966). Some studies have localized OCA2 to the surface of the melanosome (Rosemblat et al. 1994), while a more recent study did not (Chi et al. 2006). Therefore, additional studies are needed to determine the melanosomal development stage(s) where OCA2 affects pH.

SLC45A2 (OCA4, formerly known as MATP and AIM1) is a putative melanosomal membrane transport protein that is predicted to have 12 trans-membrane regions (Newton et al. 2001) localized to the melanosome (Chi et al. 2006). SLC45A2 has regions that are similar to sucrose symporters in plants, and because of this it has been hypothesized to regulate osmosis by transporting sugar across the melanosome

membrane (Newton et al. 2001). Within a mouse model, Costin et al. (2003) reported that SLC45A2 either directly or indirectly effects the trafficking of the melanosomal enzymes tyrosinase, tyrosinase-related protein 1, and tyrosinase-related protein 2 (Costin et al. 2003) similar to the effect of OCA2 mutation. In humans, Cook et al. (2009) found that there was a greater amount of tyrosinase (TYR) associated with the dark-skin allele (374L) (Cook et al. 2009) in melanocytes. Interestingly, they also found lower mRNA expression levels of samples homozygous for 374L.

Another important pigmentation gene, *SLC24A5* (formerly known as *NCKX5*), was found to cause the *golden* phenotype in zebrafish (Lamason et al. 2005). SLC24A5 was predicted to be a cation exchanger that transports Ca^{2+}/K^{+}, in exchange for Na^{+} (Lamason et al. 2005), and more recently this function has been confirmed (Ginger et al. 2008). Lamason et al. (2005) found that SLC24A5 was located on melanosomes or their precursors and hypothesized that it might function to accumulate Ca^{2+} into the melanosome. Chi et al. (2006) isolated SL24A5 in melanosomal fractions and proposed that it functioned on the surface of melanosomes. Whereas, Ginger et al. (2008) found SLC24A5 to be associated with the trans-golgi network. They also found that the allele associated with darker skin, 111A, had a higher ion exchange activity compared to the allele associated with lighter skin, 111T. They hypothesized that SLC24A5 functions in regulating Ca^{2+} concentrations in endosomes, and that this affects delivery of melanosomal proteins (such as PMEL17), and hence, melanosome maturation. SLC24A5 may explain an earlier study demonstrating high Ca^{2+} concentrations in melanosomes (Dräger 1985). In support of SLC24A5's function, the product of *TPCN2*

(two-pore segment channel 2) is another protein that studies have indicated to be involved in calcium transport in lysosome-related organelles (e.g., melanosomes) (Calcraft et al. 2009) and has also been associated with hair color (Sulem et al. 2008). In consideration of these observations, Ito and Wakamatsu (2010) have suggested that TPCN2 and SLC24A5 may act together in coordinating Ca^{2+} levels in melanosomes (Ito & Wakamatsu 2011). Hence, more studies are needed to determine the location and function of SLC24A5.

1.1.5 Melanocyte

Melanocytes of the skin (epidermal), iris (stromal), and hair follicle (follicular), are the cells that produce melanin that is externally visible. They are derived from the neural crest cells of the ectoderm layer, in contrast to other melanin producing cells (e.g., RPE) which are derived from the neural ectoderm. Epidermal melanocytes are located between the dermal and epidermal layers in the stratum basal layer of the epidermis. Morphologically, skin melanocytes have dendritic processes that are within the vicinity of approximately 36 keratinocytes (Figure 1.5). Together, the melanocyte and its associated keratinocytes are called the epidermal-melanin unit (Fitzpatrick & Breathnach 1963). As mentioned earlier, epidermal melanin is produced constitutively and transferred from the melanocyte to the keratinocytes comprising the epidermal-melanin unit by a process described in the filipodia-phagocytosis model (Singh et al. 2010).

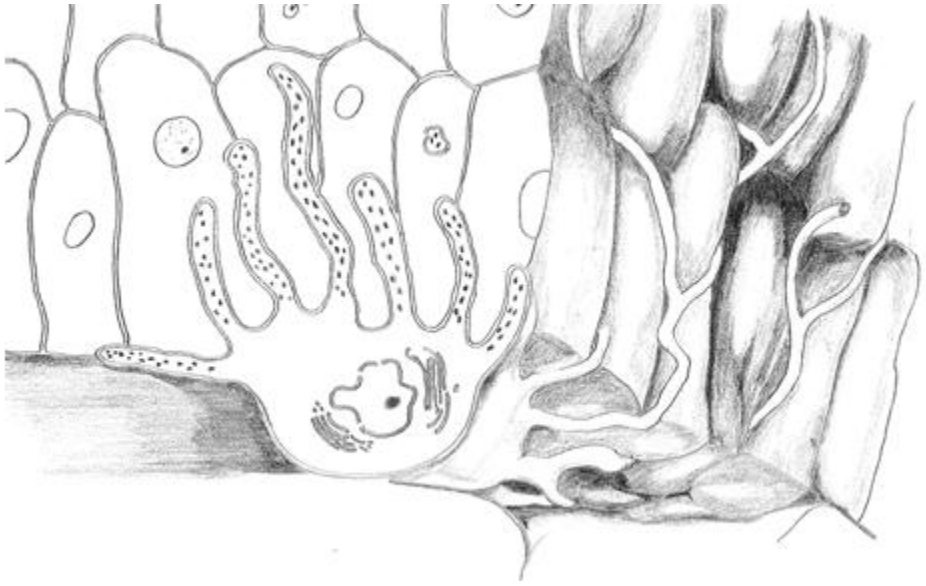

Figure 1.5 – Epidermal melanocyte and keratinocytes (adapted from Fitzpatrick et al. 1967).

Hair melanocytes, or follicular melanocytes, largely responsible for hair color, are located in the bulb of the hair shaft (reviewed in (Tobin 2008)). Similar to epidermal melanocytes, bulb melanocytes transfer their melanin to keratinocytes that get incorporated into growing strands of hair during the anagen phase of the hair cycle (phases III-VI) (Slominski et al. 2005).

Uveal melanocytes, or more specifically stromal/iridal melanocytes are located in the stroma of the eye. The melanin they produce is largely responsible for iris color (reviewed in (Sturm & Larsson 2009)). In contrast to epidermal and follicular melanoctyes, the melanin they produce is normally not transferred to surrounding cells, rather it remains within the melanocyte. Moreover, stromal melanocytes are thought to usually produce melanin up to 6 years of age based on studies of eye color changes after

which melanin production essentially stops and eye color stabilizes. However, approximately 10-15% of individuals continue to change eye color, from lighter colors to darker colors, into adulthood (Bito et al. 1997).

A key protein involved in the regulation of melanin production is located in the membrane of epidermal and follicular melanocytes is the melanocortin 1 receptor (MC1R), a seven transmembrane G-protein coupled receptor (GPCR) (reviewed in (Flanagan et al. 2000)). Ligands of MC1R include the paracrine hormones alpha-melanocyte stimulating hormone (α-MSH) and adrenocorticotropic hormone, both are produced in the keratinocytes associated with the melanocyte. They are derived from the precursor protein, proopiomelanocortin (POMC) (reviewed in (Dores 2009)). The binding of α-MSH to MC1R causes a cAMP signal cascade resulting in the activation of microphthalmia-associated transcription factor (MITF), a key regulator of at least 25 genes in melanocytes that are involved in the production of brown/black melanin (eumelanin) (reviewed in (Vachtenheim & Borovanský 2010)). The protein antagonist to signaling through MC1R is agouti-signaling protein (ASIP). The antagonistic action of ASIP results in a relative decrease in the production of eumelanin to melanin pheomelanin (reviewed in (Voisey et al. 2006)). Hence, MC1R acts as a switch between the two types of melanin for skin and hair melanoctyes. In contrast, adult stromal melanocytes have been reported not to express MC1R nor respond to its ligand, α-MSH (Li et al. 2006), *in vitro*. Lighter colored eyes have been shown to contain slightly greater amounts of pheomelanin than darker colored eyes (Wakamatsu et al. 2008),

therefore, there may be a control mechanism, but so far there are no data implicating a role for MC1R or α-MSH in eye color.

1.1.6 Genetic polymorphisms associated with normal human pigmentation

In contrast to the rest of the genome, a higher proportion of variation exists within the *MC1R* gene (nonsynonymous mutations) of Northern European populations compared with African populations (reviewed in (Makova & Norton 2005)). A study by Harding et al. (2000) provides support that MC1R is under functional constraint in African populations. However, within Northern Europeans, Harding et al (2000) and Rana et al (1999) differed in their conclusions about the amount of variation observed in *MC1R*. Harding et al. (Harding et al. 2000), using a sample size of 224 individuals suggested that the genetic diversity was due to neutrality, while Rana et al. (Rana et al. 1999), using a sample size of 121 individuals, suggested that the genetic diversity was due to positive selection. A larger study and enhanced statistical analysis may clarify whether the diversity is due to neutral or positive selection.

A couple of polymorphisms, rs6058017 and rs2424984, within *ASIP* have been associated with pigmentation variation. The rs6058017 polymorphism located within the 3' un-translated region (UTR) of the ASIP has been associated with normal human pigmentation variation of the skin (Kanetsky et al. 2002; Bonilla et al. 2005; Voisey et al. 2006), hair (Kanetsky et al. 2002), and eyes (Kanetsky et al. 2002; Frudakis et al. 2003). In particular, the G allele is associated with increased eumelanin. The G allele is postulated to decrease the stability of the mRNA transcript, resulting in a decrease in

ASIP, and consequently, a decrease in the antagonistic action of α-MSH. Additionally, a less studied polymorphism located within the vicinity of a conserved region of intron 1, rs2424984, was found to be more significantly associated with skin pigmentation variation across various populations compared to rs6058017 (Valenzuela et al. 2010).

Polymorphisms of *OCA2* have been associated with skin, hair, and eye color variation. The strongest genetic variant associated with *OCA2* and eye color variation is rs12913832. This variant lies in an evolutionary conserved region of an intron of an adjacent gene (*HERC2*) located immediately upstream of *OCA2* and has been hypothesized to be within a promoter region of *OCA2* (Sturm et al. 2008). In skin melanocytes, Cook et al. (2009) found that the non-blue eye color variant of rs12913832 was associated with increased transcript levels of *OCA2*, supporting the hypothesis that rs12913832 is located within a promoter region of *OCA2* (Cook et al. 2009). According to the National Center for Biotechnology Information (NCBI) website, the blue eye color variant has a frequency of approximately 80% in Whites, while in non-White populations, it is virtually non-existent.

Non-synonymous genetic polymorphisms of *SLC45A2* (*OCA4*) that have been associated with pigmentation variation are rs16891982 (F374L) and rs26722 (E272K) (Graf, Hodgson, & van Daal 2005; Branicki et al. 2008). The light-skin allele, 374F, decreases in frequency along a cline from northwest Europe to southeast Europe (Nakayama et al. 2002; Yuasa et al. 2006; Lucotte et al. 2010). There have been conflicting results regarding positive selection of SLC45A2. Soejima et al (2005) reported that there was positive selection, while Lao et al (2006) did not detect any

selection. Lao et al. (2006) suggested that their method of detection (extended haplotype homozygosity) was not sensitive to core variants of high-frequency, as is the case with the rs16891982 (*SLC45A2*) variant, and therefore, may have been the reason for failure to detect positive selection at *SLC45A2*.

The main genetic variant of *SLC24A5* associated with pigmentation variation is rs1426654 (A111T) (Lamason et al. 2005). The 111T allele was found to be almost fixed in European populations, while the 111A allele was found to be almost fixed in African and East Asian populations (Lamason et al. 2005). Moreover, in populations of European descent, a large region of decreased heterozygosity surrounding the 111T allele suggests that this locus has undergone positive selection (Lamason et al. 2005). This result was recently corroborated in a study that detected signatures of positive selection in the human exome (Tennessen, Madeoy, & Akey 2010).

1.1.7 Interaction

Melanin pigmentation is a polygenic trait and it is likely that some of the genes interact contributing to variation in pigmentation. Various statistical methods have been employed to detect interactions of genes associated with pigmentation. However, it should be noted that a statistical interaction may be indicative of an underlying biological interaction, and that lack of statistical detection does not necessarily imply that a biological interaction does not exist (reviewed in (Cordell 2002)). Akey et al. (2001) studied skin pigmentation variation in a Tibetan population and found that while various individual genetic markers of MC1R and OCA2 were not significantly associated with

skin pigmentation, a statistical interaction of two of the markers they tested was significant. Branicki et al. (2009) reported interactions between variants of HERC2 and MC1R for hair and skin color of a European population living in southern Poland. Valenzuela et al. (2010) reported a statistical interaction between variants of ASIP and SLC45A2. However, this interaction was detected utilizing a sample comprised of various ethnicities, and therefore is more difficult to interpret.

1.1.8 Measurements of melanin

Both indirect and more direct methods have been employed to quantify the amount of melanin in the skin, hair, and eyes to determine genetic signals for each trait. Indirect methods of measurement include qualitative categorization (self-reported, observer reported), and light reflectance spectrophotometer measurements. The Fitzpatrick Classification Scale, or variants thereof, is an indirect method that many studies have utilized to categorize skin into six different categories based on constitutive pigmentation and UVR response (erythema and facultative pigmentation) (Roberts 2009). Similar, but less standardized categorical scales have been utilized to measure hair and eye color. Another indirect, but more quantitative method of estimating melanin content of the skin and hair has been achieved by use of light reflectance spectrophotometers. Previous studies have established the wavelength of light that yields reasonable estimates of melanin content, without distinction between the components of melanin (i.e., eumelanin or pheomelanin), to be approximately 660 nm (+/- 10 nm). More recently, Shriver & Parra (2000) established that melanin content measured at this wavelength was

highly correlated with the “lightness” (or L*) scale of the CIEL*a*b* (Commission Internationale d’Eclairage) when measured across populations of varying melanin content (Shriver & Parra 2000). (The CIE convention was devised to quantify colors based on how the eye perceives colors).

Analytical chemical methods have been developed to obtain more direct measurements of melanin content (both eumelanin and pheomelanin components) of biological samples. However, these methods require destruction of the sample, and therefore, are less practical than indirect methods, and are usually not possible. Consequently, methods that indirectly estimate melanin content are often utilized to establish genetic associations with a trait. However, when indirect methods are utilized to estimate the melanin content of a trait, it is prudent that a correlation be established between the indirect and direct methods for accurate biological interpretation of the resultant genetic associations, or non-associations.

Kongshoj et al. (2006) were the first to explicitly establish a relationship between light reflectance spectrophotometry and quantitative chemical measurements of both the eumelanin and pheomelanin components of skin (Kongshoj, Thorleifsson, & Wulf 2006). Previous studies had determined correlations with spectrophotometry and chemical analysis of total melanin in the skin, they did not explicitly report the pheomelanin correlation with spectrophotometry (Alaluf et al. 2002). Kongshoj et al. (2006) showed that eumelanin was much more precisely described by spectrophotometry compared to pheomelanin. This observation was supported by a study performed by Wakamatsu et al. (2006) where they measured the components of eumelanin and pheomelanin from

cultured human melanocytes using quantitative chemical methods and compared the results with the visual phenotype of each of the cultured melanocyte cell lines (Wakamatsu et al. 2006). This suggests that previous studies that associated genetic signals with reflective spectrophotometry measurements inherently biased detection for eumelanin genetic signals compared to pheomelanin genetic signals. Therefore, measurements ascertained by reflectance spectrophotometry are of limited utility when making genetic inferences about the biology of melanin production. However, they are adequate for making genetic inferences of visual pigmentation for skin.

1.1.9 Genetic Forensics

Various classes of genetic markers have been studied for the purpose of determining an individual's identity based on analysis of DNA obtained from biological samples. Initially, markers were utilized to find potential matches from a small pool of suspects. Later, a more comprehensive set of markers was developed and an offender database comprising these markers, called the Combined DNA Index System (CODIS)-National DNA Index System (NDIS), was established for the purpose of standardizing the computer storage/retrieval and identification of genetic profiles across law enforcement agencies in the United States. CODIS utilizes 14 markers (13 autosomal microsatellites, 1 sex-linked) that uniquely identify an individual. As of December 2010, there were more than 9 million offender profiles and approximately 352,000 forensic profiles, or unmatched profiles, in the NDIS (http://www.fbi.gov/about-us/lab/codis/ndis-statistics). Over 134,500 unknown samples have produced hits within the NDIS, leaving

a significant number of non-hits. Hence, there is motivation for additional methods, such as the use of ancestry informative markers (AIMs) and/or phenotype informative SNPs, to assist in determining the identity of an unknown sample (Budowle & van Daal 2008).

AIMs are genetic markers that exhibit a high degree of variation between populations of different ancestry, and are therefore useful in reducing the pool of potential candidates based on general biogeographical appearance. Variation in AIM frequencies between populations are due to unique population histories and can be the result of random (drift) and/or non-random (selection) forces. AIMs that are a result of selection may, or may not, have functional effects on biogeographical appearance. Nevertheless, for predictive purposes, AIMs are forensically informative. Moreover, because pigmentation variants are often reflective of ancestry, many are also AIMs (e.g., rs1426654).

Genetic markers that have been associated with a forensically informative trait and are either causal, or closely linked to causal variant(s), are ideal for making predictions about physical appearance of an unidentified forensic sample. Development of prediction models of forensically informative traits, such as pigmentation, is an active area of forensic research.

1.1.10 Prediction models of normal human pigmentation

Studies have developed prediction models for normal pigmentation variation intended for forensic purposes (Duffy et al. 2007; Liu et al. 2009; Valenzuela et al. 2010; Walsh et al. 2010; Spichenok et al. 2010; Branicki, Liu, van Duijn, Draus-Barini,

Pośpiech, Walsh, Kupiec, Wojas-Pelc, & Kayser 2011a; Mengel-From et al.). Models developed by these studies vary by: population(s) studied, choice of SNPs, method of phenotype measurement, and statistical modeling/selection procedures. The studies by Duffy et al. 2007, Liu et al. 2009, S. Walsh et al. 2010, Mengel-From et al. 2010, and Branicki et al. (2011) constructed their models using a single source population from which they extrapolated predictions to other populations of different ancestry. An advantage of utilizing a population of one ancestry, in determining novel genetic variants associated with a trait for biological interpretation, is that confounding (or spurious associations) due to population stratification is minimized (Pritchard & Rosenberg 1999). However, if the objective is to determine novel genetic variants for predictive purposes, then confounding due to population stratification is arguably not an issue.

Conversely, problems can arise in extrapolating prediction models that were derived from one ethnic group and applied to other ethnic groups of different ancestry. This approach neglects phenotypes that are similar as a result of convergent evolution as has been shown with the fair skin of some East Asian populations and Europeans. Moreover, this approach cannot account for variants that are fixed (AIMs) in the training population, but not necessarily fixed in other populations. Hence, this approach likely cannot account for admixed populations (e.g., African-Americans and Hispanics). Therefore, depending on the ethnic composition and the relative percentages of each ethnic group comprising that population to be interrogated, models developed from populations of similar ancestry may have limited utility.

As mentioned earlier, the “lighter” variant of rs1426654 (SLC24A5) is essentially fixed in European populations. Whereas, in various non-European populations, including fair-complexioned East Asians populations, the “darker” variant of rs1426654 (SLC24A5) is essentially fixed (Soejima & Koda 2007). Moreover, genetic variants that are essentially fixed in European populations have been associated with skin pigmentation variation in East Asian populations, and vice-versa, suggesting convergent evolution of lighter skin.

The United States is ethnically diverse and comprised of populations of varying degrees of admixture. For example, studies indicate that the African American population contains approximately 20-30% admixture of mainly European, and some Native American, ancestry (Lee et al. 2010). Similarly, Hispanics also show varying degrees of admixture of Native American, European and African descent. Hence, the approach of developing prediction models from one ethnic group and extrapolating to ethnically diverse populations, such as to the United States population, where there is also significant recent, and non-recent admixture, may not be the optimal approach. Spurious associations can arise due to population stratification, however, the danger in utilizing spurious associations is in the interpretation of the biology of the trait of interest, but utilization of spurious associations for purely predictive purposes, is much less, if at all, an issue.

Development of forensic models for application across an ethnically diverse population poses a unique problem in genetics that has been widely unexplored. An approach to developing a forensic model that has forensic utility across a diverse ethnic

population (e.g., in the United States), while at the same time avoiding population stratification, is to first determine genetic variants associated with pigmentation within various ethnic groups. Second, develop models using those markers on a population comprised of those ethnic groups. This was the approach taken by Valenzuela et al. (2010) and Spichenok et al. (2010). Consequently, they detected the SNP rs1426654 (SLC24A5) as being a major contributor to pigmentation variation, as has been shown in bioinformatic and functional studies (Lamason et al. 2005; Ginger et al. 2008; Cook et al. 2009). In contrast, studies that used a non-ethnically diverse sample, failed to detect rs1426654 (SLC24A5) as being a major contributor to pigmentation variation.

Various statistical approaches are utilized to develop and assess the performance of prediction models. Although phenotype distributions of complex traits are continuous, the phenotypes are often categorized, or binned, to reduce "noise" in the data and increase statistical power. Categorical response variables are amenable to various forms of logistic regression.

To estimate the performance of a model, external and internal resampling methods have been developed. Some resampling methods include leave-one-out-cross validation (LOOCV), K-fold cross-validation, and the .632+ bootstrap (Molinaro, Simon, & Pfeiffer 2005). External validation is a more rigorous procedure to determine if the prediction model will generalize to populations similar to the one from which the prediction model was developed (Taylor, Ankerst, & Andridge 2008).

II. PRESENT STUDY

This thesis explores the genetic complexity of constitutive pigmentation. This includes pigmentation of the skin, eyes, and hair (discussed in the literature review). The overall approach was to measure as objectively as possible these pigmentation traits, characterize single nucleotide polymorphisms in candidate genes, and to relate phenotype using statistical methods. A cohort of 789 individuals was used to generate a prediction algorithm for the pigmentation traits (Appendix A) and to test this algorithm on an independent population (Appendix B). The results, their limitations and future directions of the field are discussed below. The references of the studies in each appendix conform to the journal format that they have been submitted to.

2.1 Conclusion

2.1.1 Modeling melanin pigmentation in skin, eyes, and hair

The pigmentation traits of human skin, eyes, and hair are complex, or polygeneic, traits that are largely influenced by melanin. Latitudinal and genetic studies suggest that melanin pigmentation has been shaped by evolutionary forces (Relethford 1997; Lamason et al. 2005; Jablonski & Chaplin 2010; Tennessen et al. 2010). Moreover, genetic studies support that lighter skin (i.e., less pigmentation) of East Asians and Europeans are due to different genetic variants, suggesting convergent evolution of phenotypes (Yuasa et al. 2006; Norton et al. 2007; Edwards et al. 2010). Melanin

pigmentation is a pronounced physical feature and is therefore a trait of interest in forensic investigations. The focus of this dissertation was to develop and validate multiple linear regression models to predict hair, skin, and eye color utilizing genetic variants of genes that have previously been associated with pigmentation variation in these traits.

A secondary focus of this dissertation was development and testing of a model selection algorithm that enabled a more consistent method of determining a set of predictors compared to the traditional stepwise regression method of model selection. The stepwise regression method selects models by either building-up a model by adding predictors one at a time and comparing a goodness-of-fit measure to determine if the predictor significantly increased the fit of the model to the data (forward stepwise regression), or by removing a predictor from a model composed of all the predictors (backward stepwise regression) and determining if the dropped predictor significantly affects how well the model fits the data. If the model's fit is not significantly affected, then the predictor is dropped. This procedure continues until predictors cannot be dropped from the model without significantly affecting the fit of the model to the data. Also, a combination of forward and backward stepwise regression procedures can be utilized for model selection.

In determining model fit, stepwise regression methods often utilize the R^2 value. However, other measures are also used such as the adjusted R^2, and Mallow's Cp statistic. These measures identify the model that 'best' fits the data, they are variants of the expression of the errors of the model (i.e., what the model does not describe). A

problem that can arise with selecting a model with minimum error, amongst a pool of models with the same number of predictors and with similar small errors, is that the small error may have occurred by chance (i.e., sampling error). Therefore, a potentially more robust approach may be to compare and select predictors that overlap in models that have the smallest errors (see algorithm presented in Appendix A). In comparing models in this manner, caution must be taken so as not to choose predictors that are highly correlated.

The extension of the algorithm presented in Appendix B enabled visual determination of predictors, this also enabled visual determination of predictors that were correlated.

2.1.2 Improvements to modeling

Validation of the melanin pigmentation models of skin, eyes, and hair described in Appendix B indicate the forensic utility of these models. Predictions may be improved by understanding more about the biology of melanin pigmentation and its relationship to other traits, or 'sub'-complex traits. To better understand the biology of melanin pigmentation, a comprehensive set of measurements of melanin pigmentation, and correlation of those measurements, may be necessary. Additionally, measurements of other traits, or 'sub'-complex traits, may be needed to improve macroscopic predictions.

Improvements to skin pigmentation measurements can enhance our ability to predict skin color. As mentioned in Appendix A, melanin pigmentation of skin has been measured using various methods. In particular, reflective spectrophotometry has been widely utilized in genetic studies because it is non-invasive and a practical method of

measuring melanin pigmentation of the skin. Also, mentioned in literature review, reflective spectrophotometry measurements of the skin measured at 660 nm (and the lightness, L*, scale of the CIEL*a*b* color convention) are more of an indication of eumelanin content rather than pheomelanin content. Additionally, measurements of skin pigmentation may be confounded at least by the varying degraded states of the melanosome in the epidermis. After melanosomes are transferred from the melanocyte to the keratinocytes of the epidermis, the keratinocytes gradually migrate to the outer most layer of the epidermis, or the stratum corneum. As the melanosome-carrying keratinocyte migrates to the stratum corneum, the melanosome is degraded. How melanin degradation affects the optical properties (i.e., reflective spectrophotometry measurements) of melanin has not been studied. Hence, studying melanin pigmentation at the resolution of the melanosome and comparing measurements to reflectance spectrophotometry measurements may provide additional insight into the biology of melanin pigmentation of the skin.

Visualization and analysis of melanin pigmentation at the resolution of the melanosome may yield additional insight into the melanogenesis of pheomelanosomes. Studies have shown that both eumelanosomes and pheomelanosomes are produced within a given melanocyte. However, the biological causes that determine the fate of a pre-melanosome forming into either a eumelanosome or pheomelanosome are not known. Melanosomal proteins have been studied by Chi et al. (2006), but without distinction to the type of melanosome. Studying melanin at the resolution of the melanosome, using proteomic methods may elucidate the biological basis for the different fates, and

therefore may enable refinement of genetic prediction models of melanin pigmentation. Alaluf et al. (2001) presented a method for visualizing intact melanosomes. Therefore, a method to sort the spherically-shaped pheomelanosomes from the other melanosomes may need to be devised to study the proteomics of these types of melanosomes.

Other high resolution methods at the level of the melanosome, such as ultrastructural measurements of the surface features of melanin have provided insight into genes that contribute to the variation in melanin. Most recently, x-ray absorption near edge spectroscopy (XANES) has indicated a novel gene involved in melanin production in mice (Haraszti et al. 2010).

Improvements to eye color measurements are more problematic. To directly quantify the amount of melanin pigmentation in relationship to eye color requires a tissue biopsy and is therefore only possible on cadaver tissue. Such methods have been developed by Wakamatsu et al. (2007). However, such measurements are less practical than methods that have been developed to precisely quantify eye color. Most recently Liu et al. (2010) developed a high-resolution digital-photograph method for measuring eye color that takes into account hue and saturation. This method enabled them to associate two novel genes, LYST and DSCR9, with eye color.

There are specific ways that hair color measurements can be improved. Recently, hair strand thickness was correlated with hair color (Vaughn et al. 2009), suggesting that in addition to melanin pigmentation, incorporation of hair thickness into genetic models may be necessary to more precisely predict hair color. Vaughn et al. (2009) showed that hair thickness was correlated with microscopic and macroscopic measurements of hair

color. Microscopic resolution refers to individual strands of hair, whereas, macroscopic resolution refers to the appearance of an individual's hair color. Moreover, Vaughn et al. (2008) suggest that hair color measurements may be more accurately described by reflectance spectrophotometry of both the b* (yellow) and L* (lightness) scales of the CIEL*a*b* color convention rather than usage of only one scale (L*) as indicated by Shriver and Parra (2000). Hence, in developing prediction models for hair color, the additional measurements of hair strand thickness along with reflective spectrophotometry measurements of both the b* and L* scales may help improve the ability to predict hair color. Naysmith et al. (2004) followed a similar procedure of measurements, however, they did not take into account hair thickness, and they limited their study to *MC1R* genetic variants.

If hair strand thickness is shown to improve modeling of hair color, then the genetic determinants governing this trait will need to be determined and incorporated into hair color models, as prior knowledge of hair strand thickness of a forensic sample will likely not be known.

2.1.3 Where is the field going?

Currently, all melanin pigmentation prediction models are based upon common single-nucleotide variants, or SNPs (minor allele frequency or MAF > 5%). With the advancement of sequencing technologies, the amount of available genetic information has and will continue to increase significantly. Of the wealth of genetic information that has become available through second generation sequencing technologies, are low-

frequency (MAF < 5%) and rare single-nucleotide variants, or rare variants. In the literature, there is ambiguity on the definition of a rare variant. Some studies define a rare variant as a nucleotide variant that has an allele frequency of less than 3%, while other studies have a more stringent definition of 1%. Studies of various populations show that an excess of low frequency and rare variants are present, suggestive and consistent with recent explosive human population growth (Durbin et al. 2010; Coventry et al. 2010; Li et al. 2010). Due to their low frequency, challenges arise in association and incorporation of rare variants into genetic models of complex traits. Most recently, studies have focused on non-synonymous rare variants in genes previously associated with a given trait of interest in relation to these challenges. Currently, methods are lacking on how to incorporate rare variants into prediction models.

2.1.4 Comments on statistical procedures

The studies of this dissertation utilized multiple linear regression (MLR) statistics to describe phenotypic variation in hair, skin, and eye color for forensic purposes. While MLR was appropriate, many other studies with the same focus utilized multinomial logistic regression statistics (Liu et al. 2010; Walsh et al. 2010; Branicki, Liu, van Duijn, Draus-Barini, Pośpiech, Walsh, Kupiec, Wojas-Pelc, & Kayser 2011b). Briefly, the objective of various modeling methods is to relate a set of independent variables (e.g., genotypes) to a response variable (e.g., phenotype) by transforming independent variables via parameter estimates. Various methods have been developed for parameter estimation and each method makes certain assumptions about the data. For MLR,

parameters are estimated by the method of ordinary least squares (OLS). OLS is a method that minimizes the sum of the squared differences, or squared error (squared residual), of predicted and actual values. To utilize OLS, several assumptions must hold, including that the independent variables are linearly related to the dependent variable and that the residuals are homoscedastic (i.e., constant variance of the residuals for all predictors).

Multinomial logistic regression (MNLR) is an extension of logistic regression (LR). The response variable of LR is binary, that is, there are two defined values that are predicted by the regression equation (e.g., blue eye color vs. not blue eye color). In contrast, the response variable of MNLR has more than two possible values (e.g., blue, green, or brown eye color). For each value, there is one prediction equation that has a binary response variable (e.g., blue eye color or not blue eye color), that is, the number of prediction equations equals the number of response variables minus one. Parameter estimates of MNLR are often calculated by the method of maximum likelihood estimation. In contrast to MLR, MNLR does not assume that the independent variable and dependent variable are linearly related. Also, the residuals do not have to be homoscedastic.

In favor of MLR in modeling complex traits, the method of fitting the parameter estimates, ordinary least-squares (OLS), is well grounded in statistical theory. Also, the resultant equation is of the same mathematical form as the standard quantitative genetics model (reviewed in (Cordell et al. 2001)). Therefore, interpreting the genetic architecture of a complex trait modeled by MLR may be more straightforward compared to other

modeling methods. Supplemental studies to this dissertation would be to develop models using multinomial logistic regression with quantitatively defined categories.

REFERENCES

Agrup, G., Hansson, C., Rorsman, H. & Rosengren, E. (1982) The effect of cysteine on oxidation of tyrosine, dopa, and cysteinyldopas. *Archives of Dermatological Research*, **272**, 103-115.

Alaluf, S., Atkins, D., Barrett, K., Blount, M., Carter, N. & Heath, A. (2002) The impact of epidermal melanin on objective measurements of human skin colour. *Pigment Cell Research / Sponsored by the European Society for Pigment Cell Research and the International Pigment Cell Society*, **15**, 119-126.

Alaluf, S., Barrett, K., Blount, M. & Carter, N. (2003) Ethnic variation in tyrosinase and TYRP1 expression in photoexposed and photoprotected human skin. *Pigment Cell Research / Sponsored by the European Society for Pigment Cell Research and the International Pigment Cell Society*, **16**, 35-42.

Barrett, H. & McElduff, A. (2010) Vitamin D and pregnancy: An old problem revisited. *Best Practice & Research. Clinical Endocrinology & Metabolism*, **24**, 527-539.

Bito, L.Z., Matheny, A., Cruickshanks, K.J., Nondahl, D.M. & Carino, O.B. (1997) Eye color changes past early childhood. The Louisville Twin Study. *Archives of Ophthalmology*, **115**, 659-663.

Boissy, R.E., Zhao, H., Oetting, W.S., Austin, L.M., Wildenberg, S.C., Boissy, Y.L., Zhao, Y., Sturm, R.A., Hearing, V.J., King, R.A. & Nordlund, J.J. (1996) Mutation in and lack of expression of tyrosinase-related protein-1 (TRP-1) in melanocytes from an individual with brown oculocutaneous albinism: a new subtype of albinism classified as "OCA3." *American Journal of Human Genetics*, **58**, 1145-1156.

Bonilla, C., Boxill, L.-A., Donald, S.A.M., Williams, T., Sylvester, N., Parra, E.J., Dios, S., Norton, H.L., Shriver, M.D. & Kittles, R.A. (2005) The 8818G allele of the agouti signaling protein (ASIP) gene is ancestral and is associated with darker skin color in African Americans. *Human Genetics*, **116**, 402-406.

Borovanský, J. & Elleder, M. (2003) Melanosome degradation: fact or fiction. *Pigment Cell Research / Sponsored by the European Society for Pigment Cell Research and the International Pigment Cell Society*, **16**, 280-286.

Branda, R.F. & Eaton, J.W. (1978) Skin color and nutrient photolysis: an evolutionary hypothesis. *Science (New York, N.Y.)*, **201**, 625-626.

Branicki, W., Brudnik, U., Draus-Barini, J., Kupiec, T. & Wojas-Pelc, A. (2008) Association of the SLC45A2 gene with physiological human hair colour variation. *Journal of Human Genetics*, **53**, 966-971.

Branicki, W., Liu, F., van Duijn, K., Draus-Barini, J., Pośpiech, E., Walsh, S., Kupiec, T., Wojas-Pelc, A. & Kayser, M. (2011a) Model-based prediction of human hair color using DNA variants. *Human Genetics*, 1-12.

Branicki, W., Liu, F., van Duijn, K., Draus-Barini, J., Pośpiech, E., Walsh, S., Kupiec, T., Wojas-Pelc, A. & Kayser, M. (2011b) Model-based prediction of human hair color using DNA variants. *Human Genetics*.

Budowle, B. & van Daal, A. (2008) Forensically relevant SNP classes. *BioTechniques*, **44**, 603-608, 610.

Bush, W.D., Garguilo, J., Zucca, F.A., Albertini, A., Zecca, L., Edwards, G.S., Nemanich, R.J. & Simon, J.D. (2006) The surface oxidation potential of human neuromelanin reveals a spherical architecture with a pheomelanin core and a eumelanin surface. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 14785-14789.

Calcraft, P.J., Ruas, M., Pan, Z., Cheng, X., Arredouani, A., Hao, X., Tang, J., Rietdorf, K., Teboul, L., Chuang, K.-T., Lin, P., Xiao, R., Wang, C., Zhu, Y., Lin, Y., Wyatt, C.N., Parrington, J., Ma, J., Evans, A.M., Galione, A. & Zhu, M.X. (2009) NAADP mobilizes calcium from acidic organelles through two-pore channels. *Nature*, **459**, 596-600.

Cheli, Y., Luciani, F., Khaled, M., Beuret, L., Bille, K., Gounon, P., Ortonne, J.-P., Bertolotto, C. & Ballotti, R. (2009) {alpha}MSH and Cyclic AMP elevating agents control melanosome pH through a protein kinase A-independent mechanism. *The Journal of Biological Chemistry*, **284**, 18699-18706.

Chen, K., Manga, P. & Orlow, S.J. (2002) Pink-eyed dilution protein controls the processing of tyrosinase. *Molecular Biology of the Cell*, **13**, 1953-1964.

Chi, A., Valencia, J.C., Hu, Z.-Z., Watabe, H., Yamaguchi, H., Mangini, N.J., Huang, H., Canfield, V.A., Cheng, K.C., Yang, F., Abe, R., Yamagishi, S., Shabanowitz, J., Hearing, V.J., Wu, C., Appella, E. & Hunt, D.F. (2006) Proteomic and bioinformatic characterization of the biogenesis and function of melanosomes. *Journal of Proteome Research*, **5**, 3135-3144.

Cook, A.L., Chen, W., Thurber, A.E., Smit, D.J., Smith, A.G., Bladen, T.G., Brown, D.L., Duffy, D.L., Pastorino, L., Bianchi-Scarra, G., Leonard, J.H., Stow, J.L. & Sturm, R.A. (2009) Analysis of cultured human melanocytes based on

polymorphisms within the SLC45A2/MATP, SLC24A5/NCKX5, and OCA2/P loci. *The Journal of Investigative Dermatology*, **129**, 392-405.

Cordell, H.J., Todd, J.A., Hill, N.J., Lord, C.J., Lyons, P.A., Peterson, L.B., Wicker, L.S. & Clayton, D.G. (2001) Statistical modeling of interlocus interactions in a complex disease: rejection of the multiplicative model of epistasis in type 1 diabetes. *Genetics*, **158**, 357-367.

Cordell, H.J. (2002) Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics*, **11**, 2463-2468.

Costin, G.-E., Valencia, J.C., Vieira, W.D., Lamoreux, M.L. & Hearing, V.J. (2003) Tyrosinase processing and intracellular trafficking is disrupted in mouse primary melanocytes carrying the underwhite (uw) mutation. A model for oculocutaneous albinism (OCA) type 4. *Journal of Cell Science*, **116**, 3203-3212.

Coventry, A., Bull-Otterson, L.M., Liu, X., Clark, A.G., Maxwell, T.J., Crosby, J., Hixson, J.E., Rea, T.J., Muzny, D.M., Lewis, L.R., Wheeler, D.A., Sabo, A., Lusk, C., Weiss, K.G., Akbar, H., Cree, A., Hawes, A.C., Newsham, I., Varghese, R.T., Villasana, D., Gross, S., Joshi, V., Santibanez, J., Morgan, M., Chang, K., Iv, W.H., Templeton, A.R., Boerwinkle, E., Gibbs, R. & Sing, C.F. (2010) Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nature Communications*, **1**, 131.

Dores, R.M. (2009) Adrenocorticotropic hormone, melanocyte-stimulating hormone, and the melanocortin receptors: revisiting the work of Robert Schwyzer: a thirty-year retrospective. *Annals of the New York Academy of Sciences*, **1163**, 93-100.

Dräger, U.C. (1985) Calcium binding in pigmented and albino eyes. *Proceedings of the National Academy of Sciences of the United States of America*, **82**, 6716-6720.

Duffy, D.L., Montgomery, G.W., Chen, W., Zhao, Z.Z., Le, L., James, M.R., Hayward, N.K., Martin, N.G. & Sturm, R.A. (2007) A three-single-nucleotide polymorphism haplotype in intron 1 of OCA2 explains most human eye-color variation. *American Journal of Human Genetics*, **80**, 241-252.

Durbin, R.M., Abecasis, G.R., Altshuler, D.L., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R.A., Hurles, M.E. & McVean, G.A. (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061-1073.

Edwards, E.A. & Duntley, S.Q. (1939) The pigments and color of living human skin. , **65**, 1-33.

Edwards, M., Bigham, A., Tan, J., Li, S., Gozdzik, A., Ross, K., Jin, L. & Parra, E.J. (2010) Association of the OCA2 polymorphism His615Arg with melanin content in east Asian populations: further evidence of convergent evolution of skin pigmentation. *PLoS Genetics*, **6**, e1000867.

Fitzpatrick, T.B. & Breathnach, A.S. (1963) [The epidermal melanin unit system.]. *Dermatologische Wochenschrift*, **147**, 481-489.

Fitzpatrick, T.B., Miyamoto, M. & Ishikawa, K. (1967) The evolution of concepts of melanin biology. *Archives of Dermatology*, **96**, 305-323.

Flanagan, N., Healy, E., Ray, A., Philips, S., Todd, C., Jackson, I.J., Birch-Machin, M.A. & Rees, J.L. (2000) Pleiotropic effects of the melanocortin 1 receptor (MC1R) gene on human pigmentation. *Human Molecular Genetics*, **9**, 2531-2537.

Frudakis, T., Thomas, M., Gaskin, Z., Venkateswarlu, K., Chandra, K.S., Ginjupalli, S., Gunturi, S., Natrajan, S., Ponnuswamy, V.K. & Ponnuswamy, K.N. (2003) Sequences associated with human iris pigmentation. *Genetics*, **165**, 2071-2083.

Fuller, B.B., Spaulding, D.T. & Smith, D.R. (2001) Regulation of the catalytic activity of preexisting tyrosinase in black and Caucasian human melanocyte cell cultures. *Experimental Cell Research*, **262**, 197-208.

Gates, R.R. & Zimmermann, A.A. (1953) Comparison of skin color with melanin content. *The Journal of Investigative Dermatology*, **21**, 339-348.

Ginger, R.S., Askew, S.E., Ogborne, R.M., Wilson, S., Ferdinando, D., Dadd, T., Smith, A.M., Kazi, S., Szerencsei, R.T., Winkfein, R.J., Schnetkamp, P.P.M. & Green, M.R. (2008) SLC24A5 encodes a trans-Golgi network protein with potassium-dependent sodium-calcium exchange activity that regulates human epidermal melanogenesis. *The Journal of Biological Chemistry*, **283**, 5486-5495.

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

Haraszti, T., Trantow, C.M., Hedberg-Buenz, A., Grunze, M. & Anderson, M.G. (2010) Spectral analysis by XANES reveals that GPNMB influences the chemical composition of intact melanosomes. *Pigment Cell & Melanoma Research*.

Harding, R.M., Healy, E., Ray, A.J., Ellis, N.S., Flanagan, N., Todd, C., Dixon, C., Sajantila, A., Jackson, I.J., Birch-Machin, M.A. & Rees, J.L. (2000) Evidence for variable selective pressures at MC1R. *American Journal of Human Genetics*, **66**, 1351-1361.

Inazu, M. & Mishima, Y. (1993) Detection of eumelanogenic and pheomelanogenic melanosomes in the same normal human melanocyte. *The Journal of Investigative Dermatology*, **100**, 172S-175S.

Iozumi, K., Hoganson, G.E., Pennella, R., Everett, M.A. & Fuller, B.B. (1993) Role of tyrosinase as the determinant of pigmentation in cultured human melanocytes. *The Journal of Investigative Dermatology*, **100**, 806-811.

Ito, S. & Wakamatsu, K. (2008) Chemistry of mixed melanogenesis--pivotal roles of dopaquinone. *Photochemistry and Photobiology*, **84**, 582-592.

Ito, S. & Wakamatsu, K. (2011) Human hair melanins: what we have learned and have not learned from mouse coat color pigmentation. *Pigment Cell & Melanoma Research*, **24**, 63-74.

Jablonski, N.G. (1999) A possible link between neural tube defects and ultraviolet light exposure. *Medical Hypotheses*, **52**, 581-582.

Jablonski, N.G. & Chaplin, G. (2000) The evolution of human skin coloration. *Journal of Human Evolution*, **39**, 57-106.

Jablonski, N.G. & Chaplin, G. (2010) Colloquium paper: human skin pigmentation as an adaptation to UV radiation. *Proceedings of the National Academy of Sciences of the United States of America*, **107 Suppl 2**, 8962-8968.

Kanetsky, P.A., Swoyer, J., Panossian, S., Holmes, R., Guerry, D. & Rebbeck, T.R. (2002) A polymorphism in the agouti signaling protein gene is associated with human pigmentation. *American Journal of Human Genetics*, **70**, 770-775.

Kobayashi, T. & Hearing, V.J. (2007) Direct interaction of tyrosinase with Tyrp1 to form heterodimeric complexes in vivo. *Journal of Cell Science*, **120**, 4261-4268.

Kongshoj, B., Thorleifsson, A. & Wulf, H.C. (2006) Pheomelanin and eumelanin in human skin determined by high-performance liquid chromatography and its relation to in vivo reflectance measurements. *Photodermatology, Photoimmunology & Photomedicine*, **22**, 141-147.

Lamason, R.L., Mohideen, M.-A.P.K., Mest, J.R., Wong, A.C., Norton, H.L., Aros, M.C., Jurynec, M.J., Mao, X., Humphreville, V.R., Humbert, J.E., Sinha, S., Moore, J.L., Jagadeeswaran, P., Zhao, W., Ning, G., Makalowska, I., McKeigue, P.M., O'donnell, D., Kittles, R., Parra, E.J., Mangini, N.J., Grunwald, D.J., Shriver, M.D., Canfield, V.A. & Cheng, K.C. (2005) SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. *Science (New York, N.Y.)*, **310**, 1782-1786.

Lee, Y.L., Teitelbaum, S., Wolff, M.S., Wetmur, J.G. & Chen, J. (2010) Comparing genetic ancestry and self-reported race/ethnicity in a multiethnic population in New York City. *Journal of Genetics*, **89**, 417-423.

Li, L., Hu, D.-N., Zhao, H., McCormick, S.A., Nordlund, J.J. & Boissy, R.E. (2006) Uveal melanocytes do not respond to or express receptors for alpha-melanocyte-stimulating hormone. *Investigative Ophthalmology & Visual Science*, **47**, 4507-4512.

Li, Y., Vinckenbosch, N., Tian, G., Huerta-Sanchez, E., Jiang, T., Jiang, H., Albrechtsen, A., Andersen, G., Cao, H., Korneliussen, T., Grarup, N., Guo, Y., Hellman, I., Jin, X., Li, Q., Liu, J., Liu, X., Sparsø, T., Tang, M., Wu, H., Wu, R., Yu, C., Zheng, H., Astrup, A., Bolund, L., Holmkvist, J., Jørgensen, T., Kristiansen, K., Schmitz, O., Schwartz, T.W., Zhang, X., Li, R., Yang, H., Wang, J., Hansen, T., Pedersen, O., Nielsen, R. & Wang, J. (2010) Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. *Nature Genetics*, **42**, 969-972.

Liu, F., van Duijn, K., Vingerling, J.R., Hofman, A., Uitterlinden, A.G., Janssens, A.C.J.W. & Kayser, M. (2009) Eye color and the prediction of complex phenotypes from genotypes. *Current Biology*, **19**, R192-R193.

Liu, F., Wollstein, A., Hysi, P.G., Ankra-Badu, G.A., Spector, T.D., Park, D., Zhu, G., Larsson, M., Duffy, D.L., Montgomery, G.W., Mackey, D.A., Walsh, S., Lao, O., Hofman, A., Rivadeneira, F., Vingerling, J.R., Uitterlinden, A.G., Martin, N.G., Hammond, C.J. & Kayser, M. (2010) Digital Quantification of Human Eye Color Highlights Genetic Association of Three New Loci. , **6**.

Lucotte, G., Mercier, G., Diéterlen, F. & Yuasa, I. (2010) A decreasing gradient of 374F allele frequencies in the skin pigmentation gene SLC45A2, from the north of West Europe to North Africa. *Biochemical Genetics*, **48**, 26-33.

Makova, K. & Norton, H. (2005) Worldwide polymorphism at the MC1R locus and normal pigmentation variation in humans. *Peptides*, **26**, 1901-1908.

Manga, P., Boissy, R.E., Pifko-Hirst, S., Zhou, B.K. & Orlow, S.J. (2001) Mislocalization of melanosomal proteins in melanocytes from mice with oculocutaneous albinism type 2. *Experimental Eye Research*, **72**, 695-710.

Mengel-From, J., Wong, T.H., Morling, N., Rees, J.L. & Jackson, I.J. Genetic determinants of hair and eye colours in the Scottish and Danish populations. , **10**, 88-88.

Mitchell, R.E. (1963) The effect of prolonged solar radiation on melanocytes of the human epidermis. *The Journal of Investigative Dermatology*, **41**, 199-212.

Molinaro, A.M., Simon, R. & Pfeiffer, R.M. (2005) Prediction error estimation: a comparison of resampling methods. *Bioinformatics*, **21**, 3301-3307.

Moyer, F.H. (1966) Genetic variations in the fine structure and ontogeny of mouse melanin granules. *American Zoologist*, **6**, 43-66.

Nakayama, K., Fukamachi, S., Kimura, H., Koda, Y., Soemantri, A. & Ishida, T. (2002) Distinctive distribution of AIM1 polymorphism among major human populations with different skin color. *Journal of Human Genetics*, **47**, 92-94.

Newton, J.M., Cohen-Barak, O., Hagiwara, N., Gardner, J.M., Davisson, M.T., King, R.A. & Brilliant, M.H. (2001) Mutations in the human orthologue of the mouse underwhite gene (uw) underlie a new form of oculocutaneous albinism, OCA4. *American Journal of Human Genetics*, **69**, 981-988.

Norton, H.L., Kittles, R.A., Parra, E., McKeigue, P., Mao, X., Cheng, K., Canfield, V.A., Bradley, D.G., McEvoy, B. & Shriver, M.D. (2007) Genetic evidence for the convergent evolution of light skin in Europeans and East Asians. *Molecular Biology and Evolution*, **24**, 710-722.

O'Neil, D. n d. "Modern Human Variation." Retrieved (http://anthro.palomar.edu/vary/vary_1.htm).

Olivares, C. & Solano, F. (2009) New insights into the active site structure and catalytic mechanism of tyrosinase and its related proteins. *Pigment Cell & Melanoma Research*, **22**, 750-760.

Orlow, S.J., Zhou, B.K., Chakraborty, A.K., Drucker, M., Pifko-Hirst, S. & Pawelek, J.M. (1994) High-molecular-weight forms of tyrosinase and the tyrosinase-related proteins: evidence for a melanogenic complex. *The Journal of Investigative Dermatology*, **103**, 196-201.

Peles, D.N., Hong, L., Hu, D.-N., Ito, S., Nemanich, R.J. & Simon, J.D. (2009) Human iridal stroma melanosomes of varying pheomelanin contents possess a common eumelanic outer surface. *The Journal of Physical Chemistry. B*, **113**, 11346-11351.

Pritchard, J.K. & Rosenberg, N.A. (1999) Use of unlinked genetic markers to detect population stratification in association studies. *American Journal of Human Genetics*, **65**, 220-228.

Prota, G. (1980) Recent advances in the chemistry of melanogenesis in mammals. *The Journal of Investigative Dermatology*, **75**, 122-127.

Puri, N., Gardner, J.M. & Brilliant, M.H. (2000) Aberrant pH of melanosomes in pink-eyed dilution (p) mutant melanocytes. *The Journal of Investigative Dermatology*, **115**, 607-613.

Rana, B.K., Hewett-Emmett, D., Jin, L., Chang, B.H., Sambuughin, N., Lin, M., Watkins, S., Bamshad, M., Jorde, L.B., Ramsay, M., Jenkins, T. & Li, W.H. (1999) High polymorphism at the human melanocortin 1 receptor locus. *Genetics*, **151**, 1547-1557.

Raposo, G. & Marks, M.S. (2007) Melanosomes--dark organelles enlighten endosomal membrane transport. *Nature Reviews. Molecular Cell Biology*, **8**, 786-797.

Relethford, J.H. (1997) Hemispheric difference in human skin color. *American Journal of Physical Anthropology*, **104**, 449-457.

Roberts, W.E. (2009) Skin type classification systems old and new. *Dermatologic Clinics*, **27**, 529-533, viii.

Rosemblat, S., Durham-Pierre, D., Gardner, J.M., Nakatsu, Y., Brilliant, M.H. & Orlow, S.J. (1994) Identification of a melanosomal membrane protein encoded by the pink-eyed dilution (type II oculocutaneous albinism) gene. *Proceedings of the National Academy of Sciences of the United States of America*, **91**, 12071-12075.

Seiji, M., Fitzpatrick, T.B. & Birbeck, M.S. (1961) The melanosome: a distinctive subcellular particle of mammalian melanocytes and the site of melanogenesis. *The Journal of Investigative Dermatology*, **36**, 243-252.

Seiji, M., Shimao, K., Birbeck, M.S. & Fitzpatrick, T.B. (1963) Subcellular localization of melanin biosynthesis. *Annals of the New York Academy of Sciences*, **100**, 497-533.

Shriver, M.D. & Parra, E.J. (2000) Comparison of narrow-band reflectance spectroscopy and tristimulus colorimetry for measurements of skin and hair color in persons of different biological ancestry. *American Journal of Physical Anthropology*, **112**, 17-27.

Simon, J.D. & Peles, D.N. (2010) The Red and the Black. *Accounts of Chemical Research*.

Singh, S.K., Kurfurst, R., Nizard, C., Schnebert, S., Perrier, E. & Tobin, D.J. (2010) Melanin transfer in human skin cells is mediated by filopodia--a model for

homotypic and heterotypic lysosome-related organelle transfer. *The FASEB Journal: Official Publication of the Federation of American Societies for Experimental Biology*, **24**, 3756-3769.

Slominski, A., Wortsman, J., Plonka, P.M., Schallreuter, K.U., Paus, R. & Tobin, D.J. (2005) Hair follicle pigmentation. *The Journal of Investigative Dermatology*, **124**, 13-21.

Soejima, M. & Koda, Y. (2007) Population differences of two coding SNPs in pigmentation-related genes SLC24A5 and SLC45A2. *International Journal of Legal Medicine*, **121**, 36-39.

Spichenok, O., Budimlija, Z.M., Mitchell, A.A., Jenny, A., Kovacevic, L., Marjanovic, D., Caragine, T., Prinz, M. & Wurmbach, E. (2010) Prediction of eye and skin color in diverse populations using seven SNPs. *Forensic Science International. Genetics*.

Srisopark, S.S. (1976) Preliminary observations on the epidermal melanocytes of Thai (Mongoloid) skin: DOPA--reaction and electron microscopic studies. *Journal of the Medical Association of Thailand = Chotmaihet Thangphaet*, **59**, 61-70.

Staricco, R.J. & Pinkus, H. (1957) Quantitative and qualitative data on the pigment cells of adult human epidermis. *The Journal of Investigative Dermatology*, **28**, 33-45.

Sturm, R.A., Duffy, D.L., Zhao, Z.Z., Leite, F.P.N., Stark, M.S., Hayward, N.K., Martin, N.G. & Montgomery, G.W. (2008) A single SNP in an evolutionary conserved region within intron 86 of the HERC2 gene determines human blue-brown eye color. *American Journal of Human Genetics*, **82**, 424-431.

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

Sulem, P., Gudbjartsson, D.F., Stacey, S.N., Helgason, A., Rafnar, T., Jakobsdottir, M., Steinberg, S., Gudjonsson, S.A., Palsson, A., Thorleifsson, G., Pálsson, S., Sigurgeirsson, B., Thorisdottir, K., Ragnarsson, R., Benediktsdottir, K.R., Aben, K.K., Vermeulen, S.H., Goldstein, A.M., Tucker, M.A., Kiemeney, L.A., Olafsson, J.H., Gulcher, J., Kong, A., Thorsteinsdottir, U. & Stefansson, K. (2008) Two newly identified genetic determinants of pigmentation in Europeans. *Nature Genetics*, **40**, 835-837.

Summers, C.G. (2009) Albinism: classification, clinical characteristics, and recent findings. *Optometry and Vision Science: Official Publication of the American Academy of Optometry*, **86**, 659-662.

Szabo, G. (1959) *Quantitative histological investigations on the melanocyte system of the human epidermis, Pigment Cell Biology*. Academic Press, New York.

Taylor, J.M.G., Ankerst, D.P. & Andridge, R.R. (2008) Validation of biomarker-based risk prediction models. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, **14**, 5977-5983.

Tennessen, J.A., Madeoy, J. & Akey, J.M. (2010) Signatures of positive selection apparent in a small sample of human exomes. *Genome Research*, **20**, 1327-1334.

Tobin, D.J. (2008) Human hair pigmentation--biological aspects. *International Journal of Cosmetic Science*, **30**, 233-257.

Truschel, S.T., Simoes, S., Setty, S.R.G., Harper, D.C., Tenza, D., Thomas, P.C., Herman, K.E., Sackett, S.D., Cowan, D.C., Theos, A.C., Raposo, G. & Marks, M.S. (2009) ESCRT-I function is required for Tyrp1 transport from early endosomes to the melanosome limiting membrane. *Traffic (Copenhagen, Denmark)*, **10**, 1318-1336.

Tsukamoto, K., Jackson, I.J., Urabe, K., Montague, P.M. & Hearing, V.J. (1992) A second tyrosinase-related protein, TRP-2, is a melanogenic enzyme termed DOPAchrome tautomerase. *The EMBO Journal*, **11**, 519-526.

Vachtenheim, J. & Borovanský, J. (2010) "Transcription physiology" of pigment formation in melanocytes: central role of MITF. *Experimental Dermatology*, **19**, 617-627.

Valenzuela, R.K., Henderson, M.S., Walsh, M.H., Garrison, N.A., Kelch, J.T., Cohen-Barak, O., Erickson, D.T., John Meaney, F., Bruce Walsh, J., Cheng, K.C., Ito, S., Wakamatsu, K., Frudakis, T., Thomas, M. & Brilliant, M.H. (2010) Predicting phenotype from genotype: normal pigmentation. *Journal of Forensic Sciences*, **55**, 315-322.

Vaughn, M.R., Brooks, E., van Oorschot, R.A.H. & Baindur-Hudson, S. (2009) A comparison of macroscopic and microscopic hair color measurements and a quantification of the relationship between hair color and thickness. *Microscopy and Microanalysis: The Official Journal of Microscopy Society of America, Microbeam Analysis Society, Microscopical Society of Canada*, **15**, 189-193.

Voisey, J., Gomez-Cabrera, M.D.C., Smit, D.J., Leonard, J.H., Sturm, R.A. & van Daal, A. (2006) A polymorphism in the agouti signalling protein (ASIP) is associated with decreased levels of mRNA. *Pigment Cell Research / Sponsored by the European Society for Pigment Cell Research and the International Pigment Cell Society*, **19**, 226-231.

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

Wakamatsu, K. & Ito, S. (2002) Advanced chemical methods in melanin determination. *Pigment Cell Research / Sponsored by the European Society for Pigment Cell Research and the International Pigment Cell Society*, **15**, 174-183.

Wakamatsu, K., Kavanagh, R., Kadekaro, A.L., Terzieva, S., Sturm, R.A., Leachman, S., Abdel-Malek, Z. & Ito, S. (2006) Diversity of pigmentation in cultured human melanocytes is due to differences in the type as well as quantity of melanin. *Pigment Cell Research / Sponsored by the European Society for Pigment Cell Research and the International Pigment Cell Society*, **19**, 154-162.

Walsh, S., Lindenbergh, A., Zuniga, S.B., Sijen, T., de Knijff, P., Kayser, M. & Ballantyne, K.N. (2010) Developmental validation of the IrisPlex system: Determination of blue and brown iris colour for forensic intelligence. *Forensic Science International. Genetics*.

Yuasa, I., Umetsu, K., Harihara, S., Kido, A., Miyoshi, A., Saitou, N., Dashnyam, B., Jin, F., Lucotte, G., Chattopadhyay, P.K., Henke, L. & Henke, J. (2006) Distribution of the F374 allele of the SLC45A2 (MATP) gene and founder-haplotype analysis. *Annals of Human Genetics*, **70**, 802-811.

Yuasa, I., Umetsu, K., Harihara, S., Miyoshi, A., Saitou, N., Park, K.S., Dashnyam, B., Jin, F., Lucotte, G., Chattopadhyay, P.K., Henke, L. & Henke, J. (2007) OCA2 481Thr, a hypofunctional allele in pigmentation, is characteristic of northeastern Asian populations. *Journal of Human Genetics*, **52**, 690-693.

APPENDIX A: PREDICTING PHENOTYPE FROM GENOTYPE: NORMAL PIGMENTATION

Predicting Phenotype from Genotype: Normal Pigmentation

Robert K. Valenzuela,[*] M.S. Miquia S. Henderson,[*] B.S. Monica H. Kim,[*] B.S. Nanibaa' A. Garrison,[*2] B.S. Jessica T. Kelch,[*] B.S. Orit Cohen-Barak,[*3] Ph.D. Drew T. Erickson,[*4] Ph.D. F. John Meaney,[*] Ph.D. J. Bruce Walsh,[†] Ph.D. Keith C. Cheng,[††] Ph.D., MD Shosuke Ito,[§] Ph.D. Kazumasa Wakamatsu,[§] Ph.D. Tony Frudakis,[**] Ph.D. Matthew Thomas,[**5] Ph.D. and Murray H. Brilliant,[*1] Ph.D.

[*]Department of Pediatrics, College of Medicine, University of Arizona, Tucson, AZ 85724, USA

[†]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85721, USA

[††]Jake Gittlen Cancer Research Foundation, Department of Pathology; Department of Pharmacology; Department of Biochemistry and Molecular Biology, The Pennsylvania State University College of Medicine, Hershey, PA 17033, USA

[§]Department of Chemistry, Fujita Health University School of Health Sciences, Toyoake, Aichi, Japan

[**]DNAPrint genomics, Inc., Sarasota, FL 34236, USA

[1]Corresponding Author: Murray H. Brilliant, Ph.D.
Phone: 520-626-3305
Fax: 520-626-7407
e-mail: mhb@peds.arizona.edu

Current Addresses:
[2]Nanibaa' A. Garrison: Department of Genetics, Stanford University School of Medicine, Stanford, California 94305, USA

[3]Orit Cohen-Barak: Pharmacology Unit, Teva Pharmaceutical Industries Ltd., Netanya, 42504, Israel

[4]Drew T. Erickson: Helen Wills Neuroscience Institute, University of California, Berkeley, CA 94720, USA

[5]Matthew Thomas: Information Security, Intelligence & Analysis, Sylint Group, Sarasota, Florida, 34230, USA

Running Head:

Normal Human Pigmentation

Corresponding Author:

Murray H. Brilliant, Ph.D.

(mailing address)

1501 N. Campbell Ave.

P.O. Box 245073

Tucson, Arizona/85724

Phone: 520-626-3305

Fax: 520-626-7407

e-mail: mhb@peds.arizona.edu

Funding

National Institute of Justice (2002-1J-CX-K010)

Conflict of interest statement

Dr. Tony Frudakis is and Dr. Matthew Thomas was an employee and shareholder of a company, DNAPrint, which performs forensic DNA analysis.

ABSTRACT

Genetic information in forensic studies is largely limited to CODIS data and the ability to match samples and assign them to an individual. However there are circumstances in which a given DNA sample does not match anyone in the CODIS database and no other information about the donor is available. In this study we determined 75 SNPs in 24 genes (previously implicated in human or animal pigmentation studies) for analysis of single- and multi-locus associations with hair, skin, and eye color in 789 individuals of various ethnic backgrounds. Using multiple linear regression (MLR) modeling, five SNPs in five genes were found to account for large proportions of pigmentation variation in hair, skin, and eyes in our across-population analyses. Thus, these models may be of predictive value to determine an individual's pigmentation type from a forensic sample, independent of ethnic origin.

KEYWORDS:

Forensic Science, Genetics, Human, Pigmentation, QTL, Phenotype

FBI CODIS statistics showed that DNA forensic profiles increased exponentially from 2001-2006 (http://www.fbi.gov/hq/lab/codis/clickmap.htm). However, the DNA forensic profile hits increased linearly, suggesting that the discrepancy between unmatched DNA profiles and hits will continue to increase as the CODIS DNA database increases. As a means of reducing the pool of suspects of an unmatched DNA profile, ancestry informative markers (AIMs) can indicate ethnicity of an unknown sample. However, admixed samples are problematic in that AIMs cannot categorize them into a particular ethnic group. Conceivably, an individual may have a majority of AIMs of one ethnic group, but depending on the amount of admixture their physical appearance may be different than what might be expected.

While AIMs can enable inferences to be made about the ethnicity of an unknown sample, they do not enable identification of an individual based on physical characteristics within an ethnic group, hence as suggested by Tully (1) phenotype prediction based on genetic tests may be a useful tool in forensic analysis. In this study, we assayed genetic markers associated with pigmentation genes (some of which are also AIMs) to determine how much variation in human pigmentation they account for. These markers have been previously studied within specific ethnic groups. We demonstrate that in ethnically-mixed populations these same markers account for a significant fraction of the pigmentation phenotype.

There are two types of melanin, eumelanin (brown/black) and pheomelanin (yellow/red) that differ in sulfur content (2). Variation in human pigmentation results from differences in the type of melanin and amount of melanin synthesized in specialized

vesicles (melanosomes) within pigment cells (melanocytes) and in the size, shape, and export of those melanosomes to the hair and skin (3). Although the rate of synthesis of melanin is much lower in the adult eye than in the skin and hair, additional background color is contained in the iris, making eye color a more complex trait. Genes previously implicated in mediating pigment variation include the melanocortin-1 receptor (*MC1R*) gene and a gene encoding its inhibitor, agouti signaling protein (*ASIP*); two genes associated with oculocutaneous albinism, *P* (*OCA2*) and *SLC45A2* (*OCA4,* formerly named MATP); and most recently, *SLC24A5*, the human orthologue of the zebrafish *golden* gene.

Multiple polymorphisms in the *MC1R* gene have been linked to red hair and fair skin (4-8) and the *ASIP* gene, encoding an inhibitor of the *MC1R* ligand, α-MSH, has been linked to skin pigmentation (9-12). A wide range of OCA phenotypes has been noted for various mutations of *OCA2* (13) and *SLC45A2*, formerly named *MATP* (14-16). Similarly, a wide range of coat color phenotypes is seen in mice with various mutations in their respective orthologous genes (*p*, (17, 18) and *Slc45a2*, formerly named *uw*: (19)). These observations suggested that variations in these genes may be associated with variation in the normal range of human pigmentation. Indeed, gene(s) associated with brown eyes and brown hair were found to map to chromosome 15q, with the *OCA2* gene as a prime candidate (20). Population studies have shown that specific polymorphisms in two genes associated with albinism, the *OCA2* gene and the *SLC45A2* gene, are strongly associated with variations in normal pigmentation of the hair (20-22), skin (22-25), and eyes (22, 26-31).

In addition to its effects on pigmentation variation, the *SLC45A2* polymorphism rs16891982 (L374F) is a useful marker of population origin (32, 33). Another marker of population origin that plays an important role in pigmentation, *SLC24A5* (or NCKX5, the human orthologue of the zebrafish *golden* gene), has been recently identified, with a coding polymorphism divergent between European/Caucasians and other human populations (32, 34). *SLC24A5* has been shown to biologically affect pigmentation in zebrafish, *Danio rerio* (34), murine and cultured human epidermal melanocytes (35).

Alleles of other pigmentation genes such as *tyrosinase-related protein 1* (*TYRP1*) and *dopachrome tautomerase* (*DCT*) have been statistically associated with human iris pigmentation (28). In association with certain alleles of other genes, specific alleles of agouti signaling protein have also been associated with human iris color (28) as well as skin color (9-11).

Previous studies have focused on the effects of a limited number of genes on hair or skin color (4, 5, 24, 34). The most comprehensive studies to date have focused on genome-wide association of SNPs with pigmentation within specific populations (8, 12, 36, 37). In this report, we set out to determine markers predictive for human pigmentation, independent of ethnic origin. We assayed 75 polymorphisms in 24 genes that were previously implicated in human or animal pigmentation studies for analysis of single- and multi-locus associations with hair, skin, and eye color in 789 individuals of various ethnic backgrounds. Multiple linear regression (MLR) modeling revealed that a surprisingly small number of markers account for large proportions of pigmentation variation in hair, skin, and eyes in our across-population analyses.

Materials and methods

Participants

Informed consent was obtained from 791 participants recruited at the University of Arizona between the ages of 18 and 40 years with no gray hair and at least one inch (measured from the roots) of un-dyed scalp hair. Participants in this study roughly mirror the ethnic composition of the student population. Phenotype data, hair samples, and buccal cell samples were collected from each participant following an IRB approved protocol. Participant hair and eye color were independently scored by an investigator; participants also indicated other relevant information, such as tanning response and ethnicity. Buccal cell samples were collected using Catch-All Sample Collection Swabs (Epicentre, Madison, WI) and processed for DNA according to the manufacturer's protocol.

Melanin Analysis

Approximately three hundred scalp hairs (1 cm at the base) were collected from each subject. Hair samples from 186 randomly selected participants were analyzed for both total melanin (combined amount of eumelanin and pheomelanin) content and the two subtypes of melanin, eumelanin and pheomelanin, following a previously published protocol (38, 39). Because not all genotypes were determined for all SNPS for all individuals, subsets of the 186 samples (54-185) were used in generating the multiple linear regression models discussed later.

Skin Reflectance

Skin reflectance was measured as others (40) have with a portable spectrophotometer (Mercury 1000, Datacolor International, Lawrenceville, NJ) fitted with a 15mm aperture. This device measures in the visible light range of 400-700 nanometers, at intervals of 20nm. Three independent reflectance measurements (measured as CIEL, L (lightness) scale of the International Committee on Illumination) of the inner aspect of the upper arm were recorded and averaged for each participant.

Eye color

Eye color was measured by matching subjects' eye color to the Kolberg Iris Color Chart® (ocularistsupplies.com) and recorded. Measurements were binned into six different color categories based on another study study (41) that correlated pigmentation content to color. Categories were binned one through six (1=blue, 2=yellow brown, 3=green, 4=packets of brown + blue/green, 5=brown, and 6=dark brown/black), where bin one corresponded to the least amount of pigmentation, and six corresponded to the highest pigmentation.

Genotyping

Some SNPs were determined by sequencing of PCR products (Fig. 1). In this case, each PCR reaction consisted of the final concentrations/quantities of the following: 1xPCR buffer, 1.5 mM $MgCl_2$, 10pmol of each primer (forward and reverse; Table 1),

0.25 mM dNTPs, 1 U Taq, and ddH_2O to 20 µl. PCR amplification was performed using a PTC-200 Thermal Cycler (MJ Research, Watertown, MA). Thermal cycle program was as follows: 3 min. @ 95˚C, 34 cycles of 30 sec. at each temperature setting (95˚C, 55˚C, and 72˚C), and a final extension of 5min. at 72˚C.

Sequencing was performed by the Genomic Analysis and Technology Core at the University of Arizona on a 3730xl DNA Analyzer (Applied Biosystems, Foster City, CA). Additionally, the first 287 DNA samples collected were genotyped at 75 SNPs in 24 genes implicated in melanin biosynthesis by DNA Print Genomics (Sarasota, FL) following a previously published protocol (27). The remaining 504 individuals were also genotyped by DNA Print Genomics (Sarasota, FL) for 37 of the 75 SNPs that were statistically significant. SNP rs12913832 (HERC2) was genotyped by TaqMan assay C__30724404 (Applied Biosystems).

Mathematical Modeling

Statistical analysis was performed using SAS (version 9.1) and JMP IN (release 5.1) statistical analysis software (SAS Institute, Cary, NC). The pool of SNPs was reduced from 75 (Fig. 1; Table 2) to 40 (within 15 genes) by choosing SNPs that were statistically significant ($p<0.05$, Table 2) by ANOVA (n=287 samples). Finally, rs1426654 (*SLC24A5*), rs6058017 (*ASIP*), and rs12913832 (HERC2) were genotyped for all individuals (n=791 samples). Two samples were dropped from the study (sample 102 and 352) because of bookkeeping inconsistencies. In accordance with standard

statistical procedure, the natural log of the ratio of eumelanin-to-pheomelanin was used to normalize the data.

A total of 40 SNPs (Fig. 1; Table 2) were used to build MLR models of three-SNPs. Initially, forward, reverse, and mixed step-wise regression methods were used to trim the number of SNPs. Different models were obtained based on the method used. To determine SNPs that were most significant in a three-SNP MLR model, we used SAS to generate all possible models of three-SNPs from the pool of 40 significant SNPs. All models were then plotted by squared regression coefficient (R^2) in descending order of value and inflections in the resulting curves were noted. To find the basis of these inflections, we plotted histograms using JMP IN that contained all SNPs that comprised all models up to the R^2 inflections (along the steepest initial slopes). In doing so it became obvious which SNPs were predominantly responsible for the inflections. This method was performed in determining SNPs for each trait.

Results

Our primary aim was to develop a forensic DNA test predictive for pigmentation phenotype. Therefore, we analyzed the data across all populations of the study. The pool of SNPs was reduced from 75 to 40 SNPs by ANOVA ($p<0.05$) (Fig. 1; Table 2). To determine the most likely SNPs for a given trait, all possible combinations of three SNP models (40 choose 3) were generated, sorted by descending R^2 value, and graphed. The three most frequent SNPs found in the range from the highest R^2 model to the first inflection of the graph, were used to construct an MLR model for each trait. These

models accounted for significant variation in each of the four measured traits: scalp-hair total melanin (76.3%), natural log of the ratio of eumelanin-to-pheomelanin (43.2%), skin reflectance (45.7%), and eye color (74.8%); Table 3.

Scalp-hair total melanin

The R^2 curve generated from the three-SNP models showed SNPs rs16891982 (SLC45A2), rs1426654 (SLC24A5), and rs12913832 (*HERC2*) to be the most frequent SNPs (310 or 3.1%) above the first inflection of the R^2 curve (Fig. 2A). These three SNPs yielded an R^2 of 76.3% (n=143 samples).

Natural log of the ratio of eumelanin-to-pheomelanin

The R^2 curve generated from the three-SNP models showed SNPs rs16891982 (SLC45A2), rs1426654 (SLC24A5), and rs1805007 (*MC1R*) to be the most frequent SNPs (430 or 4.4%) above the first inflection of the R^2 curve (Fig. 2B). These three SNPs yielded an R^2 of 43.2% (n=162 samples).

Skin Reflectance

The R^2 curve generated from the three-SNP models showed SNPs rs16891982 (SLC45A2), rs1426654 (SLC24A5), and rs2424984 (*ASIP*) to be the most frequent SNPs (750 or 7.6%) above the first inflection of the R^2 curve (Fig. 2C). These three-SNPs yielded an R^2 value of 45.7% (n=447 samples). An interaction term of rs2424984 (*ASIP*)

and rs16891982 (*SLC45A2*) increased the R^2 value of the model by approximately 4% to 49.6% (n=447 samples).

Eye color

The R^2 curve generated from the three-SNP models showed SNP rs12913832 (HERC2) to be the most frequent SNP (710 or 7.2%) above the first inflection of the R^2 curve (Fig. 2D). rs12913832 yielded an R^2 value of 74.8% (n=484 samples). The next two inflection points showed SNPs rs12913832 (*HERC2*), rs16891982 (SLC45A2), and rs1426654 (*SLC24A5*) to be the most frequent SNPs. These three SNPs yielded an R^2 value of 76.4% (n=353 samples).

Discussion

Using MLR, we have shown that a large proportion of variation in hair, skin, and eye color across diverse human populations can be accounted for by a small number of SNPs. Starting from an initial candidate group of 75 SNPs, we found by ANOVA that 40 of these SNPs were significantly ($p<0.05$) associated with hair pigmentation, skin reflectance, or eye color phenotype. Initially, we trimmed the number of SNPs by forward, backward, and/or a combination of both step-wise regression methods. However, each method yielded a different model. In an attempt to circumvent this problem we modeled all possible combinations of significant SNPs. The R^2 values of the top models differed on average by less than 3-10 thousandths of a percent or smaller. The question arose as to which was the best model. To answer this question we chose

SNPs that were most frequent in the highest R^2 three-SNP-models. This method ultimately trimmed the SNPs to five SNPs (three coding, and two non-coding) in five genes that accounted for most of the variance (76.3% for hair total melanin; 43.2% for hair eumelanin-to-pheomelanin ratio; 45.7% for skin CIEL; and 76.4% for eye color).

Total Hair Melanin and Eumelanin-to-Pheomelanin Ratio

Previous studies have examined the overall color of hair in relationship to various genetic markers. Overall hair color is the result of at least two parameters: total melanin and the ratio of eumelanin-to-pheomelanin. These can be measured objectively by chemical analysis (2). We found that although both parameters are associated with SNPs from *SLC45A2* and *HERC2*, they differ in the third most significant genetic contributor (*SLC24A5* for total hair melanin, and *MC1R* for the ratio of eumelanin-to-pheomelanin).

A high proportion of phenotypic variance of total hair melanin (76.3%) can be accounted for by three-SNPs: rs1426654 (*SLC24A5*), rs16891982 (*SLC45A2*), and rs12913832 (*HERC2*) (Fig. 2A). While *SLC24A5* is an AIM, it plays an important role in pigmentation (34). And in studies carefully controlled for population stratification, the coding SNP rs1426654 is indeed a determinant for normal human pigmentation variation (42). Similarly, rs16891982 (*SLC45A2*) has been shown to be an AIM and has been shown to be associated with pigmentation variation in mice (19) and humans (22, 32) in studies controlled for population stratification (42).

For analysis of total hair melanin, SNP rs12913832 (*HERC2*) was the third most significant contributor in a three-SNP model. Although this SNP lies within the HERC2

gene, it may be part of a promoter region for the adjacent *OCA2* gene (37, 43-46). *OCA2* has been associated with albinism in diverse populations (13), and has also been associated with hair color by linkage analysis (20).

For the natural log of the ratio of eumelanin-to-pheomelanin, a high proportion of phenotypic variance (43.2%) can be accounted for by three SNPs. Two of these are in common with total hair melanin: rs1426654 (*SLC24A5*) and rs16891982 (*SLC45A2*) (Fig. 2B). SNP rs1805007 (*MC1R*) was the third most significant contributor in a three SNP model. *MC1R* has been shown to be a major determinant in whether eumelanin or pheomelanin is produced (21, 47). *MC1R* variants that decrease the protein's functionality have been shown to be associated with an increased incidence of red hair color in humans (4, 5, 48). Chemically, this translates to increased pheomelanin and decreased eumelanin production.

Different studies have analyzed *MC1R* in different ways. Many analyzed specific populations to determine which SNPs were associated with variation in hair color, other studies focused on specific populations and on red hair. Moreover, the statistical analyses employed were different and different variants have been studied. In this study, we analyzed for significance by 1) one-way ANOVA, and 2) a three-SNP model based on frequency of contributors in the models with the highest R^2 values; Table 4.

Although we did not examine all known SNPs of *MC1R*, some of the SNPs we examined have been studied in relationship to red hair color. Valverde *et al.* (1995) found rs2228479 (V60L) in combination with other non-synonymous *MC1R* SNPs in British and Irish to be associated with red hair color. In contrast, two studies (5, 7) did

not find an association of rs2228479 with red hair. Looking across populations of all hair colors, we did not find rs2228479 to be significant by ANOVA. Additionally, Flanagan *et al.* (2000) and Branicki *et al.* (2007) found that rs1805007 (R151G) in combination with rs1805008 (R160W) to be associated with red hair color. Moreover, Sulem *et al.* (2007) found both of these SNPs to be associated with hair color in Icelandic and Dutch populations. In our three-SNP MLR model, rs1805007 (R151G) was the third most important genetic contributor for the ratio of pigmentation in hair color. We note that our analysis of the ratio of hair pigmentation did not focus on any particular hair color.

Skin Reflectance (CIEL)

A high proportion of phenotypic variance of skin reflectance (45.7%) can be accounted for by three-SNPs: rs1426654 (*SLC24A5*), rs16891982 (*SLC45A2*), and rs2424984 (*ASIP*) (Fig. 2C). Studies have examined SNPs rs1426654 (*SLC24A5*) and rs16891982 (*SLC45A2*) and found both to be associated with normal human pigmentation within various ethnic groups and allele frequencies across various ethnic groups (24, 32, 34, 42). Thus, rs1426654 (*SLC24A5*) and rs16891982 (*SLC45A2*) mediate pigmentation variation and they are also AIMs. Lamason *et al*. found SNP rs1426654 to be an AIM in determining European vs. non-European ethnic origin (34). Stokowski *et al.* (2007) controlled for population stratification and showed rs1426654 (*SLC24A5*) to be significantly associated with a dichotomously defined skin-reflectance in South Asians. Although it is an AIM, it does not clearly distinguish between Europeans and Sri Lankans

(32). In contrast, SNP rs16891982 (*SLC45A2*) does distinguish between Europeans and Sri Lankans (32).

The third most significant genetic contributor was SNP rs2424984 (*ASIP*). *ASIP* has been shown to be associated with skin pigmentation (12), namely for rs6058017 (9, 10). Although we found rs6058017 to be significant by ANOVA, we did not find it to be a better predictor in skin reflectance than rs2424984. For both a single SNP and a three-SNP model, rs2424984 was a better predictor for skin reflectance.

Other studies have found *MC1R* (21, 25, 43, 47), *OCA2*, and *DCT* (49) to be associated with normal skin pigmentation. Similarly, we found SNPs within the above genes to be associated by ANOVA with normal skin pigmentation. However, none were found to be significant contributors in a three-SNP MLR model. This does not imply that these genes are not important in pigmentation, it simply means that our method was unable to detect their significance. This may be due to a variety of factors such as sampling error due to small sample size, different populations studied, and/or different methods of measuring skin reflectance.

Eye Color

A high proportion of phenotypic variance of eye color (74.8%) can be accounted for by one SNP rs12913832 (*HERC2*) (Fig. 2D). Previous studies have examined eye color in relationship to various genetic markers. *OCA2* (8, 20, 26, 28, 29), *SLC45A2* (22), and *MC1R* (8, 50) have been statistically associated with variation in eye color.

Duffy et al. (2007) found three-SNPs within intron 1 of *OCA2* that when considered as a haplotype-diplotype explained about 74% of eye color variation. However, there are significant differences between our study and theirs. Among the differences were: populations studied, SNPs genotyped, and binning of eye colors. The population studied was Northern European, whereas, we analyzed across various ethnic populations and we genotyped additional SNPs within genes other than *OCA2*.

Most recently, genome-wide-association studies have shown that intronic SNPs of a gene 5' to OCA2, *HERC2*, have given the highest association with eye color (37, 43, 45). Studies suggest that HERC2 contains a promoter region for OCA2 (37, 43, 45, 46, 51). To date, SNP rs12913832 (*HERC2*) in intron 86, was shown to have the highest association and most likely causative SNP in determining eye color (45), explaining 68% of the variance between blue eyed and brown eye color in an Anglo-Celtic population. Therefore, rs12913832 is likely the causative SNP for European/Caucasians. We note that rs12913832 varies in Caucasian/European populations (where eye color is varied), but only one allele is found in non-Caucasian/European populations (where most non-Caucasian/Europeans have brown/dark eye color). Regardless of whether rs12913832 is causative or an AIM, it is extremely predictive for eye color both within, and across, populations.

A three SNP model of rs12913832 (*HERC2*), rs16891982 (*SLC45A2*), rs1426654 (*SLC24A5*) is only marginally better at explaining the variance (76%) than rs12913832 (*HERC2*) alone. Like, rs12913832 (*HERC2*), SNPs rs16891982 (*SLC45A2*), rs1426654 (*SLC24A5*) are AIMs that distinguish European/Caucasians from non-

European/Caucasians. However, individually rs16891982 (*SLC45A2*) accounts for 38% of eye color variance and rs1426654 (*SLC24A5*) accounts for 34% of eye color variance, significantly less than rs12913832 (*HERC2*) that accounts for 74% of the variance.

Statistical Interactions

Statistical interaction of SNPs rs12910433 (*OCA2*) and rs2228479 (*MC1R*) have been associated with skin reflectance in a Tibetan population (40). We did not find this interaction, however, we did find an interaction between SNPs rs16891982 (*SLC45A2*) and rs2424984 (*ASIP*) for the across population skin reflectance MLR model. These differences may reflect our use of an ethnically-diverse sample; thus, the statistical product of these two genes might be more predictive of skin reflectance across populations. Other interactions, such as between rs12913832 (*HERC2*) and rs1800407 (*OCA2*), may exist in explaining variation in eye color (45), but we cannot confirm this from our data.

Conclusion

We found 5 SNPs in 5 genes that were informative for normal human pigmentation (*SLC45A2*, *SLC24A5*, *MC1R*, *HERC2*, and *ASIP*). Three of these (*SLC45A2*, *SLC24A5* , and *MC1R*) were coding. Clearly, variations within these proteins can result in functional variation that contributes to the pigmentation variation. In contrast, the most significant SNPs found in *OCA2-HERC2* and *ASIP* were not coding. We genotyped most of all known coding SNPs within these genes that showed allele

frequency differences across populations, and none were as significantly correlated to pigmentation as the informative non-coding SNPs. This suggests that the regulation of the expression of *OCA2* and *ASIP* may underlie pigmentation variation. Although our results demonstrate that relatively few SNPs in a relatively few genes control a significant proportion of normal human pigmentation variation, it is certain that additional polymorphisms in these and other genes account for the remaining variation.

Although the polymorphisms that we report are in genes known to regulate pigmentation, some or all of the variance in pigmentation explained by them may be because they are also AIMs. Nevertheless, they are predictive markers for normal human pigmentation variation across various ethnic backgrounds. We note that these models of human constitutive pigmentation phenotype have significant implications for forensic science. These results suggest that assays can be developed, independent of ethnicity, to predict hair, skin, and eye color from DNA samples. Preliminary analysis of an independent sample set (n=261 samples), has validated the predictive utility of these models (Valenzuela et al, in preparation). Additionally, recent genome-wide-association studies have uncovered a variety of other candidate SNPs that contribute to phenotypic variation in pigmentation. These include SNPs rs12896399 (*SLC24A4*), rs12203592 (*IRF4*), rs1540771 (*6p25*.3), and rs35264875 (*TPCN2*) (12, 43, 46). Future studies are needed to evaluate these and other SNPs, as they are uncovered, to further refine these predictive models such that they can accurately predict the pigmentation phenotype of the donors of otherwise unknown forensic samples.

Acknowledgements

We thank Domonique Smith and Benjamin Metelits for their technical help with this project. The work was supported by a grant from the National Institute of Justice.

Additional Information and Reprint Requests

Murray H. Brilliant, Ph.D.

Department of Pediatrics

University of Arizona College of Medicine

1501 N. Campbell Ave.

Tucson, Arizona 85724

E-mail: mhb@peds.arizona.edu

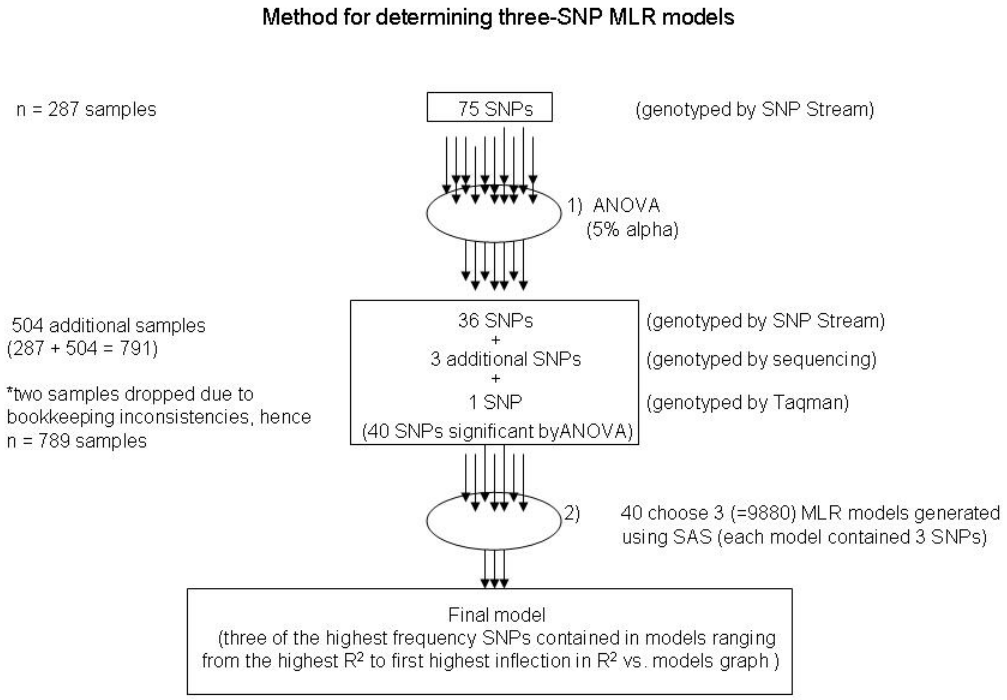

Fig. 1-Flow diagram of method for determining three-SNP MLR models

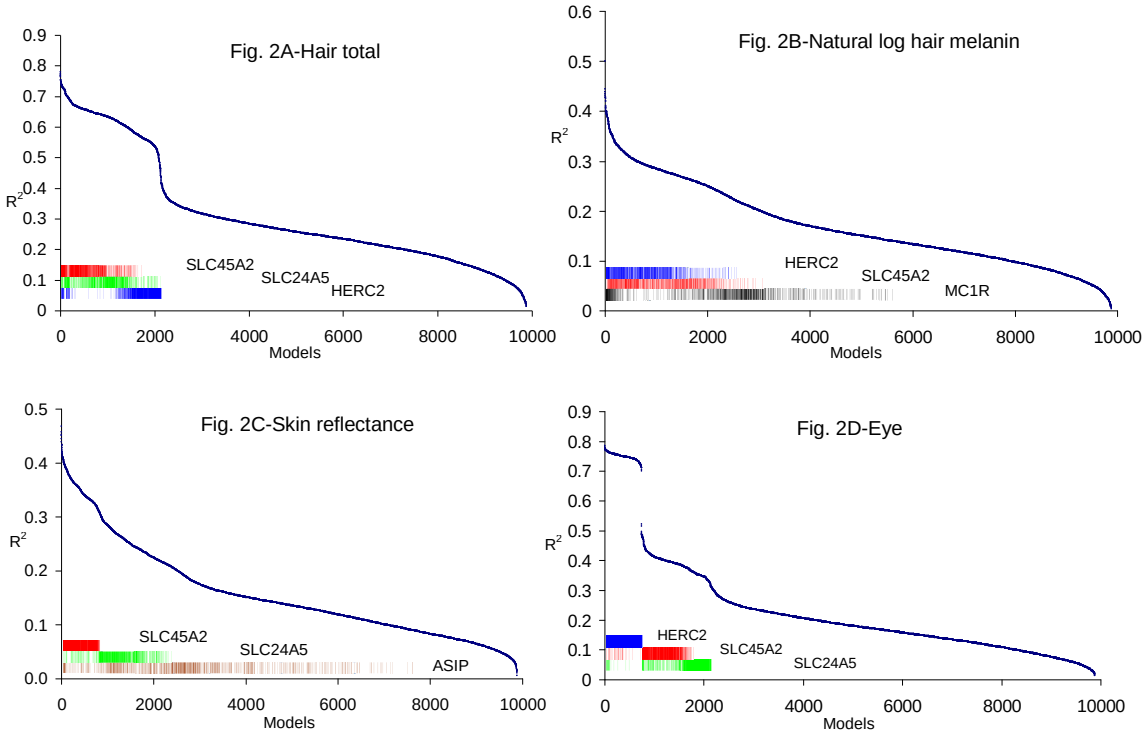

Fig. 2- Three-SNP MLR models for four pigmentation traits across populations. In each panel, the horizontal-axis depicts all 9,880 combinations of significant SNPs in a three-SNP MLR model (i.e., 40 choose 3). The vertical-axis is the R2 value for each model. The three-SNPs that occurred most frequently in the models with the highest R2 values are indicated by the colored bar insets. Each model contained rs16891982 (*SLC45A2*) in red. 2A) Hair total melanin's second and third SNPs were rs1426654 (*SLC24A5*) in green, and rs12913832 (*HERC2*) in blue, model R2 value of 76.3%. 2B) Natural log of hair melanin ratio's second and third SNPs were rs1426654 (*SLC24A5*) and rs1805007 (*MC1R*) in black, model R2 value of 43.2%. 2C) Skin Reflectance's second and third SNPs were rs1426654 (*SLC24A5*) and rs2424984 (*ASIP*) in brown, model R2 value of 45.7%. 2D) Eye color's SNP was rs12913832 (*HERC2*) in blue, model R2 value of 74.8%. The second and third SNPs rs16891982 (*SLC45A2*) and rs1426654 (*SLC24A5*) only marginally increased the model R2 value to 76.4%.

Table 1-Primers used to amplify *OCA2* exons 9 and 13, and sequences containing SNPs rs1426654 and rs6058017.

rs1800401, and rs1800402 (*OCA2*)	Primer Name	Primer sequence
forward	MHB581	(5'-GGGCTGAATTGTTCCATTTG-3')
reverse	MHB582	(5'-TCTCACGGATCTCAAGCCTC-3')
sequencing	MHB583	(5'-TGGCTGATACAGAGGGAGGT-3')
rs1800407 (*OCA2*)		
forward	MHB593	(5'-GGCTCCCTGTTCTTAAAGTC-3')
reverse	MHB594	(5'-TGAGAATGGAACCTGGAGCC-3')
sequencing	MHB595	(5'-AAGGCTGCCTCTGTTCTACG-3')
rs1426654 (*SLC24A5*)		
forward	MHB1755	(5'-TCATAAAGAAGCAAAACATTGGA-3')
reverse	MHB1756	(5'-AGCACAGATGCCAAGGAGAT-3')
sequencing	MHB1757	(5'-TGCCAATATCTCCCTTTGTG-3')
rs6058017 (*ASIP*)		
forward	MHB1836	(5'-GGCTTCGATGAAGAAAGTGG-3')
reverse	MHB1837	(5'-AGGGGATAGCCTCGTTCCTA-3')
sequencing	MHB1838	(5'-ATGGGACTTCAGGGAGACCT-3')

TABLE 2- 75 SNPs within 24 genes.

Gene	SNP	Total hair melanin	Natural log of ratio of melanins	Skin Reflectance (CIEL)	Eye Color
AP3D1	rs2238600				
AP3D1	rs4806830				
ASIP	rs2424984	*	*	*	*
ASIP	rs6058017	*	*	*	*
ASIP	rs2296151				
CYP2C8	rs1341164				
CYP2D6	rs1058174				
CYP4B1	rs1572603			*	
DCT	rs1325611	*	*	*	*
DCT	rs1407995	*	*	*	*
GPR143	rs3044	*	*	*	*
GPR143	rs3810741				
GPR143	rs2521667				
HERC2	rs12913832	*	*	*	*
HPS1	rs1804689				
HPS1	rs2296430				
HPS1	rs3830015				
HPS3	rs2689234				
HPS4	rs1894704		*	*	
HPS4	rs3752589				*
HPS4	rs3752590				*
HPS4	rs739289		*		
HPS5	rs2305564				
HPS6	rs4917959				
LYST	rs3768051				
MAOA	rs979605				
MC1R	rs1805007		*	*	
MC1R	rs1805008	*	*	*	*
MC1R	rs3212346	*	*		
MC1R	rs3212355			*	
MC1R	rs3212357	*	*	*	*
MC1R	rs3212370				
MC1R	rs3212368				
MC1R	rs3212366				
MC1R	rs3212363				
MC1R	rs3212364				
MC1R	rs2228479				
MC1R	rs1805005				
MC1R	rs3212353				
MC1R	rs3212352				
MC1R	rs3212351				
MLPH	rs2292885	*		*	
MLPH	rs1343768				
MYO18A	rs11080078				*
MYO5A	rs1724630			*	
MYO5A	rs752864			*	
MYO5A	rs2290332	*	*	*	
MYO7A	rs2276289	*		*	
MYO7A	rs3737454	*		*	
MYO7A	rs948962				
MYO7A	rs1320703				
MYO7A	rs2276288				
OCA2	rs1037208			*	
OCA2	rs10852218		*	*	
OCA2	rs11638265	*		*	*
OCA2	rs1800404	*	*	*	*
OCA2	rs1800407	*		*	
OCA2	rs1800410	*	*	*	*
OCA2	rs1800411	*	*	*	*
OCA2	rs1800414	*			
OCA2	rs1900758	*	*	*	*
OCA2	rs749846	*	*		*
OCA2	rs1800402				
OCA2	rs737051				
OCA2	rs1800415				
OCA2	rs2305253				
OCA2	rs1800401				
OCA2	rs2044627				
SLC24A5	rs1426654	*	*	*	*
SLC45A2	rs16891982	*	*	*	*
SLC45A2	rs2287949		*	*	*
SLC45A2	rs26722	*	*	*	*
SLC45A2	rs40132	*	*	*	*
TYRP1	rs2733832	*	*	*	*
TYRP1	rs683				

* Significant SNPs by ANOVA ($p<0.05$) (40 SNPs within 16 genes) with respect to each of the four phenotypes for the first 287 samples.

Note: rs3212363 (*MC1R*) was genotyped for all individuals, it was not significant across all populations.

TABLE 3- Across populations R^2 values of individual SNPs, interaction terms, full model, and sample size of each ethnic group (self-reported).

	Individual R^2 values % (sample size; p value)	Caucasian	African-American	Hispanic	South Asian	East Asian	Native American	Admixture	not listed
Scalp-Hair Total Melanin									
rs16891982 (*SLC45A2*)	63.7 (n=172, p=7.2x10-38)	102	18	14	4	10	5	19	
rs1426654 (*SLC24A5*)	62.3 (n=166, p=2.76x10-35)	101	18	11	2	8	5	21	
rs12913832 (*HERC2*)	54.0 (n=176, p=6.69x10-30)	108	16	13	4	10	5	20	
Model	76.3 (n=143)	86	14	10	2	8	5	18	
ln(Eumelanin/Pheomelanin)									
rs16891982 (*SLC45A2*)	24.7 (n=172, 3.88x10-11)	102	18	14	4	10	5	19	
rs12913832 (*HERC2*)	26.5 (n=176, p=2.79x10-12)	108	16	13	4	10	5	20	
rs1805007 (*MC1R*)	14.0 (n=186, p=1.41x10-7)	113	19	14	4	10	5	21	
Model	43.2 (n=162)	97	15	13	4	10	5	18	
Average Skin Reflectance									
rs16891982 (*SLC45A2*)	32.1 (n=487, p=2.28x10-41)	267	19	83	23	42	14	30	9
rs1426654 (*SLC24A5*)	23.9 (n=528, p=7.09x10-32)	296	20	88	22	41	15	33	13
rs2424984 (*ASIP*)	15.6 (n=538, p=2.18x10-20)	300	20	89	25	44	15	32	13
Model	45.7 (n=447)	244	19	74	20	39	14	28	9
Model+rs16891982*rs2424984	49.6 (n=447)								
Eye Color									
rs12913832 (*HERC2*)	74.8 (n=484, p=1.09x10-144)	253	19	70	39	48	15	28	12
rs16891982 (*SLC45A2*)	38.4 (n=397, p=3.55x10-42)	191	21	56	35	47	13	24	10

rs1426654 (*SLC24A5*)	34.3 (n=490, p=3.64x10-45)	260	21	69	36	46	15	30	13
Model	76.4 (n=353)	171	16	49	32	43	13	20	9

TABLE 4- ANOVA results for *MC1R* SNPs with respect to hair total melanin and natural log of eumelanin-to-pheomelanin.

MC1R SNP	Total melanin (From first 287 individuals	n	ratio	Total melanin (From full data set)	n	ratio
rs3212357	†	141	†	†	219	†
rs3212346	*	138	†	†	191	†
rs1805008	*	141	†	§	218	†
rs1805007	§	141	†	†	219	†
rs3212355	§	137	§	†	214	§
rs3212370	‡	141	‡			
rs3212368	‡	141	‡			
rs3212364	‡	141	‡			
rs3212352	‡	141	‡			
rs3212351	‡	141	‡			
rs3212366	§	137	§			
rs3212353	§	136	§			
rs1805005	§	124	§			
rs2228479	§	122	§			

* = $p<0.05$
† = $p<0.01$
‡ = no variation
§ = $p>0.05$

References

1. Tully G: Genotype versus phenotype: Human Pigmentation. Forensic Sci Intl: Genetics 2007;1:105-10.
2. Wakamatsu K, Ito S: Advanced chemical methods in melanin determination. Pigment Cell Res 2002;15:174-83.
3. Nordlund JJ, Ortonne J. The normal color of human skin. In: Nordlund JJ, Boissy RE, Hearing VJ, King RA, Oetting WS, Ortonne J, editors. The pigmentary system. second edition. Malden: Blackwell Publishing; 2006. p. 504-20.
4. Valverde P, Healy E, Jackson I, Rees JL, Thody AJ: Variants of the melanocyte-stimulating hormone receptor gene are associated with red hair and fair skin in humans. Nat Genet 1995;11:328-30.
5. Flanagan N, Healy E, Ray A, Philips S, Todd C, Jackson IJ, et al.: Pleiotropic effects of the melanocortin 1 receptor (MC1R) gene on human pigmentation. Hum Mol Genet 2000;9:2531-7.
6. Rees JL: The genetics of sun sensitivity in humans. Am J Hum Genet 2004;75:739-51.
7. Branicki W, Brudnik U, Kupiec T, Wolanska-Nowak P, Wojas-Pelc A: Determination of phenotype associated SNPs in the MC1R gene. J Forensic Sci 2007;52:349-54.

8. Sulem P, Gudbjartsson DF, Stacey SN, Helgason A, Rafnar T, Magnusson KP, et al.: Genetic determinants of hair, eye and skin pigmentation in Europeans. Nat Genet 2007.

9. Bonilla C, Boxill LA, Donald SA, Williams T, Sylvester N, Parra EJ, et al.: The 8818G allele of the agouti signaling protein (ASIP) gene is ancestral and is associated with darker skin color in African Americans. Hum Genet 2005;116:402-6.

10. Kanetsky PA, Swoyer J, Panossian S, Holmes R, Guerry D, Rebbeck TR: A polymorphism in the agouti signaling protein gene is associated with human pigmentation. Am J Hum Genet 2002;70:770-5.

11. Voisey J, Gomez-Cabrera Mdel C, Smit DJ, Leonard JH, Sturm RA, van Daal A: A polymorphism in the agouti signalling protein (ASIP) is associated with decreased levels of mRNA. Pigment Cell Res 2006;19:226-31.

12. Sulem P, Gudbjartsson DF, Stacey SN, Helgason A, Rafnar T, Jakobsdottir M, et al.: Two newly identified genetic determinants of pigmentation in Europeans. Nat Genet 2008.

13. Oetting WS, King RA: Molecular basis of albinism: mutations and polymorphisms of pigmentation genes associated with albinism. Hum Mutat 1999;13:99-115.

14. Inagaki K, Suzuki T, Shimizu H, Ishii N, Umezawa Y, Tada J, et al.: Oculocutaneous albinism type 4 is one of the most common types of albinism in Japan. Am J Hum Genet 2004;74:466-71.

15. Newton JM, Cohen-Barak O, Hagiwara N, Gardner JM, Davisson MT, King RA, et al.: Mutations in the human orthologue of the mouse underwhite gene (uw) underlie a new form of oculocutaneous albinism, OCA4. Am J Hum Genet 2001;69:981-8.

16. Rundshagen U, Zuhlke C, Opitz S, Schwinger E, Kasmann-Kellner B: Mutations in the MATP gene in five German patients affected by oculocutaneous albinism type 4. Hum Mutat 2004;23:106-10.

17. Johnson DK, Stubbs LJ, Culiat CT, Montgomery CS, Russell LB, Rinchik EM: Molecular analysis of 36 mutations at the mouse pink-eyed dilution (p) locus. Genetics 1995;141:1563-71.

18. Lyon MF, King TR, Gondo Y, Gardner JM, Nakatsu Y, Eicher EM, et al.: Genetic and molecular analysis of recessive alleles at the pink-eyed dilution (p) locus of the mouse. Proc Natl Acad Sci U S A 1992;89:6968-72.

19. Sweet HO, Brilliant MH, Cook SA, Johnson KR, Davisson MT: A new allelic series for the underwhite gene on mouse chromosome 15. J Hered 1998;89:546-51.

20. Eiberg H, Mohr J: Assignment of genes coding for brown eye colour (BEY2) and brown hair colour (HCL3) on chromosome 15q. Eur J Hum Genet 1996;4:237-41.

21. Sturm RA, Teasdale RD, Box NF: Human pigmentation genes: identification, structure and consequences of polymorphic variation. Gene 2001;277:49-62.

22. Graf J, Hodgson R, van Daal A: Single nucleotide polymorphisms in the MATP gene are associated with normal human pigmentation variation. Hum Mutat 2005;25:278-84.

23. Duffy DL, Box NF, Chen W, Palmer JS, Montgomery GW, James MR, et al.: Interactive effects of MC1R and OCA2 on melanoma risk phenotypes. Hum Mol Genet 2004;13:447-61.

24. Nakayama K, Fukamachi S, Kimura H, Koda Y, Soemantri A, Ishida T: Distinctive distribution of AIM1 polymorphism among major human populations with different skin color. J Hum Genet 2002;47:92-4.

25. Shriver MD, Parra EJ, Dios S, Bonilla C, Norton H, Jovel C, et al.: Skin pigmentation, biogeographical ancestry and admixture mapping. Hum Genet 2003;112:387-99.

26. Duffy DL, Montgomery GW, Chen W, Zhao ZZ, Le L, James MR, et al.: A three-single-nucleotide polymorphism haplotype in intron 1 of OCA2 explains most human eye-color variation. Am J Hum Genet 2007;80:241-52.

27. Frudakis T, Terravainen T, Thomas M: Multilocus OCA2 genotypes specify human iris colors. Hum Genet 2007.

28. Frudakis T, Thomas M, Gaskin Z, Venkateswarlu K, Chandra KS, Ginjupalli S, et al.: Sequences associated with human iris pigmentation. Genetics 2003;165:2071-83.

29. Rebbeck TR, Kanetsky PA, Walker AH, Holmes R, Halpern AC, Schuchter LM, et al.: P gene as an inherited biomarker of human eye color. Cancer Epidemiol Biomarkers Prev 2002;11:782-4.

30. Sturm RA, Frudakis TN: Eye colour: portals into pigmentation genes and ancestry. Trends Genet 2004;20:327-32.

31. Zhu G, Evans DM, Duffy DL, Montgomery GW, Medland SE, Gillespie NA, et al.: A genome scan for eye color in 502 twin families: most variation is due to a QTL on chromosome 15q. Twin Res 2004;7:197-210.

32. Soejima M, Koda Y: Population differences of two coding SNPs in pigmentation-related genes SLC24A5 and SLC45A2. Int J Legal Med 2007;121:36-9.

33. Yuasa I, Umetsu K, Harihara S, Kido A, Miyoshi A, Saitou N, et al.: Distribution of the F374 allele of the SLC45A2 (MATP) gene and founder-haplotype analysis. Ann Hum Genet 2006;70:802-11.

34. Lamason RL, Mohideen MA, Mest JR, Wong AC, Norton HL, Aros MC, et al.: SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. Science 2005;310:1782-6.

35. Ginger RS, Askew SE, Ogborne RM, Wilson S, Ferdinando D, Dadd T, et al.: SLC24A5 encodes a trans-Golgi network protein with potassium-dependent sodium-calcium exchange activity that regulates human epidermal melanogenesis. J Biol Chem 2008;283:5486-95.

36. Stokowski RP, Krishna Pant PV, Tony D, Fereday A, Hinds DA, Jarman C, et al.: A Genomewide Association Study of Skin Pigmentation in a South Asian Population. Am J Hum Genet 2007;81.

37. Kayser M, Liu F, Janssens AC, Rivadeneira F, Lao O, van Duijn K, et al.: Three genome-wide association studies and a linkage analysis identify HERC2 as a human iris color gene. Am J Hum Genet 2008;82:411-23.

38. Ito S, Wakamatsu K: An improved modification of permanganate oxidation of eumelanin that gives a constant yield of pyrrole-2,3,5-tricarboxylic acid. Pigment Cell Res 1994;7:141-4.

39. Wakamatsu K, Ito S, Rees JL: The usefulness of 4-amino-3-hydroxyphenylalanine as a specific marker of pheomelanin. Pigment Cell Res 2002;15:225-32.

40. Akey JM, Wang H, Xiong M, Wu H, Liu W, Shriver MD, et al.: Interaction between the melanocortin-1 receptor and P genes contributes to inter-individual variation in skin pigmentation phenotypes in a Tibetan population. Hum Genet 2001;108:516-20.

41. Wakamatsu K, Hu DN, McCormick SA, Ito S: Characterization of melanin in human iridal and choroidal melanocytes from eyes with various colored irides. Pigment Cell Melanoma Res 2008;21:97-105.

42. Stokowski RP, Pant PV, Dadd T, Fereday A, Hinds DA, Jarman C, et al.: A genomewide association study of skin pigmentation in a South Asian population. Am J Hum Genet 2007;81:1119-32.

43. Sulem P, Gudbjartsson DF, Stacey SN, Helgason A, Rafnar T, Magnusson KP, et al.: Genetic determinants of hair, eye and skin pigmentation in Europeans. Nat Genet 2007;39:1443-52.

44. Lehman AL, Silvers WK, Puri N, Wakamatsu K, Ito S, Brilliant MH: The underwhite (uw) locus acts autonomously and reduces the production of melanin. J Invest Dermatol 2000;115:601-6.

45. Sturm RA, Duffy DL, Zhao ZZ, Leite FP, Stark MS, Hayward NK, 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 2008;82:424-31.

46. Han J, Kraft P, Nan H, Guo Q, Chen C, Qureshi A, et al.: A genome-wide association study identifies novel alleles associated with hair color and skin pigmentation. PLoS Genet 2008;4:e1000074.

47. Rees JL: Genetics of hair and skin color. Annu Rev Genet 2003;37:67-90.

48. Beaumont KA, Shekar SN, Cook AL, Duffy DL, Sturm RA: Red hair is the null phenotype of MC1R. Hum Mutat 2008.

49. Lao O, de Gruijter JM, van Duijn K, Navarro A, Kayser M: Signatures of positive selection in genes associated with human skin pigmentation as revealed from analyses of single nucleotide polymorphisms. Ann Hum Genet 2007;71:354-69.

50. Kanetsky PA, Ge F, Najarian D, Swoyer J, Panossian S, Schuchter L, et al.: Assessment of polymorphic variants in the melanocortin-1 receptor gene with cutaneous pigmentation using an evolutionary approach. Cancer Epidemiol Biomarkers Prev 2004;13:808-19.

51. Lehman AL, Nakatsu Y, Ching A, Bronson RT, Oakey RJ, Keiper-Hrynko N, et al.: A very large protein with diverse functional motifs is deficient in rjs (runty, jerky, sterile) mice. Proc Natl Acad Sci U S A 1998;95:9436-41.

APPENDIX B: PREDICTION MODEL VALIDATION: NORMAL HUMAN PIGMENTATION VARIATION

Prediction Model Validation: Normal Human Pigmentation Variation

Robert K. Valenzuela[ab], Shosuke Ito[c], Kazumasa Wakamatsu[c], and Murray H. Brilliant[ab†]

[a]Department of Pediatrics, College of Medicine, University of Arizona, Tucson, AZ 85724, USA.

[b]Center for Human Genetics, Marshfield Clinic Research Foundation, Marshfield, WI 54449, USA.

[c]Department of Chemistry, Fujita Health University School of Health Sciences, Toyoake, Aichi, Japan.

Robert K. Valenzuela: rkv1@email.arizona.edu

Shosuke Ito: sito@fujita-hu.ac.jp

Kazumasa Wakamatsu: kwaka@fujita-hu.ac.jp

[†]Corresponding Author: Tel.: +1 715-207-9493; Fax: +1 715-389-4950

E-mail: Brilliant.Murray@mcrf.mfldclin.edu

Word count: 158 abstract; 4206 text; 14 references; 11 figures; 3 tables

Running Head: Model Validation: Normal Human Pigmentation

Abstract

In a past study, we developed multiple linear regression (MLR) models that employed three single nucleotide polymorphisms (SNPs) that predicted a significant proportion of variation in pigmentation phenotypes from a large population cohort (n=789, training sample). Multiple linear regression models were developed for skin reflectance, eye, and two aspects of hair color (log of the ratio of eumelanin-to-pheomelanin and total melanin). In this report, using an independent cohort (n=242 , test sample), we 1) externally cross-validated the prediction models, and 2) tested and refined the algorithm presented in the study by Valenzuela and colleagues, (2010). Relative shrinkage was moderate for skin reflectance (23.4%), eye color (19.4%), and the log of the ratio of eumelanin-to-pheomelanin in hair (37.3%), and largest for total melanin (67%) in hair. Independent construction of predictive models using our algorithm for the test sample set yielded the same or similar models as the training sample set. Two of the three SNPs composing the models were the same, with some variability in the third SNP of the model.

Keywords: Forensic Science, Genetics; Human; Pigmentation; Prediction Models; QTL

Introduction

According to the Federal Bureau of Investigation (FBI) Laboratory's Combined DNA Index System (CODIS) – National DNA Index System (NDIS) statistics (http://www.fbi.gov/hq/lab/codis/clickmap.htm), there are significantly more unmatched profiles than there are matched profiles. Ancestry informative markers (AIMs) can be helpful in reducing the pool of suspects. However, a more efficient means of reducing a pool of suspects is to predict an unmatched profile's phenotype based on their genetic information. Forensically informative phenotypes include skin, eye, and hair color. The appearances of these traits are largely influenced by pigmentation, which is a quantitative trait controlled by many genetic loci.

To make prediction models, genetic variants associated with the trait of interest must first be found. To avoid spurious associations due to population stratification, genetic variants are typically determined within a single population. While this approach is beneficial in avoiding spurious associations, the variants found may not be significantly associated with the trait of interest in other populations due to differing evolutionary histories. For instance, studies have provided evidence that suggest that the fairer skin of East Asians and Europeans is due to different genetic variants within genes associated with pigmentation – an example of convergent evolution (1-3).

In our previous paper, we addressed the problem of constructing forensic models for skin, eye, and hair color by developing models using an ethnically diverse sample. Moreover, we presented a novel algorithm to determine the prediction models. To determine the performance of the models we developed (4), in this report we have

externally cross-validated the pigmentation prediction models by using an independent and ethnically diverse sample (test sample). We also corroborated the results of this algorithm (ie, the models determined by the training sample) by applying our method to the test sample. Finally, we refined the algorithm and present a procedure that allows dynamic analysis of various SNP models and their R^2-curve inflections. This dynamic analysis allows us to observe the individual components (SNPs) of each possible model. This has facilitated the identification of more robust genetic models for describing pigmentation variation across various ethnicities.

Materials and Methods

Phenotype data, hair samples, and buccal cell samples were collected from each participant following Institutional Review Board approval of the protocol. Participants, phenotype measurements, and mathematical modeling have been described elsewhere (4), with the exception of the hair-melanin chemical analysis. The test sample's hair eumelanin chemical analysis was performed using a minor variation (an alkaline peroxide oxidation method rather than the acidic permanganate oxidation method that was used for the training sample) of the chemical analysis described elsewhere (5). Briefly, sample homogenate (100 µL) was taken in a 10 ml screw-capped conical test tube, to which 375 µL 1 mol/L K_2CO_3 and 25 µL 30% H_2O_2 (final concentration: 1.5%) were added (6), and then mixed vigorously at room temperature for 20 hr. The residual H_2O_2 was decomposed by the addition of 50 µL 10% Na_2SO_3, and the mixture was then acidified with 140 µL 6 mol/L HCl. After vortex-mixing, the reaction mixture was centrifuged at 4,000 g for 1

min, and an aliquot (80 µL) of the supernatant was directly injected into the HPLC system. H_2O_2 oxidation products were analyzed with the HPLC system consisting of a JASCO 880-PU liquid chromatograph (JASCO Co., Tokyo, Japan), a Shiseido C_{18} column (Shiseido Capcell Pak MG; 4.6 x 250 mm; 5 µm particle size) and a JASCO UV detector. The mobile phase was 0.1 mol/L potassium phosphate buffer (pH 2.1): methanol, 99: 1 (v/v). Analyses were performed at 45˚C at a flow rate of 0.7 mL/min. Absorbance of the eluent was monitored at 269 nm. The results of the two different chemical analysis methods have been shown to be highly correlated (unpublished data, Wakamatsu *et al.*). The test sample's data was transformed to match the training sample's data. SNPs were genotyped using the SNPlex™ Genotyping System (Applied Biosystems).

Design parameters of the SNPlex™ Genotyping System did not allow all of the significant SNPs of Valenzuela *et al*. (4) study to be genotyped, and additional SNPs that have been subsequently shown to be associated with human pigmentation were genotyped, so that the genotyping between the training and test sets was not identical. However, the test set was typed for the most significant SNPs from the training sample. The relationship between the SNPs genotyped for the training and test samples are illustrated in Figure 1.

The inflection-point method for choosing the SNPs of the prediction models presented in Valenzuela *et al.* (4) was refined by assigning the highest R^2-value model (all models sorted in descending values of R^2) a value of 1, and each subsequent model

was numbered consecutively (i=1…$\begin{pmatrix} total-number-of-SNPs \\ number-of-SNPs-in-model \end{pmatrix}$). For a given SNP, if it was present in a given model, then it was assigned a value of one, otherwise, it was assigned a value of zero (let presence/absence be called state). A function was chosen that weights the state of a given SNP more heavily for the highest R^2 values as compared to lower R^2 values and such that lower R^2-value models were dependant on higher R^2-value models. The preliminary function was as such,

$$f(i) = \frac{\sum_{i=1}^{i} state_i}{i}$$

When all MLR models are generated all SNPs are equally represented, hence each SNP is present a constant number of times. More precisely, if *x* is the pool size of independent variables, and *y* is the number of independent variables in a model, then each independent variable is represented a constant=$\frac{(x-1)!}{(x-y)!(y-1)!}$ number of times when $\begin{pmatrix} x \\ y \end{pmatrix}$ models are generated. Consequently, all independent variable functions must attain a value of $\frac{y}{x}$ at the last row number. Additionally, a value of $\frac{cons\tan t}{row}$ is attained once an independent variable has reached full representation.

In general, an increasing presence of a SNP as a function of row number was reflected in its SNP-curve as a positive slope; in contrast, its diminishing presence was

reflected in its SNP-curve as a negative slope (Figure 2). As a function of increasing row number, the faster a SNP exhausts its representation (ie, approaches its asymptote), then the more important its contribution is to higher R^2 models (this is one way to view "prominent" contributors). Also, the greater the presence of an independent variable at higher R^2 values, then the more "important" it is as a contributor. If a SNP was present at each consecutive row, beginning from the first row, then the slope of its SNP-curve was zero, with a function value of one until its non-presence in a model.

Two parameters were varied in R^2-SNP curve generation/comparison for each trait: sample (training sample/test sample), and model size (3 SNPs/2 SNPs). SNP curves of different R^2 curves were compared by varying one parameter while the other parameter was held fixed.

Statistical Analysis

Cross validation was performed by taking the difference in R^2 values of the training and test samples (i.e., shrinkage = $R^2_{training} - R^2_{test}$). Relative shrinkage was calculated by taking the ratio of shrinkage to the training sample's R^2 (i.e., 1- $\frac{R^2_{test}}{R^2_{training}}$). All R^2 values were calculated by using the beta estimates of the training sample. Statistical values and models were calculated by using SAS (version 9.1) and JMP (release 8.0) statistical software packages (SAS Institute, Cary, North Carolina). All plots were graphed using R statistical freeware package (version 2.10.1) (7).

Results

To determine the predictive ability of the models generated by Valenzuela *et al*. (4), we externally cross-validated the models. Ethnic composition of the external sample set is listed in table 1. External cross-validation was performed by taking the difference, or shrinkage, of corresponding R^2 values of each trait for each sample (Table 2). The R^2 values of the test sample were calculated by using the beta estimates derived from the training sample set's prediction models (Table 3).

We also tested the algorithm presented in Valenzuela *et al*. (4) by applying the algorithm to each sample set and comparing the results for each corresponding trait for each sample set. We generated three- and two-SNP R^2 curves (ie, 29-choose-3 and 29-

choose-2, respectively) from which we determined three-SNP prediction models. All possible combinations of models were generated by using a pool of 29 SNPs (Figure 1) that were common to both sample sets and were found to be significant in each sample set by one-way ANOVA ($p < 0.05$; Table 4). We also generated SNP curves (see Materials and Methods) so that we could compare curves of a given trait between samples.

External Validation

Skin Reflectance: The model derived from the training sample set for the average skin reflectance was composed of SNPs rs16891982 (*SLC45A2*), rs1426654 (*SLC24A5*), and rs2424984 (*ASIP*); together they yielded an R^2 value of 45.7% (n=447). Applying this model's beta estimates to the test sample set yielded an R^2 value of 35.0% (n=186); hence, the shrinkage was 10.7%, with a relative shrinkage of 23.4%.

Applying the algorithm to the training sample set and the test sample set, both the three- (row inflection 395, Figure 3; and row inflection 219 Figure 4) and two-SNP R^2 curves resulted in the same three SNPs: rs16891982 (*SLC45A2*), rs1426654 (*SLC24A5*), and rs2424984 (*ASIP*). The corresponding SNP curves between the two sample sets were similar. In particular, inflections in the SNP curve of rs16891982 (*SLC45A2*) were often mirrored by inflections in the SNP curve of rs1426654 (*SLC24A5*) indicating that for many of the high R^2 models, either one or the other of the two SNPs was present. The exhaustion of rs16891982 (*SLC45A2*) resulted in a noticeable inflection in all skin reflectance R^2 curves.

Eye Color: The model derived from the training sample set for eye color was composed of SNPs rs12913832 (*HERC2*), rs16891982 (*SLC45A2*), and rs1426654 (*SLC24A5*); together they yielded an R^2 value of 76.4% (n=353). Applying this model's beta estimates to the test sample yielded an R^2 value of 61.6% (n=204); hence, the shrinkage was 14.8%, with a relative shrinkage of 19.4%.

Applying the algorithm to the training sample set and the test sample set, both the three- (row inflection 438, Figure 5; and row inflection 464, Figure 6) and two-SNP R^2 curves resulted in the same three SNPs: rs12913832 (*HERC2*), rs16891982 (*SLC45A2*), and rs1426654 (*SLC24A5*). The corresponding SNP curves between the two sample sets were similar. In both sample sets, the SNP curve of rs12913832 (*HERC2*) was the highest frequency SNP until exhaustion, marked by a major inflection of the R^2 curve. The order that the SNP curves of rs16891982 (*SLC45A2*) and rs1426654 (*SLC24A5*) reached exhaustion varied between sample sets.

Eumelanin-to-Pheomelanin Ratio: The model derived from the training sample set for the natural logarithm of the ratio of eumelanin-to-pheomelanin was composed of SNPs rs16891982 (*SLC45A2*), rs12913832 (*HERC2*), and rs1805007 (*MC1R*); together yielding an R^2 value of 43.2% (n=162). Applying this model's beta estimates to the test sample yielded an R^2 value of 27.1%; hence, the shrinkage was 16.1%, with a relative shrinkage of 37.3%.

Applying the algorithm to the training sample set, the three-SNP R^2 curve of the training sample set resulted in the three SNPs (row inflection 162, Figure 7): rs12913832 (*HERC2*), rs16891982 (SLC45A2), and rs1805007 (*MC1R*). The analogous inflection of

the three-SNP R^2 curve of the test sample set resulted in the three SNPs (row inflection 149; Figure 8): rs12913832 (*HERC2*), rs16891982 (SLC45A2), and rs1426654 (*SLC24A5*). Closer inspection of the test sample's three-SNP R^2 curve revealed an inflection at row 35 (Figure 9). The three highest frequency SNPs at row 35 inflection were the same as the training sample set's at row 162: rs12913832 (*HERC2*), rs16891982 (SLC45A2), and rs1805007 (*MC1R*). Inflections in the SNP curve of rs12913832 (*HERC2*) were often mirrored by inflections in the SNP curve of rs16891982 (*SLC45A2*) for all R^2 curves; however, the SNP curves varied substantially between sample sets.

Total Hair Melanin: The model derived from the training sample set for hair total melanin was composed of SNPs rs16891982 (*SLC45A2*), rs1426654 (*SLC24A5*), and rs12913832 (*HERC2*); together yielding an R^2 value of 76.3% (n=143). Applying this model's beta estimates to the test sample yielded an R^2 value of 25.2% (n=164); hence, the shrinkage was 51.1% with a relative shrinkage of 67.0%.

Applying the algorithm to the training sample set, both the three- and two-SNP R^2 curves resulted in the same three SNPs (row inflection 180, Figure 10): rs16891982 (*SLC45A2*), rs1426654 (*SLC24A5*), and rs12913832 (*HERC2*). However, applying the algorithm to the test sample set resulted in four SNPs (row inflection 398, Figure 11) rs16891982 (*SLC45A2*), rs1426654 (*SLC24A5*), rs12913832, and rs1800404 (OCA2); the latter two SNPs were of equal frequency. The test sample set's two-SNP R^2 curve resulted in SNP rs16891982 (SLC45A2); all other SNPs were of equal frequency. The SNP curves of rs16891982 (*SLC45A2*) and rs1426654 (*SLC24A5*), and consequently, corresponding R^2 curves, varied considerably between the samples. In the training sample

set, rs16891982 (*SLC45A2*) was present in fewer high-R^2 models as compared to the test sample. However, their inflections were mirror images of each other in both sample sets.

Discussion

In this report, using an independent test sample (n=242) we externally cross-validated the pigmentation prediction models derived from the training sample (n=789) that we presented in Valenzuela *et al*. (4). The relative shrinkage was modest for skin reflectance (23.4%), eye color (19.4%), and the ratio of eumelanin-to-pheomelanin of hair (37.3%), but was largest for hair total melanin (67.0%). We also refined the model building algorithm we presented in Valenzuela *et al*. (4) by adding SNP curves (see Materials and Methods) and tested the model building algorithm by applying it to both the training and test samples. The SNP curves gave us a better understanding of the behavior of the most prominent SNPs with respect to the R^2 curve inflections and in relationship to each other. We determined three-SNP models as we did in Valenzuela *et al*. (4), from three- and two-SNP models. In doing so, we found that the same third most prominent SNP, as was determined in Valenzuela *et al*. (4), could often be determined from the two-SNP model R^2 curves. Applying the algorithm to each sample set resulted in the same two SNPs, with variability in the third SNP, when comparing between sample sets for a given trait (total melanin, eumelanin-to-pheomelanin ratio, skin reflectance, and eye color).

Skin Reflectance

The shrinkage result for skin reflectance was 10.7%, with a relative shrinkage of 23.4%. We used the same method in measuring skin reflectance in the training and test samples, and therefore the modest shrinkage was within the range of our expectation.

The algorithm yielded the same three SNPs in both sample sets: rs16891982 (*SLC45A2*), rs1426654 (*SLC24A5*), and rs2424984 (*ASIP*). The mirror-like behavior of rs16891982 (*SLC45A2*) and rs1426654 (*SLC24A5*) was likely a result of their correlation (chi-square test, $correlation_{training}$=244.733; $correlation_{test}$=115.302).

Eye Color

The shrinkage result for eye color was 14.8% with a relative shrinkage of 19.4%. We used the same method in measuring eye color in the training and test samples, and therefore, the modest shrinkage was within the range of our expectation.

The algorithm yielded the same three SNPs in both sample sets: rs12913832 (*HERC2*), rs16891982 (*SLC45A2*), and rs1426654 (*SLC24A5*). The SNP curves were similar in behavior between samples. However, in the training sample rs16891982 (SLC45A2) reached exhaustion before rs1426654 (SLC24A5) did; whereas in the test sample, rs1426654 (SLC24A5) reached exhaustion first. The variability in SNP curves may be the result of experimental error in measurement, as we used an eye chart to record eye color, and/or it could be due to sampling error.

Eumelanin-to-Pheomelanin Ratio

The shrinkage result for eumelanin-to-pheomelanin ratio was 16.1% with a relative shrinkage of 37.3%, substantially less than the shrinkage result of hair total melanin. The shrinkage was probably less pronounced for the natural log of the ratio of eumelanin-to-pheomelanin simply because taking the natural log of the ratio of eumelanin-to-pheomelanin decreased the spread or variation for values greater than unity (all ratios were greater than unity). Therefore, the variation in the different chemical analyses would be more easily detected in total melanin.

The algorithm yielded the same three common SNPs in both sample sets (for both the three- and two-SNP R^2 curves): rs12913832 (*HERC2*), rs1805007 (*MC1R*), and rs16891982 (*SLC45A2*). However, in the test sample, SNPs rs16891982 (*SLC45A2*), and rs1426654 (*SLC24A5*) were of equal frequency for the "third" SNP (two-SNP R^2 curve). In comparing the SNP curves of rs16891982 (*SLC45A2*) and rs1426654 (*SLC24A5*) in the test sample, rs16891982 (*SLC45A2*) was clearly more prominent than rs1426654 (*SLC24A5*) after the major inflection. The mirror-like behavior of rs12913832 (*HERC2*) and rs16891982 (*SLC45A2*) was likely a result of their correlation (chi-square test, $correlation_{training}$=74.8; $correlation_{test}$=66.8).

Total Hair Melanin

The shrinkage result for total hair melanin was 51.1% with a relative shrinkage of 67.0%. This was likely the result of using different chemical analysis methods for the training sample and the test sample. Although the chemical analysis methods of eumelanin were highly correlated, the variance in results of the two chemical analysis

methods increased at higher melanin values (personal correspondence, Wakamatsu *et al*.). If the difference in chemical analysis was indeed the reason for the variation in the models, then as expected the models predicting capabilities were highly contingent upon the method of data ascertainment.

The algorithm yielded the same three SNPs in both sample sets: rs16891982 (*SLC45A2*), rs1426654 (*SLC24A5*), and rs12913832 (*HERC2*). However, there was a marked difference in the behavior of SNP curve rs16891982 (*SLC45A2*) between samples. In the training sample, SNP curve rs16891982 (*SLC45A2*) slope varied between positive and negative, indicating that it was present in many, but not all, of the highest R^2 models; whereas in the test sample, the SNP curve of rs16891982 (*SLC45A2*) had a constant slope of zero, indicating that it was present in all of the highest R^2 models. The variance in SNP curve rs16891982 (*SLC45A2*) between samples was also reflected in the R^2 curves. The variability was most likely attributable to the different chemical analysis methods used between samples. The mirror-like behavior of rs16891982 (*SLC45A2*) and rs1426654 (*SLC24A5*) was likely a result of their correlation (Pearson's chi-square test; df=4; $correlation_{training}$=124.288; $correlation_{test}$=114.991). However, although the prominent SNP curves varied between samples, their relationship within a sample set remained unchanged, consequently supporting the argument that differences in algorithm results were due to the different chemical analysis methods.

Extending the comparison of determining the three most prominent SNPs from the two-SNP R^2, we also compared the SNPs determined by our algorithm to the three most significant SNPs, as determined by one-way ANOVA. We found that the third most

prominent SNP, as determined from either two- or three-SNP R^2-curves, were not always the same as the three most significant SNPs as determined by one-way ANOVA. In particular, the third most prominent SNP of the skin reflectance model, as determined by the algorithm, was rs2424984 (*ASIP*). However, as determined by one-way ANOVA, it was the fourth most statistically significant SNP, while rs12913832 (*HERC2*) was the third most statistically significant SNP (training sample). Similarly, the third most prominent SNP of the natural log of the ratio of eumelanin-to-pheomelanin model, as determined by the algorithm, was rs1805007 (*MC1R*). However, as determined by one-way ANOVA, rs1805007 (*MC1R*) was the fourth most prominent SNP, while rs1426654 (*SLC24A5*) was the third most prominent SNP (training sample).

To determine if differential-missing SNP data could be attributed to the non-correspondence of the third SNP between the algorithm and ranking by one-way ANOVA, we selected the 10 most significant SNPs as determined by one-way ANOVA and removed all individuals with missing genotype information, such that all SNPs contributed the same amount of genetic information in all models. Applying the algorithm yielded the same prominent SNPs for skin reflectance (training sample). Interestingly, however, one-way ANOVA of the non-missing data set yielded a different ranking of the SNPs, such that rs2424984 (*ASIP*) was the seventh most significant SNP rather than the fourth most significant SNP, as was the case in the missing genotype data set.

Conclusion

The results demonstrate the utility of our algorithm for consistently selecting the same independent variables of a given trait for building prediction models. Additionally, the refinement of our algorithm, by adding curves of each independent variable (SNP curves), gave us insight into the behavior of prominent SNPs as a function of row number in relationship to inflections of the R^2 curve, to each other (namely, covariance/co-inheritance), and between samples. Moreover, by comparing the algorithm results of two- and three-SNP R^2-curves, we found that the third most prominent SNP, as determined by the two-SNP R^2-curve, was often the same third SNP as determined by the three-SNP R^2-curve. Our results suggest that the third most prominent SNP may be inferred from the two-SNP R^2-curve. We note that a weakness to our algorithm (4) is that SNPs not significant by one-way ANOVA are excluded from the analysis; therefore, significant genetic interactions of non-significant single SNPs, as detected by the method presented by Akey *et al.* (8) may not be detected.

Our skin reflectance model had a relatively low R^2 value (45.7%, training sample set) and a relative shrinkage of 23.4% when applied to the test sample set. The shrinkage was modest, hence, our model does have forensic utility; however, additional SNPs, not present in our pool of SNPs, likely will account for additional phenotypic variability. We have chosen SNPs that have been previously associated with macroscopic measurements of mouse/human pigmentation. To determine additional genetic associations of the macroscopic measurement (skin reflectance), microscopic (and perhaps chemical analysis) measurements are likely necessary. For example, Szabo *et al*. (9) showed that

morphological differences exist in melanosome structure for various ethnicities. Conceivably, microscopic differences of pigment granules, or other differences, exist within ethnicities as well. Our measurements did not take into account these microscopic measurements, nor has any other study of which we are aware. Microscopic resolution may be necessary to determine SNPs that account for additional variation in skin reflectance. In other words, statistically significant genetic signals may be lost by grouping objects of similar macroscopic measurement that differ microscopically. Accounting for the genetic variations associated with these microscopic differences may enable development of models with increased predictive capabilities with relatively few SNPs.

Similarly, for eye color, we took macroscopic measurements. However, in contrast to skin color, our model had a relatively high R^2 value (76.4%, training sample) and a relative shrinkage of 19.4% when applied to the test sample. The shrinkage was modest, thus forensically useful, suggesting that much of the variation in eye color is determined by relatively few SNPs, and that the SNPs from our SNP pool captured that variation. More precise measurements (10; 11) of intermediate eye colors are necessary in order to determine associated genetics signals, and therefore, to develop a prediction model that accurately describes intermediate eye color.

In contrast, our hair melanin models are based on a sub-phenotype (melanin) of hair color. Our ratio of eumelanin-to-pheomelanin model had a relatively low R^2 value (43.2%, training sample set) and a relative shrinkage of 37.3% when applied to the test sample set. In contrast, our total hair melanin model had a relatively high R^2 value

(76.3%, training sample set) and a relative shrinkage of 67.0% when applied to the test sample set. As we mentioned in the discussion, the large shrinkage in total melanin model was likely due to the different chemical analyses. Less shrinkage was likely observed in the ratio of eumelanin-to-pheomelanin model because of the natural log transformation of the data. Because of the different chemical methodologies employed, we cannot determine whether our hair models are forensically useful or not. Clearly, the efficacy of the model is highly contingent upon the method of measurement. Had the same chemical analyses been performed in both the training and test samples, and had validation resulted in a modest to relatively small shrinkage, our models would be predictive but not forensically useful in their present form. For our hair melanin models to be forensically useful, a mathematical relationship between melanin content and hair color needs to be established. Although hair color is largely influenced by melanin content, other (sub-quantitative) traits, such as hair thickness (12) and rate of growth, likely contribute to hair color.

Our initial choice of SNPs may be one of the reasons for the low R^2 value (43.2%) of the hair ratio of melanins (training data set). We chose SNPS from genes that have previously been associated with pigmentation; however, the ratio of melanins may be governed by other genes that are not detectable when eumelanin and pheomelanin are measured as a sum, but may be detectable when measured as a ratio. This is not surprising as the ratio of melanin, to our knowledge, has not been investigated at this level of detail and associated with genetic variants on a genome-wide scale. However,

our choice of SNPs did enable development of the total melanin model that had a relatively high R^2 value.

Our model building method, as with any model building method, strives to develop robust prediction models. These models are merely a starting place to *predict* normal human pigmentation variation, independent of ethnic origin. Other studies have developed prediction models for eye, skin, and hair color (13-18). However, with the exception of the study by Spinochek et al. (2010), these studies trained their models utilizing a population of exclusively of European descent. Not surprisingly, their models are lacking a major melanin associated SNP, rs1426654 (SLC24A5). Studies have indicated the functional efficacy of rs1426654 in human epidermal melanocytes (19). The 111T allele of rs1426654 is nearly fixed in populations of European descent (20), and therefore is classified as an AIM (ancestry informative marker). Hence, not only is rs1426654 an AIM, evidence suggest that it is a functional AIM for pigmentation.

While models produced utilizing a population of one ethnic background may predict a low error-rate for samples of that same background, their performance may be limited to their training population. Whereas, the approach we and others (Spichenok et al., 2010) have utilized may have more realistic predictive outcomes for a heterogeneous population such as the United States population. Various ethnicities comprise the United States population, including a significant percentage of admixed populations. The African American population contains approximately 20-30% admixture of mainly European, and some Native American, ancestry (21). Similarly, Hispanics also show varying degrees of admixture of Native American, European and African descent. Both

variants of SNP rs1426654 are present at high frequency (and not fixed) in these admixed populations, highlighting the necessity to take into consideration this SNP in forensic prediction. Utilizing a training sample of various ethnic compositions has enabled us to develop prediction models independent of ethnicity. Finally, in building prediction models using a population of diverse ethnic backgrounds allows for the possibility of converging phenotypes to be modeled (given that the 'convergent' SNPs are tested).

The SNPs we have chosen for our models explain a fair to a large amount of the phenotypic variation in our test sample comprised of various ethnicities. However, other forms of genetic variation not examined here (ie, rare variants, copy number variants, inversions, and/or combinations of variations) are likely present and may contribute to additional phenotypic variation.

Conflict of interest statement

The authors have no actual or potential conflict of interest to declare.

Acknowledgements

We thank Dr. JB Walsh (Department of Ecology & Evolutionary Biology University of Arizona, Tucson, Arizona, USA), Dr. LJ Baier (Diabetes Molecular Genetics Section Phoenix Epidemiology and Clinical Research Branch (NIDDK NIH), Phoenix, Arizona, USA), Dr. YL Muller (Diabetes Molecular Genetics Section Phoenix Epidemiology and Clinical Research Branch (NIDDK NIH) Phoenix, Arizona, USA), Dr. O. Cohen-Barak (Pharmacology Unit, Teva Pharmaceutical Industries Ltd., Netanya,

Israel), Dr. DT Erickson (Ernest Gallo Clinic and Research Center, UCSF, San Francisco, California, USA), Christine Klassen (College of Medicine, University of Arizona, Tucson, Arizona, USA), Rongji Chen (University of Arizona, Tucson, Arizona, USA), Jason Fabian (University of Arizona, Tucson, Arizona, USA), and Justin Garrison (University of Arizona, Tucson, Arizona, USA) for their technical help with this project. We also would like to thank Dr. JB Walsh (University of Arizona, Tucson, Arizona, USA) Dr. RP Erickson (University of Arizona, Tucson, Arizona, USA) and Dr. SJ Shrodi (Center for Human Genetics, Marshfield Clinic Research Foundation, Marshfield, Wisconsin, USA) for reviewing this manuscript prior to submission. We also thank the Marshfield Clinic Research Foundation's Office of Scientific Writing and Publication for editorial assistance of this manuscript.

Funding

The work was supported by a grant from the National Institute of Justice (2002-1J-CX-K010)

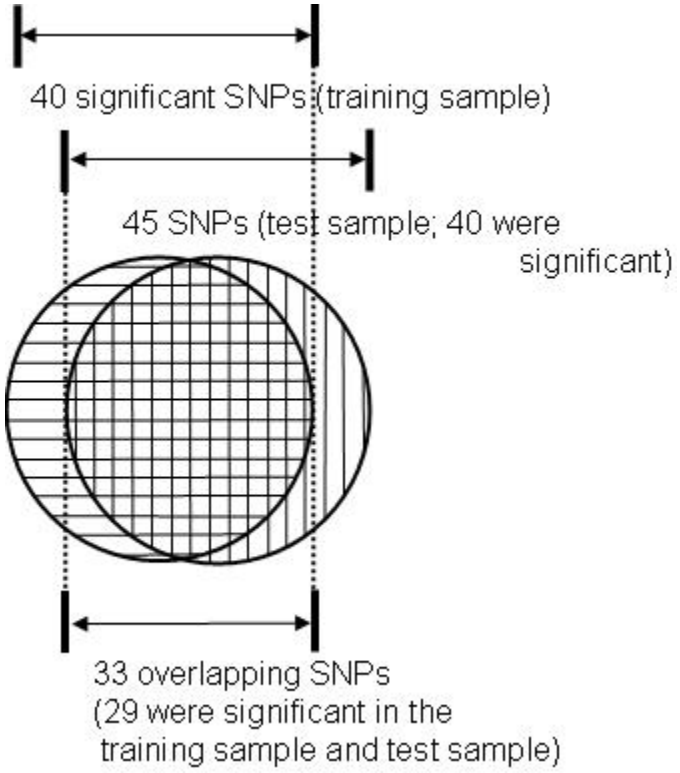

Figure 1. Relationship between SNP pools for the training and the test samples.
Note: Significance was determined by ANOVA (alpha < 0.05).

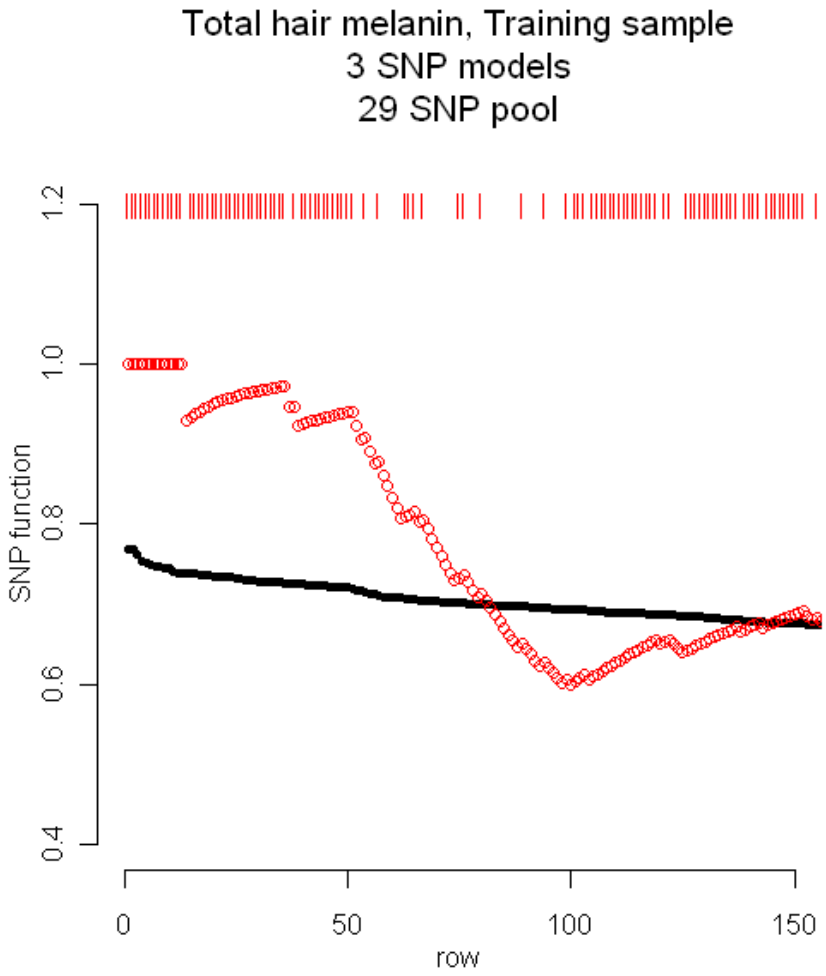

Figure 2. SNP-function (red curve) behavior in relationship to presence/absence (red tics) of SNP rs16891982 (*SLC45A2*) in a three SNP model. The R2 curve is black.

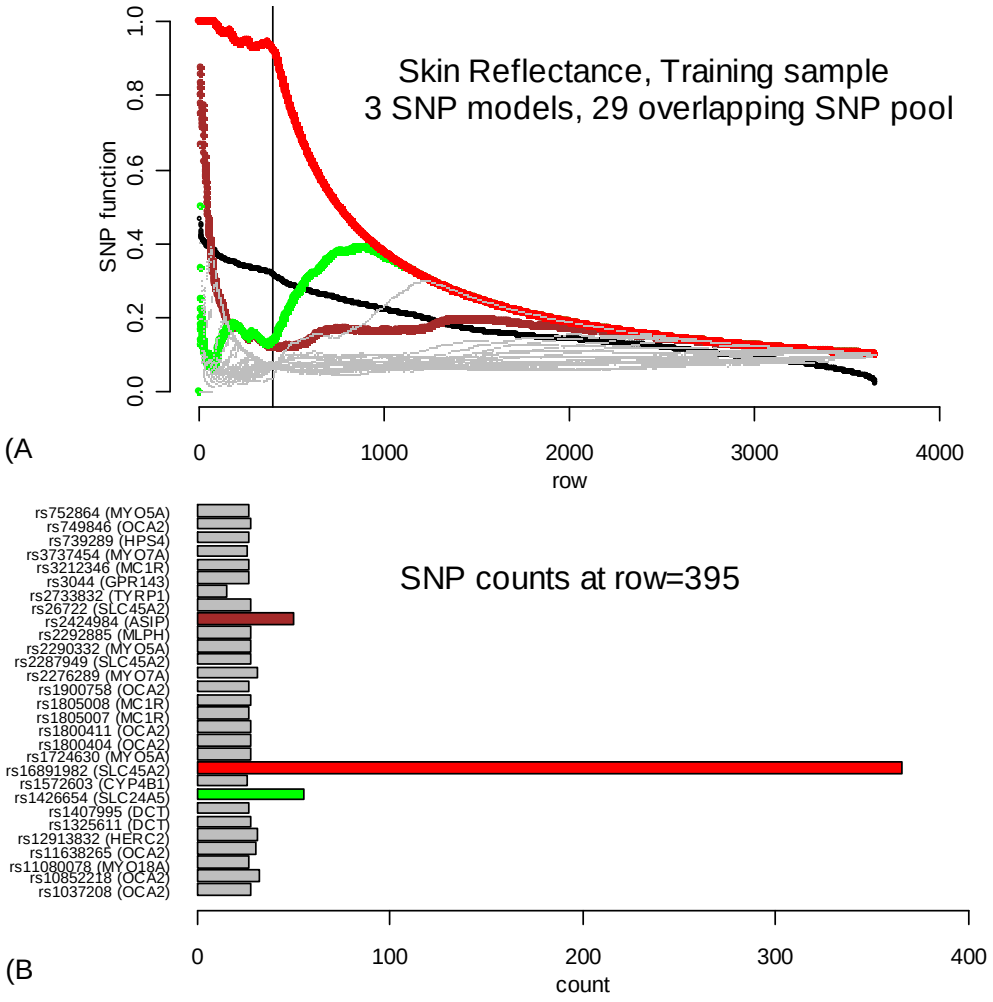

Figure 3. (**A**) Three-SNP multiple linear regression (MLR) models for skin reflectance across populations (training sample set). The horizontal-axis depicts all 3654 combinations (ie, 29-choose-3) of significant SNPs in a three-SNP MLR model. The vertical-axis is the R2 value for each model (black curve or three-SNP R2 curve) and also the SNP function value for each SNP curve. The R2 curve inflection (row 395) is indicated by a vertical black line. The SNP curves of the three highest frequency SNPs at row 395 are indicated by colors (rs16891982 (*SLC45A2*), red; rs1426654 (*SLC24A5*), green; rs2424984 (*ASIP*), brown). (**B**) Bar plot of the SNPs that were present in all models from row 1 to row 395.

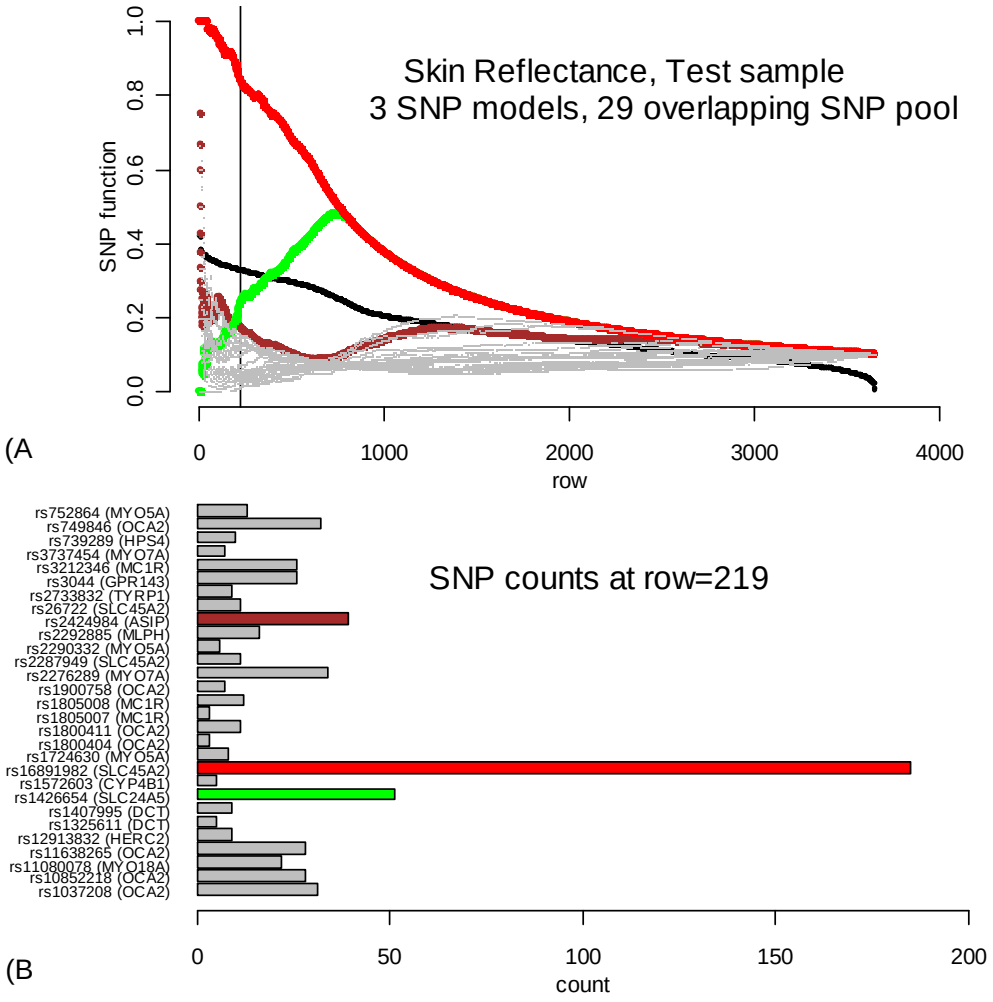

Figure 4. (**A**) Three-SNP multiple linear regression (MLR) models for skin reflectance across populations (test sample set). The horizontal-axis depicts all 3654 combinations (ie, 29-choose-3) of significant SNPs in a three-SNP MLR model. The vertical-axis is the R2 value for each model (black curve or three-SNP R2 curve) and also the SNP function value for each SNP curve. The R2 curve inflection (row 395) is indicated by a vertical black line. The SNP curves of the three highest frequency SNPs at row 219 are indicated by colors (rs16891982 (*SLC45A2*), red; rs1426654 (*SLC24A5*), green; rs2424984 (*ASIP*), brown). (**B**) Bar plot of the SNPs that were present in all models from row 1 to row 219.

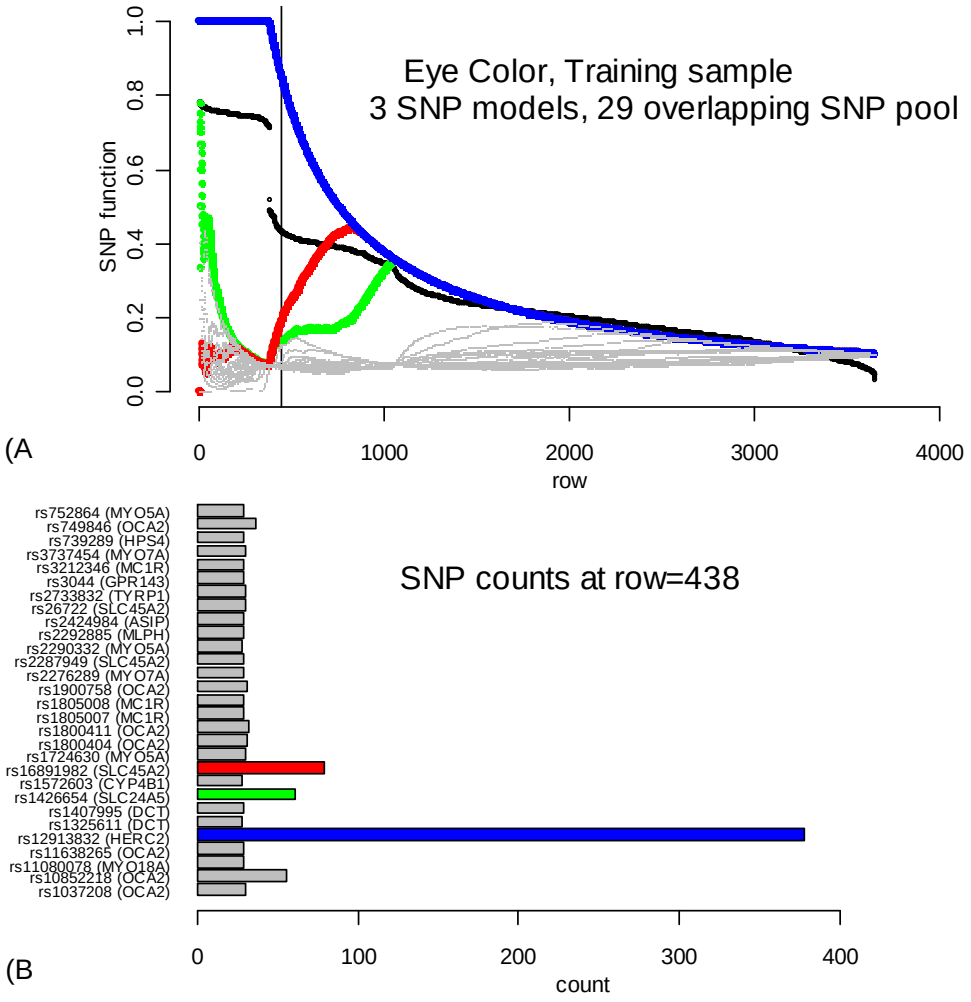

Figure 5. (**A**) Three-SNP multiple linear regression (MLR) models for eye color across populations (training sample set). The horizontal-axis depicts all 3654 combinations (ie, 29-choose-3) of significant SNPs in a three-SNP MLR model. The vertical-axis is the R2 value for each model (black curve or three-SNP R2 curve) and also the SNP function value for each SNP curve. The R2 curve inflection (row 438) is indicated by a vertical black line. The SNP curves of the three highest frequency SNPs at row 438 are indicated by colors (rs12913832 (*HERC2*), blue; rs16891982 (*SLC45A2*), red; rs1426654 (*SLC24A5*), green). (**B**) Bar plot of the SNPs that were present in all models from row 1 to row 438.

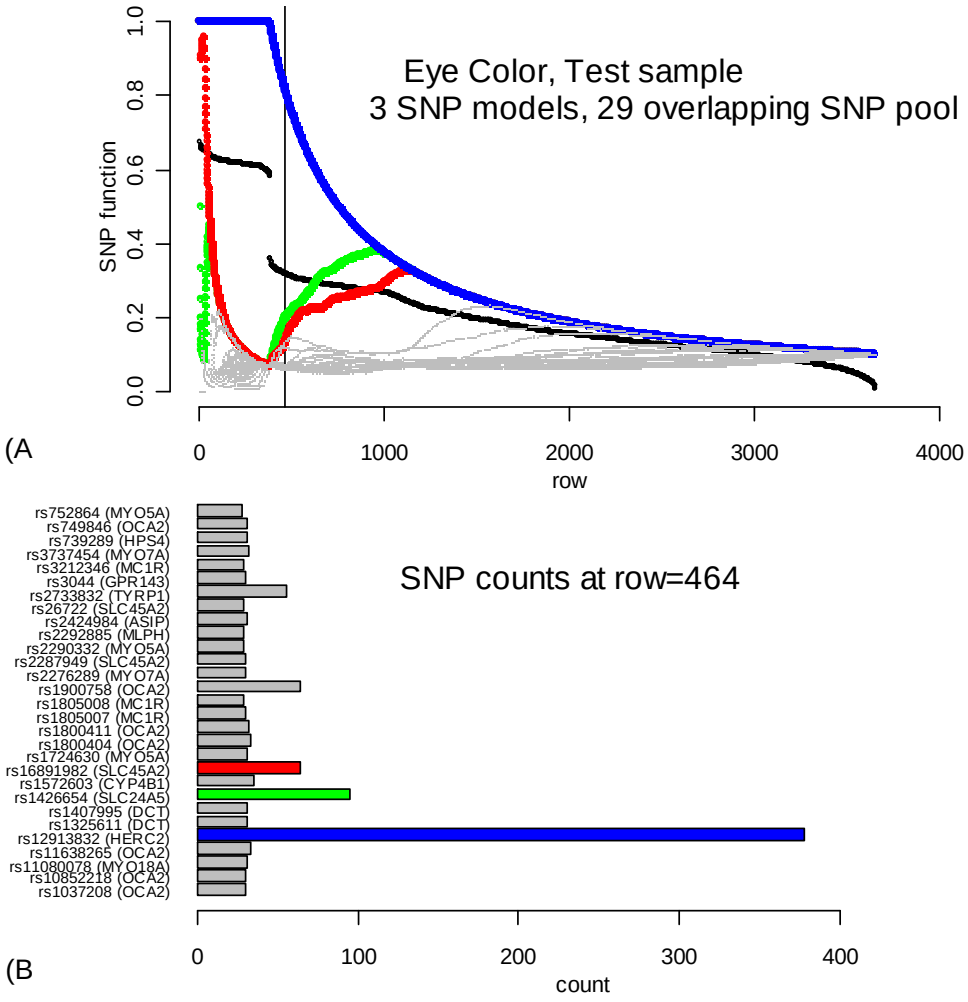

Figure 6. (**A**) Three-SNP multiple linear regression (MLR) models for eye color across populations (test sample set). The horizontal-axis depicts all 3654 combinations (ie, 29-choose-3) of significant SNPs in a three-SNP MLR model. The vertical-axis is the R2 value for each model (black curve or three-SNP R2 curve) and also the SNP function value for each SNP curve. The R2 curve inflection (row 464) is indicated by a vertical black line. The SNP curves of the three highest frequency SNPs at row 464 are indicated by colors (rs12913832 (*HERC2*), blue; rs16891982 (*SLC45A2*), red; rs1426654 (*SLC24A5*), green). (**B**) Bar plot of the SNPs that were present in all models from row 1 to row 464.

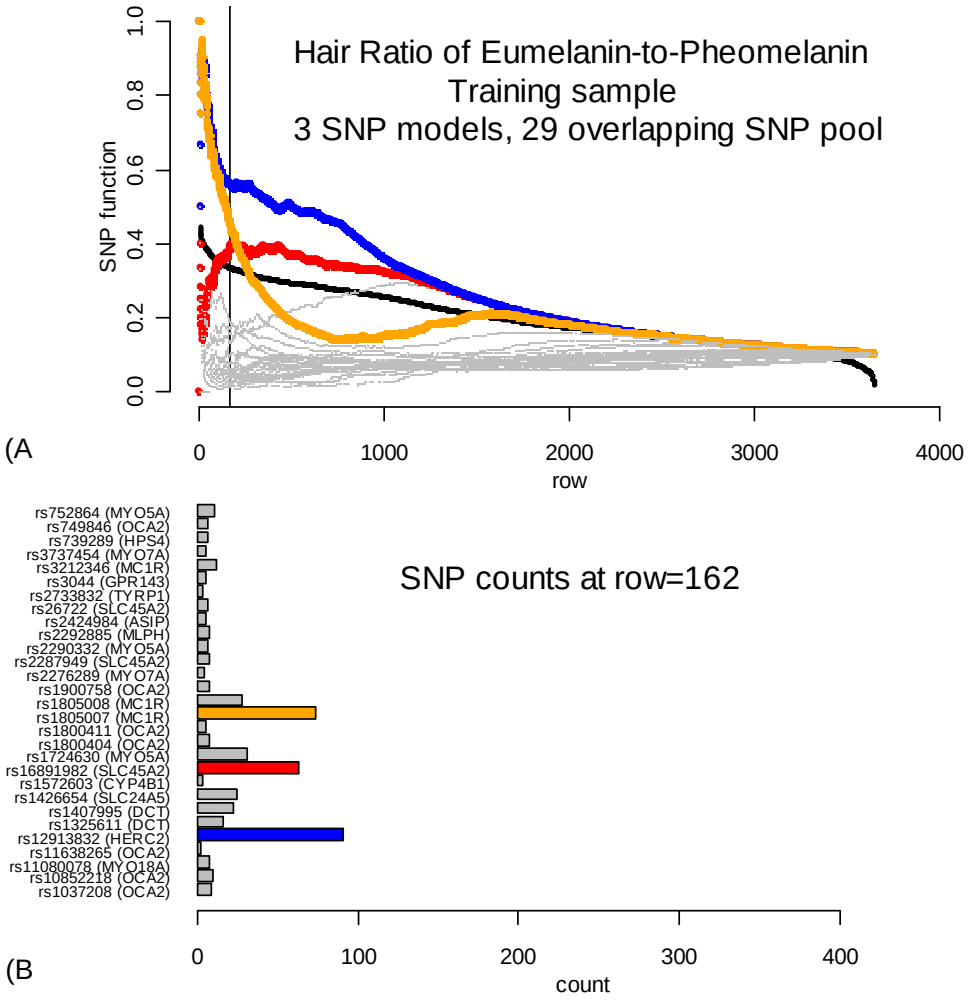

Figure 7. (**A**) Three-SNP multiple linear regression (MLR) models for natural log of hair melanin ratio's across populations (training sample set). The horizontal-axis depicts all 3654 combinations (ie, 29-choose-3) of significant SNPs in a three-SNP MLR model. The vertical-axis is the R2 value for each model (black curve or three-SNP R2 curve) and also the SNP function value for each SNP curve. The R2 curve inflection (row 162) is indicated by a vertical black line. The SNP curves of the three highest frequency SNPs at row 162 are indicated by colors (rs12913832 (*HERC2*), blue; rs1805007 (*MC1R*), orange ; rs16891982 (*SLC45a2*), red).
(**B**) Bar plot of the SNPs that were present in all models from row 1 to row 162.

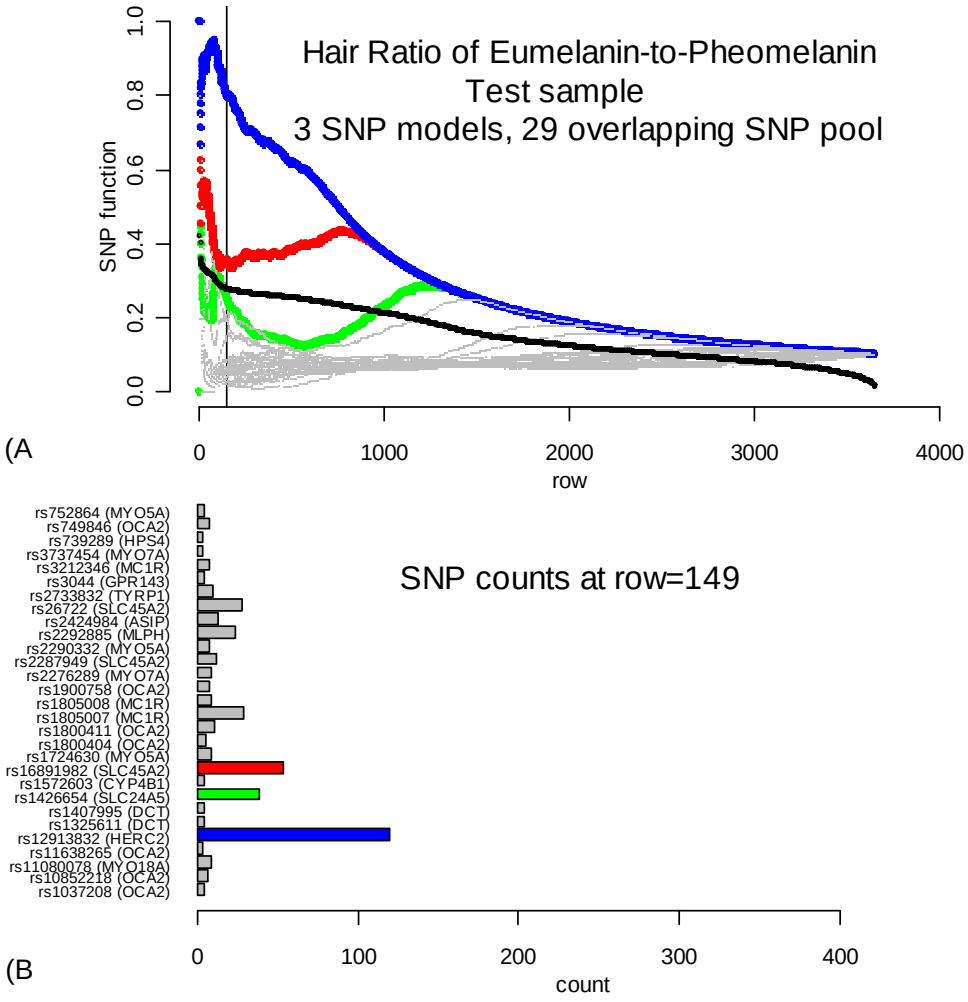

Figure 8. (**A**) Three-SNP multiple linear regression (MLR) models for natural log of the hair ratio's melanins across populations (test sample set). The horizontal-axis depicts all 3654 combinations (ie, 29-choose-3) of significant SNPs in a three-SNP MLR model. The vertical-axis is the R2 value for each model (black curve or three-SNP R2 curve) and also the SNP function value for each SNP curve. The R2 curve inflection (row 149) is indicated by a vertical black line. The SNP curves of the three highest frequency SNPs at row 149 are indicated by colors (rs12913832 (*HERC2*), blue; rs16891982 (*SLC45A2*), red; rs1426654 (*SLC24A5*), green).
(**B**) Bar plot of the SNPs that were present in all models from row 1 to row 149.

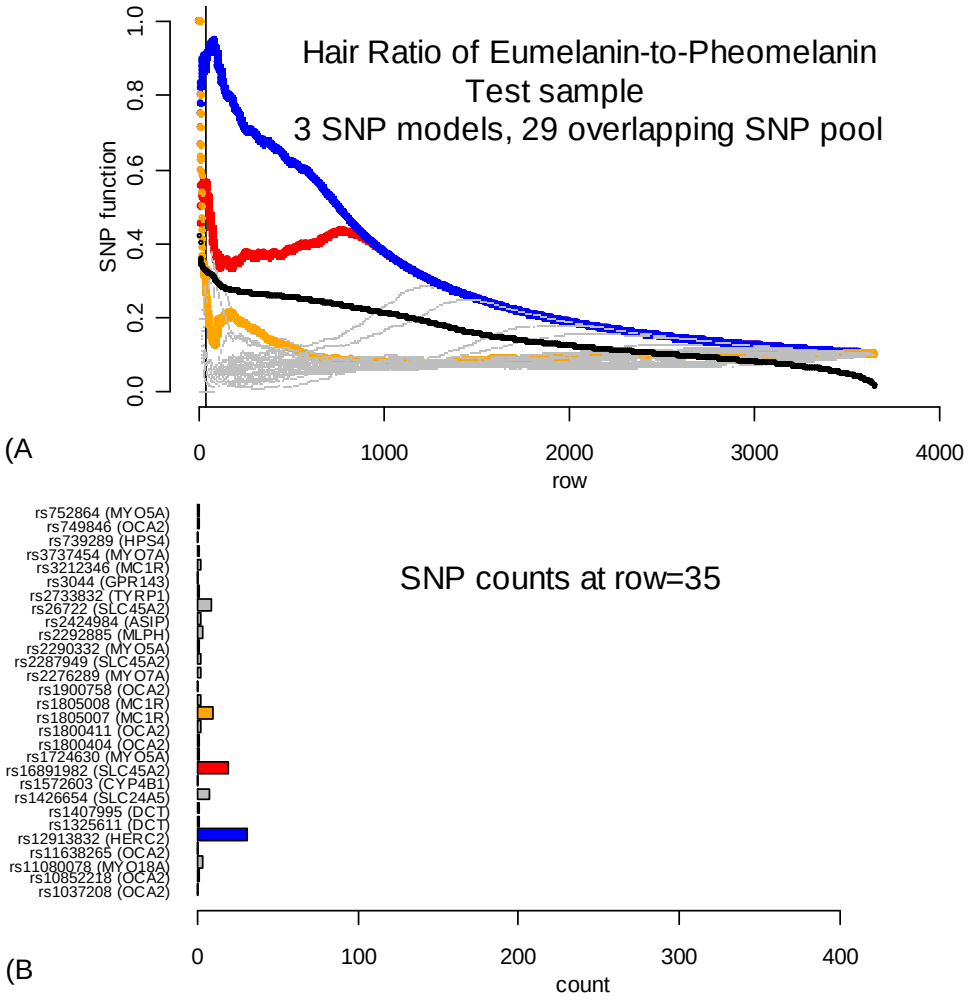

Figure 9. (**A**) Three-SNP multiple linear regression (MLR) models for natural log of the hair ratio's melanins across populations (test sample set). The horizontal-axis depicts all 3654 combinations (ie, 29-choose-3) of significant SNPs in a three-SNP MLR model. The vertical-axis is the R2 value for each model (black curve or three-SNP R2 curve) and also the SNP function value for each SNP curve. The R2 curve inflection (row 35) is indicated by a vertical black line. The SNP curves of the three highest frequency SNPs at row 35 are indicated by colors (rs12913832 (*HERC2*), blue; rs16891982 (*SLC45A2*), red; rs1805007 (*MC1R*), orange).
(**B**) Bar plot of the SNPs that were present in all models from row 1 to row 35.

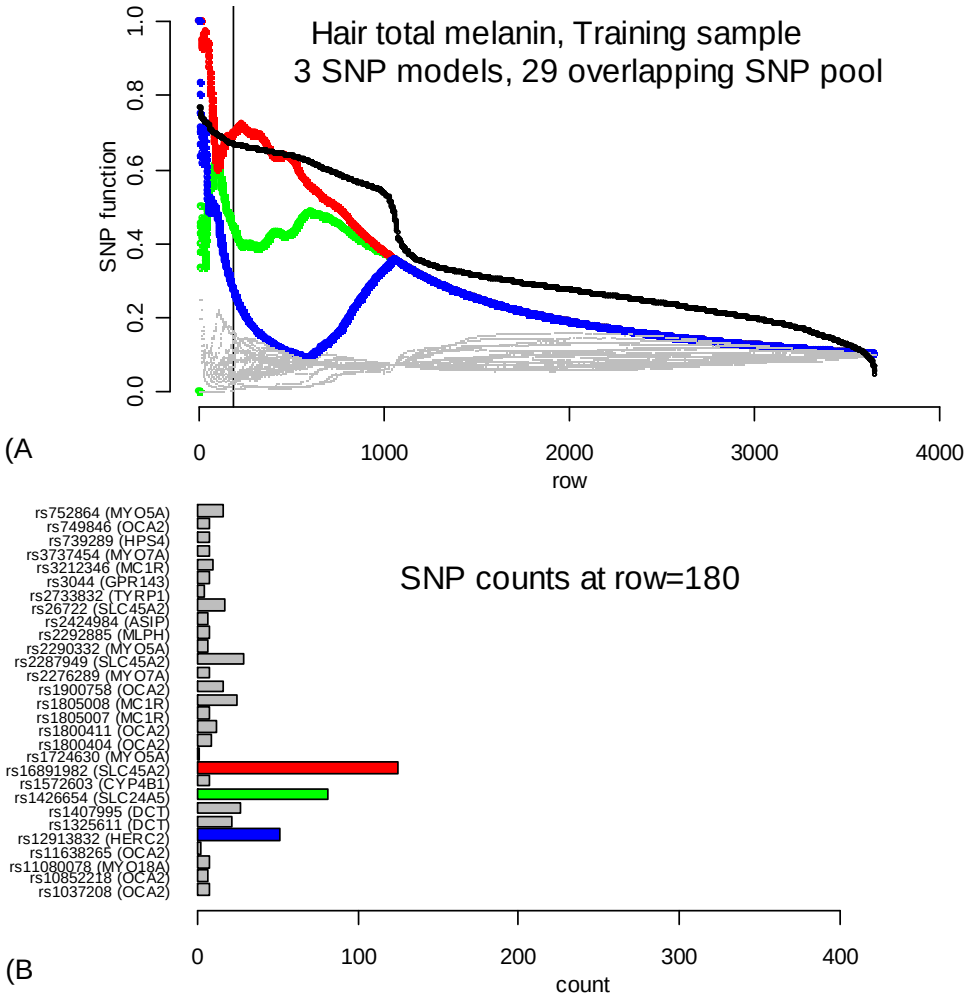

Figure 10. (**A**) Three-SNP multiple linear regression (MLR) models for hair total melanin across populations (training sample set). The horizontal-axis depicts all 3654 combinations (ie, 29-choose-3) of significant SNPs in a three-SNP MLR model. The vertical-axis is the R2 value for each model (black curve or three-SNP R2 curve) and also the SNP function value for each SNP curve. The R2 curve inflection (row 180) is indicated by a vertical black line. The SNP curves of the three highest frequency SNPs at row 180 are indicated by colors (rs12913832 (*HERC2*), blue; rs16891982 (*SLC45A2*), red; rs1426654 (*SLC24A5*), green). (**B**) Bar plot of the SNPs that were present in all models from row 1 to row 180.

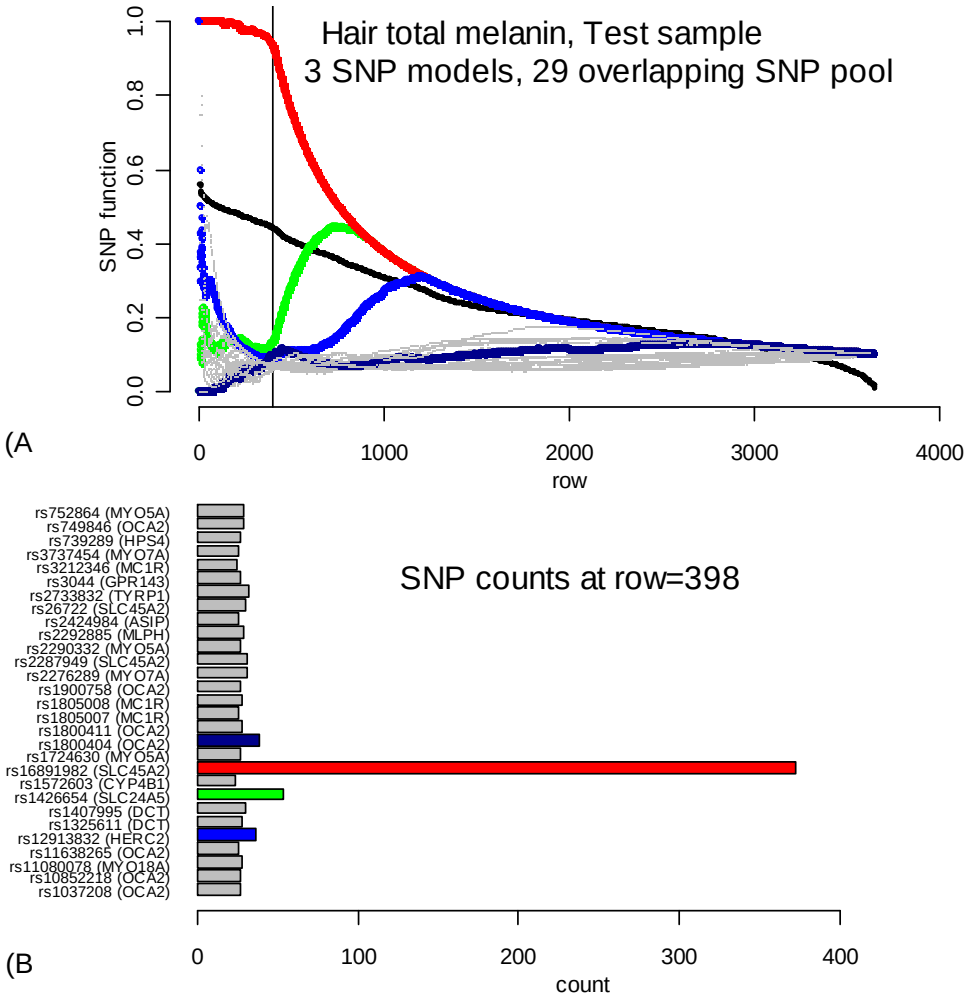

Figure 11. (**A**) Three-SNP multiple linear regression (MLR) models for hair total melanin across populations (test sample set). The horizontal-axis depicts all 3654 combinations (ie, 29-choose-3) of significant SNPs in a three-SNP MLR model. The vertical-axis is the R2 value for each model (black curve or three-SNP R2 curve) and also the SNP function value for each SNP curve. The R2 curve inflection (row 398) is indicated by a vertical black line. The SNP curves of the four highest frequency SNPs at row 398 are indicated by colors (rs16891982 (*SLC45A2*), red; rs1426654 (*SLC24A5*), green; rs12913832 (*HERC2*), blue; rs1800404 (*OCA2*), dark blue).
(**B**) Bar plot of the SNPs that were present in all models from row 1 to row 180.

Table 1. Test cohort. Sample size of each ethnic group (self-reported) utilized for model validation.

	Caucasian	African-American	Hispanic	South Asian	East Asian	Native American	Admixture	not listed	TOTAL
Average Skin Reflectance	101	9	34	0	17	8	10	7	186
Eye Color	110	12	38	0	17	8	11	8	204
Hair ratio of Eumelanin-to-Pheomelanin	92	6	36	0	14	8	8	6	170
Total hair melanin	90	6	34	0	14	7	7	6	163

Table 2. Cross validation results. R^2 values of the training sample and the test sample using the training sample's beta estimates.

	Training sample		Test sample			
	R^2	**sample size**	**R^2**	**sample size**	**Shrinkage**	**Relative Shrinkage (%)**
Skin reflectance	45.7	447	35.0	186	10.7	23.4
Eye color	76.4	353	61.6	204	14.8	19.4
Hair ratio of Eumelanin-to-Pheomelanin	43.2	162	27.1	170	16.1	37.3
Total hair melanin	76.3	143	25.2	163	51.1	67.0

Table 3. Prediction models' beta estimates derived from the training sample.

Skin reflectance			rs16891982			rs1426654			rs2424984	
	intercept	GG	GC	CC	AA	AG	GG	TT	CT	CC
	59.8	3.0	-0.1	-2.9	1.1	-1.7	0.5	3.2	3.0	-6.2
Eye color			rs12913832			rs16891982			rs1426654	
	intercept	AA	GA	GG	GG	GC	CC	AA	AG	GG
	4.4	1.2	0.5	-1.8	-0.4	0.1	0.2	-0.3	0.2	0.1
Hair ratio of Eumelanin-to-Pheomelanin			rs16891982			rs12913832			rs1805007	
	intercept	GG	GC	CC	AA	GA	GG	CC	TC	CC
	4.6	-0.4	0.3	0.2	0.7	0.1	-0.8	0.8	-0.8	0.8
Total hair melanin			rs16891982			rs1426654			rs12913832	
	intercept	GG	GC	CC	AA	AG	GG	AA	GA	GG
	12011.2	-3096.2	163.4	2932.7	-2196.4	953.4	1243.1	2347.4	-115.4	-2232.0

TABLE 4. 52 SNPs within or close to 22 genes

Gene	SNP	40 significant SNPs (training sample)	45 SNPs (test sample)	40 Significant SNPs (test sample)	33 overlapping SNPs	29 significant SNPs (training and test samples)	Total hair melanin	Natural log of ratio of melanins	Skin reflectance (CIEL)	Eye Color
AMACR	rs13289		+	+					*	
ASIP	rs2424984	+	+	+	+	+			*	
ASIP	rs6058017	+								
CYP4B1	rs1572603	+	+	+	+	+				*
DCT	rs1325611	+	+	+	+	+	*	*	*	*
DCT	rs1407995	+	+	+	+	+	*	*		*
GPR143	rs3044	+	+	+	+	+	*	*	*	*
HERC2	rs1129038		+	+			*	*	*	*
HERC2	rs12913832	+	+	+	+	+	*	*	*	*
HERC2	rs1667394		+	+			*	*	*	*
HERC2	rs916977		+	+			*	*	*	*
HPS3	rs2689234	+	+		+					
HPS4	rs1894704	+								
HPS4	rs3752589	+								
HPS4	rs3752590	+	+		+					
HPS4	rs739289	+	+	+	+	+				*
IRF4	rs12203592		+	+			*		*	*
MC1R	rs1805007	+	+	+	+	+		*		
MC1R	rs1805008	+	+	+	+	+		*		
MC1R	rs3212346	+	+	+	+	+	*		*	
MC1R	rs3212355	+								
MC1R	rs3212357	+								
MLPH	rs2292885	+	+	+	+	+		*	*	
MYO18A	rs11080078	+	+	+	+	+				*
MYO5A	rs1724630	+	+	+	+	+	*	*	*	*
MYO5A	rs2290332	+	+	+	+	+	*	*		*
MYO5A	rs752864	+	+	+	+	+	*			
MYO7A	rs2276289	+	+	+	+	+	*		*	*
MYO7A	rs3737454	+	+	+	+	+			*	*
near ASIP	rs1015362		+	+				*	*	
near KITLG	rs12821256		+	+			*			
near SLC24A4	rs12896399		+	+			*	*	*	
near TYRP1	rs1408799		+	+			*	*	*	*
OCA2	rs1037208	+	+	+	+	+			*	
OCA2	rs10852218	+	+	+	+	+			*	

OCA2	rs11638265	+	+	+	+	+	*	*	*	*
OCA2	rs1375164		+	+			*	*	*	*
OCA2	rs1800404	+	+	+	+	+	*	*	*	*
OCA2	rs1800407	+	+		+					
OCA2	rs1800410	+								
OCA2	rs1800411	+	+	+	+	+	*	*	*	*
OCA2	rs1800414	+								
OCA2	rs1900758	+	+	+	+	+	*	*	*	*
OCA2	rs749846	+	+	+	+	+	*	*		*
SLC24A5	rs1426654	+	+	+	+	+	*	*	*	*
SLC45A2	rs16891982	+	+	+	+	+	*	*	*	*
SLC45A2	rs2287949	+	+	+	+	+	*		*	*
SLC45A2	rs26722	+	+	+	+	+	*	*	*	*
SLC45A2	rs40132	+	+		+					
TPCN2	rs35264875		+							
TYR	rs1393350		+	+			*	*	*	*
TYRP1	rs2733832	+	+	+	+	+	*	*	*	*

***** Significant SNPs (test sample) by ANOVA ($p<0.05$)

References

1. Edwards M, Bigham A, Tan J, Li S, Gozdzik A, Ross K, et al. Association of the OCA2 polymorphism His615Arg with melanin content in east Asian populations: further evidence of convergent evolution of skin pigmentation. PLoS Genet 2010;6(3) e1000867.

2. Norton HL, Kittles RA, Parra E, McKeigue P, Mao X, Cheng K, et al. Genetic evidence for the convergent evolution of light skin in Europeans and East Asians. Mol. Biol. Evol 2007 Mar;24(3):710-722.

3. Yuasa I, Umetsu K, Harihara S, Miyoshi A, Saitou N, Park KS, et al. OCA2 481Thr, a hypofunctional allele in pigmentation, is characteristic of northeastern Asian populations. J. Hum. Genet 2007;52(8):690-693.

4. Valenzuela RK, Henderson MS, Walsh MH, Garrison NA, Kelch JT, Cohen-Barak O, et al. Predicting phenotype from genotype: normal pigmentation. J. Forensic Sci 2010 Mar;55(2):315-322.

5. Wakamatsu K, Fujikawa K, Zucca FA, Zecca L, Ito S. The structure of neuromelanin as studied by chemical degradative methods. J. Neurochem 2003 Aug;86(4):1015-1023.

6. Wakamatsu K, Ohtara K, Ito S. Chemical analysis of late stages of pheomelanogenesis: conversion of dihydrobenzothiazine to a benzothiazole structure. Pigment Cell Melanoma Res 2009 Aug;22(4):474-486.

7. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2009.Available from: http://www.R-project.org

8. Akey JM, Wang H, Xiong M, Wu H, Liu W, Shriver MD, et al. Interaction between the melanocortin-1 receptor and P genes contributes to inter-individual variation in skin pigmentation phenotypes in a Tibetan population. Hum. Genet 2001 Jun;108(6):516-520.

9. Szabó G, Gerald AB, Pathak MA, Fitzpatrick TB. Racial differences in the fate of melanosomes in human epidermis. Nature 1969 Jun;222(5198):1081-1082.

10. Sturm RA, Larsson M. Genetics of human iris colour and patterns. Pigment Cell Melanoma Res 2009 Oct;22(5):544-562.

11. Liu F, Wollstein A, Hysi PG, Ankra-Badu GA, Spector TD, Park D, et al. Digital Quantification of Human Eye Color Highlights Genetic Association of Three New Loci. PLoS Genet 2010 May;6(5) e1000934.

12. Vaughn MR, Brooks E, van Oorschot RAH, Baindur-Hudson S. A comparison of macroscopic and microscopic hair color measurements and a quantification of the relationship between hair color and thickness. Microsc. Microanal 2009 Jun;15(3):189-193.

13. Duffy DL, Montgomery GW, Chen W, Zhao ZZ, Le L, James MR, et al. A three-single-nucleotide polymorphism haplotype in intron 1 of OCA2 explains most human eye-color variation. Am. J. Hum. Genet 2007 Feb;80(2):241-252.

14. Liu F, van Duijn K, Vingerling JR, Hofman A, Uitterlinden AG, Janssens ACJW, et al. Eye color and the prediction of complex phenotypes from genotypes. Current Biology 2009;19(5):R192-R193.

15. Walsh S, Lindenbergh A, Zuniga SB, Sijen T, de Knijff P, Kayser M, et al. Developmental validation of the IrisPlex system: Determination of blue and brown iris colour for forensic intelligence. Forensic Sci Int Genet 2010 Oct;DOI:10.1016/j.fsigen.2010.09.008

16. Mengel-From J, Wong TH, Morling N, Rees JL, Jackson IJ. Genetic determinants of hair and eye colours in the Scottish and Danish populations. BMC Genet 2011;10:88-88.

17. Spichenok O, Budimlija ZM, Mitchell AA, Jenny A, Kovacevic L, Marjanovic D, et al. Prediction of eye and skin color in diverse populations using seven SNPs. Forensic Sci Int Genet 2010 Nov;DOI:10.1016/j.fsigen.2010.09.008

18. Branicki W, Liu F, van Duijn K, Draus-Barini J, Pośpiech E, Walsh S, et al. Model-based prediction of human hair color using DNA variants. Human Genetics 2011; DOI 10.1007/s00439-010-0939-8.

19. Ginger RS, Askew SE, Ogborne RM, Wilson S, Ferdinando D, Dadd T, et al. SLC24A5 encodes a trans-Golgi network protein with potassium-dependent sodium-calcium exchange activity that regulates human epidermal melanogenesis. J. Biol. Chem 2008 Feb;283(9):5486-5495.

20. Lamason RL, Mohideen M-APK, Mest JR, Wong AC, Norton HL, Aros MC, et al. SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. Science 2005 Dec;310(5755):1782-1786.

21. Lee YL, Teitelbaum S, Wolff MS, Wetmur JG, Chen J. Comparing genetic ancestry and self-reported race/ethnicity in a multiethnic population in New York City. J. Genet 2010 Dec;89(4):417-423.

APPENDIX C: HUMAN SUBJECTS APPROVAL

THE UNIVERSITY OF ARIZONA.

Human Subjects
Protection Program

1618 E Helen St.
P.O. Box 245137
Tucson, AZ 85724-5137
Tel: (520) 626-6721
http://irb.arizona.edu

27 January 2009

Robert Valenzuela, PhD Student
Advisor: Murray Brilliant, PhD
College of Medicine
PO Box 245079

Project NO. 09-0025-00 Data Analysis of Human Pigmentation Genes

Dear Mr. Valenzuela:

We received documents concerning your above cited project. Regulations published by the U.S. Department of Health and Human Services [45 CFR Part 46.101(b) (4)] exempt this type of research from review by our Institutional Review Board.

Exempt status is granted with the understanding that no further changes or additions will be made to the procedures followed (copies of which we have on file) without the review and approval of the Human Subjects Committee and your College or Departmental Review Committee. Any research related physical or psychological harm to any subject must also be reported to each committee.

Thank you for informing us of your work. If you have any questions concerning the above, please contact this office.

Sincerely,

Elizabeth Boyd, Ph.D.
Assistant Vice-President, Research Compliance & Policy
Office of Responsible Conduct for Research

cc: Departmental/College Review Committee
EB:mm