Hum Genet (2011) 129:443–454
DOI 10.1007/s00439-010-0939-8

ORIGINAL INVESTIGATION

Model-based prediction of human hair color using DNA variants

Wojciech Branicki · Fan Liu · Kate van Duijn · Jolanta Draus-Barini · Ewelina Pośpiech · Susan Walsh · Tomasz Kupiec · Anna Wojas-Pelc · Manfred Kayser

Received: 23 August 2010 / Accepted: 7 November 2010 / Published online: 4 January 2011

Abstract Predicting complex human phenotypes from genotypes is the central concept of widely advocated personalized medicine, but so far has rarely led to high accuracies limiting practical applications. One notable exception, although less relevant for medical but important for forensic purposes, is human eye color, for which it has been recently demonstrated that highly accurate prediction is feasible from a small number of DNA variants. Here, we demonstrate that human hair color is predictable from DNA variants with similarly high accuracies. We analyzed in Polish Europeans with single-observer hair color grading 45 single nucleotide polymorphisms (SNPs) from 12 genes previously associated with human hair color variation. We found that a model based on a subset of 13 single or compound genetic markers from 11 genes predicted red hair color with over 0.9, black hair color with almost 0.9, as well as blond, and brown hair color with over 0.8 prevalence-adjusted accuracy expressed by the area under the receiver characteristic operating curves (AUC). The identified genetic predictors also differentiate reasonably well between similar hair colors, such as between red and blond-red, as well as between blond and dark-blond, highlighting the value of the identified DNA variants for accurate hair color prediction.

Introduction

The concept of personalized medicine assumes that prediction of phenotypes based on genome information can enable better prognosis, prevention and medical care which can be tailored individually (Brand et al. 2008; Janssens and van Duijn 2008). However, practical application of genome-based information to medicine requires the disease risk to be predicted with high accuracy, while knowledge on genetics of common complex diseases is still insufficient to allow their accurate prediction solely from DNA data (Alaerts and Del-Favero 2009; Chung et al. 2010; Ku et al. 2010; McCarthy and Zeggini 2009). Another potential application for prediction of phenotypes from genotypes is forensic science. Knowledge gained on externally visible characteristics (EVC) from genotype data obtained by examination of crime scene samples may be used for investigative intelligence purposes, especially in suspectless cases (Kayser and Schneider 2009). The idea is based on using DNA-predicted EVC information to encircle a perpetrator in a larger population of unknown suspects. Such approach could also be useful in cases pertaining identification of human remains by extending anthropological findings on physical appearance of an identified individual. However, the genetic understanding of human

Electronic supplementary material The online version of this article (doi:10.1007/s00439-010-0939-8) contains supplementary material, which is available to authorized users.

W. Branicki · J. Draus-Barini · E. Pośpiech · T. Kupiec
Section of Forensic Genetics, Institute of Forensic Research,
Westerplatte 9, 31-033 Kraków, Poland

W. Branicki
Department of Genetics and Evolution, Institute of Zoology,
Jagiellonian University, Ingardena 6, 30-060 Kraków, Poland

F. Liu · K. van Duijn · S. Walsh · M. Kayser (✉)
Department of Forensic Molecular Biology,
Erasmus MC University Medical Center Rotterdam,
PO Box 2040, Rotterdam 3000 CA, The Netherlands
e-mail: m.kayser@erasmusmc.nl

A. Wojas-Pelc
Department of Dermatology,
Collegium Medicum of the Jagiellonian University,
Kopernika 19, 31-501 Kraków, Poland

appearance is still in its infancy. One notable exception is eye (iris) color, where previous candidate gene studies and especially recent genome-wide association studies (GWAS) revealed 15 genes involved (Eiberg et al. 2008; Frudakis et al. 2003; Graf et al. 2005; Han et al. 2008; Kanetsky et al. 2002; Kayser et al. 2008; Liu et al. 2010; Rebbeck et al. 2002; Sulem et al. 2007). One of them, *HERC2*, harbors genetic variation most strongly associated with human eye color variation (Eiberg et al. 2008; Kayser et al. 2008; Liu et al. 2010; Sturm et al. 2008). Moreover, a recent systematic study investigating the predictive value of eye color associated single nucleotide polymorphisms (SNPs) (Liu et al. 2009) found that a model with 15 SNPs from 8 genes predicts categorized blue and brown eye color with high accuracies with a subset of only 6 SNPs covering most of the predictive information. The IrisPlex system employing these six SNPs and a prediction model based on data from thousands of Europeans was recently developed and validated for DNA prediction of human eye color in forensic applications (Walsh et al. 2010a, b). Furthermore, a recent GWAS on quantitative eye color explained about 50% of continuous eye color variation by genetic factors (Liu et al. 2010).

The recent progress on DNA prediction of human eye color raises expectations for the DNA prediction of other human pigmentation traits, such as hair color. Inheritance of one particular hair color in humans i.e., red hair, has already been explained to a significant degree. Valverde et al. (1995) have found that red hair color is mainly associated with polymorphisms in the *MC1R* gene. This information has been confirmed since in many other studies performed on various population samples (Box et al. 1997; Han et al. 2008; Harding et al. 2000; Flanagan et al. 2000; Kanetsky et al. 2004; Pastorino et al. 2004; Rana et al. 1999; Sulem et al. 2007). *MC1R* SNPs are fairly indicative for red hair and thus have already been implemented in forensic science (Branicki et al. 2007; Grimes et al. 2001), but the practical application of red hair color prediction (without the ability to predict other hair colors) strongly depends on the population it is applied to, given the strong differences in red hair color frequency between populations. Additional data on red hair color inheritance came from a recent GWAS in Icelanders, which revealed two SNPs in the *ASIP* gene, representing a *MC1R* antagonist, as significantly associated with red hair color (Sulem et al. 2008). Moreover, a position in the 3′-UTR of the *ASIP* gene was previously associated with dark hair color in European populations (Kanetsky et al. 2002; Voisey et al. 2006) using a candidate gene approach. The candidate gene approach also delivered two non-synonymous SNPs in *SLC45A2* (MATP) with association to dark hair color in another study with several confirmation studies (Branicki et al. 2008a; Fernandez et al. 2008; Graf et al. 2005). Via several large GWASs various SNPs in/nearby genes in addition to *MC1R*, *ASIP* and *SLC45A2* have been found with association to human hair color variation, such as *OCA2*, *HERC2*, *SLC24A4*, *KITLG*, *TYR*, *TPCN2*, *TYRP1*, *IRF4*, *EXOC2*, *KIF26A*, and *OBSCN* (Han et al. 2008; Sulem et al. 2007, 2008). Furthermore, two recently published studies tested for hair color association of SNPs from a large number of candidate genes, not only confirmed some previously known hair color genes, such as *KITLG*, *OCA2*, *MC1R*, *TYRP1*, *TYR*, *SLC45A2*, *HERC2*, *ASIP*, but additionally reported association with quantitative measures of hair color of SNPs in additional genes, such as *SLC24A5*, *MYO5A*, *MYO7A*, *MLPH*, *GPR143*, *DCT*, *HPS3*, *GNAS*, *PRKARIA*, *ERCC6*, and *DTNBP1* in one or both studies (Mengel-From et al. 2009; Valenzuela et al. 2010). In the present study, we tested in Polish Europeans with single-observer hair color phenotype data the predictive power of 45 SNPs from 12 genes previously implicated with replicated evidence in human hair color variation.

Materials and methods

Subjects and hair color phenotyping

Samples were collected in years 2005–2009 from unrelated Europeans living in southern Poland. The study was approved by the Ethics Committee of the Jagiellonian University, number KBET/17/B/2005 and the Commission on Bioethics of the Regional Board of Medical Doctors in Krakow number 48 KBL/OIL/2008. All participants gave informed consent. Samples were collected from patients attending dermatological consultations at the Department of Dermatology of the Jagiellonian University Hospital. Hair color phenotypes were collected by a combination of self-assessment and professional single observer grading. A single dermatologist interviewed and assessed all the subjects. The questionnaire included basic information, such as gender and age as well as data concerning pigmentation phenotype. The study included 385 individuals (39.0% male) after genetic and phenotypic quality control. Categorical hair color was assessed by examination of the scalp in majority of cases. In a rare number of elderly volunteers and volunteers with dyed hair at the time of inspection we used ID photographs in combination with self-assessment for establishing natural hair color phenotypes. Hair color was classified into 7 categories: blond (16.4%), dark-blond (37.7%), brown (9.4%), auburn (3.1%), blond-red (11.2%), red (10.6%), and black (11.7%). For some analyses, we grouped blond and dark-blond into one blond group (54.1%) and auburn, blond-red, and red into one red group (24.9%) resulting into 4 categories. Notably, the frequency of red hair color in our study population is higher than expected in the general Polish population because of an enrichment of red hair colored individuals in the sampling process. This was done to demonstrate prediction accuracy of red hair, similar to other hair color, as red hair normally is relatively rare in the Polish population. Hence, the color distribution of the selected samples in our study does not represent that of the general Polish population (a point not relevant for the purpose of our study).

SNP ascertainment and genotyping

This study was based on 45 SNPs from 12 genes (Table 1), including *SLC45A2*, *IRF4*, *EXOC2*, *TYRP1*, *TPCN2*, *TYR*, *KITLG*, *SLC24A4*, *OCA2*, *HERC2*, *MC1R*, and *ASIP* that were associated with human hair color variation in several previous studies (Duffy et al. 2007; Graf et al. 2005; Han et al. 2008; Kanetsky et al. 2002, 2004; Valverde et al. 1995; Sulem et al. 2007, 2008). A subset of 25 SNPs were genotyped via mass spectrometry using Sequenom multiplexing (see Supplementary Table S1 for additional details about markers and methods). Multiplex assay design was performed with the software MassARRAY Assay Design version 3.1.2.2 (Sequenom Inc., San Diego, USA). The settings were iPLEX and high multiplexing. For the rest of the settings the default was used. The results were two 7-plexes and one 11-plex (see Supplementary Table 1). A 2 ng of dried genomic DNA in 384-well plates (Applied Biosystems) was amplified in a reaction volume of 5 μl containing 1× PCR Buffer, 1.625 mM $MgCl_2$, 500 μM dNTPs, 100 nM each PCR primer, 0.5 U PCR enzyme (Sequenom). The reaction was incubated in a GeneAmp PCR System 9700 (Applied Biosystems) at 94°C for 4 min followed by 45 cycles of 94°C for 20 s, 56°C for 30 s, 72°C for 1 min, followed by 3 min at 72°C. To remove the excess dNTPs, 2 μl SAP mix containing 1× SAP Buffer and 0.5 U shrimp alkaline phosphatase (Sequenom) was added to the reaction. This was incubated in a GeneAmp PCR System 9700 (Applied Biosystems) at 37°C for 40 min followed by 5 min at 85°C for deactivation of the enzyme. Then 2 μl of Extension mix is added containing a concentration of adjusted extend primers varying between 3.5 and 7 μM for each primer, 1× iPLEX buffer (Sequenom), iPLEX termination mix (Sequenom) and iPLEX enzyme (Sequenom). The extension reaction was incubated in a GeneAmp PCR System 9700 (Applied Biosystems) at 94°C for 30 s followed by 40 cycles of 94°C for 5 s, 5 cycles of 52°C for 5 s, and 80°C for 5 s, then 72°C for 3 min. After the extension reaction it is desalted by the addition of 6 mg Clean Resin (Sequenom) and 16 μl water, rotating the plate for 15 min and centrifuging. The extension product was spotted onto a G384 + 10 SpectroCHIP (Sequenom) with the MassARRAY Nanodispenser model rs1000 (Sequenom). The chip is then transferred into the MassARRAY Compact System (Sequenom) where the data are collected, using TyperAnalyzer version 4.0.3.18 (Sequenom), SpectroACQUIRE version 3.3.1.3 (Sequenom), GenoFLEX version 1.1.79.0 (Sequenom), and MassArrayCALLER version 3.4.0.41 (Sequenom). The data were checked manually after the data collection. Vaguely positioned dots produced by TyperAnalyzer 3.4 (Sequenom) and all the wells that 50% or more failed SNPs were excluded from analysis. Blanks (2%), controls (2%), and duplicates (9%) were checked for any inconsistencies and false positives. Furthermore, SNPs in the *MC1R* gene were analyzed by amplification and cycle sequencing of the complete *MC1R* exon using the procedure described in Branicki et al. (2007). Selected polymorphisms within genes *ASIP* (rs6058017), *OCA2* (rs1800407, rs1800401, rs7495174, rs4778241, rs4778138), *SLC45A2* (rs16891982, rs26722), and *HERC2* (rs916977) were analyzed using minisequencing procedure and SNaPshot multiplex kit. Detailed information concerning protocols and primers were described elsewhere (Branicki et al. 2007; 2008a, b, 2009; Brudnik et al. 2009). Briefly, the PCR reaction was composed of 5 μl of Qiagen multiplex PCR kit (Qiagen, Hilden, Germany), 1 μl of primer premix, 2 μl of Q solution, and 2 μl (approximately 5 ng) of template DNA. The temperature profile was as follows: {94° for 15 min (94° for 30 s, 58–60°C for 90 s, and 72° for 90 s) × 32, 72°/10 min} 4°C/∞. The PCR products were purified with a mixture of ExoI and SAP enzymes (Fermentas, Vilnius, Lithuania) and subjected to multiplex minisequencing reactions with a SNaPshot multiplex kit (Applied Biosystems, Foster City, CA, USA). A 2 μl of SNaPshot kit was combined with 1 μl of extension primers premix, 1 μl of the purified PCR product, and nuclease-free water up to 10 μl. The products of extension reactions were purified with SAP enzyme and analyzed on ABI 3100 Avant genetic analyzer (Applied Biosystems, Foster City, CA, USA).

Statistic analysis

The frequency of hair color was compared between male and females using cross tabulation. Mean age was compared between color categories using one-way analysis of variance. *MC1R* polymorphisms are largely recessive when considered individually, but also interact with each other through the genetic mechanism known as "compound heterozygosity". The high-penetrance variants, traditionally coded as "**R**", include Y152OCH, N29insA, D84E (rs1805006), R142H (rs11547464), R151C (rs1805007), R160W (rs1805008), D294H (rs1805009), and low-penetrance variants, as "**r**", include V60L (rs1805005), V92M (rs2228479), I155T (rs1110400), R163Q (rs885479). We followed the tradition and used the **R** and **r** variables by combining the information of all known causal variants in

Table 1 Single SNP hair color association in a Polish sample

Variant	Chr	Position	Gene	A	B	MAF	Color	OR	95% Lower CI	95% Upper CI	*P* val	Other
rs16891982	5	33987450	*SLC45A2*	G	C	0.02	Black	5.11	1.79	14.55	0.002	Yes
rs28777	5	33994716	*SLC45A2*	A	C	0.02	Black	7.05	2.23	22.25	0.001	Yes
rs26722	5	33999627	*SLC45A2*	G	A	0.02	Black	5.53	1.64	18.68	0.006	Yes
rs12203592	6	341321	*IRF4*	C	T	0.08	Black	2.35	1.22	4.54	0.011	Yes
rs9378805	6	362727	*IRF4*	A	C	0.45					0.103	
rs4959270	6	402748	*EXOC2*	C	A	0.46	Black	0.56	0.35	0.91	0.020	Yes
rs1408799	9	12662097	*TYRP1*	C	T	0.29					0.097	
rs2733832	9	12694725	*TYRP1*	T	C	0.40					0.177	
rs683	9	12699305	*TYRP1*	A	C	0.34					0.099	
rs35264875	11	68602975	*TPCN2*	A	T	0.23					0.158	
rs3829241	11	68611939	*TPCN2*	G	A	0.37					0.183	
rs2305498	11	68623490	*TPCN2*	G	A	0.27					0.230	
rs1011176	11	68690473	*TPCN2*	T	C	0.36					0.096	
rs1042602	11	88551344	*TYR*	C	A	0.28					0.255	
rs1393350	11	88650694	*TYR*	G	A	0.25	Brown	1.70	1.02	2.82	0.041	Yes
rs12821256	12	87852466	*KITLG*	T	C	0.09					0.052	
rs12896399	14	91843416	*SLC24A4*	G	T	0.44					0.064	
rs4904868	14	91850754	*SLC24A4*	C	T	0.46	Blond	0.64	0.43	0.97	0.037	
rs2402130	14	91870956	*SLC24A4*	A	G	0.16	D-blond	0.62	0.39	0.96	0.033	
rs1800407	15	25903913	*OCA2*	C	T	0.07	Red	3.23	1.07	9.76	0.038	
rs1800401	15	25933648	*OCA2*	C	T	0.06					0.105	
rs16950821	15	25957102	*OCA2*	C	T	0.12					0.170	
rs7174027	15	26002360	*OCA2*	C	T	0.12					0.146	
rs4778138	15	26009415	*OCA2*	T	C	0.16	Brown	1.80	1.02	3.19	0.043	Yes
rs4778241	15	26012308	*OCA2*	G	T	0.18					0.078	
rs7495174	15	26017833	*OCA2*	T	C	0.05					0.069	
rs12913832	15	26039213	*HERC2*	C	T	0.22	Black	3.33	1.99	5.57	4.3E−06	Yes
rs7183877	15	26039328	*HERC2*	C	A	0.07					0.124	
rs11635884	15	26042564	*HERC2*	T	C	0.01					0.135	
rs916977	15	26186959	*HERC2*	C	T	0.15	Red	0.34	0.18	0.65	0.001	Yes
rs8039195	15	26189679	*HERC2*	T	C	0.11	Red	0.30	0.14	0.64	0.002	Yes
MC1R_R	16		*MC1R*	wt	R	0.31	Red	12.64	7.03	22.74	2.5E−17	Yes
MC1R_r	16		*MC1R*	wt	r	0.20	Red	2.50	1.35	4.31	0.003	Yes
rs1805005	16	89985844	*MC1R*	G	T	0.08	Blond	2.99	1.52	5.86	0.001	Yes
Y152OCH	16	89986122	*MC1R*	A	C	0.00					0.982	
N29insA	16	89985753	*MC1R*	–	insA	0.01	Red	53.60	1.29	2221.72	0.036	
rs1805006	16	89985918	*MC1R*	C	A	0.00					0.476	
rs2228479	16	89985940	*MC1R*	G	A	0.09	Red	0.43	0.19	0.97	0.043	
rs11547464	16	89986091	*MC1R*	G	A	0.02	Red	3.35	1.04	10.76	0.042	
rs1805007	16	89986117	*MC1R*	C	T	0.11	Red	6.69	3.50	12.79	9.3E−09	Yes
rs1110400	16	89986130	*MC1R*	T	C	0.02					0.314	
rs1805008	16	89986144	*MC1R*	C	T	0.16	Red	5.69	3.31	9.78	3.2E−10	Yes
rs885479	16	89986154	*MC1R*	G	A	0.03	Blond	2.90	1.21	6.96	0.017	
rs1805009	16	89986546	*MC1R*	G	C	0.01	Red	31.85	2.61	388.28	0.007	Yes
rs1015362	20	32202273	*ASIP*	C	T	0.30	B-red	1.67	1.02	2.75	0.043	
rs6058017	20	32320659	*ASIP*	A	G	0.13					0.211	
rs2378249	20	32681751	*ASIP*	A	G	0.18	Red	2.34	1.14	4.82	0.021	Yes

MAF minor allele frequency, *Color* the most significantly associated color, *OR* the allelic odds ratio for the minor B allele, shown only if $P < 0.05$, *P* val the *P* value adjusted for age and gender, *Other* if the SNP is also associated with other colors with $P < 0.05$

the *MC1R* gene in the association and the prediction analyses. The **R** variant was defined by the total number of the high-penetrance variants in the *MC1R* gene so that each individual has three possible genotype states, homozygote wildtype for all variants (wt/wt), heterozygote for one high-penetrant variant (wt/**R**), and homozygote for at least one or compound heterozygote for at least two high-penetrant variants (**R/R**). The **r** variant was defined similarly by considering the low-penetrant variants in the *MC1R* gene (wt/wt, wt/**r**, **r/r**). All ascertained SNPs including the **R** and **r** variants were tested for association with each hair color category (binary coded 0, 1) using logistic regression adjusted for gender and age. We derived the allelic Odds Ratios (ORs), where the SNP genotypes were coded using 0, 1, or 2 number of the minor alleles (Table 1). We also derived the genotypic ORs, where the homozygote minor alleles and heterozygote genotypes were compared with homozygote wildtypes (Supplementary Table S2).

We used a multinomial logistic regression model for the prediction analysis, and the modeling details follow closely the previous study of eye color (Liu et al. 2009). Consider hair color, *y*, to be four categories blond, brown, red, and black, which are determined by the genotype, *x*, of *k* SNPs, where *x* represents the number of minor alleles per *k* SNP. Let $\pi 1$, $\pi 2$, $\pi 3$, and $\pi 4$ denote the probability of blond, brown, red, and black, respectively. The multinomial logistic regression can be written as

$$\begin{aligned}\operatorname{logit}(\Pr(y = \text{blond}|x_1 \ldots x_k)) &= \ln\left(\frac{\pi_1}{\pi_4}\right)\\ &= \alpha_1 + \sum \beta(\pi_1)_k x_k\end{aligned}$$

$$\begin{aligned}\operatorname{logit}(\Pr(y = \text{brown}|x_1 \ldots x_k)) &= \ln\left(\frac{\pi_2}{\pi_4}\right)\\ &= \alpha_2 + \sum \beta(\pi_2)_k x_k\end{aligned}$$

$$\operatorname{logit}(\Pr(y = \text{red}|x_1 \ldots x_k)) = \ln\left(\frac{\pi_3}{\pi_4}\right) = \alpha_3 + \sum \beta(\pi_3)_k x_k$$

where α and β can be derived in the training set.

Hair color of each individual in the testing set can be probabilistically predicted based on his or her genotypes and the derived α and β,

$$\pi_1 = \frac{\exp(\alpha_1 + \sum \beta(\pi_1)_k x_k)}{1 + \exp(\alpha_1 + \sum \beta(\pi_1)_k x_k) + \exp(\alpha_2 + \sum \beta(\pi_2)_k x_k) + \exp(\alpha_3 + \sum \beta(\pi_3)_k x_k)}$$

$$\pi_2 = \frac{\exp(\alpha_2 + \sum \beta(\pi_2)_k x_k)}{1 + \exp(\alpha_1 + \sum \beta(\pi_1)_k x_k) + \exp(\alpha_2 + \sum \beta(\pi_2)_k x_k) + \exp(\alpha_3 + \sum \beta(\pi_3)_k x_k)}$$

$$\pi_3 = \frac{\exp(\alpha_3 + \sum \beta(\pi_3)_k x_k)}{1 + \exp(\alpha_1 + \sum \beta(\pi_1)_k x_k) + \exp(\alpha_2 + \sum \beta(\pi_2)_k x_k) + \exp(\alpha_3 + \sum \beta(\pi_3)_k x_k)}$$

$$\pi_4 = 1 - \pi_1 - \pi_2 - \pi_3.$$

Categorically, the color category with the max(π_1, π_2, π_3, π_4) was considered as the predicted color.

We evaluated the performance of the prediction model in the testing set using the area under the receiver operating characteristic (ROC) curves, or AUC (Janssens et al. 2004). AUC is the integral of ROC curves which ranges from 0.5 representing total lack of prediction to 1.0 representing perfect prediction. Cross-validations were conducted 1,000 replicates; in each replicate 80% individuals were used as the training set and the remaining samples were used as the testing set. The average accuracy estimates of all replicates were reported. Because of a relatively small sample size and rare *MC1R* polymorphisms with large effects, the cross-validation may give conservative estimates of the prediction accuracy. Thus, we report both the results with and without cross-validations, i.e. using the whole sample for training and prediction.

The selection of SNPs in the final model was based on the contribution of each SNP to the predictive accuracy using a step-wise analysis by iteratively including the next largest contributor to the model. The contribution of each SNP was measured by the gain of total AUC of the models with and without that SNP. The MC1R, **R** and **r**, and the *OCA2* SNP, rs1800407, were always included in the prediction model due to their known biological function. The *HERC2* SNP, rs12193832, was also always included because of its known extraordinary large effect on all human pigmentation traits.

Because the sample size included in the current study is relatively small, we estimated the effect of sample size on the accuracy of the prediction analysis using the data from a previously published study of eye color (Liu et al. 2009), in which a larger sample was available ($N = 6{,}168$). A sample of *n* individuals was randomly bootstrapped 1,000 times from the 6,168 participants of the Rotterdam Study, for whom the eye color information and genotypes of the six most important eye color SNPs were available. For each bootstrap, a binary logistic regression model was built in a randomly selected subsample (80% of *n* individuals) using the six most eye color predictive SNPs from Liu et al. (2009) as the predictors and the blue eye color (yes, no) as the binary outcome. The logistic model was then used to predict the blue color in the remaining sample (20% of n individuals), based on which an AUC value was derived. The mean, the 95% upper, and the 95% lower AUC values of the 1,000 bootstraps were reported. The bootstrap analysis was conducted for various n ranging from 100 to 800 (Supplementary Figure S1).

Further, we conducted a prediction analysis using the multinomial LASSO regression model implemented in the R library glmnet v1.1-4 (Friedman et al. 2010). The cross-validations of LASSO analysis were also conducted 1,000 replicates based on the 80–20% split.

Results and discussion

First, we tested the genotyped SNPs for hair color association in our study sample. Although variation in *MC1R* is usually attributed to red hair color (Branicki et al. 2007; Grimes et al. 2001; Valverde et al. 1995), the compound variant MC1R-R in our study was significantly associated with all but one (auburn) hair color category, albeit its association was strongest with red hair (allelic OR: 12.6; 95% CI: [7.0–22.7]; $P = 2.5\times10^{-17}$; Table 1). The lack of association of the MC1R-R variant with auburn hair color may be caused by the small sample size of the auburn category and/or problems with correct classification of this hair color as reported elsewhere (Mengel-From et al. 2009). Furthermore, MC1R-R showed a clear recessive effect and a compound-heterozygote effect in that the **R/R** genotype carriers were much more likely to have red hair (genotypic OR: 262.2; 95% CI: [65.2–1,055.3]; $P = 4.5 \times 10^{-15}$) than the **wt/R** carriers (genotypic OR: 5.6; 95% CI: [2.5–12.6]; $P = 4.0 \times 10^{-5}$; Supplementary Table S2). The stronger association of *MC1R* SNPs with red hair than with non-red hair colors as observed here was also found previously (Han et al. 2008; Sulem et al. 2007). The SNP rs12913832 in the *HERC2* gene was significantly associated with all hair color categories, most significantly with brown (allelic OR for T vs. C: 3.5; 95% CI: [2.0–6.1]; $P = 1.3 \times 10^{-5}$) and black (allelic OR: 3.3; 95% CI: [2.0–5.6]; $P = 4.3 \times 10^{-6}$; Table 1) hair. The T allele of rs12913832 showed a dominant effect on darker hair color in that the heterozygote carriers had a further increased OR of black hair (genotypic OR: 8.6; 95% CI: [3.9–18.9]; $P = 7.2 \times 10^{-8}$; Supplementary Table S2). This SNP was associated with total hair melanin in a recent study (Valenzuela et al. 2010). A previous study found *HERC2* SNPs significantly associated with non-red, but not with red, hair colors (Sulem et al. 2007), and another one reported *HERC2* association only with dark hair color (Mengel-From et al. 2009). However, an additional study found *HERC2* association with all hair colors, albeit reported stronger association with non-red hair colors than with red hair (Han et al., 2008), in agreement with our findings. Additional SNPs in *MC1R* and *HERC2* were also significantly associated with several hair colors (Table 1). Except for *MC1R* and *HERC2* genes, no significant evidence of a dominant or a recessive effects on hair color was found for any other gene studied (Supplementary Table S2). SNPs in *SLC45A2* (rs28777 allelic OR for C vs. G: 7.05; 95% CI: [2.2–22.3]; $P = 0.001$), *IRF4* (rs12203592 allelic OR for T vs. C: 7.05; 95% CI: [2.2–22.3]; $P = 0.01$), and *EXOC2* (rs4959270 allelic OR for A vs. C: 0.56; 95% CI: [0.35–0.91]; $P = 0.02$) were most significantly associated with black hair color (Table 1), in line with the previous reports (Han et al. 2008; Mengel-From et al. 2009). Further, an association of *SLC45A2* with total hair melanin was reported (Valenzuela et al. 2010). SNPs in the *ASIP* gene were associated with red (rs2378249, $P = 0.02$), dark blond (rs2378249, $P = 0.02$), and blond-red (rs1015362, $P = 0.04$; Table 1). Significant *ASIP* association with red hair was reported previously (Sulem et al. 2008), as well as with total hair melanin (Valenzuela et al. 2010). The *OCA2* gene was most significantly associated with brown hair color (rs4778138, $P = 0.03$), confirming previous findings of *OCA2* involvement in hair color variation (Han et al. 2008; Mengel-From et al. 2009; Valenzuela et al. 2010), although one previous GWAS did not find significant evidence (Sulem et al. 2007). The *TYR* gene was significantly associated with brown (rs1393350, $P = 0.02$) and the *SLC24A4* gene with blond (rs4904868, $P = 0.04$) and dark blond (rs2402130, $P = 0.03$). These results are largely consistent with previous findings (Sulem et al. 2007; Han et al. 2008; Mengel-From et al. 2009). Overall, at least one SNP in 9 out of the 12 genes studied showed significant association with certain hair color categories in our sample (Table 1). For three genes (*TYRP1*, *TPCN2*, and *KITLG*) the SNPs tested did not reveal statistically significant hair color association (but see below for the predictive effects of two of these genes), although these genes have been implicated in human hair color variation elsewhere (Sulem et al. 2007, 2008; Valenzuela et al. 2010; Mengel-From et al. 2009). This discrepancy may be influenced by the relatively small sample size in our study and the putatively smaller effect size of these three genes relative to the other genes studied.

The main goal of this study, however, was to investigate the predictive value of hair color associated SNPs as established in previous, and (mostly) confirmed in the present study. DNA-based prediction accuracies for hair color categories were evaluated by means of the area under the ROC curves (AUC), ranging from 0.5 (random) to 1 (perfect) prediction. Our model revealed that 13 single or combined (MC1R-R and MC1R-r) genetic variants from

Table 2 Parameters of the prediction model based on multinomial logistic regression in a Polish sample

SNP	Gene	Effect	4 Hair color categories				7 Hair color categories						
		Allele	Rank	b1	b2	b3	Rank	b1	b2	b3	b4	b5	b6
Constant				1.70	−1.26	−2.62		0.50	1.29	−1.26	−1.93	−4.31	−6.96
MC1R_R	*MC1R*	R	1	1.11	0.55	4.09	1	1.70	0.98	0.54	1.84	5.41	6.02
rs12913832	*HERC2*	T	2	−1.75	0.10	−2.49	2	−2.58	−1.55	0.10	−0.89	−4.68	−3.21
rs12203592	*IRF4*	T	3	−1.29	−1.15	−1.13	4	−1.43	−1.24	−1.14	−1.03	−1.16	−1.19
rs1042602	*TYR*	A	4	0.39	0.30	1.20	3	0.53	0.35	0.30	1.07	1.08	1.67
rs4959270	*EXOC2*	A	5	0.77	0.85	1.15	5	0.56	0.81	0.84	1.24	0.92	1.11
rs28777	*SLC45A2*	C	6	−1.69	−13.89	0.10	12	−13.78	−1.31	−11.09	−7.84	2.02	−2.91
rs683	*TYRP1*	C	7	0.10	0.58	−0.02	10	−0.21	0.21	0.57	−0.49	−0.09	0.24
rs1800407	*OCA2*	T	8	0.49	−1.14	0.19	8	1.01	0.44	−1.12	−10.32	1.02	1.26
MC1R_r	*MC1R*	r	9	0.46	0.55	0.61	6	0.74	0.40	0.55	−0.53	1.22	1.69
rs2402130	*SLC24A4*	G	10	−0.48	−0.09	−0.54	9	−0.22	−0.57	−0.09	−0.61	−0.61	−0.73
rs12821256	*KITLG*	C	11	0.69	0.01	0.87	11	0.45	0.72	−0.02	0.71	0.30	0.94
rs16891982	*SLC45A2*	C	12	−0.82	−11.78	−3.48	13	−0.62	−0.84	−9.09	−9.55	−6.77	−0.97
rs2378249	*ASIP*	G	13	−0.18	−0.16	0.40	7	0.17	−0.29	−0.16	−0.54	0.43	1.03

b1, b2, b3 in the 4 categories are the betas for blond, brown, and red, all versus black; b1 to b6 in the 7 categories are the betas for blond, d-blond, brown, auburn, b-red, and red, all versus black; rank, prediction rank with 1 having the highest and 13 having the lowest rank in the prediction analysis

all, but one (*TPCN2*) of the 12 genes investigated contribute independently to the AUC value (Table 2) for 4 (Fig. 1a) and 7 hair color categories (Fig. 1b). As may be expected from the association results, MC1R_R has the most predictive power on red hair (AUC 0.86–0.88), and its predictive effect on non-red hair colors was considerably lower (AUC 0.63–0.68, Fig. 1). The *HERC2* SNP rs12913832, when added to MC1R_R in the model, contributed most of all other genetic predictors to the accuracy for predicting all color categories (ΔAUC 0.08 for blond, 0.12 for brown, 0.03 for red, and 0.13 for black, Fig. 1). Adding the remaining 11 independent genetic predictors provides accuracy increase and usually with decreasing effects while increasing the number of markers (Fig. 1). Notably, some SNPs without statistically significant hair color association in our study ($P > 0.05$) did provide independent information toward hair color prediction (such as rs1042602 in *TYR*, rs683 in *TYRP1*, and rs12821256 in *KITLG*). Only the non-synonymous SNPs from the *TPCN2* gene tested did not contribute to the prediction model and did not show a statistically significant association with any hair color category. Rs35264875 and rs3829241 in *TPCN2* had been discovered recently as significantly associated with blond versus brown hair color in Icelandic and replicated in Icelandic and Dutch people (Sulem et al. 2008). Predicting each color type separately using binary logistic regression yield slightly lower accuracy compared to the multinomial model (Supplementary Table S3).

Overall, hair color prediction with 13 DNA components from 11 genes showed very good accuracy without cross-

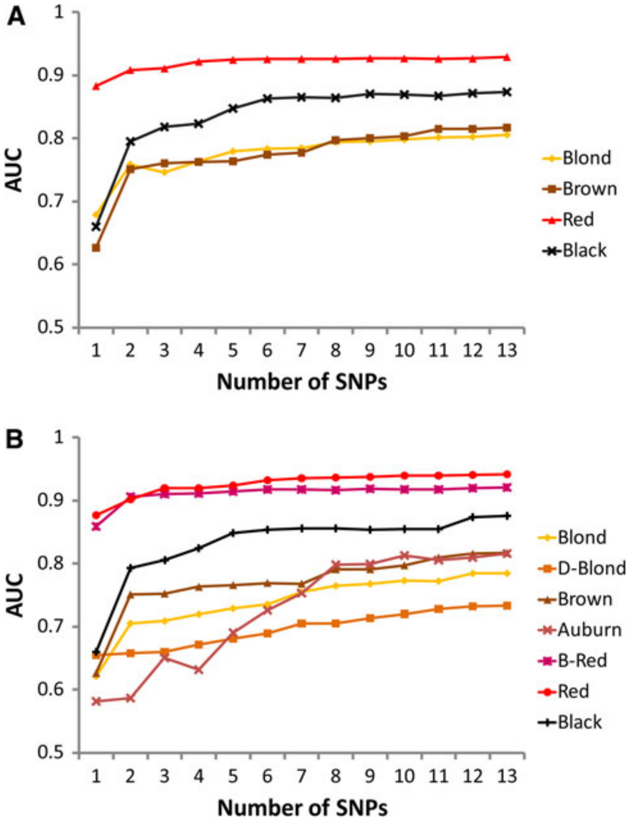

Fig. 1 Accuracy of hair color prediction using DNA variants in a Polish sample. AUC was plotted against the number of SNPs included in the multinomial logistic model for predicting 4 (**a**) and 7 (**b**) hair color categories. SNP annotation and prediction ranks are provided in Table 2

Table 3 Hair color prediction accuracy using 13 genetic markers in a Polish sample

Accuracy	4 Hair color categories				7 Hair color categories						
	Blond	Brown	Red	Black	Blond	D-blond	Brown	Auburn	B-red	Red	Black
Using all sample											
AUC	0.81	0.82	0.93	0.87	0.78	0.73	0.82	0.82	0.92	0.94	0.88
Sensitivity	0.88	0.08	0.78	0.31	0.17	0.80	0.14	0.00	0.53	0.66	0.38
Specificity	0.55	0.99	0.95	0.97	0.96	0.53	0.98	1.00	0.95	0.94	0.95
PPV	0.70	0.38	0.84	0.58	0.48	0.51	0.45	0.00	0.56	0.59	0.49
NPV	0.80	0.91	0.93	0.91	0.86	0.82	0.92	0.97	0.94	0.96	0.92
Average of 1,000 cross-validations											
AUC	0.75	0.72	0.90	0.78	0.70	0.66	0.73	0.64	0.85	0.90	0.81
Sensitivity	0.83	0.05	0.74	0.24	0.15	0.71	0.08	0.00	0.41	0.44	0.29
Specificity	0.52	0.98	0.93	0.96	0.93	0.51	0.97	1.00	0.92	0.93	0.94
PPV	0.67	0.21	0.77	0.45	0.33	0.46	0.20	0.00	0.42	0.43	0.43
NPV	0.72	0.91	0.91	0.90	0.85	0.75	0.91	0.97	0.92	0.93	0.91

AUC the area under the ROC curves, *PPV* positive predictive value, *NPV* negative predictive value

validation, such as AUC for blond = 0.81, brown = 0.82, red = 0.93, black = 0.87 in the 4 category model (Table 3; Fig. 1a), and AUC for blond = 0.78, d-blond = 0.73, brown = 0.82, auburn = 0.82, b-red = 0.92, red = 0.94, black = 0.88 (Table 3; Fig. 1b) when considering 7 categories. The mean accuracies derived from 1,000 cross-validations are somewhat lower for all hair color categories (least so for red), likely because of sample size effects as the rare alleles with large effects are not well captured in the training sets (Table 3).

In general, the sensitivities for predicting brown, red, and black colors were considerably lower than the respective specificities, except for blond in the 4 categories and dark blond in the 7 categories (Table 3). The very low sensitivities for brown may reflect uncertainties in distinguishing between the dark-blond and brown colors on one side, and between the auburn, red and blond-red colors on the other side during phenotyping, as well as an additional sample size effect for auburn representing the smallest hair color group in our study ($N = 12$). However, the final model showed a good power to discriminate highly similar hair color categories, such as red and blond-red, as well as between blond and dark-blond (Table 3), underlining the value of the genetic markers involved in our hair color prediction model.

The ROC curves from the final model (Fig. 2) provide practical guides for the choices between desired false positive thresholds (1-specificity) and expected true positive rates (sensitivity) for predicting all color categories. For example, if the desired false positive threshold is 0.2 (in other words, if we use the predicted probability of $P > 0.8$ as the threshold for prediction, thus we know that we have at least 80% chance to be correct), then the expected true positive rates (or sensitivities) are 0.61 for blond (meaning that if a person has blond hair, our model provides a 61% chance to predict him/her as blond), 0.69 for brown, 0.78 for black, and 0.88 for red. Notably, incorrect predictions fall more frequently in the neighboring category than in a more distant category, so the predictive information can still provide useful information.

We noticed that the prediction accuracies for the blond and brown colors were somewhat lower than those for black and red colors. One reason for this difference may be in the environmental rather than genetic contribution to hair color variation. Hair color changes in some individuals during adolescence and such change is most often from blond to brown (Rees 2003). Since in our study we used adult individuals, those volunteers who had experienced such specific hair color change when being younger were grouped most likely in the brown hair category, although they may have blond associated genotypes. Consequently, these individuals would have lowered the prediction accuracy for brown relative to the brown-haired individuals who have not changed from blond. Our study design did not allow recording age-dependent hair color change, but this factor may be considered and tested in future studies. Although, volunteers in the red hair color group of our study was significantly younger at time of sampling than people in any other hair color category groups ($P < 0.01$), including age in the prediction modeling had only very little impact on the accuracy (AUC change <0.01). The age difference is most likely due to our targeted sampling procedure in which the red hair color category was over-sampled in young individuals (see material section for further details). In this study, gender was not significantly associated with any hair color and had no significant impact on hair color prediction accuracy.

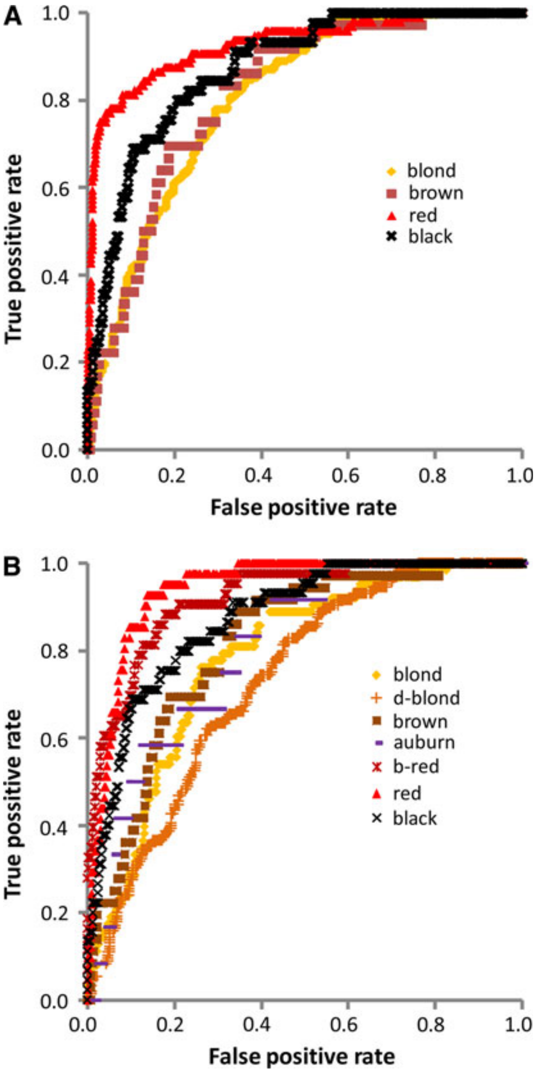

Fig. 2 ROC curves for the final model including 13 DNA predictors for 4 (**a**) and 7 (**b**) hair color categories in a Polish sample

Model-based hair color prediction analysis was also performed in a previous study using SNPs from *MC1R*, *OCA2/HERC2*, *SLC24A4*, *TYR*, *KITLG* and a marker from the region 6p25.3 (Sulem et al. 2007) that is close to the *IRF4* and *EXOC2* hair color candidate genes (Han et al. 2008). However, the prediction approach used by Sulem et al. (2007) is not directly comparable with ours; they applied a two-step approach and the steps did not only differ in the predicted hair colors, but also in the genetic markers used. First, they predicted red hair by only using the two most important red hair associated polymorphisms in *MC1R* (rs1805007 and rs1805008) and found that from those Icelandic individuals (used for replication) who were predicted with >50% probability to have red hair, about 70% indeed had red hair. To make these previous findings more comparable with ours, we performed red hair prediction in our data by using only rs1805007 and rs1805008 as used by Sulem et al. (2007) and received an AUC of 0.83. Notably, this value is considerably lower than the one we received for red hair using all markers analyzed in the present study (0.93 or 0.94). Hence, we can conclude that the additional SNPs we used in our full model, in particular the additional *MC1R* SNPs, improved red hair color prediction accuracy in our study. In a second step, Sulem et al. (2007) used associated SNPs from all 6 loci to predict blond, dark blond/light brown, and brown/black hair color categories. They found in their Icelandic replication set that among the individuals for whom brown hair color was predicted with >50% probability, about 60% indeed had brown/black hair. However, their prediction results were much less convincing for blond since, from the individuals predicted to be blond with only >40% probability (the highest threshold reported for blond), less than 50% were indeed blond, but about 50% were dark blond/brown and a few percentage were dark or red. Performing AUC prediction in our samples only with the SNPs used by Sulem et al. (2007) resulted in AUC values of 0.69 for blond, 0.71 for brown, and 0.75 for black. Again, AUCs for all non-red hair color categories as achieved in the present study considerably exceed those estimated from the markers used by Sulem et al. (2007), which demonstrates the extra value of the additional markers we included in our model for accurate prediction also of non-red hair colors. A recently published candidate gene study employed linear regression modeling using SNPs from hair color candidate genes and found that three SNPs in *HERC2*, *SLC45A2* and *SLC24A5* together explain 76% of total hair melanin in the study population (Valenzuela et al. 2010).

It has been shown that the least absolute shrinkage and selection operator (LASSO) approach (Tibshirani 1996) can be used to estimate marker effects of thousands of SNPs in linkage disequilibrium (LD) (Usai et al. 2009). Because some of the SNPs included in our study were in LD, we additionally performed the multinomial LASSO regression and compared the prediction results with those from our multinomial logistic regression model. The AUC estimates from LASSO using all samples (AUC blond = 0.88, brown = 0.89, red = 0.96, black = 0.96) are slightly higher than the ones from the multinomial logistic regression (Table 3). However, the average AUC values from the 1,000 cross-validations of the LASSO approach (AUC blond = 0.66, brown = 0.62, red = 0.86, and black = 0.76) are considerably lower than the ones obtained from all samples with the same approach, and are also lower than the results from the multinomial logistic regression (Table 3). This may indicate that there is a potential over-fitting problem in the LASSO method and our data.

Because the sample size used in this study is relatively small ($N = 385$), we estimated the effect of the total sample size on the accuracy of pigmentation prediction using a bootstrap analysis of the eye color data published previously (Liu et al. 2009), in which a AUC value of 0.91 was obtained for predicting blue eye color based on a large population sample ($N = 6{,}168$). As evident from Supplementary Figure S1, if the total sample size is smaller than 300 individuals, the AUC value for blue eye color tends to be under-estimated with large confidence intervals. For example, with only 100 samples the mean AUC value from 1,000 bootstrap analyses was considerably lower (AUC = 0.85, 95% CI: [0.6–1.0]; Figure S1) than the value of 0.91 as achieved with thousands of samples (Liu et al. 2009). However, this effect quickly diminishes when the sample size increases, and with about 350 samples the mean AUC value was close (AUC = 0.90, 95% CI: [0.80–0.97]; Figure S1) to the value obtained from thousands of samples, and only increased marginally until 800 samples. From this example of blue eye color we may extrapolate that the AUCs for hair color obtained from the 385 samples included in the present study (which are similar to the AUC obtained for blue eye color) are unlikely to change drastically when more individuals are added to the hair color model.

Many genetic studies on hair color (as well as eye and skin color) use phenotypic information provided by self-assessment, i.e. questionnaires filled out by the individual participants (e.g. Sulem et al. 2007, 2008; Han et al. 2008), which may be expected not to be completely reliable. To avoid hair color phenotype uncertainties potentially generated by such multiple-observer approach, we performed single-observer hair color grading in the present study. Some studies applied quantitative measures of hair color (Valenzuela et al. 2010; Mengel-From et al. 2009; Shekar et al. 2008). However, it is not clear how these methods as well as self-assessment and single-observer hair color categorization compare to each other and what the impact on DNA-based prediction accuracies is. On the one hand Vaughn et al. (2008) in a phenotypic study found some differences between single-observer hair color grading and spectrophotometric measurement, but the sample size was low (with about 100 individuals). On the other hand Shekar et al. (2008) in a genetic study could not confirm the utility of spectrophotometric estimation in relation to hair color rating. The single-observer grading approach we applied in the present study was found to be more accurate than using self-assessed hair color grading (Vaughn et al. 2008).

In conclusion, we demonstrated that human hair color categories can be accurately predicted from a relatively small number of DNA variants. The prediction accuracies achieved here for red and black hair color were in the similarly high precision range as previously obtained for blue and brown eye color, for which practical applications has already been implemented (Walsh et al. 2010a, b). Slightly lower prediction accuracies obtained here for blond and brown hair color, which were still higher than previously observed for non-blue/non-brown eye color (Liu et al. 2009), may be influenced by age-dependent hair color change during adolescence, which shall be investigated in more detail in future studies. Although our example of using eye color to monitor the effect of sample size to the AUC-based prediction accuracy of pigmentation traits indicate that the sample size used here for hair color prediction is large enough to obtain a reasonably accurate prediction model, our results may be further replicated in a larger study. Furthermore, it shall be tested in future studies if and to what extent SNPs from other genes with recently reported hair color association not used here add to the hair color prediction accuracy as presented. Overall, we evidently present hair color as the third externally visible characteristic that can be reliably predicted from DNA data after iris color (Liu et al. 2009; Walsh et al. 2010a, b; Valenzuela et al. 2010; Mengel-From et al. 2010), and human age, the latter demonstrated recently using quantification of T-cell DNA rearrangement (Zubakov et al. 2010). We therefore expect DNA-based hair color prediction, e.g. using the markers suggested here, to be used in future practical applications, such as in the forensic context. Furthermore, our study demonstrates that markers not statistically significantly associated with a trait in a study population can still independently contribute to the trait prediction in the same population, a notion that shall be considered in the design of future genetic prediction studies, including for diseases risks.

Acknowledgments We would like to thank all sample donors for having participated in this project. The study was supported by grant from the Ministry of Science and Higher Education in Poland no ON301115136 (fund of science in years 2009–2012) to WB and received additional support by funds from the Netherlands Forensic Institute to MK, and by a grant from the Netherlands Genomics Initiative/Netherlands Organization for Scientific Research (NWO) within the framework of the Forensic Genomics Consortium Netherlands to MK.

Note added in proof Because Valenzuela et al. (2010) recently showed that rs1426654 in SLC24A5 is one of three SNPs that in their study occurred most frequently with highest R2 values in multiple linear regression models for total hair melanin, and we initially did not consider this marker in our study design, we subsequently genotyped rs1426654 in our study population using SNaPshot technology. We found that of the 385 Polish individuals studied, 380 (98.7%) were homozygote for the derived allele AA (Thr111 ? Thr111) and none gave the ancestral homozygote GG. Since the five heterozygote AG carriers showed various hair color categories (1x blond, 2x dark-blond, 2x black), our data are inconclusive with an hypothesis that rs1426654 (SLC24A5) is associated with hair color variation (P>0.99), and we find rs1426654 not informative for DNA-based hair color estimation in our studied population. Our data, together with previously reported very high frequencies of the derived A allele in European populations and very high frequencies of the ancestral G allele in Africans and East Asians [Lamason et al. (2005)

Science 310:1782–1786; Soekima and Koda (2007) Int J Legal Med 121:36–39; Dimisianos et al. (2009) Exp Dermatol 18(2):175–7], suggest that the findings for rs1426654 in Valenzuela et al. (2010) using individuals of various ethnic backgrounds may in respect of hair melanin (most likely also eye color at least) simply reflect biogeographic ancestry rather than genuine hair (or eye) color association.

References

Alaerts M, Del-Favero J (2009) Searching genetic risk factors for schizophrenia and bipolar disorder: learn from the past and back to the future. Hum Mutat 30:1139–1152

Box NF, Wyeth JR, O'Gorman LE, Martin NG, Sturm RA (1997) Characterization of melanocyte stimulating hormone receptor variant alleles in twins with red hair. Hum Mol Genet 6:1891–1897

Brand A, Brand H, Schulte in den Bäumen T (2008) The impact of genetics and genomics on public health. Eur J Hum Genet 16:5–13

Branicki W, Brudnik U, Kupiec T, Wolańska-Nowak P, Wojas-Pelc A (2007) Determination of phenotype associated SNPs in the *MC1R* gene. J Forensic Sci 52:349–354

Branicki W, Brudnik U, Draus-Barini J, Kupiec T, Wojas-Pelc A (2008a) Association of the *SLC45A2* gene with physiological human hair colour variation. J Hum Genet 53:966–971

Branicki W, Brudnik U, Kupiec T, Wolańska-Nowak P, Szczerbińska A, Wojas-Pelc A (2008b) Association of polymorphic sites in the *OCA2* gene with eye colour using the tree scanning method. Ann Hum Genet 72:184–192

Branicki W, Brudnik U, Wojas-Pelc A (2009) Interactions between *HERC2*, *OCA2* and *MC1R* may influence human pigmentation phenotype. Ann Hum Genet 73:160–170

Brudnik U, Branicki W, Wojas-Pelc A, Kanas P (2009) The contribution of melanocortin 1 receptor gene polymorphisms and the agouti signaling protein gene 8818A>G polymorphism to cutaneous melanoma and basal cell carcinoma in a Polish population. Exp Dermatol 18:167–174

Chung CC, Magalhaes WC, Gonzalez-Bosquet J, Chanock SJ (2010) Genome-wide association studies in cancer—current and future directions. Carcinogenesis 31:111–120

Duffy DL, Montgomery GW, Chen W, Zhao ZZ, Le L, James MR, Hayward NK, Martin NG, Sturm RA (2007) A three-single-nucleotide polymorphism haplotype in intron 1 of *OCA2* explains most human eye-color variation. Am J Hum Genet 80:241–252

Eiberg H, Troelsen J, Nielsen M, Mikkelsen A, Mengel-From J, Kjaer KW, Hansen L (2008) Blue eye color in humans may be caused by a perfectly associated founder mutation in a regulatory element located within the *HERC2* gene inhibiting *OCA2* expression. Hum Genet 123:177–187

Fernandez LP, Milne RL, Pita G, Avilés JA, Lázaro P, Benítez J, Ribas G (2008) *SLC45A2*: a novel malignant melanoma associated gene. Hum Mutat 29:1161–1167

Flanagan N, Healy E, Ray A, Philips S, Todd C, Jackson IJ, Birch-Machin MA, Rees JL (2000) Pleiotropic effects of the melanocortin 1 receptor (*MC1R*) gene on human pigmentation. Hum Mol Genet 9:2531–2537

Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. J Stat Softw 33:1–22

Frudakis T, Thomas M, Gaskin Z, Venkateswarlu K, Chandra KS, Ginjupalli S, Gunturi S, Natrajan S, Ponnuswamy VK, Ponnuswamy KN (2003) Sequences associated with human iris pigmentation. Genetics 165:2071–2083

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

Grimes EA, Noake PJ, Dixon L, Urquhart A (2001) Sequence polymorphism in the human melanocortin 1 receptor gene as an indicator of the red hair phenotype. Forensic Sci Int 122:124–129

Han J, Kraft P, Nan H, Guo Q, Chen C, Qureshi A, Hankinson SE, Hu FB, Duffy DL, Zhao ZZ, Martin NG, Montgomery GW, Hayward NK, Thomas G, Hoover RN, Chanock S, Hunter DJ (2008) A genome-wide association study identifies novel alleles associated with hair color and skin pigmentation. PLoS Genet 4:e1000074

Harding RM, Healy E, Ray AJ, Ellis NS, Flanagan N, Todd C, Dixon C, Sajantila A, Jackson IJ, Birch-Machin MA, Rees JL (2000) Evidence for variable selective pressures at *MC1R*. Am J Hum Genet 66:1351–1361

Janssens AC, van Duijn CM (2008) Genome-based prediction of common diseases: advances and prospects. Hum Mol Genet 17:166–173

Janssens AC, Pardo MC, Steyerberg EW, van Duijn CM (2004) Revisiting the clinical validity of multiplex genetic testing in complex diseases. Am J Hum Genet 74:585–588; author reply 588–589

Kanetsky PA, Swoyer J, Panossian S, Holmes R, Guerry D, Rebbeck TR (2002) A polymorphism in the agouti signaling protein gene is associated with human pigmentation. Am J Hum Genet 70:770–775

Kanetsky PA, Ge F, Najarian D, Swoyer J, Panossian S, Schuchter L, Holmes R, Guerry D, Rebbeck TR (2004) Assessment of polymorphic variants in the melanocortin-1 receptor gene with cutaneous pigmentation using an evolutionary approach. Cancer Epidemiol Biomarkers Prev 13:808–819

Kayser M, Schneider PM (2009) DNA-based prediction of human externally visible characteristics in forensics: motivations, scientific challenges, and ethical considerations. Forensic Sci Int Genet 3:154–161

Kayser M, Liu F, Janssens AC, Rivadeneira F, Lao O, van Duijn K, Vermeulen M, Arp P, Jhamai MM, van Ijcken WF et al (2008) Three genome-wide association studies and a linkage analysis identify *HERC2* as a human iris color gene. Am J Hum Genet 82:411–423

Ku CS, Loy EY, Pawitan Y, Chia KS (2010) The pursuit of genome-wide association studies: where are we now? J Hum Genet 55:195–206

Liu F, van Duijn K, Vingerling JR, Hofman A, Uitterlinden AG, Janssens AC, Kayser M (2009) Eye color and the prediction of complex phenotypes from genotypes. Curr Biol 19:192–193

Liu F, Wollstein A, Hysi PG, Ankra-Badu GA, Spector TD, Park D, Zhu G, Larsson M, Duffy DL, Montgomery GW et al (2010) Digital quantification of human eye color highlights genetic association of three new loci. PLoS Genet 6:e1000934

McCarthy MI, Zeggini E (2009) Genome-wide association studies in type 2 diabetes. Curr Diabetes Rep 9:164–171

Mengel-From J, Wong TH, Morling N, Rees JL, Jackson IJ (2009) Genetic determinants of hair, eye colours in the Scottish, Danish populations. BMC Genet 10:88

Mengel-From J, Borsting C, Sanchez JJ, Eiberg H, Mohrling N (2010) Human eye color and *HERC2, OCA2,* and *MATP*. Forensic Sci Int Genet 4:323–328

Pastorino L, Cusano R, Bruno W, Lantieri F, Origone P, Barile M, Gliori S, Shepherd GA, Sturm RA, Bianchi-Scarra G (2004) Novel *MC1R* variants in Ligurian melanoma patients and controls. Hum Mutat 24:103

Rana BK, Hewett-Emmett D, Jin L, Chang BH, Sambuughin N, Lin M, Watkins S, Bamshad M, Jorde LB, Ramsay M, Jenkins T, Li WH (1999) High polymorphism at the human melanocortin 1 receptor locus. Genetics 151:1547–1557

Rebbeck TR, Kanetsky PA, Walker AH, Holmes R, Halpern AC, Schuchter LM, Elder DE, Guerry D (2002) P gene as an inherited biomarker of human eye color. Cancer Epidemiol Biomarkers Prev 11:782–784

Rees JL (2003) Genetics of hair and skin color. Annu Rev Genet 37:67–90

Shekar SN, Duffy DL, Frudakis T, Sturm RA, Zhao ZZ, Montgomery GW, Martin NG (2008) Linkage and association analysis of spectrophotometrically quantified hair color in Australian adolescents: the effect of *OCA2* and *HERC2*. J Invest Dermatol 128:2807–2814

Sturm RA, Duffy DL, Zhao ZZ, Leite FP, Stark MS, Hayward NK, Martin NG, Montgomery GW (2008) A single SNP in an evolutionary conserved region within intron 86 of the *HERC2* gene determines human blue-brown eye color. Am J Hum Genet 82:424–431

Sulem P, Gudbjartsson DF, Stacey SN, Helgason A, Rafnar T, Magnusson KP, Manolescu A, Karason A, Palsson A, Thorleifsson G et al (2007) Genetic determinants of hair, eye and skin pigmentation in Europeans. Nat Genet 39:1443–1452

Sulem P, Gudbjartsson DF, Stacey SN, Helgason A, Rafnar T, Jakobsdottir M, Steinberg S, Gudjonsson SA, Palsson A, Thorleifsson G et al (2008) Two newly identified genetic determinants of pigmentation in Europeans. Nat Genet 40:835–837

Tibshirani R (1996) Regression shrinkage and selection via the Lasso. J Roy Stat Soc B 58:267–288

Usai MG, Goddard ME, Hayes BJ (2009) LASSO with cross-validation for genomic selection. Genet Res 91:427–436

Valenzuela RK, Henderson MS, Walsh MH, Garrison NA, Kelch JT, Cohen-Barak O, Erickson DT, John Meaney F, Bruce Walsh J, Cheng KC, Ito S, Wakamatsu K, Frudakis T, Thomas M, Brilliant MH (2010) Predicting phenotype from genotype: normal pigmentation. J Forensic Sci 55(2):315–322

Valverde P, Healy E, Jackson I, Rees JL, Thody AJ (1995) Variants of the melanocyte-stimulating hormone receptor gene are associated with red hair and fair skin in humans. Nat Genet 11:328–330

Vaughn M, van Oorschot R, Baindur-Hudson S (2008) Hair color measurement and variation. Am J Phys Anthropol 137:91–96

Voisey J, Gomez-Cabrera Mdel C, Smit DJ, Leonard JH, Sturm RA, van Daal A (2006) A polymorphism in the agouti signaling protein (*ASIP*) is associated with decreased levels of mRNA. Pigment Cell Res 19:226–231

Walsh S, Lindenbergh A, Zuniga SB, Sijen T, de Knijff P, Kayser M, Ballantyne KN (2010a) Developmental validation of the IrisPlex system: determination of blue and brown iris colour for forensic intelligence. Forensic Sci Int Genet. doi:10.1016/j.fsigen.2010.09.008

Walsh S, Liu F, Ballantyne KN, van Oven M, Lao O, Kayser M (2010b) IrisPlex: a sensitive DNA tool for accurate prediction of blue and brown eye colour in the absence of ancestry information. Forensic Sci Int Genet. doi:10.1016/j.fsigen.2010.02.004

Zubakov D, Liu F, van Zelm MC, Vermeulen J, Oostra BA, van Duijn CM, Driessen GJ, van Dongen JJM, Kayser M, Langerak AW (2010) Estimating human age from T cell DNA rearrangements. Curr Biol 20(22):R970