Genome-wide association meta-analysis of individuals of European ancestry identifies new loci explaining a substantial fraction of hair color variation and heritability

Pirro G. Hysi, [1,2]

Ana Valdes, [1,3]

Fan Liu, [4,5,6]

Nicholas A. Furlotte, [7]

David M. Evans, [8,9]

Veronique Bataille, [1]

Alessia Visconti, [1]

Gibran Hemani, [9]

George McMahon, [9]

Susan M. Ring, [9]

George Davey Smith, [9]

David L. Duffy, [10]

Gu Zhu, [10]

Scott D. Gordon, [10]

Sarah E. Medland, [10]

Bochao D. Lin, [11]

Gonneke Willemsen, [11]

Jouke Jan Hottenga, [11]

Dragana Vuckovic, [12,13]

Giorgia Girotto, [12,13]

Ilaria Gandin, [13]

Cinzia Sala, [14]

Maria Pina Concas, [15]

Marco Brumat, [13]

Paolo Gasparini, [12,13]

Daniela Toniolo, [14]

Massimiliano Cocca, [13]

Antonietta Robino, [15]

Seyhan Yazar, [16]

Alex W. Hewitt, [16,17,18]

Yan Chen, [4,5]

Changqing Zeng, [4]

Andre G Uitterlinden, [19,20]

M. Arfan Ikram, [20]

Merel A. Hamer, [21]

Cornelia M. van Duijn, [20]

Tamar Nijsten, [21]

David A. Mackey, [16]

Mario Falchi, [1]

Dorret I Boomsma, [11]

Nicholas G. Martin, [10]

David A. Hinds, [7]

Manfred Kayser, [6]✉

Email m.kayser@erasmusmc.nl

Timothy D. Spector, [1]✉

Email tim.spector@kcl.ac.uk

The International Visible Trait Genetics Consortium
The International Visible Trait Genetics Consortium

[1] King's College London Department of Twins Research and Genetic Epidemiology, London, UK

[2] Department of Ophthalmology, King's College London, London, UK

[3] Division of Rheumatology, Orthopaedics and Dermatology, School of Medicine, University of Nottingham, Nottingham, UK

[4] Key Laboratory of Genomic and Precision Medicine, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, PR China

[5] University of Chinese Academy of Sciences, Beijing, PR China

[6] Department of Genetic Identification, Erasmus MC University Medical Center Rotterdam, Rotterdam, The Netherlands

[7] 23andMe, Inc., Mountain View, CA, USA

[8] University of Queensland Diamantina Institute, Translational Research Institute, Brisbane, QC, Australia

[9] MRC Integrative Epidemiology Unit, University of Bristol, Bristol, UK

[10] QIMR Berghofer Medical Research Institute, Brisbane, Australia

[11] Netherlands Twin Register, Department of Biological Psychology, Vrije Universiteit, Amsterdam, The Netherlands

[12] Experimental Genetics Division, Sidra Medical and Research Centre, Doha, Qatar

[13] Department of Medical Sciences, University of Trieste, Trieste, Italy

[14] Division of Genetics of Common Disorders, S. Raffaele Scientific Institute, Milan, Italy

[15] Institute for Maternal and Child Health IRCCS "Burlo Garofolo", Trieste, Italy

[16] Centre for Ophthalmology and Visual Science, University of Western Australia, Lions Eye Institute, Perth, WA, Australia

[17] Centre for Eye Research Australia, University of Melbourne, Department of Ophthalmology, Royal Victorian Eye and Ear Hospital, Melbourne, Australia

[18] School of Medicine, Menzies Research Institute Tasmania, University of Tasmania, Hobart, Australia

[19] Department of Internal Medicine, Erasmus MC University Medical Center Rotterdam, Rotterdam, The Netherlands

[20] Department of Epidemiology, Erasmus MC University Medical Center Rotterdam, Rotterdam, The Netherlands

[21] Department of Dermatology, Erasmus MC University Medical Center Rotterdam, Rotterdam, The Netherlands

Received: 17 December 2016 / Accepted: 7 March 2018

Abstract

Hair color is one of the most recognizable visual traits in European populations and is under strong genetic control. Here we report the results of a genome-wide association study meta-analysis of almost 300,000 participants of European descent. We identified 123 autosomal and one X-chromosome loci significantly associated with hair color; all but 13 are novel. Collectively, single-nucleotide polymorphisms associated with hair color within these loci explain 34.6% of red hair, 24.8% of blond hair, and 26.1% of black hair heritability in the study

populations. These results confirm the polygenic nature of complex phenotypes and improve our understanding of melanin pigment metabolism in humans.

AQ1

Genome-wide meta-analysis identifies >100 loci associated with hair color variation in humans of European ancestry. These loci explain a large portion of the heritability of these traits & provide insights into pathways regulating hair pigmentation.

A list of members and affiliations appears in the Supplementary Note.

These authors contributed equally: Pirro G. Hysi, Ana Valdes, Fan Liu.

These authors jointly supervised this work: Manfred Kayser, Timothy D. Spector.

Main

Human pigmentation refers to coloration of external tissues due to variations in quantity, ratio, and distribution of the two main types of the pigment melanin: eumelanin and pheomelanin[1]. Most melanin is produced by melanosomes[2, 3], large organelles specialized in melanin synthesis and transportation located mainly in the epidermis, hair, and iris, as well as the central nervous system. Early humans had darkly pigmented skin[4, 5], which protected against high ultraviolet radiation (UVR) and its consequences, such as skin cancer[6] and folate depletion[7]. European and Asian populations evolved lighter skin pigmentation[8, 9] as they migrated toward northern latitudes with lower UVR[4]. The lighter pigmentation maximizes UVR absorption needed to maintain adequate vitamin D levels. In Europeans, pigmentation of skin, hair, and/or eyes has characteristic geographic distributions because of natural selection[10] and perhaps genetic drift; a role for sexual selection has been debated[11, 12, 13].

AQ2

AQ3

AQ4

Hair color is one of the most prominent traits in humans. Twin studies suggest that up to 97% of variation in hair color may be explained by heritable factors[14], and genome-wide association studies (GWAS)[15, 16, 17, 18, 19, 20] have identified several chromosomal regions associated with hair color and related pigmentation

traits[21]. Except for red hair, known variants have a relatively low predictive value[22], and the heritability gap remains relatively large[14], which suggests that many hair color genes remain undiscovered.

Here we report the results of a meta-analysis of two GWAS carried out in two large discovery cohort studies: 157,653 research participants from the 23andMe, Inc. customer base[18] and 133,238 individuals from the UK Biobank (UKBB). Participants in both studies self-reported the natural color of their hair in adulthood (Supplementary Fig. 1 and Supplementary Note). For the purpose of this work, each hair color category collected (black, dark brown, light brown, red, and blond) was assigned numerical values ranging from lowest (blond) to highest (black). These codes were used as the outcome variable in linear regression-based GWAS analyses. To minimize population admixture and stratification, the analyses were restricted to individuals of European ancestry (Supplementary Figs. 2 and 3) and adjusted for the first ten principal components of the genotype matrix, as well as for age and sex.

The analyses confirmed a strong association between hair color and principal components, especially in the less ethnically homogeneous 23andMe dataset, which includes participants of more varied European origin, in line with the known north–south cline in hair color variation and other regional differences in hair color across Europe[12] (Supplementary Table 1). The strongest associations in both groups were with sex (Table 1). Women were more likely to report blond (OR = 1.20 and OR = 1.29 in the 23andMe and UKBB participants, respectively) or red hair (OR = 1.72 and OR = 1.40, respectively) than any other color and three to five times less likely to report black hair (OR = 0.35 and OR = 0.20, respectively) compared to men.

Table 1

Effect of sex on the hair color phenotypes in the 23andMe and UK Biobank cohorts

23andMe	Odds ratio	Standard error	95% confidence interval	
			Lower	Upper
Blond (all)	1.202	0.024	1.174	1.230
Red	1.721	0.014	1.675	1.768
Light brown	1.116	0.013	1.088	1.145

23andMe cohort, n = 157,653 independent participants; UKBB cohort, n = 133,238 independent participants.

23andMe	Odds ratio	Standard error	95% confidence interval	
			Lower	Upper
Dark brown	0.663	0.011	0.650	0.677
Black	0.348	0.030	0.329	0.369
UKBB	**Odds ratio**	**Standard error**	**95% confidence interval**	
			Lower	**Upper**
Blond	1.285	0.018	1.241	1.330
Red	1.395	0.026	1.325	1.469
Light brown	1.101	0.011	1.077	1.125
Dark brown	0.993	0.011	0.971	1.015
Black	0.195	0.033	0.182	0.208

23andMe cohort, $n = 157{,}653$ independent participants; UKBB cohort, $n = 133{,}238$ independent participants.

AQ5

Genomic inflation factors[23] (λ_{GC}) from the 23andMe and the UKBB GWAS were 1.147 and 1.146, respectively, in line with expectations of high power to detect large polygenic effects in these large samples[24] (Supplementary Fig. 4). Meta-analyzed GWAS results reached conventional genome-wide significance ($P < 5 \times 10^{-8}$) in many regions, primarily clustering around 123 distinct autosomal genomic single-nucleotide polymorphisms (SNPs) and one X-chromosomal locus (Fig. 1 and Supplementary Table 2), mostly new (Table 2). In line with power expectations (Supplementary Fig. 5), 75 of these regions were significant genome-wide in at least one of the two cohorts and always at least nominally significant ($P < 0.01$) in the other.

Fig. 1

Manhattan plot of the inverse variance meta-analysis for association with hair color of the 23andMe and UKBB cohorts (meta-analysis $n = 290{,}891$).

The unadjusted significance of association (y axis) for each SNP on different chromosomes is shown in alternating navy and green along the x axis, with polymorphisms reaching significance at GWAS level ($P < 5 \times 10^{-8}$) depicted in red. Values on the y axis were truncated at $P = 10^{-500}$.

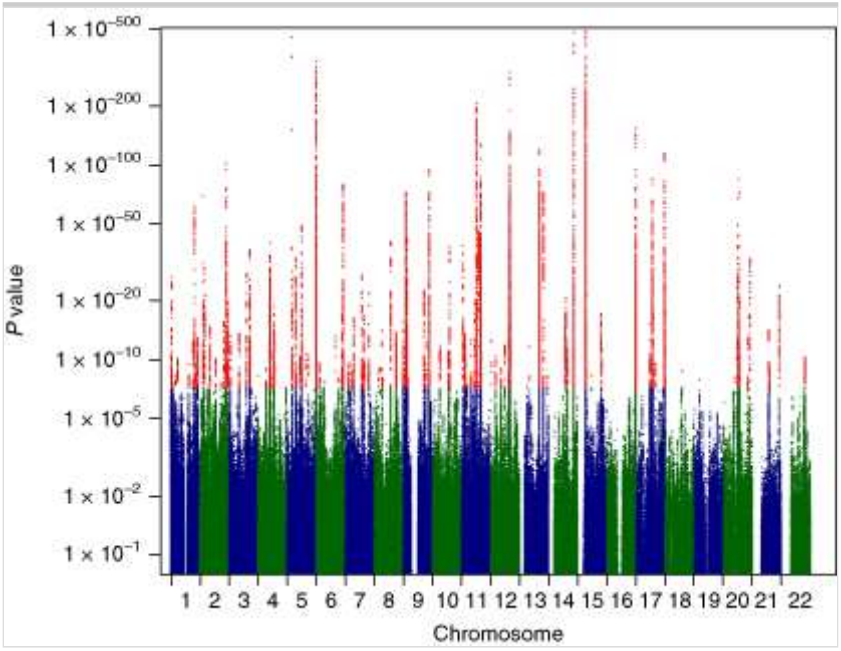


Table 2

A selection of genes newly associated with hair color

						UK Biobank			
Chr	**Position (Build 37)**	**SNP ID**	**Ref. allele**	**Freq**	**Nearest gene**	***n***	**Beta**	**s.e.m.**	
1	8207579	rs80293268	G	0.047	*SLC45A1*	132,221	0.194	0.009	1 × 9
1	205181062	rs2369633	T	0.089	*DSTYK*	132,887	−0.071	0.007	9 × 2
2	28613302	rs71443018	G	0.039	*FOSL2*	126,428	0.133	0.010	2 × 3
9	126808006	rs58979150	T	0.108	*LHX2*	132,883	0.089	0.006	1 × 4

The selection was based on the strength of their effect, which is defined as the standardized Biobank, 23andMe, and their meta-analysis, as well as for the meta-analysis results from the these results, and effect sizes (beta) are given in s.d. units. A, C, T, and G under the 'Refere for which the effect size and allele frequencies (freq.) are reported. Frequencies are given fo frequencies in the 23andMe and UK Biobank. Associations with *P* values of less than 10

						UK Biobank			
Chr	**Position (Build 37)**	**SNP ID**	**Ref. allele**	**Freq**	**Nearest gene**	***n***	**Beta**	**s.e.m.**	
13	78391757	rs1279403	T	0.406	*EDNRB*	133,238	−0.086	0.004	1 1
15	48426484	rs1426654	G	0.021	*SHC4*	133,238	0.188	0.069	(
17	39551099	rs117612447	T	0.029	*KRT31*	133,238	0.063	0.011	2 × 8
20	52661068	rs73132911	T	0.046	*BCAS1*	132,836	0.089	0.009	6 × 2

The selection was based on the strength of their effect, which is defined as the standardized Biobank, 23andMe, and their meta-analysis, as well as for the meta-analysis results from the these results, and effect sizes (beta) are given in s.d. units. A, C, T, and G under the 'Refere for which the effect size and allele frequencies (freq.) are reported. Frequencies are given fo frequencies in the 23andMe and UK Biobank. Associations with *P* values of less than 10^{-10}

Previously known pigmentation loci were all strongly associated in the meta-analysis results: *HERC2* (rs12913832), *IRF4* (rs12203592), and *MC1R* (rs1805007), as well as others, showed some of the strongest statistical evidence for association ever published for human complex traits. Strong associations were found for genes whose mutations reportedly cause impairment of pigmentation, such as Waardenburg (*EDNRB*, rs1279403, $P < 10^{-100}$; *MITF*, rs9823839, $P < 10^{-100}$), Hermansky–Pudlak (*HPS5*, rs201668750, $P = 4.68 \times 10^{-11}$), trichomegaly (*FGF5*, rs7681907, $P = 5.684 \times 10^{-25}$), or Ablepharon macrostomia (*TWIST2*, rs11684254, $P = 1.233 \times 10^{-20}$) syndromes. Many polymorphisms significantly ($P < 5 \times 10^{-8}$) associated with hair color in our meta-analysis had existing entries in the GWAS catalog[21]. In previous publications, they were associated with several phenotypes, including most known pigmentation loci (Supplementary Table 3).

Among the associated loci, some of the strongest effects were observed for two solute carrier 45 A family members (*SLC45A1*, rs80293268, $P < 10^{-100}$; and *SLC45A2*, rs16891982, $P < 10^{-100}$); polymorphisms near a third solute carrier gene

were also significantly associated with the trait (rs60086398 upstream of *SLC7A1*, $P = 4.93 \times 10^{-8}$). In addition, forkhead box family genes (*FOXO6*, rs3856254, $P = 4.0 \times 10^{-9}$; and *FOXE1*, rs3021523, $P = 4.23 \times 10^{-23}$) and sex-determining region Y (SRY)-box genes (*SOX5*, rs9971729, $P = 8.8 \times 10^{-17}$; and *SOX6*, rs1531903, $P = 9.1 \times 10^{-16}$) were among those highlighted in our results. An additional locus, located on chromosome X in the second intron of the collagen type IV alpha 6 gene, was also significantly associated (*COL46A*, rs1266744, $P = 5.03 \times 10^{-12}$). Chromosome Y information was not analyzed. Notably, given the observed strong association of hair color with sex, there was no particular difference in effect sizes observed for these loci among men and women in either cohort (Supplementary Table 4 and Supplementary Fig. 6); only one SNP significantly associated with hair color in the meta-analysis showed significant ($P = 1.6 \times 10^{-8}$) interaction with sex in the 23andMe (Supplementary Table 5), but much weaker interaction in the UK Biobank cohort ($P = 0.04$). As reported before[10], some hair color genes are subject to significant natural selection (Supplementary Table 6); SNPs associated with hair color in our meta-analysis tended to have lower selection score centiles and higher than average evidence for natural selection within European populations ($P = 0.04$) and compared to Africans (Supplementary Fig. 7).

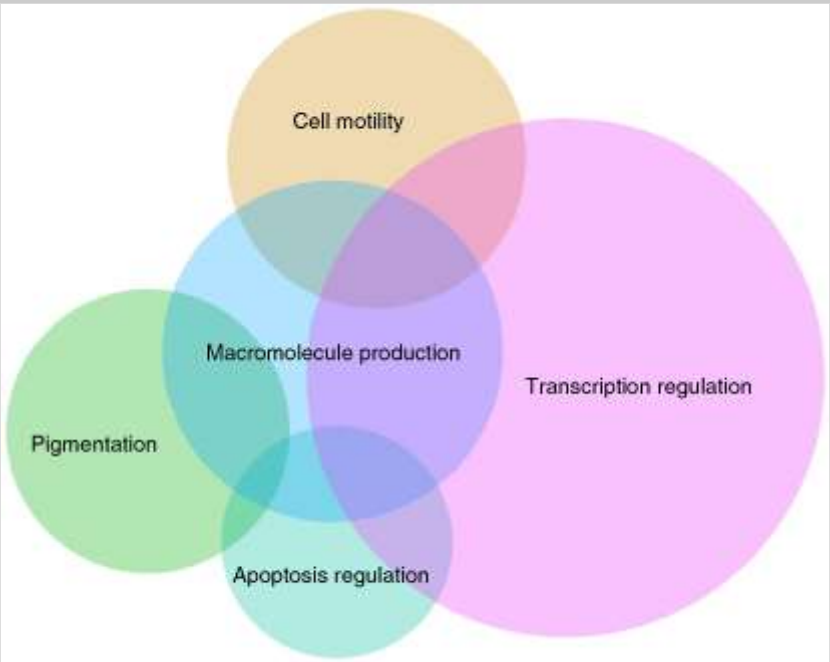
To further validate the results and to introduce a testing dataset, we collected GWAS summary statistics from ten additional cohorts with 27,865 European participants from the International Visible Trait Genetics (VisiGen) Consortium[25] and meta-analyzed them. For 114 of the 123 autosomal loci highlighted by the discovery GWAS meta-analysis, the direction of the association was the same as observed in the meta-analysis; despite the lower statistical power of the replication due to smaller sample sizes, most leading SNP loci from the discovery meta-analysis (75 of the 123 autosomal regions) replicated at least at a nominal level and the same direction of association ($P < 0.05$). For 35 of these loci, the association was stronger even after correction for multiple testing (Supplementary Table 2).

Next, we assessed the potential relationship of the most associated polymorphisms and expression of the genes nearest to them. In line with most previous GWAS[26], the majority of these polymorphic loci had expression quantitative trait loci (eQTL) effects in several tissues. The strongest associations were observed with transcript of the *CBWD1* (rs478882, $P = 1.30 \times 10^{-30}$), *PPM1A* (rs7154748, $P = 3.30 \times 10^{-14}$), and *RALY* genes (rs6059655 being associated with *ASIP* gene expression, $P = 6.0 \times 10^{-9}$) in sun-exposed skin tissues (Supplementary Table 7). As expected, genes showing the strongest association in the meta-analysis were significantly enriched

for several Gene Ontology entries, especially pigmentation and melanin biosynthetic and metabolic processes (Fig. 2 and Supplementary Table 8).

Fig. 2

Gene Ontology Biological Processes annotations for genes adjacent to the SNPs showing the strongest associations with hair color via GWAS meta-analysis in the 23andMe and UKBB cohorts.



Conditional analysis of the discovery cohorts identified 258 SNPs independently associated with hair color (Supplementary Table 9). These SNPs explain overall 20.68% of the hair color heritability (using ordinal categories) and 34.58% (s.e. = 3.64%) of the population liability scale[27] heritability for red hair (vs. any other color, assuming population prevalence is as in the UKBB at 0.047), 24.80% for blond hair (s.e.m. = 2.49%, assuming a prevalence of 0.11) and 26.12% (s.e.m. = 3.11%) of the black hair heritability (prevalence 0.046; Table 3).

Table 3

Phenotypic variance explained by the identified autosomal loci significantly associated with hair color

	Current heritability estimates					Previous estimates	
Phenotype	$V(G)/V_p$	**s.e.m.**	$V(G)/V_{p\_L}$	**s.e.m.**	**Prevalence**	$V(G)/V_p$	**s.e.m.**

	Current heritability estimates					Previous estimates	
Phenotype	$V(G)/V$	**s.e.m.**	$V(G)/V$	**s.e.m.**	**Prevalence**	$V(G)/V$	**s.e.m.**
Blond	0.094	0.009	0.248	0.025	0.113	0.058	0.022
Red	0.074	0.008	0.346	0.036	0.046	0.069	0.069
Black	0.056	0.007	0.261	0.031	0.047	0.005	0.005

The current estimates are given as the ratio of the genetic variance, $V(G)$, over the phenotypic variance (V_p) and scaled over the population prevalence, $V(G)/V_{p\_L}$, (estimated in the UKBB cohort, n = 133,238), on the right. The estimates of genetic variance explained by known SNPs before this study were taken from previous publications. The phenotypes in this table were compared with all other hair colors. Since 80% of the participants reported some shade of brown hair color (dark or light), the heritabilities for these two phenotypes were considered baseline and were not calculated.

AQ6

Finally, we modeled hair color prediction in two cohorts (QIMR n = 7,283; RS n = 7,724) using the 258 independently associated SNPs from the discovery GWAS meta-analysis (Supplementary Table 9) together with previously reported SNP predictors for hair color from the HIrisPlex system[28]. We split the data into model building (80%) and validation (20%) sets to assure that marker discovery, model building, and model validation were independently executed. In both cohorts, prediction accuracies were high for black (QIMR area under the curve (AUC) = 0.91, RS AUC = 0.81) and red (AUC = 0.87 and 0.84, respectively) hair colors, but lower for blond (AUC = 0.79 and 0.74, respectively) and brown (AUC = 0.76 and 0.64, respectively; Supplementary Table 10 and Supplementary Fig. 8). Using the same datasets, these new models outperformed the previous HIrisPlex model[22] (QIMR and RS black AUC = 0.82 vs. 0.77, red AUC = 0.87 vs. 0.83, blond AUC = 0.67 vs. 0.65, brown AUC = 0.66 vs. 0.57; Supplementary Table 10).

AQ7

Our work has identified over 100 new genetic loci involved in hair pigmentation in Europeans and raises several questions. First, the observation of higher prevalence of lighter hair colors among women (Supplementary Fig. 9) follows previous findings based on objective quantitative measurement of hair color[29, 30], suggesting that sex is truly associated with hair color, independent of socially driven self-reporting bias. Second, although hair pigmentation spans a spectrum from very bright (blond) to very dark (black), the genetic mechanisms do not always follow this linear scale, as red hair color often has unique predisposing genetic factors[16, 17]. However, our results explain even higher portions of

heritability than before[14] for all hair colors, not just for the extremes of the light–dark hair color spectrum. Third, hair color is a trait that follows special distribution patterns and therefore is prone to issues of population structure bias that may be controlled for in several ways. A comparison of different methodologies (Supplementary Fig. 10) shows that our approach is roughly equivalent with others. Fourth, annotation of genetic regions based on physical distances and association probability most likely underestimates the number of regions involved in hair pigmentation. For example, the involvement of *OCA2* and *HERC2* genes in human pigmentation is not simply due to linkage disequilibrium[31], yet because of their proximity, both loci in our study were assigned to the same association region. This would, however, not affect the conditional analysis at a marker level, which discriminates separate effects.

In conclusion, this large GWAS meta-analysis has improved our knowledge of the genetic control of human hair pigmentation by bringing the number of known loci into the hundreds. The newly identified genetic loci explain substantial portions of the hair color phenotypic variability and can guide future research into better understanding the functional mechanisms linking these genes to pigmentation variation. Our findings may also be useful in the future both for better molecular understanding of human pigmentation, including its DNA-based prediction as relevant in forensic and anthropological applications, and for the diseases that result from biological impairment of pigmentation, including the development of treatment strategies.

URLs

Description of the hair color phenotyping in the UK Biobank participants: https://biobank.ctsu.ox.ac.uk/crystal/field.cgi?id = 1747.

Description of the genotyping procedures for the UK Biobank participants: http://www.ukbiobank.ac.uk/wp-content/uploads/2014/04/Affymetrix-UKB_WCSGAX-Genotype-Data-Generation.pdf.

Genotype imputation and genetic association studies of UK Biobank Interim Data Release, May 2015: http://www.ukbiobank.ac.uk/wp-content/uploads/2014/04/imputation_documentation_May2015.pdf.

Methods

The 23andMe cohort

All research participants were drawn from the customer base of 23andMe, Inc., a consumer genetics company. This cohort has been described in detail previously[32]. All participants included in the analyses provided informed consent and answered surveys online according to our human subjects protocol, which was reviewed and approved by Ethical & Independent Review Services, a private institutional review board (http://www.eandireview.com). Hair color phenotypes were used to create an ordinal trait with values: 0, light blond; 1, dark blond; 2, red; 3, light brown; 4, dark brown; 5, black. DNA extraction and genotyping were performed on saliva samples by CLIA-certified and CAP-accredited clinical laboratories of Laboratory Corporation of America. Samples were genotyped on one of four genotyping platforms. The V1 and V2 platforms were variants of the Illumina HumanHap550 + BeadChip, including about 25,000 custom SNPs selected by 23andMe for a total of about 560,000 SNPs. The V3 platform was based on the Illumina OmniExpress + BeadChip, with custom content to improve the overlap with our V2 array, with a total of about 950,000 SNPs. The V4 platform in current use is a fully custom array, including a lower-redundancy subset of V2 and V3 SNPs, with additional coverage of lower-frequency coding variations, and about 570,000 SNPs. Samples that failed to reach a 98.5% call rate were re-analyzed. For the GWAS only, participants with > 97% European ancestry, as determined through an analysis of local ancestry, were included. For the purposes of ethnic categorization, an algorithm first partitioned phased genomic data into short windows of about 100 SNPs and used a support vector machine (SVM) to classify individual haplotypes into one of 31 reference populations. The SVM classifications then fed into a hidden Markov model (HMM) that accounts for switch errors and incorrect assignments, and gives probabilities for each reference population in each window. The reference population data are derived from public datasets (the Human Genome Diversity Project, HapMap, and 1,000 Genomes), as well as 23andMe customers who have reported having four grandparents from the same country. A maximal set of unrelated individuals was chosen for each analysis using a segmental identity-by-descent (IBD) estimation algorithm[33]. Individuals were defined as related if they shared more than 700 cM IBD, including regions where the two individuals share either one or both genomic segments identical-by-descent. This level of relatedness corresponds approximately to the minimal expected sharing between first cousins in an outbred population.

Participant genotype data were imputed against the September 2013 release of 1,000 Genomes Phase 1 reference haplotypes, phased with ShapeIt2[34]. We phased and imputed data for each genotyping platform separately. We phased using an internally developed phasing tool that implements the Beagle haplotype graph-

based phasing algorithm[35], modified to separate the haplotype graph construction and phasing steps.

SNPs with Hardy–Weinberg equilibrium $P < 10^{-20}$, call rate < 95%, or with large allele frequency discrepancies compared to European 1,000 Genomes reference data were excluded from imputation. Imputation was done against all-ethnicity 1,000 Genomes haplotypes (excluding monomorphic and singleton sites) using Minimac[36]. For the X chromosome, separate haplotype graphs were built for the non-pseudoautosomal region and each pseudoautosomal region, and these regions were phased separately. Males and females were imputed together using Minimac2[36], as with the autosomes, treating males as homozygous pseudo-diploids for the non-pseudoautosomal region.

Association test results were computed by linear regression, assuming additive allelic effects. For tests, imputed dosages rather than best-guess genotypes were used. Covariates for age, gender, and the top five principal components to account for residual population structure were also included into the model. Results for the X chromosome were computed similarly, with male genotypes coded as if they were homozygous diploid for the observed allele.

HLA allele dosages were imputed from SNP genotype data using HIBAG[37]. We imputed alleles for HLA-A, -B, -C, -DPB1, -DQA1, -DQB1, and -DRB1 loci at four-digit resolution. To test associations between HLA allele dosages and phenotypes, we performed logistic or linear regression using the same set of covariates as that used in the SNP-based GWAS for that phenotype. We performed separate association tests for each imputed allele.

The UK Biobank

The UK Biobank database includes 502,682 participants who were aged from 49–69 years when recruited between 2006 and 2010 from across the UK to take part in the project. The study was approved by the National Research Ethics Committee (REC reference 11/NW/0382). The participants filled out several questionnaires about their lifestyle, environmental risk factors, and medical history and gave their informed consent[38]. The participants were invited, through a computerized questionnaire, to answer the question “What best describes your natural hair color? (If your hair color is grey, the color before you went grey)”. The participants’ answers were used to create a hair color variable with values: 1, blond; 2, red; 3, light brown; 4, dark brown; 5, black; and other codes for “other,” “don’t know,” or “prefer not to answer.” The later three values were removed from analyses.

Extracted DNA was then processed in the approximate order received to produce genotype data using the Affymetrix Axiom platform, as described elsewhere (see URLs). Details on genotyping procedure and quality control can be found elsewhere (see URLs). Phasing on the autosomes was carried out using a version of the SHAPEIT2[34] program modified to allow for very large sample sizes. The new algorithm uses a divisive clustering algorithm to identify clusters of haplotypes and calculates Hamming distances only between pairs of haplotypes within each cluster. Only haplotypes within each cluster are used as candidates for the surrogate family copying states in the HMM model. Imputation was carried out using the same algorithm as is implemented in the IMPUTE2 program. More detailed information on the imputation procedure followed can be found elsewhere (see URLs). Linear models were built for the main GWAS analysis: haircolor ~genotype + age + sex + PC1:10 + genotyping platform.

Replication cohorts (the VisiGen Consortium)

Subjects were individuals of European descent participating in any of the 10 studies from the International Visible Trait Genetics Consortium[39] (VisiGen). The VisiGen participants were phenotyped through self-report, and each phenotypic category was assigned a unique numerical value within each cohort. The self-reported hair color categories were, however, highly heterogeneous across the 10 studies. Therefore, all participants (n = 27,865) were genotyped, imputed, and analyzed separately by each participating center, using standard techniques described in the Supplementary Note.

Meta-analyses

Results from each participating cohort (23andMe and UK Biobank) were standardized to allow for the different scales (six categories of hair color in the former but only five in the latter) and minimize differences arising from the slightly different categorizations of the same phenotype. Both weighted z-scores and inverse variance analyses (the latter using standardized linear regression coefficients and standard errors) were calculated using Metal[40]. The results obtained from both methods were similar, and inverse-variance results are reported throughout the manuscript for the discovery cohort. Association effect sizes, standard errors, and probabilities were taken from the association analyses software. When the significance of the association exceeded the range of float numbers determined by the system and software, the probabilities were calculated using Mathematica 11.1 computational algebra (Wolfram Research Inc.). Meta-analyses of the replication

cohorts were calculated using the weighted *z*-score method to reflect the fact that the phenotypic definitions were not harmonized across the different participating cohorts (please refer to the Supplementary Note for more detailed population description and phenotypic definitions).

Conditional and explained heritability analyses

The program GCTA[41] was used for conditional analyses[42] to identify independent effects within associated loci as well as to calculate the phenotypic variance explained[43] by all polymorphisms, genotyped or imputed, associated with the trait after the conditional analyses. The threshold of significance was set at $P < 5 \times 10^{-8}$ and the colinearity threshold was $r^2 = 0.8$. These estimates were derived from the UKBB cohort.

Natural selection

Results of three statistical tests for natural selection were obtained from the 1,000 Genomes selection browser[44]. Results from three selected tests are reported: iHS[45] and two cross-population comparison (XP–EHH tests, CEU vs. YRI, and CEU vs. CHB) based on the extended haplotype homozygosity test[46]. The absolute test scores and rank scores ($-\log_{10}$ of the centile of the absolute test score across the genome) for each SNP of interest are reported.

Prediction analyses

Using the independently associated SNPs identified in the discovery GWAS meta-analysis (Supplementary Table 9) together with previously reported hair color predicting SNPs from the HIrisPlex System[22, 28], we performed hair color prediction modeling in the QIMR, RS, and combined QIMR + RS datasets. For this, we randomly split the QIMR, RS, and combined QIMR + RS data into 80% training sets and 20% validation sets, respectively. This approach assures the use of totally independent datasets for predictive marker discovery, model building, and model validation. Of the 258 independently associated SNPs in the discovery GWAS meta-analysis (see Supplementary Table 9), five SNPs failed imputation quality control in the RS and one in the QIMR and were therefore not included in the models. In the combined QIMR + RS analysis, we used the overlapping set of SNPs. The performance of the prediction models was evaluated in the respective validation datasets using the area under the receiver operating characteristic (ROC) curve (AUC). AUC is the integral of ROC curves, which ranges from 0.5, representing total lack of prediction, to 1.0, representing perfect prediction.

Prediction analyses were conducted in R version 3.2.3 using relevant packages, including lars, nnet, and pROC.

Effects of variants on gene expression

The potential eQTL effects of the variants of interest were evaluated in all 57 tissues available at the GTEx[47] portal (see URLs). Associations with the levels of expressions of adjacent genes were assessed for all variants identified through the conditional analysis.

Gene set enrichment analyses

Gene set enrichment analyses were carried out on summary results obtained from the meta-analysis of the UK Biobank and 23andMe subjects using Magenta software[48].

Data availability

This work used data from two primary sources. The original datasets can be accessed as follows.

For UK Biobank, data can be accessed through the UK Biobank Access management (https://www.ukbiobank.ac.uk/register-apply/). The hair color data accession codes are 1747.0.0, 1747.1.0, and 1747.2.0. The participants' age UK Biobank accession code is 21022, for sex 31.0.0, and the precomputed principal components used here are 22009.0.1 through 22009.0.10.

For the 23andMe participants, requests for summary statistics can be accessed at https://researchers.23andme.org/collaborations. There are no accession codes available.

For the TwinsUK datasets, access can be requested through http://www.twinsuk.ac.uk/data-access/, and access to the secondary source of data through the corresponding authors.

Methods

Methods, including statements of data availability and any associated accession codes and references, are available at https://doi.org/10.1038/s41588-018-0100-5.

Electronic supplementary material

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-018-0100-5.

Acknowledgements

This research has been conducted using the UK Biobank Resource under Application Number 12052.
The ALSPAC work is supported by a Medical Research Council program grant (MC_UU_12013/4 to D.M.E). The UK Medical Research Council and the Wellcome Trust (grant refs: 092731 and 102215/2/13/2) and the University of Bristol provide core support for ALSPAC. D.M.E. is supported by an Australian Research Council Future Fellowship (FT130101709). This publication is the work of the authors and D.M.E. will serve as guarantor for the contents of this paper. ALSPAC GWAS data was generated by Sample Logistics and Genotyping Facilities at the Wellcome Trust Sanger Institute and LabCorp (Laboratory Corporation of America) using support from 23andMe.
The ERF Study was supported by the joint grant from the Netherlands Organization for Scientific Research (NWO, 91203014), the Center of Medical Systems Biology (CMSB), Hersenstichting Nederland, Internationale Stichting Alzheimer Onderzoek (ISAO), Alzheimer Association project number 04516, Hersenstichting Nederland project number 12F04(2).76, and the Interuniversity Attraction Poles (IUAP) program. As a part of EUROSPAN (European Special Populations Research Network), ERF was supported by European Commission FP6 STRP grant number 018947 (LSHG-CT-2006-01947) and also received funding from the European Community's Seventh Framework Programme (FP7/2007-2013)/grant agreement HEALTH-F4-2007-201413 by the European Commission under the program "Quality of Life and Management of the Living Resources" of 5th Framework Programme (no. QLG2-CT-2002-01254). High-throughput analysis of the ERF data was supported by joint grant from Netherlands Organization for Scientific Research and the Russian Foundation for Basic Research (NWO-RFBR 047.017.043).
The INGI research was supported by funds from Compagnia di San Paolo, Torino, Italy; Fondazione Cariplo, Italy and Ministry of Health, Ricerca Finalizzata 2008 and CCM 2010 and Telethon, Italy. Additional support was provided by the Italian Ministry of Health (RF 2010 to PG), FVG Region, and Fondo Trieste.

The NTR study was supported by multiple grants from the Netherlands Organization for Scientific Research (NWO: 016-115-035, 463-06-001, 451-04-034), ZonMW (31160008, 911-09-032); from the Institute for Health and Care Research (EMGO +); and from the Biomolecular Resources Research Infrastructure (BBMRI-NL, 184.021.007), European Research Council (ERC-230374). Genotyping was made possible by grants from NWO/SPI 56-464-14192, Genetic Association Information Network (GAIN) of the Foundation for the National Institutes of Health, Rutgers University Cell and DNA Repository (NIMH U24 MH068457-06), the Avera Institute, Sioux Falls (USA), and the National Institutes of Health (NIH R01 HD042157-01A1, MH081802, Grand Opportunity grants 1RC2 MH089951 and 1RC2 MH089995). B.D.L. is supported by a PhD grant (201206180099) from the China Scholarship Council.
QIMR funding was provided by the Australian National Health and Medical Research Council (241944, 339462, 389927, 389875, 389891, 389892, 389938, 442915, 442981, 496739, 552485, 552498), the Australian Research Council (A7960034, A79906588, A79801419, DP0770096, DP0212016, DP0343921), the FP-5 GenomEUtwin Project (QLG2-CT-2002-01254), and the US National Institutes of Health (NIH grants AA07535, AA10248, AA13320, AA13321, AA13326, AA14041, MH66206). Statistical analyses were carried out on the Genetic Cluster Computer, which is financially supported by the Netherlands Scientific Organization (NWO 480-05-003). S.E.M. and D.L.D. are supported by the National Health and Medical Research Council (NHMRC) Fellowship Scheme. The 20-year follow-up of Generation 2 of the Raine Study was funded by Australian National Health and Medical Research Council (NHMRC) project grant 1021105, Lions Eye Institute, the Australian Foundation for the Prevention of Blindness, and the Ophthalmic Research Institute of Australia. S.Y. is supported by NHMRC Early Career Fellowship (CJ Martin - Overseas Biomedical Fellowship). The Rotterdam Study is supported by the Netherlands Organization of Scientific Research NWO Investments (nr. 175.010.2005.011, 911-03-012). This study is funded by the Research Institute for Diseases in the Elderly (014-93-015; RIDE2), the Netherlands Genomics Initiative (NGI)/Netherlands Organization for Scientific Research (NWO) project nr. 050-060-810. The Rotterdam Study is supported by the Erasmus MC and Erasmus University Rotterdam; the Netherlands Organization for Scientific Research (NWO); the Netherlands Organization for Health Research and Development (ZonMw); the Research Institute for Diseases in the Elderly (RIDE) the Netherlands Genomics Initiative (NGI); the Ministry of Education, Culture and Science; the Ministry of Health Welfare and Sport; the European Commission (DG XII); and the Municipality of Rotterdam. The generation and management of GWAS genotype data for the Rotterdam Study were executed by the Human

Genotyping Facility of the Genetic Laboratory of the Department of Internal Medicine, Erasmus MC.
The TwinsUK study was funded by the Wellcome Trust (105022/Z/14/Z); European Community's Seventh Framework Programme (FP7/2007-2013). The study also receives support from the National Institute for Health Research (NIHR)-funded BioResource, Clinical Research Facility and Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust in partnership with King's College London. SNP genotyping was performed by the Wellcome Trust Sanger Institute and National Eye Institute via NIH/CIDR.

AQ8

Author contributions

P.G.H., A. Valdes, and F.L. jointly wrote the manuscript, coordinated meta-analyses, and performed prediction modeling. N.A.F., D.M.E., V.B., A. Visconti, G.H., G.M., S.M.R., D.L.D., G.Z., S.D.G., S.E.M., B.D.L., G.W., J.J.H., D.V., G.G., I.G., C.S., M.P.C., M.B., D.T., M.C., A.R., S.Y., A.W.H., Y.C., C.Z., A.G.U., M.A.H., T.N., M.F., and D.A.H. each conducted part of the analyses described in this work. G.D.S., P.G., C.M.v.D., M.A.I., D.A.M., D.I.B., N.G.M., and M.F. contributed populations samples and data used for analyses. M.K. and T.D.S. jointly coordinated the work and participated in manuscript preparation.

Competing interests N.A.F. and D.A.H. are employees of the 23andMe Inc. consumer genetics company.

AQ9

Integrated supplementary information

Supplementary Figure 1

Distribution of self-reported hair color in the UK Biobank and 23andMe cohorts.

Plots show the densities for each hair color category in the UK Biobank (top) and 23andMe cohorts (bottom).

Supplementary Figure 2

Plot of the first five principal components in UK Biobank and 23andMe subjects included in the analyses.

The top-left panel shows the participants of European descent that were included in the analyses (blue) in the backdrop of all multi-ethnic UK Biobank and 23andMe participants. The plots for any two of the first principal components for the subjects of European descent (included in the analyses) are also given.

Supplementary Figure 3

Violin plots of the first principal components for the UK Biobank and 23andMe participants.

Each violin plot shows the distribution of principal components computed in the UK Biobank subjects (top) and the 23andMe cohort (bottom).

Supplementary Figure 4

Expected variation of genomic control inflation factor (λ) as a function of sample size.

The expected λ were calculated using the same number of loci, LD structure and trait heritability observed in the 23andMe cohort (blue) and the UK Biobank cohort (red) in absence of factors artificially inflating association. These genomic inflation factors observed in the discovery cohort are a reflection of the power of these large sample analyzed in presence of polygenicity; they would be merely the equivalent of a $\lambda_{GC} = 1.0186$ for the 23andMe cohort and $\lambda_{GC} = 1.0221$ for the UKBB cohort had the effective sample sizes been smaller ($n = 20{,}000$) and the underlying genetic effects remained the same.

Supplementary Figure 5

Power to detect genome-wide significant associations.

Plot generated using a sample of $n = 140{,}000$ subjects at $\alpha = 10^{-7}$. Power is shown for the most representative range of allele frequencies detected in GWAS (MAF = 0.10, red line; MAF = 0.30, blue line). The value on the *y*-axis denotes the probability that a true association would be detected at the pre-defined α; for example, for a locus with a MAF = 0.10 with an effect size of 0.035 s.d. per each copy of the risk allele over the trait, there is a 60% probability to replicate the same association at a genome-wide level of significance.

Supplementary Figure 6

Correlation of effect sizes between men and women participating in the UK Biobank and the 23andMe cohorts.

The black line denotes parity of effects, and each point represents one of the GWAS associated SNPs shown in Supplementary Table 2.

Supplementary Figure 7

Evidence of natural selection for the SNPs significantly associated with hair color.

Three different natural selection tests were selected (see Online Methods), and the values plotted here represent the centile rank of the natural selection score within European populations (iHS test, red), compared to YRI Africans (XP-EHH CEU vs. YRI, green) and to CHB Chinese populations (blue) using 1000 Genomes data. The plot shows an enrichment for higher ranks (lower centiles) of selection for some SNPs within European populations (two tailed Wilcoxon test $P = 0.04$), evidence for stronger selection for these SNPs in Europeans compared to Africans (two-tailed Wilcoxon $P = 0.014$), but less significant compared to Chinese (two tailed Wilcoxon $P = 0.056$).

Supplementary Figure 8

Hair color prediction in participants of the Rotterdam and QIMR studies.

The prediction model is based on multinomial logistic regression including 20 HIrisPlex SNPs, a polygenic score for blond-brown-black (233 SNPs), and a polygenic score for red (25 SNPs); see **Supplementary Note** for marker and model specifications.

Supplementary Figure 9

Sex-specific prevalence of hair color categories in the UK Biobank and 23andMe cohorts.

In both cohorts, there is a higher prevalence of blond, red and light brown hair colors among women and a higher prevalence of black and dark brown hair colors among men.

Supplementary information

Supplementary Figures Reporting Summary

Supplementary Figures 1–9

Supplementary Note and Supplementary Tables

Supplementary Note 1 and Supplementary Tables 1–10

References

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

2. Randhawa, M. et al. Evidence for the ectopic synthesis of melanin in human adipose tissue. *FASEB J.* **23**, 835–843 (2009).

3. Sturm, R. A., Teasdale, R. D. & Box, N. F. Human pigmentation genes: identification, structure and consequences of polymorphic variation. *Gene* **277**, 49–62 (2001).

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

5. Jablonski, N. G. & Chaplin, G. Colloquium paper: human skin pigmentation as an adaptation to UV radiation. *Proc. Natl Acad. Sci. USA* **107Suppl 2**, 8962–8968 (2010).

6. Greaves, M. Was skin cancer a selective force for black pigmentation in early hominin evolution? *Proc. Biol. Sci.* **281**, 20132955 (2014).

7. Branda, R. F. & Eaton, J. W. Skin color and nutrient photolysis: an evolutionary hypothesis. *Science* **201**, 625–626 (1978).

8. Norton, H. L. et al. Genetic evidence for the convergent evolution of light skin in Europeans and East Asians. *Mol. Biol. Evol.* **24**, 710–722 (2007).

9. Wilde, S. et al. Direct evidence for positive selection of skin, hair, and eye pigmentation in Europeans during the last 5,000y. *Proc. Natl Acad. Sci. USA* **111**, 4832–4837 (2014).

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

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

12. Frost, P. European hair and eye color - a case of frequency-dependent sexual selection? *Evol. Hum. Behav.* **27**, 85–103 (2006).

13. Madrigal, L. & Kelly, W. Human skin-color sexual dimorphism: a test of the sexual selection hypothesis. *Am. J. Phys. Anthropol.* **132**, 470–482 (2007).

14. Lin, B. D. et al. Heritability and genome-wide association studies for hair color in a Dutch twin family based sample. *Genes (Basel)* **6**, 559–576 (2015).

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

16. Han, J. et al. A genome-wide association study identifies novel alleles associated with hair color and skin pigmentation. *PLoS Genet.* **4**, e1000074 (2008).

17. Sulem, P. et al. Two newly identified genetic determinants of pigmentation in Europeans. *Nat. Genet.* **40**, 835–837 (2008).

18. Eriksson, N. et al. Web-based, participant-driven studies yield novel genetic associations for common traits. *PLoS Genet.* **6**, e1000993 (2010).

19. Kenny, E. E. et al. Melanesian blond hair is caused by an amino acid change in TYRP1. *Science* **336**, 554 (2012).

20. Zhang, M. et al. Genome-wide association studies identify several new loci associated with pigmentation traits and skin cancer risk in European Americans. *Hum. Mol. Genet.* **22**, 2948–2959 (2013).

21. MacArthur, J. et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45**(D1), D896–D901 (2017).

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

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

24. Yang, J. et al. Genomic inflation factors under polygenic inheritance. *Eur. J. Hum. Genet.* **19**, 807–812 (2011).

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

26. Nicolae, D. L. et al. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet.* **6**, e1000888 (2010).

27. Lee, S. H., Wray, N. R., Goddard, M. E. & Visscher, P. M. Estimating missing heritability for disease from genome-wide association studies. *Am. J. Hum. Genet.* **88**, 294–305 (2011).

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

29. 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. *BMC Genet.* **10**, 88 (2009).

30. Shekar, S. N. et al. Spectrophotometric methods for quantifying pigmentation in human hair-influence of *MC1R* genotype and environment. *Photochem. Photobiol.* **84**, 719–726 (2008).

31. Visser, M., Kayser, M. & Palstra, R. J. *HERC2* rs12913832 modulates human pigmentation by attenuating chromatin-loop formation between a long-range enhancer and the *OCA2* promoter. *Genome Res.* **22**, 446–455 (2012).

32. Pickrell, J. K. et al. Detection and interpretation of shared genetic influences on 42 human traits. *Nat. Genet.* **48**, 709–717 (2016).

33. Henn, B. M. et al. Cryptic distant relatives are common in both isolated and cosmopolitan genetic samples. *PLoS One* **7**, e34267 (2012).

34. Delaneau, O., Zagury, J. F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods* **10**, 5–6 (2013).

35. Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* **81**, 1084–1097 (2007).

36. Fuchsberger, C., Abecasis, G. R. & Hinds, D. A. Minimac2: faster genotype imputation. *Bioinformatics* **31**, 782–784 (2015).

37. Zheng, X. et al. HIBAG-HLA genotype imputation with attribute bagging. *Pharmacogenomics J.* **14**, 192–200 (2014).

38. Allen, N. et al. UK Biobank: current status and what it means for epidemiology. *Health Policy Technol.* **1**, 123–126 (2012).

39. Keating, B. et al. First all-in-one diagnostic tool for DNA intelligence: genome-wide inference of biogeographic ancestry, appearance, relatedness, and sex with the Identitas v1 Forensic Chip. *Int. J. Legal Med.* **127**, 559–572 (2013).

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

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

42. Yang, J. et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* **44**, 369–375 (2012). S1–S3.

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

44. Pybus, M. et al. 1000 Genomes Selection Browser 1.0: a genome browser dedicated to signatures of natural selection in modern humans. *Nucleic Acids Res.* **42**, D903–D909 (2014).

45. Voight, B. F., Kudaravalli, S., Wen, X. & Pritchard, J. K. A map of recent positive selection in the human genome. *PLoS Biol.* **4**, e72 (2006).

46. Sabeti, P. C. et al. Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**, 913–918 (2007).

47. Consortium, G. T. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).

48. Segrè, A. V., Groop, L., Mootha, V. K., Daly, M. J. & Altshuler, D. Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. *PLoS Genet.* **6**, e1001058 (2010).