**RESEARCH ARTICLE SUMMARY**

**HUMAN EVOLUTION**

**Loci associated with skin pigmentation identified in African populations**

Nicholas G. Crawford, Derek E. Kelly,* Matthew E. B. Hansen,* Marcia H. Beltrame,* Shaohua Fan,* Shanna L. Bowman,* Ethan Jewett,* Alessia Ranciaro, Simon Thompson, Yancy Lo, Susanne P. Pfeifer, Jeffrey D. Jensen, Michael C. Campbell, William Beggs, Farhad Hormozdarian, Sununguko Wata Mpoloka, Gaonyadiwe George Mokone, Thomas Nyambo, Dawit Wolde Meskel, Gurja Belay, Jake Haut, NISC Comparative Sequencing Program, Harriet Rothschild, Leonard Zon, Yi Zhou, Michael A. Kovacs, Mai Xu, Tongwu Zhang, Kevin Bishop, Jason Sinclair, Cecilia Rivas, Eugene Elliot, Jiyeon Choi, Shenghao A. Li, Belynda Hicks, Shawn Burgess, Christian Abnet, Dawn E. Watkins-Chow, Elena Oceana, Yun S. Song, Eleazar Eskin, Kevin M. Brown, Michael S. Marks,† Stacie K. Loftus,† William J. Pavan,† Meredith Yeager,† Stephen Chanock,† Sarah A. Tishkoff†;

**INTRODUCTION:** Variation in pigmentation among human populations may reflect local adaptation to regional light environments, because dark skin is more photoprotective, whereas pale skin aids the production of vitamin D. Although genes associated with skin pigmentation have been identified in European populations, little is known about the genetic basis of skin pigmentation in Africans.

**RATIONALE:** Genetically and phenotypically diverse African populations are informative for mapping genetic variants associated with skin pigmentation. Analysis of the genetics of skin pigmentation in Africans informs upon melanocyte biology and the evolution of skin pigmentation in humans.

**RESULTS:** We observe extensive variation in skin pigmentation in Africa, with lowest melanin levels observed in southern African San hunter-gatherers and highest levels in East African Nilo-Saharan pastoralists. A genome-wide association study (GWAS) of 1570 Africans identified variants significantly associated with skin pigmentation, which clustered in four genomic regions that together account for almost 30% of the phenotypic variation.

GWAS and functional assays illuminate the genetic basis of pigmentation in Africa. A GWAS identified four genomic regions associated with skin pigmentation in Africa. Functional assays in melanocytes, zebrafish, and mice characterized their impact on skin pigmentation. Evolutionary genetic analyses revealed that most derived variants evolved before the origin of modern humans. Ma, million years ago.

The most significantly associated single-nucleotide polymorphisms were at SLC24A5, a gene associated with pigmentation in Europeans. We show that SLC24A5 was introduced into East Africa >5 thousand years ago (ka) and has risen to high frequency.

The second most significantly associated region is near the gene MFSD12. In vitro and in vivo analyses, we show that MFSD12 codes for a lysosomal protein that modifies pigmentation in human melanocytes, with decreased MFSD12 expression associated with darker pigmentation. We also show that genetic knockouts of MFSD12 orthologs affect pigmentation in both zebrafish and mice.

A third highly associated region encompasses a cluster of genes that play a role in ultraviolet (UV) response and DNA damage repair. We find the strongest associations in a regulatory region upstream of DDB1, the gene encoding damage-specific DNA binding protein 1, and that these variants are associated with increased expression of DDB1. The alleles associated with light pigmentation swept to near fixation outside of Africa due to positive selection, and we show that these lineages coalesce ~60 ka, corresponding with the time of migration of modern humans out of Africa.

The fourth significantly associated region encompasses the OCA2 and HERC2 loci. We identify previously uncharacterized variants at HERC2 associated with the expression of OCA2. These variants arose independently from eye and skin pigmentation–associated variants in non-Africans. We also identify variants at OCA2 that are correlated with alternative splicing; alleles associated with light pigmentation are correlated with a shorter transcript, which lacks a trans-membrane domain.

**CONCLUSION:** We identify previously uncharacterized genes and variants associated with skin pigmentation in ethnically diverse Africans. These genes have diverse functions, from repairing UV damage to playing important roles in melanocyte biology. We show that both dark and light pigmentation alleles arose before the origin of modern humans and that both light and dark pigmented skin has continued to evolve throughout hominin history. We show that variants associated with dark pigmentation in Africans are identical by descent in South Asian and Australo-Melanesian populations. This study sheds light on the evolutionary history, and adaptive significance, of skin pigmentation in humans.

The list of author affiliations is available in the full article online.

*These authors contributed equally to this work.
†These authors contributed equally to this work.
‡Corresponding author. Email: tishkoff@pennmedicine.upenn.edu.

Cite this article as N. G. Crawford et al., Science 358, eaan8433 (2017). DOI: 10.1126/science.aan8433
HUMAN EVOLUTION

Loci associated with skin pigmentation identified in African populations

Nicholas G. Crawford,1 Derek E. Kelly,1,2* Matthew E. B. Hansen,1 Marcia H. Beltrame,1 Shaohua Fan,1 Shanna L. Bowman,3,4† Simon Thompson,1 Yancy Lo,1 Susanne P. Pfeifer,7 Jeffrey D. Jensen,7 Michael C. Campbell,1,8 William Beggs,1 Farhad Hormozdari,9,10 Sununguko Wata Mpoloka,11 Gaonyadive George Mokone,12 Thomas Nyambo,13 Dawit Wolde Meskel,14 Gurja Belay,14 Jake Haut,1 NISCI Comparative Sequencing Program,† Harriet Rothschild,15 Leonard Zon,15,16 Yi Zhou,15,17 Michael A. Kovač,18 Mai Xu,18 Tongwu Zhang,18 Kevin Bishop,19 Jason Sinclair,19 Cecilia Rivas,20 Eugene Elliot,20 Jiyeon Choi,18 Shengchao A. Li,21,22 Belynda Hicks,21,22 Shawn Burgess,19 Christian Abnet,21 Dawn E. Watkins-Chow,20 Elena Oceana,23 Yun S. Song,5,6,24,25,26 Eleazar Eskini,27 Kevin M. Brown,18 Michael S. Marks,5,6,24,25,26 Stacie K. Loftus,20,† William J. Pavan,20,† Meredith Yeager,21,22 Stephen Chanock,21,† Sarah A. Tishkoff1,28,§

Despite the wide range of skin pigmentation in humans, little is known about its genetic basis in global populations. Examining ethnically diverse African genomes, we identify variants in or near SLC24A5, MFSD12, DDB1, TMEM138, OCA2, and HERC2 that are significantly associated with skin pigmentation. Genetic evidence indicates that the light pigmentation variant at SLC24A5 was introduced into East Africa by gene flow from non-Africans. At all loci, variants associated with dark pigmentation in Africans are identical by descent in South Asian and Australo-Melanesian populations. Functional analyses indicate that MFSD12 encodes a lysosomal protein that affects melanogenesis in zebrafish and mice, and that mutations in melanocyte-specific regulatory regions near DDB1/TMEM138 correlate with expression of ultraviolet response genes under selection in Eurasians.

Skin pigmentation is highly variable within Africa

To identify genes affecting skin pigmentation in Africa, we used a DSm II ColorMeter to quantify light reflectance from the inner arm as a proxy for melanin levels in 2101 ethnically and genetically diverse Africans living in Ethiopia, Tanzania, and Botswana (table S1 and figs. S1 and S2) (8). Skin pigmentation levels vary extensively among Africans, with darkest pigmentation observed in Nilo-Saharan-speaking pastoralist populations in eastern Africa and lightest pigmentation observed in San hunter-gatherer populations from southern Africa (Fig. 2 and table S1).

A locus associated with light skin color in Europeans is common in East Africa

We genotyped 1570 African individuals with quantified pigmentation levels using the Illumina Infinium Omni5 Genotyping array. After quality control, we retained ~4.2 million biallelic single-nucleotide polymorphisms (SNPs) for analysis. A genome-wide association study (GWAS) analysis with linear mixed models, controlling for age, sex, and genetic relatedness (9), identified four regions with multiple significant associations (P < 5 × 10−8) (Fig. 1, fig. S3, and tables S2 and S3).

We then performed fine-mapping using local imputation of high-coverage sequencing data from a subset of 135 individuals and data from the Thousand Genomes Project (TGP) (Fig. 3 and table S3) (80). We ranked potential causal variants

*These authors contributed equally to this work. †National Institutes of Health Intramural Sequencing Center (NISC) Comparative Sequencing Program collaborators and affiliations are listed in the supplementary materials. §These authors contributed equally to this work. Corresponding author. Email: tishkoff@pennmedicine.upenn.edu
within each locus using CAVIAR, a fine-mapping method that accounts for linkage disequilibrium (LD) and effect sizes (Table 1) (11). We characterized global patterns of variation at these loci using whole-genome sequences from West African, Eurasian, and Australo-Melanesian populations (10, 12, 13).

The SNPs with strongest association with skin color in Africans were on chromosome 15 at or near the solute carrier family 24 member 5 (SLC24A5) gene (Figs. 1 and 3 and tables S2 and S3). A functional nonsynonymous mutation within SLC24A5 (rs1426654) (14) was significantly associated with skin color (F test, P = 5.5 × 10^{-12}) and was identified as potentially causal by CAVIAR (Table 1). The rs1426654 (A) allele is at high frequency in European, Pakistani, and Indian populations (Fig. 1) and is a target of selection in Europeans, Central Asians, and North Indians (15–18). In Africa, this variant is common (28 to 50% frequency) in populations from Ethiopia and Tanzania with high Afro-Asiatic ancestry (19, 20) and is at moderate frequency (5 to 11%) in San and Bantu-speaking populations from Botswana with low levels of East African ancestry and recent European admixture (Fig. 1 and figs. S2 and S4) (21, 22). We observe a signature consistent with positive selection at SLC24A5 in Europeans based on extreme values of Tajima’s D statistic (fig. S5).

On the basis of coalescent analysis with sequence data from the Simons Genomic Diversity Project (SGDP) (13), the time to most recent common ancestor (TMRCA) of most Eurasian lineages containing the rs1426654 (A) allele is 29 thousand years ago (ka) (95% critical interval [CI], 28 to 31 ka), consistent with previous studies (15, 17) (Fig. 4). Haplotype analysis indicates that the rs1426654 (A) variant in Africans is on the same extended haplotype background as Europeans (Fig. 5 and fig. S6), likely reflecting gene flow from western Eurasia over at least the past 3 to 9 ky (23). The rs1426654 (A) variant is at high frequency (28%) in Tanzanian populations, suggesting a lower bound (~5 ka) for introduction of this allele into East Africa, the time of earliest migration from Ethiopia into Tanzania (24). Furthermore, the frequency of the rs1426654 (A) variant in eastern and southern Africans exceeds the inferred proportion of non-African ancestry (figs. S2 and S4). Estimates of genetic differentiation (FST) at the rs1426654 SNP between the West African Yoruba (YRI) and Ethiopian Amhara populations is 0.76, among the top 0.01% of values on chromosome 15 (table S4). These results are consistent with selection for the rs1426654 (A) allele in African populations following introduction, although complex models of demographic history cannot be ruled out.

**Fig. 1. Correlations between allele frequencies at loci associated with pigmentation and UV exposure in global populations.** (A) Global variation in skin pigmentation indicated by melanin index (MI). These data were integrated with MI data for global populations from (1.105). (B) Mean erythemal dose rate. (C) Manhattan plot of −log10 transformed P values from GWAS of skin pigmentation with the Illumina Omni5 SNP array. (D) Quantile-quantile (QQ) plot of observed versus expected P values from the GWAS. In both (C) and (D), significant SNPs at P < 5 × 10^{-8} are highlighted in purple. (E to L) Allele frequencies of genetic variants associated with skin pigmentation in global populations. African populations included are Ethiopia Nilo-Saharan (1), Ethiopia Omotic (2), Ethiopia Semitic (3), Ethiopia and Tanzania Cushitic (4), Tanzania Nilo-Saharan (5), Tanzania Hadza (6), Tanzania Sandawe (7), Botswana Bantu (8), Botswana San/Bantu admixed (9), and Botswana San (10). The Melanesian (MEL) samples are from (12), and the Australian Aboriginal and Papua New Guinean samples (merged) are from the SGDP (PNG) (13). All other populations are from the TGP (10). Non-African populations in the Americas are indicated: CEU (European ancestry), ASW (African-American Southwest U.S.), and ACB (African Caribbean in Barbados).

### A lysosomal transporter protein associated with skin pigmentation

The region with the second strongest genetic association with skin pigmentation contains the major facilitator superfamily domain containing 12 (MFSD12) gene on chromosome 19 (Figs. 1 and 3 and tables S2 and S3). MFSD12 is homologous to other genes containing MFSD domains, conserved throughout vertebrates, which function as transmembrane solute transporters (25). MFSD12 mRNA levels are low in depigmented skin of vitiligo patients (26), likely due to autoimmune-related destruction of melanocytes.

The MFSD12 locus is in a region with extensive recombination, enabling us to fine-map eight potentially causal SNPs (Table 1 and table S3) that cluster in two regions: one within MFSD12 and the other ~7600 to 9000 base pairs (bp) upstream of MFSD12 (Fig. 3). Many SNPs are in predicted regulatory regions active in melanocytes and/or keratinocytes (Table 1 and Fig. 3) and show enhancer activity in luciferase expression assays in a WM88 melanoma cell line (Table 1, table S5, and fig. S7). Within MFSD12, the two SNPs that CAVIAR identifies as having the highest probability of being causal are rs56203814 (F test, P = 3.6 × 10^{-22}), a synonymous variant within exon 9, and rs10424065 (F test, P = 5.1 × 10^{-20}), located within intron 8. They are 130 bp apart, are in strong LD, and affect gene expression in luciferase expression assays (1.5 to 2.7x higher expression than wild type).
expression than the minimal promoter; fig. S7). The SNPs upstream of MFSD12 with highest probability of being causal are rs112332856 ($F$ test, $P = 3.8 \times 10^{-16}$) and rs6510760 ($F$ test, $P = 6.5 \times 10^{-15}$). They are 346 bp apart, are in strong LD, and affect gene expression in luciferase expression assays (4.0 to 19.7× higher expression than the minimal promoter; fig. S7).

The derived rs65203814 (T) and rs10424065 (T) alleles associated with dark pigmentation are present only in African populations (or those of recent African descent) and are most common in East African populations with Nilo-Saharan ancestry (Fig. 1 and fig. S4). Coalescent analysis of the SGDP data set indicates that the TMRCA for the derived rs6510760 (A) allele is 996 ka [95% CI, 0.82 to 1.2 million years ago (Ma); Fig. 4].

We do not detect evidence for positive selection at MFSD12 using Tajima’s $D$ and IHS statistics (figs. S5 and S8; as expected if selection were ancient [28]). However, levels of genetic differentiation are elevated when comparing East African Nilo-Saharan and western European (CEU) populations (for example, $F_{ST} = 0.85$ for rs112332856, top 0.05% on chromosome 19), consistent with differential selection at this locus (table S4) [29].

MFSD12 is within a cluster of 10 genes with high expression levels in primary human melanocytes relative to primary human keratinocytes [30], with MFSD12 as the most differentially expressed (90×; table S6). The genomic region (chr19:3541782-3581062) encompassing MFSD12 and neighboring gene HMG20B (a transcription factor common in melanocytes) has numerous deoxyribonuclease (DNase) I hypersensitive sites (DHS) and is enriched for H3K27ac enhancer marks in melanocytes (top 0.1% genome-wide; Fig. 3), suggesting that this region may regulate expression of genes critical to melanocyte function [31].

Analyses of gene expression using RNA sequencing (RNA-seq) data from 106 primary melanocyte cultures (table S7) indicate that African ancestry is significantly correlated with decreased MFSD12 gene expression [Pearson correlation coefficient (PCC), $P = 5.0 \times 10^{-7}$; fig. S9]. We observed significant associations between genotypes at rs6510760 and rs112332856 with expression of HMG20B [Bonferroni-adjusted $P$ ($P_{adj} < 4.9 \times 10^{-3}$) and MFSD12 ($P_{adj} < 3.4 \times 10^{-5}$) (fig. S9)]. In each case, the alleles associated with dark pigmentation correlate with decreased gene expression. Allele-specific expression (ASE) analysis indicates that individuals heterozygous for either rs6510760 or rs112332856 show increased allelic imbalance, relative to homozygotes, for MFSD12 (Mann-Whitney-Wilcoxon test, $P = 4.9 \times 10^{-8}$ and $1.3 \times 10^{-3}$, respectively), consistent with regulation of gene expression in cis. A haplotype containing the rs6510760 (A) and rs112332856 (C) alleles (associated with dark pigmentation) are common in all sub-Saharan Africans except the San, as well as in South Asian and Australo-Melanesian populations (Fig. 1 and fig. S4). Haplotype analysis places the rs6510760 (A) allele [and linked rs112332856 (C) allele] in Australo-Melanesians on similar haplotype backgrounds relative to central and eastern Africans (Fig. 5 and fig. S6), suggesting that they are identical by descent from an ancestral African population. Coalescent analysis of the SGDP data set indicates that the TMRCA for the derived rs6510760 (A) allele is 996 ka [95% CI, 0.82 to 1.2 million years ago (Ma); Fig. 4].

The derived rs65203814 (T) and rs10424065 (T) alleles associated with dark pigmentation are present only in African populations (or those of recent African descent) and are most common in East African populations with Nilo-Saharan ancestry (Fig. 1 and fig. S4). Coalescent analysis of the SGDP data set indicates that the TMRCA for the derived rs6510760 (A) allele is 996 ka [95% CI, 0.82 to 1.2 million years ago (Ma); Fig. 4].
MFSD12 suppresses eumelanin biogenesis in melanocytes from lysosomes

We silenced expression of the mouse ortholog of MFSD12 (Mfsd12) using short hairpin RNAs (shRNAs) in immortalized melan-Ink4a mouse melanocytes derived from C57BL/6J-Ink4a mice (32), which almost exclusively make eumelanin (Fig. 6). Reduction of Mfsd12 mRNA by ~80% with two distinct lentivirally encoded shRNAs (Fig. 6A) caused a 30 to 50% increase in melanin content compared to control cells (Fig. 6B), and a higher percentage of melanosomes per total cell area in most cells compared to cells transduced with nontarget shRNA (Fig. 6, C and D). A fraction of MFSD12-depleted cells harbored large clumps of melanin in autophagosome-like structures (fig. S10). These data suggest that MFSD12 suppresses eumelanin content in melanocytes and may offset autophagy.

We assessed the localization of human MFSD12 isoform c (RefSeq NM_174983.4) tagged at the C terminus with the hemagglutinin (HA) epitope (MFSD12-HA). By immunofluorescence microscopy, MFSD12-HA localized to punctate structures throughout the cell. Surprisingly, these puncta, like those labeled by the endogenous lysosomal membrane protein LAMP2, but not the melanosomal enzyme TYRP1, overlapped only weakly with pigmented melanosomes (Fig. 6, E to G; quantified in Fig. 6H). Instead, MFSD12-HA colocalized with LAMP2 (Fig. 6E; quantified in Fig. 6H), indicating that the MFSD12 protein localizes to late endosomes and/or lysosomes in melanocytes and not to eumelanosomes.

MFSD12 influences pigmentation in zebrafish xanthophore pigment cells

We targeted transmembrane domain 2 (TMD2) in the highly conserved zebrafish ortholog of mfsd12a with CRISPR-Cas9 (Fig. 7). We focused on mfsd12a because its paralog mfsd12b is predicted to be a pseudogene (33). Although pigmentation was not qualitatively altered in melanophores, the cells that make eumelanin, compound heterozygotes of mfsd12a alleles exhibited reduced staining of xanthophores, the cells responsible for pteridine-based yellow pigmentation in wild-type zebrafish (Fig. 7, A and B) (34, 35). This was not due to a failure of the xanthophores to develop in mfsd12a mutants, because green fluorescent protein (GFP)-labeled xanthophores were robust along the lateral line in both wild-type and mfsd12a mutant zebrafish (Fig. 7, C and D). Together, these results suggest that MFSD12 influences xanthophore pigment production in pterinosomes.

Functional characterization of MFSD12 in mice

CRISPR-Cas9 was used to generate an Mfsd12 null allele in a wild-type mouse background (Fig. 7E and fig. S11). Four founders were observed with a uniformly gray coat color, rather than the expected agouti coat color (fig. S11, A and B). These four gray founders harbored deletions at

Fig. 3. Genomic context of GWAS loci. Plot of \( -\log_{10}(P\text{ value}) \) versus genomic position for variants near the four regions with most strongly associated SNPs from GWAS, including annotations for genes, MITF ChIP-seq (chromatin immunoprecipitation sequencing) data for melanocytes (48), a CTFC ChIP-seq track for NHEK keratinocytes, and H3K27ac, DNase I hypersensitive sites (DHS), and chromHMM tracks for melanocytes and keratinocytes from the Roadmap Epigenomics data set (30). Genome-wide significant variants are highlighted in red. Circles, squares, and triangles denote noncoding, synonymous, and nonsynonymous variants, respectively. (A) SLC24A5 locus. (B) MFSD12 locus. (C) DDB1/TMEM138 locus. (D) OCA2/HERC2 locus.
the targeted site (fig. S1C). Microscopic observation revealed a lack of pheomelanin, resulting in white, rather than yellow, banding of hairs in Mfsd12 mutants (Fig. 7F).

The Mfsd12 knockout coat color appeared phenotypically similar to that of grizzled (gr) mice, an allele previously mapped to a syntenic ~2-Mb region overlapping Mfsd12 (36). Like our CRISPR-Cas9 Mfsd12 knockout, homozygous gr/gr mice are characterized by a gray coat resulting from dilution of yellow pheomelanin pigment from the subterminal agouti band of the hair shaft. Exome sequencing of an archived gr/gr DNA sample, subsequently confirmed by Sanger sequencing in an independent colony, identified a 9-bp in-frame deletion within exon 2 of Mfsd12 (fig. S12) as the sole mutation affecting a coding sequence in this mapped candidate region. The deleted amino acids for the gr/gr allele, Mfsd12 p.Leu163->ala165del, are in the cytoplasmic loop between the transmembrane domains TM4 and TM5 within a highly conserved MFS domain (fig. S13).

These results indicate that mutation of Mfsd12 is responsible for the gray coat color of gr/gr mutant mice, and that loss of MFSD12 reduces pheomelanin within the hair of agouti mice. Together, these results indicate that MFSD12 plays a conserved role in vertebrate pigmentation. Depletion of MFSD12 increases eumelanin content in a cell-autonomous manner in skin melanocytes, consistent with the lower levels of MFSD12 expression observed in melanocytes from individuals with African ancestry. Because MFSD12 localizes to lysosomes and not to eumelanosomes, this may reflect an indirect effect through modified lysosomal function. By contrast, loss of MFSD12 has the opposite effect through modified pheomelanosomes, which may have an indirect effect on function of pheomelanosomes, a potentially different intracellular origin of eumelanosomes (38). Although disruption of MFSD12 alone accounts for changes in pigmentation, the role of neighboring loci such as HMG20B on pigmentation remains to be explored.

**Skin pigmentation-associated loci that play a role in UV response are targets of selection**

Another genomic region associated with pigmentation encompasses a ~195-kb cluster of genes on chromosome 11 that play a role in UV response and melanoma risk, including the damage-specific DNA binding protein 1 (DDB1) gene (Figs. 1 and 3 and table S3). DDB1 (complexed with DDB2 and XPC) functions in DNA repair (39); levels of DDB1 are regulated by UV exposure and MC1R signaling, a regulatory pathway of pigmentation (40). DDB1 is a component of CUL4-RING E3 ubiquitin ligases that regulate several cellular and developmental processes (41); it is critical for follicle maintenance and female fertility in mammals (42) and for plastid size and fruit pigmentation in tomatoes (43). Knockouts of DDB1 orthologs are lethal in both mouse and fruitfly development (44), and DDB1 only exhibits rare (<1% frequency) nonsynonymous mutations in the TGP data set. Genetic variants near DDB1 were associated with human pigmentation in an African population with high levels of European admixture (7).

Because of extensive LD in this region, CAVIAR identified 33 SNPs predicted to be causal (Table 1). The most strongly associated SNPs are located in a region conserved across vertebrates flanked by TMEM138 and TMEM216 (45) ~36 to 44 kb upstream of DDB1 and are in high LD within this cluster ($r^2 > 0.7$ in East Africans) (Fig. 3, Table 1, and table S3). Among these, the most significantly associated SNP is rs7948623 (P test, $P = 2.2 \times 10^{-13}$), located ~7600 bp downstream of TMEM138, which shows enhancer activity in WM88 melanoma cells (91.9 to 140.8× higher than the minimal promoter; fig. S7 and table S5) and interacts with the promoters of DDB1 and neighboring genes in MCF-7 cells (Table 1 and Fig. 3) (46, 47).

A second group of tightly linked SNPs (LD $r^2 > 0.7$ in East Africans) with predicted high probability of containing causal variants spans a ~185-kb region encompassing DDB1 and TMEM138 (Table 1 and Fig. 3). Two SNPs that tag this LD block are rs1377457 ($P$ test, $P = 1.5 \times 10^{-8}$), located ~7600 bp downstream of TMEM138, and rs148172827 ($P$ test, $P = 1.8 \times 10^{-8}$), an insertion/deletion polymorphism at TKFC (trikinase and FMN cyclase) located in an enhancer active in WM88 melanoma cells (67.6 to 76.2× higher than the minimal promoter; fig. S7 and table S5), which overlaps an MIF5 binding site in melanocytes (30, 48); both SNPs interact with the promoters of DDB1 and neighboring genes in MCF-7 cells (Table 1 and Fig. 3) (46, 47). SNPs within introns of DDB1 (rs12298370, rs7934735, rs1230664, rs12275843, and rs7120594) also tag this LD block (Table 1 and Fig. 3).

RNA-seq data from 106 primary melanocyte cultures indicate that African ancestry is significantly
correlated with increased \textit{DDB1} gene expression (PCC, \( P = 2.6 \times 10^{-5} \); fig. S9). Association tests using a permutation approach indicated that, of the 35 protein-coding genes with a transcription start site within 1 Mb of rs7948623, expression of \textit{DDB1} is most strongly associated with a SNP in an intron of \textit{DDB1}, rs7120594, at marginal statistical significance after correction for ancestry and multiple testing (\( P_{\text{adj}} = 0.06 \); fig. S9). The allele associated with dark pigmentation at rs7120594 correlates with increased \textit{DDB1} expression. We did not have the power to detect an association between expression of \textit{DDB1} and SNPs in LD with rs7948623 due to low minor allele frequencies (~2%). The role of \textit{DDB1} and neighboring loci...
in human pigmentation remains to be further explored.

The derived rs7948623 (T) allele near TMEM138 (associated with dark pigmentation) is most common in East African Nilo-Saharan populations and is at moderate to high frequency in South Asian and Australo-Melanesian populations (88 to 96%), and is at moderate to high frequency in South Asian and Australo-Melanesian populations (12 to 66%) (Fig. 1 and fig. S4). The derived (T) allele (associated with light pigmentation) is nearly fixed in European, East Asian, and Native American populations.

In South Asians and Australo-Melanesians, the alleles associated with darker pigmentation reside on haplotypes closely related, or identical, to those observed in Africa (Fig. 5 and fig. S6), suggesting that they are identical by descent. The TMRCAs for the derived dark allele at rs7948623 and the derived light allele at rs1230664 are estimated to be older than 600 and 250 ka, respectively (Fig. 4).

Consistent with a selective sweep, we see an excess of rare alleles (and extreme negative Tajima’s D values) and high levels of homozygosity extending ~350 to 550 kb in Europeans and Asians, respectively (figs. S5 and S14). We observe extreme negative Tajima’s D values in East African Nilo-Saharans and San over a shorter distance (115 and 100 kb, respectively) (fig. S5).

A haplotype extending greater than 195 kb is common in Eurasians and rare in Africans (Fig. 5) and tags the alleles associated with light skin pigmentation. The TMRCa of a large number of haplotypes carrying the rs7948623 (A) allele in non-Africans, associated with light pigmentation, is 60 ka (95% CI, 58 to 62 ka), close to the inferred time of the migration of modern humans out of Africa (Fig. 4) (49). These results, combined with large $F_{ST}$ values between Africans and Europeans at SNPs tagging the extended haplotype near $DDB1$ (for example, $F_{ST} = 0.98$ between Nilo-Saharans and CEU at rs7948623, within the top 0.01% of values...
and San (Fig. 1 and fig. S4), consistent with a previous observation (56). Haplotype (Fig. 5) and coalescent analyses (Fig. 4 and fig. S6) show two divergent clades, one enriched for the rs1800404 (C) allele and the other for the rs1800404 (T) allele. Coalescent analysis indicates that the TMRCA of all lineages is 1.7 Ma (95% CI, 1.5 to 2.0 Ma), and the TMRCA of lineages containing the derived (T) allele is 629 ka (95% CI, 426 to 848 ka) (Fig. 4). The deep coalescence of lineages, and the positive Tajima’s D values in this region in both African and non-African populations (fig. S5), is consistent with balancing selection acting at this locus.

The SNP with highest probability of being causal in region 3 is rs4932620 (F test, \(P = 3.2 \times 10^{-5}\)) located within intron 11 of HERC2 (Fig. 3, Table 1, and table S3). This SNP is 917 bp from rs916977, a SNP associated with blue eye color in Europeans (57, 58), and is in strong LD (\(r^2 = 1.0\) in most East African populations), with SNPs extending into region 2 of HERC2 (Table 1). The derived rs4932620 (T) allele associated with dark skin pigmentation is most common in Ethiopian populations with high levels of Nilo-Saharan ancestry and is at moderate frequency in other African, Hadza, and Tanzania Nilo-Saharan populations (Fig. 1 and fig. S4). Haplotype analysis indicates that the rs4932620 (T) allele in South Asians and Australo-Melanesians is on the same or similar haplotype background as in Africans (Fig. 5 and fig. S6), suggesting that it is identical by descent. The TMRCA of haplotypes containing the rs4932620 (T) allele is 247 ka (95% CI, 158 to 345 ka) (Fig. 4).

We also observe an LD block of SNPs within region 2 of HERC2 that are associated with skin pigmentation, although they do not reach genome-wide significance (table S3). These are in a region with enhancer activity in Europeans (50). For example, SNP rs6497271 (F test, \(P = 1.8 \times 10^{-6}\)), which is located 437 bp from SNP rs12913832, has been associated with skin color in Europeans (50) and is in a consensus SOX2 motif (a transcription factor that modulates levels of MITF in melanocytes) (Fig. 3) (59). The ancestral rs6497271 (A) allele associated with dark pigmentation is on haplotypes in South Asians and Australo-Melaneseans similar or identical to those in Africans (Fig. 5 and fig. S6), suggesting that they are identical by descent. The derived (G) allele associated with light skin pigmentation is most common in Europeans and San and dates to 921 ka (95% CI, 703 ka to 1.2 Ma) (Figs. 1 and 4 and figs. S4 and S6). SNPs associated with pigmentation at all three regions show high allelic differentiation when comparing East African Nilo-Saharans and CEU (\(F_{ST} = 0.72\) to 0.85, top 0.5% on chromosome 15) (table S4).

Analyses of RNA-seq data from 106 primary melanocyte cultures indicate that African ancestry is significantly correlated with increased OCA2 gene expression (\(P_{adj} = 6.1 \times 10^{-6}\)) (fig. S9). A permutation approach identified significant associations between OCA2 expression and SNPs within an LD block between rs4932620 extending across regions 2 and 3 (\(P_{adj} = 2.2 \times 10^{-5}\)).
Table 1. Annotations of candidate causal SNPs from GWAS. Top candidate causal variants for the four regions identified based on analysis with CAVIAR (1). For each variant, the genomic position (Location), RSID, and Ancestral-Derived alleles are shown, with the allele associated with dark pigmentation in bold. Beta and standard error [Beta(SE)] and the \( P \) values from the GWAS (F test, linear mixed model) are given. For functional genomic data, nearest genes are given and variants overlapping DHS sites for melanocytes (ED59) (DHS melanocytes) and/or other cell types (DHS other) available from Roadmap Epigenomics are indicated with X (30, 92). Variants intersecting enhancer regions tested by luciferase assays were labeled with Y (significant enhancer activity) or N (no enhancer activity) (fig. S7). Chromatin interactions with nearby genes measured in MCF-7 or K562 cell lines as identified by ChIA-PET are listed with gene names (Chromatin interactions) (46, 47). SNPs that are in strong LD (\( r^2 > 0.7 \) in East Africans) are numerically labeled in the column titled LD block.

<table>
<thead>
<tr>
<th>Location</th>
<th>RSID</th>
<th>Ancestral-Derived</th>
<th>Beta(SE)</th>
<th>P</th>
<th>DHS melanocytes</th>
<th>DHS other</th>
<th>Luciferase activity</th>
<th>LD block</th>
<th>Nearest gene</th>
<th>Chromatin interactions</th>
</tr>
</thead>
<tbody>
<tr>
<td>15:48452896</td>
<td>rs2413887</td>
<td>T&gt;C</td>
<td>7.70(0.44)</td>
<td>4.9 \times 10^{-62}</td>
<td>X</td>
<td>X</td>
<td>N</td>
<td>1</td>
<td>CTXN2</td>
<td>MYEF2</td>
</tr>
<tr>
<td>15:48426484</td>
<td>rs1426654</td>
<td>G&gt;A</td>
<td>7.69(0.44)</td>
<td>5.5 \times 10^{-62}</td>
<td>X</td>
<td>X</td>
<td>N</td>
<td>1</td>
<td>SLC2A5</td>
<td></td>
</tr>
<tr>
<td>15:48392165</td>
<td>rs1834640</td>
<td>G&gt;A</td>
<td>5.6(0.44)</td>
<td>3.2 \times 10^{-61}</td>
<td>X</td>
<td>1</td>
<td>SLC2A5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15:48400199</td>
<td>rs2675345</td>
<td>G&gt;A</td>
<td>7.62(0.44)</td>
<td>6.7 \times 10^{-61}</td>
<td>X</td>
<td>1</td>
<td>SLC2A5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15:48460188</td>
<td>rs8028919</td>
<td>G&gt;A</td>
<td>-4.95(0.41)</td>
<td>5.0 \times 10^{-32}</td>
<td>X</td>
<td>1</td>
<td>MYEF2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19:3545022</td>
<td>rs10424065</td>
<td>C&gt;T</td>
<td>4.48(0.48)</td>
<td>5.1 \times 10^{-10}</td>
<td>X</td>
<td>X</td>
<td>Y</td>
<td>2</td>
<td>MFSID12</td>
<td>CACTIN, MFSID12</td>
</tr>
<tr>
<td>19:3544892</td>
<td>rs56203814</td>
<td>C&gt;T</td>
<td>-3.83(0.50)</td>
<td>3.6 \times 10^{-10}</td>
<td>X</td>
<td>Y</td>
<td>2</td>
<td>MFSD12</td>
<td>CACTIN, MFSID12</td>
<td></td>
</tr>
<tr>
<td>19:3566631</td>
<td>rs111317445</td>
<td>C&gt;T</td>
<td>3.51(0.42)</td>
<td>1.7 \times 10^{-6}</td>
<td>N</td>
<td>3</td>
<td>HMGB2B</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19:3547955</td>
<td>rs10144812</td>
<td>C&gt;T</td>
<td>4.38(0.53)</td>
<td>3.8 \times 10^{-10}</td>
<td>X</td>
<td>Y</td>
<td>2</td>
<td>MFSID2</td>
<td>CACTIN, FZRI, MFSID2</td>
<td></td>
</tr>
<tr>
<td>19:3565599</td>
<td>rs11232856</td>
<td>T&gt;C</td>
<td>3.52(0.43)</td>
<td>3.8 \times 10^{-10}</td>
<td>X</td>
<td>X</td>
<td>Y</td>
<td>3</td>
<td>HMGB2B</td>
<td>MFSID2</td>
</tr>
<tr>
<td>19:3565253</td>
<td>rs6510760</td>
<td>G&gt;A</td>
<td>3.54(0.45)</td>
<td>6.5 \times 10^{-10}</td>
<td>X</td>
<td>Y</td>
<td>3</td>
<td>MFSD12</td>
<td>MFSID2</td>
<td></td>
</tr>
<tr>
<td>19:3545150</td>
<td>rs73527942</td>
<td>T&gt;G</td>
<td>-3.58(0.47)</td>
<td>4.8 \times 10^{-14}</td>
<td>X</td>
<td>X</td>
<td>Y</td>
<td>2</td>
<td>MFSID2</td>
<td>CACTIN, MFSID12</td>
</tr>
<tr>
<td>19:3547685</td>
<td>rs142317543</td>
<td>C&gt;T</td>
<td>6.99(0.92)</td>
<td>5.0 \times 10^{-14}</td>
<td>X</td>
<td>Y</td>
<td>2</td>
<td>MFSID2</td>
<td>CACTIN, FZRI, MFSID2</td>
<td></td>
</tr>
<tr>
<td>19:3566513</td>
<td>rs7254463</td>
<td>C&gt;T</td>
<td>2.90(0.50)</td>
<td>9.0 \times 10^{-9}</td>
<td>X</td>
<td>N</td>
<td>3</td>
<td>HMGB2B</td>
<td>MFSID2</td>
<td></td>
</tr>
<tr>
<td>19:3565357</td>
<td>rs7246261</td>
<td>C&gt;T</td>
<td>2.71(0.47)</td>
<td>1.1 \times 10^{-8}</td>
<td>X</td>
<td>X</td>
<td>3</td>
<td>HMGB2B</td>
<td>MFSID2</td>
<td></td>
</tr>
<tr>
<td>19:3565909</td>
<td>rs6510761</td>
<td>T&gt;C</td>
<td>2.79(0.50)</td>
<td>2.2 \times 10^{-8}</td>
<td>X</td>
<td>X</td>
<td>Y</td>
<td>3</td>
<td>MFSD12</td>
<td>MFSID2</td>
</tr>
<tr>
<td>11:6117717</td>
<td>rs7948623</td>
<td>A&gt;T</td>
<td>-2.94(0.44)</td>
<td>2.2 \times 10^{-11}</td>
<td>X</td>
<td>X</td>
<td>Y</td>
<td>4</td>
<td>TMEM138, TKFC, DDB1, TMEM138</td>
<td></td>
</tr>
<tr>
<td>11:6148456</td>
<td>rs397709980</td>
<td>G&gt;A</td>
<td>-2.90(0.43)</td>
<td>2.4 \times 10^{-11}</td>
<td>N</td>
<td>X</td>
<td>4</td>
<td>TMEM16</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11:6152830</td>
<td>rs4453253</td>
<td>C&gt;T</td>
<td>-2.85(0.43)</td>
<td>5.4 \times 10^{-11}</td>
<td>X</td>
<td>Y</td>
<td>4</td>
<td>TMEM16</td>
<td>CYB561A3, TKFC, DDB1, TMEM138, TMEM16</td>
<td></td>
</tr>
<tr>
<td>11:61153401</td>
<td>rs4939520</td>
<td>C&gt;T</td>
<td>-2.79(0.43)</td>
<td>1.4 \times 10^{-10}</td>
<td>X</td>
<td>N</td>
<td>4</td>
<td>TMEM16</td>
<td>CYB561A3, TKFC, DDB1, TMEM138, TMEM16</td>
<td></td>
</tr>
</tbody>
</table>

continued on next page
Alles in this LD block associated with dark pigmentation correlate with increased OCA2 expression. We did not observe associations between the candidate causal variants in region 1 and OCA2 expression despite a high minor allele frequency (34%). However, we observe a significant association between a haplotype tagged by rs1800404 and alternative splicing resulting in inclusion/exclusion of exon 10 (linear regression t test, P = 9.1 × 10^-6). Exon 10 encodes the amino acids encompassing the third transmembrane domain of OCA2 and is the location of several albinism-associated OCA2 mutations (60, 61), raising the possibility that the shorter transcript encodes a nonfunctional channel. Comparing splice junction usage across individuals, we estimate that each additional copy of the light rs1800404 (T) allele reduces inclusion of exon 10 by ~20% (95% CI, 17.9 to 21.5%; fig. S9). Therefore, homozygotes for the light rs1800404 (T) allele are expected to produce ~60% functional OCA2 protein (compared to individuals with albinism who produce no functional OCA2 protein).

**Skin pigmentation is a complex trait**

To estimate the proportion of pigmentation variance explained by the top eight candidate SNPs at SLC24A5, MFSD12, DDB1/TMEM138, and OCA2/HERC2, we used a linear mixed model with two genetic random effect terms: one based on the genome-wide kinship matrix and the other based on the candidate causal variants in region 1 and OCA2/HERC2 (3.9%; SE, 2.9%). Thus, ~29% of the additive heritability of skin pigmentation in Africans is due to variation at these four regions. This observation indicates that the genetic architecture of skin pigmentation is simpler (that is, fewer genes of stronger effect) than other complex traits, such as height (62). In addition, most candidate causal variants are in noncoding regions, indicating the importance of regulatory variants influencing skin pigmentation phenotypes.

**Evolution of skin pigmentation in modern humans**

Skin pigmentation is highly variable within Africa. Populations such as the San from southern Africa are the most lightly pigmented among Africans, whereas the East African Nilo-Saharan populations are the most darkly pigmented in the world (Fig. 1). Most alleles associated with light and dark pigmentation in our data set are estimated to have originated before the origin of modern humans ~300 ka (27). In contrast to the lack of variation at MC1R, which is under purifying selection in Africa (63), our results indicate that both light and dark alleles at MFSD12, DDB1, OCA2, and HERC2 have been segregating in the hominin lineage for hundreds of thousands of years (Fig. 4). Furthermore, the ancestral allele is associated with light pigmentation in about half of the predicted causal SNPs; Neandertal and Denisovan genome sequences, which diverged from modern human sequences 804 ka (64), contain the ancestral allele at all loci. These observations are consistent with the hypothesis that darker pigmentation is a derived trait that originated in the genus Homo within the past ~2 million years (Myr) after human ancestors lost most of their protective body hair, although these ancestral hominins may have been moderately, rather than darkly, pigmented (65, 66). Moreover, it appears that both light and dark pigmentation have continued to evolve over hominid history.

Individuals from South Asia and Australo-Melanesian share variants associated with dark pigmentation at MFSD12, DDB1/TMEM138, OCA2, and HERC2 that are identical by descent from Africans. This raises the possibility that other phenotypes shared between Africans and some South Asian and Australo-Melanesian populations may also be due to genetic variants identical by descent from African populations rather than convergent evolution (67). This observation is consistent with a proposed southern migration route out of Africa ~80 ka (68). Alternatively, it is possible that light and dark pigmentation alleles segregated in a single African source population (69, 69) and that alleles associated with dark pigmentation were maintained outside of Africa only in the South Asian and Australo-Melanesian populations due to selection.

By studying ethnically, genetically, and phenotypically diverse Africans, we identify novel pigmentation loci that are not highly polymorphic in European populations. The loci identified in this study appear to affect multiple phenotypes. For
example, DDB1 influences pigmentation (43), cellular response to the mutagenic effect of UVR (40), and female fertility (42). Thus, some of the pigmentation-associated variants identified here may be maintained because of pleiotropic effects on other aspects of human physiology.

It is important to note that genetic variants that do not reach genome-wide significance in our study might also affect the pigmentation phenotype. The 1000 most strongly associated SNPs exhibit enrichment for genes involved in pigmentation and melanocyte physiology in the mouse phenotype database and in ion transport and pyrimidine metabolism in humans (table S8). Future research in larger numbers of ethnically diverse Africans may reveal additional loci associated with skin pigmentation and will further shed light on the evolutionary history, and adaptive significance, of skin pigmentation in humans.

Materials and methods

Individuals in the study were sampled from Ethiopia, Tanzania and Botswana. Written informed consent was obtained from all participants, and research/ethics approval and permits were obtained from all relevant institutions (70). To measure skin pigmentation we used a DSM II ColorMeter to quantify reflectance from the inner under arm. Red reflectance values were corrected to a standard melanin index score (70, 77). DNA was extracted from whole blood using a salting out procedure (PureGen). A total of 1570 samples were genotyped on the Illumina Omni5M SNP array (5M dataset) that includes ~4.5 million SNPs. Genotypes were clustered and called in Genome Studio software. Variant positions are reported in hg39/37 coordinates. The overall completion rate was 98.8%. Each individual’s sex was verified based on X chromosome inbreeding coefficients. We used Beagle 4.0 (72) to phase the Illumina 5M SNP array data merged with SNPs from the TGP dataset that were filtered to exclude related individuals.

High coverage (>30x) Illumina Sequencing was performed on a subset of the genotyped individuals (N = 135). Variants were called following the approach described in (13). Adapter sequences were trimmed with trimadap. Reads were aligned using bwa mem to the human reference sequence build 37 (hg19). After alignment we marked duplicate reads prior to calling variants with GATK HaplotypeCaller (73). To select high quality variants we employed a two-set filtering strategy. First, we used the GATK variant quality score recalibration to score variant sites. We used TGP, OMM, and our curated genotypes from the Illumina Omni 5M SNP array as training data. After recalibration we discarded sites with the lowest scores. In addition, we discarded sites in low-complexity regions listed in (74) and duplicate regions identified with Delly (75).

We performed local imputation around each of the regions showing significant associations with skin pigmentation from GWAS using the Illumina Omni 5M SNP dataset. We extracted array genotypes within 1 Mb (500 kb upstream and 500 kb downstream) of top GWAS variants from each region and phased them using SHAPEIT2 (76, 77). The reference panel came from two datasets: filtered variants from the 135 African genomes and TGP (10). After phasing, imputation was performed using Minimac3 (78). Imputation performed very well at most loci (R² > 0.91 with MAF ≥ 0.05) (table S3).

To identify SNPs associated with pigmentation, GWAS was performed first on the Illumina Omni 5M SNP dataset, and independently with imputed variants at candidate regions, using linear mixed models implemented in EMMAX software (30). Sex and sex were included as covariates, and we corrected for genetic relatedness with an IBS kinship matrix. We used CAVAR to identify variants in the imputed dataset most likely to be causal (11). Ontology enrichment for genes near the top 1000 most strongly associated variants from the 5M dataset was obtained using the annotation tool GREAT (79).

We estimated the contribution to the variation in melanin index from the top candidate causal variants with a restricted maximum likelihood (REML) analysis implemented in the Genome-wide Complex Trait Analysis (GCTA) software (80). The variance parameters for two genetic relationship matrices (GRMs) are estimated: one GRM is constructed (81) from genome-wide background variants with MAF > 0.01, and one GRM is constructed from the set of 8 top pigmentation-associated variants. The contribution of each locus to the melanin index variation is estimated similarly, using all genome-wide significant (P < 5 × 10⁻⁸) variants within each locus to construct the pigmentation-associated GRM (table S3). REML iterations are based on maximizing the Average Information matrix.

To test for neutrality in the regions flanking our top GWAS variants we calculated Tajima’s D, FST, and extended haplotype homozygosity using iHS (82–84). Tajima’s D was measured along chromosomes 11 and 15 using 50-kb sliding windows. Due to a high recombination rate observed near the MFSD2 locus, we used 10-kb windows in that region (chromosome 19). Vcftools was used to calculate both Tajima’s D and FST (85). To calculate extended haplotype homozygosity (iHS) we used Selscan (86). Unstandardized iHS scores were normalized within 100 kb bins according to the frequency of the derived allele. We then identified the signals of positive selection by calculating the proportion of SNPs with iHS > 2 in non-overlapping windows (83). To identify outlier windows we calculated 5th and 95th percentiles. Population differentiation was assessed with Weir and Cockerham’s fixation index FST (82) between each pair of populations. Outliers were identified using empirical P values.

Median joining haplotype networks (87) for the Illumina Omni 5M SNP dataset were constructed and visualized at genomic regions of interest using NETWORK (88). In addition, we constructed genealogies of regions flanking candidate causal SNPs using a hierarchical clustering approach with sequence data from the Simons Genome Diversity Project (13). Briefly, we considered a single copy of each chromosome from each of the 279 individuals from (33). We inferred recombination breakpoints within a symmetric window surrounding each locus using the program kwearg (89) and identified the longest shared haplotype between each pair of sequences in which no recombination events occurred. We then computed the expected coalescence time between each pair of sequences, conditional on the observed number of mutations in the non-recombining region. Genealogies were constructed by applying the WPGMA hierarchical clustering algorithm to the estimated pairwise coalescence times. Our estimator accounts for recombination events and the population size history. However, simulation studies indicate that accounting for time-varying population size has relatively little effect on our estimates when the size changes according to previously inferred histories for human populations (70, 90). Because the true population sizes and relationships among the populations we considered are complex and imprecisely known, we assumed a constant population size of N = 10⁴ in our analyses. The robustness analysis presented in (70) describes how our time estimates would change under different demographic histories and selective pressures.

To identify candidate causal GWAS variants altering gene expression we visualized and intersected variants with chromHMM tracks (91), DNase I hypersensitivity peaks, H3K27ac signal tracks for keratinocytes and melanocytes (30), CTCF signal tracks from keratinocytes (92), and ChiP-seq signal tracks from MITF (48). Variants were intersected with chromatin annotations using bedtools (93). Functional consequences of variants were also assessed using deepSEA (94) and deltaSVM (95). The effect of genetic variants on transcription factor binding was predicted using the MEF suite (96) for all transcription factors in the JASPAR 2016 CORE Vertebbrate motif set (97).

To test for associations between gene expression, genetic variation, and ancestry we used eQTL and allele-specific expression (ASE) analyses on transcriptomes and genotype data from primary cultures of human melanocytes, isolated from foreskin of 106 individuals of assorted ancestries. All 106 individuals were genotyped on Illumina OmniExpress arrays and genotypes were subsequently imputed using the Michigan Imputation Server (79) based on the TGP reference panel and using SHAPEIT for phasing (77). RNA sequencing was performed to a mean depth of 87 million reads per sample. STAR (98) was used for aligning reads, and RSEM (99) was used to quantify the gene expression. Quantile normalization was applied in all samples to get the final RSEM value. To account for hidden factors driving expression variability, a set of covariates were further identified using the PEER method (100) and applied to calculate the normalized expression matrices. Principal components analysis was performed using genotypic data to capture population structure and ancestry using the struct.pca module of GLU (https://github.com/bioinformed/glu-genetics). Using the normalized
expression values and principal components, Pearson correlation between gene expression levels and ancestry was calculated, and associations between GWAS variant genotypes and gene expression levels were evaluated using ordinary least squares regression.

To identify associations between our GWAS candidate causal variants and expression of nearby genes (using the 106 melanocyte transcriptomes), we first found all protein-coding genes with transcription start site (TSS) within 1 Mb of the top GWAS variant for each locus and RSEM values greater than 0.5 in the primary melanocyte cultures. Pearson correlation was used to measure the association between ancestry and gene expression.

For each locus, we tested whether any genes with a transcription start site within 1 Mb of the top SNP had an eQTL amongst the set of pigmentation QTls using an additive linear model with the first two principal components of ancestry as covariates. To identify significant variant-gene associations we used a permutation approach for each locus independently (70). This was repeated for each gene, and focal variants across all genes were adjusted for multiple testing using the Bonferroni correction (101).

We also carried out allele-specific expression (ASE) analyses for each significant eQTL SNP. Sites with at least 30 mapped reads, ≤5% mapping bias, and ENCODE 125 bp mappability score with at least 30 mapped reads, <5% mapping bias, (ASE) analyses for each significant eQTL SNP. Sites heterozygous coding variant amongst the melanocyte associations we used a permutation approach for each locus independently (70). We repeated for each gene, and focal variants across all genes were adjusted for multiple testing using the Bonferroni correction (101).

For OC42 we tested for an association between inclusion rates of exon 10, which contains our top candidate causal variant in the region, rs1800404, and individual genotypes at rs1800404. For each melanocyte transcriptome, reads spanning the exon 9 to exon 10 and exon 9 to exon 11 junctions were extracted from the splice-junction files output by STAR. For each individual, a percent splice in (PSI) value was calculated. To estimate the effect of variation at rs1800404 on exon 10 inclusion, ordinary least squares regression was carried out between PSI and dosage of the alternative allele for rs1800404 across individuals. A two-sided t test was used to calculate a P value.

To test the functional impact of a subset of GWAS variants on gene expression, predicted regulatory sequences containing variants were cloned into a pGL4.23 firefly luciferase reporter vector. Vectors were transfected into a WM88 melanocytic melanoma cell line and co-transfected with renilla luciferase control vector (pRL-CMV) in a dual luciferase assay. Relative luciferase activity (firefly/renilla luminescence ratio) is presented as fold change compared to cells transfected with the empty pGL4.23 vector. Data were analyzed with a modified Kruskal-Wallis Rank Sum test and pairwise comparisons between groups were performed using the Conover method. P values were corrected for multiple comparisons with the Benjamini-Hochberg method using the R package PMCMR, and P values less than 0.05 were considered significant.

We characterized the function of MFSD12 in vitro in immortalized melanocytes and in vivo in both zebrafish and mice. Immunolabeled melanin-ink4a melanocytes from C57BL/6J/lnk4a-Atf3−/− mice were cultured as described (32). To deplete MFSD12, cells were infected with recombinant lentiviruses—generated by transient transfection in HEK293T cells—to express MFsd12-targeted shRNAs or non-target controls. Cells resistant to puromycin (also encoded by the lentiviruses) were analyzed 5 to 7 days after infection. Knockdown efficiency of MFsd12 mRNA in cells expressing MFsd12-specific shRNAs relative to non target shRNA was quantified by reverse transcription/quantitative PCR (detected with SYBR Green; Applied Biosystems) relative to Tubb4b (encoding β-tubulin) as a reference gene. Melanin content in cell lysates was determined by a spectrophotometric assay as described (102), and melanin coverage in intact cells was determined by bright field microscopy and analysis using the “Analyze Particles” plug-in in ImageJ (National Institutes of Health). To analyze MFSD12-HA localization, melanin-ink4a melanocytes were transiently transfected with MFSD12-HA expression plasmids and analyzed 48 hours later by bright field and immunofluorescence microscopy as described (103) using the TA99 monoclonal antibody to TYRP1 (American Type Culture Collection) to detect melanosomes, rabbit anti-LAMPA2 (Abcam) to detect lysosomes, and rat anti-HA (Roche) to detect the transgene. Percent signal overlap in the cell periphery was determined from background-subtracted, thresholded, binary images using the “Analyze Particles” plug-in in ImageJ. Statistical significance was determined using unpaired, two-tailed student’s t tests: P < 0.05; *P < 0.10; **P < 0.001; ***P < 0.0001. **** Details are provided in (70).

Zebrafish mutagenesis using CRISPR-Cas9 was performed to target mfsd12a (70). Compound heterozygous mutant fish for analysis were generated from F1 incrosses of mutant founder fish. For methylene blue staining, embryos were collected following fertilization and placed in zebrafish system water containing 0.0002% methylene blue and analyzed at 6 dpf. For GFP expression, embryos were injected with 25 pg Tg(α:5:PALM-GFP) and 80 pg tol2 mRNA and GFP expression was evaluated in mosaic injected fish at 5 dpf. Larvae were anesthetized in sub-lethal 1x tricaine solution and placed in 100 μl of a low melt agarose solution (0.8%).

In mice two targets for CRISPR-Cas9 cleavage were selected within exon 2 of Mfsd12 to generate a 134 bp deletion resulting in a null allele of Mfsd12. A mixture of Cas9 mRNA (Trillink BioTechnologies) and each of the two synthesized gRNAs was used for pronuclear injection into C57BL/6J × FVB/N F1 hybrid zygotes. Mutation carrying mice were viable and presented with gray coat color distinct from littermates. Hairs were plucked from postnatal day 18 mice and individual awl hairs were mounted in permount and imaged with a stereomicroscope (Zeiss SteReO Discovery.V12) at the base of the bub apical yellow band where the switch from eumelanin to pheomelanin is visible.

To characterize Mfsd12 in grizzled mice, Illuuna generated whole genome sequences of grizzled, JIGR/DN (gr/gr) reads were mapped using bwa mem to GRCh38/mm10 (available at SRA Accession SRR5572237). Sequence variants between JIGR/DN (gr/gr) and C57BL6/J reference genome within the gr/gr candidate region were identified using SnpEff (104). Validation of a 12-bp deletion within Mfsd12 was performed using samples from an independently maintained gr/gr colony provided by the laboratory of Dr. Margit Burmeister.

REFERENCES AND NOTES


84. F. Tajima, Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123, 585–595 (1989); pmid: 2513255


ACKNOWLEDGMENTS

We thank J. Akey and R. McCoy for Melanesian genotype data; A. Clark, C. Brown, and Y. S. Park for critical review of the manuscript; members of the Tishkoff laboratory for helpful discussion; A. Weaverathna at Wistar Institute, Philadelphia, for providing the WM88 melanocytic cell line; D. Parichy for the axic1

palmGFP plasmid; G. Xu and R. Yang at University of Pennsylvania for technical assistance; M. Burmeister at University of Michigan for grizzled mouse samples; L. Garrett at the Embryonic Stem Cell and Transgenic Mouse Core [National Genome Research Institute (NGRI)]; R. Sood in the Zebrasfish Core (NGHR); and the African participants. We acknowledge the contribution of the staff members of the Cancer Genomics Research Laboratory [National Cancer Institute (NCI)], the NIH Intramural Sequencing Center, the NCI Center for Cancer Research Sequencing Facility, the Yale University Skin SPORE Specimen Resource Core, and the Botswana–University of Pennsylvania Partnership. This work used computational resources of the NIH High-Performance Computing (HPC) Biowulf cluster. This research was funded by the following grants: NIH grants R01DK104339-0 and R01GM13657-01 and NSF grant BCS-1317217 to S.A.T., NIH grant RO1 AR48155 from the National Institute of Arthritis and Musculoskeletal and Skin Diseases (NIAMS) to M.S.M., NIH grant RO1 AR06318 from NIMDS to E.O., NIH grants 5R24OD007870-04 and IU54DK10805-01 to L.Z. and Y.S.S., NIH grant RO1-GM044402 to Y.S.S., and NIH grant K22 GM081259 from NIGMS to S.B. M.H.B. was partly supported by a "Science Without Borders" fellowship from CNPq, Brazil. Y.S.S. is a Chan Zuckerberg Biobank investigator. This work was supported in part by the Center of Excellence in Environmental Toxicology (NIH P30-ES013508, T32-ES029851 to M.E.B.H.) (National Institute of Environmental Health Sciences), the Intramural Program of the NHGRI, and the Division of Cancer Epidemiology and Genetics, NCI, NIH, federal funds from NCI under contract HHSN26120080001E. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does the mention of trade names, commercial products, or organizations imply endorsement by the U.S. government. L.Z. is a founder and stockholder of Fate Therapeutics, Marauder Therapeutics, and Scholar Rock. Data are available at dbGAP accession number phs001396.v1.p1 and SRA BioProject PRJNA392485. In memory of G. Lema and E. Kimaro, who made important contributions to this project.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/358/6365/eaan8433/suppl/DC1

Materials and Methods

Figs. S1 to S21

Tables S1 to S28

NIH Comparative Sequencing Program Collaborator List

References (106–133)

27 May 2017; accepted 3 October 2017

Published online 12 October 2017

10.1126/science.aar8433

Crawford et al., Science 358, eaan8433 (2017) 17 November 2017