BMC
Evolutionary Biology

RESEARCH ARTICLE   **Open Access**

Exploring signatures of positive selection in pigmentation candidate genes in populations of East Asian ancestry

Jessica L Hider[1], Rachel M Gittelman[2], Tapan Shah[3], Melissa Edwards[1], Arnold Rosenbloom[3], Joshua M Akey[2] and Esteban J Parra[1*]

Abstract

Background: Currently, there is very limited knowledge about the genes involved in normal pigmentation variation in East Asian populations. We carried out a genome-wide scan of signatures of positive selection using the 1000 Genomes Phase I dataset, in order to identify pigmentation genes showing putative signatures of selective sweeps in East Asia. We applied a broad range of methods to detect signatures of selection including: 1) Tests designed to identify deviations of the Site Frequency Spectrum (SFS) from neutral expectations (Tajima's D, Fay and Wu's H and Fu and Li's D* and F*), 2) Tests focused on the identification of high-frequency haplotypes with extended linkage disequilibrium (iHS and Rsb) and 3) Tests based on genetic differentiation between populations (LSBL). Based on the results obtained from a genome wide analysis of 25 kb windows, we constructed an empirical distribution for each statistic across all windows, and identified pigmentation genes that are outliers in the distribution.

Results: Our tests identified twenty genes that are relevant for pigmentation biology. Of these, eight genes (*ATRN, EDAR, KLHL7, MITF, OCA2, TH, TMEM33* and *TRPM1,*) were extreme outliers (top 0.1% of the empirical distribution) for at least one statistic, and twelve genes (*ADAM17, BNC2, CTSD, DCT, EGFR, LYST, MC1R, MLPH, OPRM1, PDIA6, PMEL (SILV)* and *TYRP1*) were in the top 1% of the empirical distribution for at least one statistic. Additionally, eight of these genes (*BNC2, EGFR, LYST, MC1R, OCA2, OPRM1, PMEL (SILV)* and *TYRP1*) have been associated with pigmentary traits in association studies.

Conclusions: We identified a number of putative pigmentation genes showing extremely unusual patterns of genetic variation in East Asia. Most of these genes are outliers for different tests and/or different populations, and have already been described in previous scans for positive selection, providing strong support to the hypothesis that recent selective sweeps left a signature in these regions. However, it will be necessary to carry out association and functional studies to demonstrate the implication of these genes in normal pigmentation variation.

Background

The major out of Africa migrations of anatomically modern humans took place within the last 60,000-50,000 years [1]. As a result of these migrations, humans encountered novel environments, with varying climates, pathogens and foods and adapted to these new conditions via natural selection. One of the climatic factors showing clear geographic patterns is ultraviolet (UV) radiation, which is more intense and shows less seasonality in equatorial and tropical areas than in high latitude regions [2]. Skin pigmentation, which is primarily determined by the amount, type and distribution of cutaneous melanin, also shows a clear latitudinal gradient, and this has been explained as the result of natural selection [3-8]. In agreement with this hypothesis, genome-wide scans have shown that many genes involved in the pigmentation pathway show signatures of positive selection [9-21]. Interestingly, most of the putative selection signatures have been identified in European and East Asian populations, indicating that the majority of the selective sweeps took place after the out-of-Africa migration of modern humans. Although some of the pigmentation

* Correspondence: esteban.parra@utoronto.ca
[1]Department of Anthropology, University of Toronto at Mississauga, Mississauga, Ontario, Canada
Full list of author information is available at the end of the article

genes show signatures of selection that are shared between European and East Asian populations (e.g. *KITLG*) [12,18] many genes show positive selection signals in only one population (e.g. *SLC24A5* and *SLC45A2* in Europe, *DCT* in East Asia) or independent signals in European and East Asian groups (*OCA2*). These findings support an evolutionary model in which the most important changes in pigmentary traits occurred after the migration out-of-Africa and the separation of the lineages that gave rise to contemporary European and East Asian populations [5,12,22-25].

Most of the surveys of signatures of selection published to date have been based on data from the HapMap project or the Human Genome Diversity Project (HGDP), which primarily captured common genetic variants. The recent availability of the 1000 Genomes project Phase I data, which includes full genome sequences (based on a combination of low-coverage whole genome sequencing and targeted deep exome sequencing) for more than 1,000 individuals from 14 populations, has opened new opportunities to study genetic variation in our species [26]. In particular, the improved representation of rare variants and the decreased bias in variant detection would be expected to increase the power of some of the tests used to identify selective sweeps. In this study, we applied a range of genome-wide methods to detect signatures of selection in the 1000 Genomes Phase I dataset. The methods employed include: 1) Tests designed to identify deviations of the Site Frequency Spectrum (SFS) from neutral expectations (Tajima's D, Fay and Wu's H and Fu and Li's D* and Fu's F*), 2) Tests focused on the identification of high-frequency haplotypes with extended Linkage Disequilibrium (LD) (iHS and Rsb) and 3) Tests based on genetic differentiation between populations (LSBL). The main goal of the study was to identify pigmentation genes that have been the target of positive selection in East Asia. To date, the overwhelming majority of the genetic association studies focused on pigmentary traits have been carried out in European populations and as a result the last decade has brought a much better understanding of the genetic basis of normal pigmentation variation in this group [6,12,27,28]. In contrast to the long list of genes that have been associated with pigmentary traits in European populations, there is very limited knowledge concerning the genes involved in pigmentary traits in East Asia. Notable exceptions are the genes *OCA2* and *MC1R*, which harbor non-synonymous mutations, rs1800414 (His615Arg) in OCA2 and rs885479 (Arg163Gln) in MC1R, that have been associated with skin pigmentation in East Asian populations [23,29,30]. These polymorphisms are present in high frequency in East Asian populations, and are absent or present at low frequencies in European and African populations,

suggesting again that there has been convergent evolution towards depigmentation in Europe and East Asia. Additional research efforts in East Asian populations, or admixed populations showing a substantial East Asian contribution [31], will be necessary in order to elucidate the genetic architecture of pigmentation in East Asian populations, and more generally, the evolutionary events responsible for the pigmentary changes that took place after the out-of-Africa migration of modern humans. By identifying pigmentation genes showing putative signatures of selective sweeps in East Asia, we will be able to prioritize a list of genes for subsequent association studies in East Asian samples characterized with quantitative methods (e.g. skin reflectometry).

Methods
Samples
All the statistical analyses were completed using the 1,000 Genomes Phase 1 data, which includes approximately 38 million Single Nucleotide Polymorphisms (SNPs) [26]. Indels were excluded from all the analyses. The 1000 genomes data set includes samples Japanese from Tokio (JPT), Han Chinese from Beijing (CHB) and Southern Han Chinese (CHS).

Statistics used to identify putative signatures of positive selection
1-Statistics based on the Site Frequency Spectrum (SFS)
These statistics compare different estimators of the population mutation rate $\ominus = 4N\mu$, which have the same expectation under neutrality.

Tajima's D [32]. This test compares estimates of $\ominus$ derived from the average number of pairwise differences (π) and the number of segregating sites (S).
Fu and Li's D* [33]. This test compares estimates of $\ominus$ derived from the number of segregating sites (S) and the number of singleton mutations (η_s, alleles appearing only once in the sample).
Fu's F* [34]. This test compares estimates of $\ominus$ derived from the average number of pairwise differences (π) and the number of singleton mutations (η_s).

For these three tests, negative values indicate an excess of rare polymorphisms, and positive values an excess of intermediate-frequency alleles with respect to neutral expectations.
Fay and Wu's H [35]. In contrast to the previous three tests, H requires information on allele state (ancestral vs. derived). This test compares estimates of $\ominus$ derived from the average number of pairwise differences (π) with another estimate derived from the frequency of derived alleles at segregating sites ($\ominus_H$). H is negative when

derived alleles are found at high frequency, with respect to neutral expectations.

These four statistics were estimated for non-overlapping 25 kilobase windows, using a Python script. The statistics were calculated independently in three East Asian samples from the 1000 Genomes Phase 1 panel: Han Chinese from Beijing (CHB) (97), Southern Han Chinese (CHS) (100) and Japanese from Tokyo (JPT) (89). Variants that did not pass the 1000 genomes project filtering metrics were masked, as well as any variants in which the ancestral allele could not be determined. Windows with less than 10 markers were excluded from further analyses.

2-Tests based on genetic differentiation
We estimated genetic differentiation using the Locus Specific Branch Length (LSBL), as described in Shriver et al., 2004 [36]. In this case, we focused on the identification of regions with high East Asian LSBL values, indicating strong differentiation between East Asia and Europe/Africa. For these analyses, we used the combined 1000 Genomes Phase I East Asian (Han Chinese from Beijing, Southern Han Chinese and Japanese from Tokyo, N = 286), European (Tuscans from Italy, British, Finnish, Iberians, and Utah residents with Western European ancestry, N = 379) and African (Yoruba from Nigeria and Luhya from Kenya, N = 185) samples. East Asian LSBL values were estimated from the East Asian-African, East Asian-European and African-European pairwise F_{ST} distances for each locus using the formula LSBL(Eas) = (Eas-Eur F_{ST} + Eas-Afr F_{ST} −Afr-Eur F_{ST})/2. F_{ST} values were calculated with the program VCFTOOLS using Weir and Cockerham (1984) unbiased estimator [37]. Negative F_{ST} values were converted to zero. Using a combination of shell and python scripts, we created non-overlapping windows of 25 kilobases, and reported for each window the maximum LSBL. Windows with less than 10 markers were excluded from further analyses.

3-Long-range haplotype tests
We employed two approaches based on haplotype diversity. For these tests, we restricted the analyses to markers with minor allele frequencies equal or higher than 5%. The statistics are based on the combined 1000 Genomes Phase I East Asian samples (iHS test), and the combined 1000 Genomes Phase I East Asian, European and African samples (Rsb tests). More details about the statistics are described below.

iHS (Integrated Haplotype Score) iHS compares integrated EHH (Extended Haplotype Homozygosity) values between alleles at a given SNP. EHH quantifies the breakdown of LD at increasing distances from each allele (ancestral or derived). Large negative iHS values are indicative of unusually long haplotypes carrying the derived allele and large positive values are associated with long haplotypes carrying the ancestral allele [38]. iHS values were estimated using the program rehh [39].

Rsb Rsb is a standardized ratio of iES (Integrated EHHS) from two populations. iES integrates the area under the curve of site-specific EHH (EHHS) [11]. Extreme values of Rsb indicate slower haplotype homozygosity decay in one population versus another. This test was designed to identify potential sweeps that have occurred only in one population. Given that we are primarily interested in identifying sweeps that are specific to East Asian populations, we focused on the comparison between East Asian and European populations, and East Asian and African populations. In this particular situation, extreme positive values of Rsb will indicate longer haplotypes in East Asian populations than in European or African populations. Rsb(Eas-Eur) and Rsb(Eas-Afr) were estimated using the program rehh [39].

After obtaining the iHS, Rsb(Eas-Eur), and Rsb(Eas-Afr) statistics for each locus, we used a combination of shell and python scripts to report the results for non-overlapping windows of 25 kilobases, indicating for each window the maximum absolute value of iHS (or the maximum value of Rsb). Windows with less than 10 markers were excluded from further analyses.

Construction of empirical distribution of p-values based on results for 25 kb windows and identification of putative pigmentation genes under positive selection.
Based on the results obtained in the analyses of 25 kb windows (e.g. values obtained for each of the SFS statistics, and maximum values for LSBL, iHS and Rsb, see above for additional information), we sorted the windows in descending order based on the values of the relevant statistics, and identified pigmentation genes that are outliers in the empirical distribution (top 0.1% or 1% of the distribution), following the approach detailed below:

1-Identification of extreme outliers with empirical p-values < 0.001 and annotation of the relevant windows using the DAVID database
We used the ENSEMBL genome browser (http://useast. ensembl.org/index.html) to identify genes overlapping with the top 0.1% of the 25 kb windows for each statistic. These genes were then annotated using the DAVID database (Database for Annotation, Visualization and Integrated Discovery) [40] in order to identify genes involved in the pigmentation pathway. Briefly, we used the ENSEMBL gene IDs retrieved from the ENSEMBL genome browser as input to perform a functional annotation of each gene using DAVID Functional Annotation

Tool. This tool provides different types of annotations for each gene, including annotations based on functional categories, gene ontology (e.g. GOTERM, PANTHER), pathways (e.g. BIOCARTA, KEGG_PATHWAY), protein domains (e.g. INTERPRO) and protein interactions.

2-Identification of known pigmentation genes with empirical p-values < 0.01

We prepared a list of known pigmentation genes that 1) Have been associated with pigmentary traits in association studies or 2) Have been reported as outliers in previous scans of positive selection in human populations. The list included the following genes:

ADAM17, ADAMTS20, AP3D1, ASIP, ATRN, BLOC1S6 (PLDN), BNC2, CTSD, DCT, DRD2, DTNBP1, EDAR, EDN2, EGFR, HPS1, IRF4, KIT, KITLG, LYST, MATP (SLC45A2), MC1R, MITF, MLPH, MYO5A, MYO7A, OCA2/HERC2, OPRM1, PAX3, PDIA6, PMEL (SILV), POMC, PPARD, RAB27A, RAD50, RGS19, SLC24A4, SLC24A5, TYR, TYRP1, TP53BP1, TRPM1 and *TPCN2*.

We retrieved the results of the 8 statistics analyzed in this study for all the 25 Kb windows overlapping the aforementioned genes, and identified the genes with windows in the top 1% of the empirical distributions.

Therefore, all the genes reported in the Results and Discussion section are outliers that show statistics in the top 1% of the empirical distribution, and in some cases, the top 0.1% of the empirical distribution.

Results and discussion

We carried out genome-wide scans for signatures of selection in East Asians, with a major focus on the identification of pigmentation genes that have been under positive selection in this population group. We used three types of statistics: Statistics based on the Site Frequency Spectrum (SFS) (D, D*, F* and H), statistics based on genetic differentiation (LSBL) and long-range haplotype tests (iHS and Rsb) (See the Materials and Methods section for more details about each statistic). Importantly, these statistics are based on different characteristics of the data and are powered to identify different types of selective sweeps. For example, tests based on the SFS are primarily powered to identify older selective events and recently completed sweeps, whereas long-range haplotype tests are more useful to identify more recent events (<30,000 years ago) and incomplete or partial sweeps [41]. All statistical analyses were based on the 1,000 Genomes Phase 1 reference samples, which include approximately 38 million SNPs. For the statistical analyses, we created non-overlapping windows of 25 kb, which were used to construct an empirical distribution for each statistic. We identified genes located within the top 0.1% of the windows for each statistic and these genes were then annotated using the DAVID

database in order to select genes that may potentially be involved in the pigmentation pathway. In addition to these extreme outliers, we also explored if a list of genes that have been previously associated with pigmentary traits in association studies or reported as outliers in previous scans of positive selection in human populations were located in the top 1% of the empirical distributions.

The top 0.1% 25Kb windows identified for the different statistics (957 windows) are depicted in Additional file 1: Table S1, and the basic annotations for the genes retrieved using the DAVID database (422 genes) are provided as Additional file 2: Table S2. Many of the genes reported to harbor signatures of positive selection in previous studies including East Asian populations, based on a wide range of methods, such as the Composite of Multiple Signals (CMS) test [42], the XP-EHH test [15,17] the iHS test [17,43], the LRH test [43], the Rsb test [11] and the XP-CLR test [20] are also outliers in our study. Overall, 79 genes described in these studies were also identified in our analyses, and these genes are highlighted in red in Additional file 2: Table S2. It is important to note that among the genes reported in Additional file 2: Table S2, many are located on the same genomic regions. In fact, several genomic regions are characterized by the presence of large numbers of outlier windows, such as the *EDAR* region on chromosome 2, or a genomic region on chromosome 17 characterized by extreme values for several SFS statistics (Additional file 1: Table S1 and Additional file 2: Table S2). Presumably, these are genomic regions that have been under strong and relatively recent positive selection, and the selective sweeps left a strong signature in these regions, encompassing several genes. Further analyses would be needed to determine which genes were targets of positive selection in these regions. Our primary goal in this study has been to identify pigmentation genes showing putative signals of positive selection.

Table 1 shows the list of outlier genes that are relevant for pigmentation biology. There are 20 genes in the list. Of these, 8 genes (*ATRN, EDAR, KLHL7, MITF, OCA2, TH, TMEM33* and *TRPM1*,) were extreme outliers (top 0.1% of the empirical distribution) for at least one statistic, and 12 genes (*ADAM17, BNC2, CTSD, DCT, EGFR, LYST, MC1R, MLPH, OPRM1, PDIA6, PMEL (SILV)* and *TYRP1*) were in the top 1% of the empirical distribution for at least one statistic. Most of the genes are outliers for more than one statistic, and show multiple significant windows. It is important to note that, with the exception of *TH, KLHL7* and *CTSD*, these genes have already been described in genome-wide scans of signatures of selection in previous studies [12-17,20,21,38, 44-49], lending strong support to the hypothesis that positive selection has substantially shaped the patterns of variation of these genes. Additionally, eight of these

Table 1 Pigmentation genes that are outliers based on the empirical distribution of different tests of positive selection

Gene	iHS	LSBL	Rsb-Eas-Afr	Rsb-Eas-Eur	Tajima D	H	D*	F*
Top 0.1%								
EDAR	+(0.0019, 2)	++ (9.3E-06, 5)	++ (9.6E-04, 2)	+(0.0049, 1)	++(6.2E-04, 3) CHB			
MITF					++ (9.9E-04, 4) CHB/CHS/JPT	+(0.0037, 3) CHB/CHS/JPT	++ (2.1E-04, 5) CHS/JPT	++ (1.8E-04, 6) CHS/JPT
ATRN					+(0.0077, 2) CHB		++ (9.6E-04, 3) CHB	++ (9.4E-04, 6) CHB/CHS
OCA2		++(5.6E-05, 5)						
TRPM1		+(0.0012, 6)	+(0.0019, 5)		++(8.9E-04, 1) CHS			
TH	++(7.9E-04, 1)		+(0.0073, 1)	+(0.0028, 1)				
KLHL7							++(8.81E-04, 5) CHB/CHS	+(0.0010, 5) CHB/CHS
TMEM33					++(1.72E-04, 9) CHB/CHS/JPT		+(0.0029, 3) CHB	+(0.0017, 3) CHB
Top 1%								
PMEL (SILV)							+(0.0051, 2) CHS	+(0.0072, 2) CHS
BNC2					+ (0.0026, 2) CHB/CHS	+ (0.0056, 2) CHS/JPT	+ (0.0051, 1) CHS	+ (0.0039, 1) CHS
EGFR		+ (0.0029, 3)	+ (0.0056, 1)	+ (0.0028, 1)				
LYST			+ (0.0019, 1)				+ (0.0072, 1) CHS	+ (0.0057, 1) CHS
DCT		+ (0.0099, 1)					+(0.0064, 1) CHB	+(0.0067, 1) CHB
OPRM1					+ (0.0079, 1) JPT			
TYRP1						+(0.0081, 1) CHS		
MC1R	+(0.0073, 1)		+ (0.0027, 1)					
MLPH			+(0.0015, 4)			+(0.0014, 3) CHB/CHS/JPT		
ADAM17			+(0.0015, 2)	+(0.0044, 2)	+(0.0021, 3) JPT			
CTSD							+(0.0045, 1) JPT	+(0.0044, 1) JPT
PDIA6			+(0.0069, 1)				+(0.0072, 1) JPT	

The analysis was based on windows of 25 kilobases. In each cell, we indicate if the gene is in the top 0.1% (++) or 1% (+) of the empirical distribution for the relevant statistics. In parenthesis, we report the smallest p-value observed for any of the windows overlapping the gene, and the number of windows with p-values < 0.01. For the Site Frequency Spectrum tests (D, H, D*, and F*) we also indicate the East Asian populations in which we identified outlier 25 kb windows.

genes (*BNC2, EGFR, LYST, MC1R, OCA2, OPRM1, PMEL (SILV)* and *TYRP1*) have been associated with pigmentary traits in association studies. The *OCA2* gene is of particular interest, because different haplotypes are associated with pigmentary traits in Europeans and East Asians. Variants located in the nearby *HERC2* gene, which affect the transcription of the *OCA2* gene, are strongly associated with blue eye color in European populations [50-55]. Another non-synonymous variant, which is common in East Asian populations but absent or very rare in Europe, has been associated with skin pigmentation

in East Asia [23,29]. Several polymorphisms in the *MC1R* gene show a strong association with red hair/fair skin in European populations (Asp84Glu, Arg151Cys, Arg160Trp and Asp294His), and other variants also show a weak association with these traits (Val60Leu, Val92Met and Arg163Gln) [56]. Interestingly, the derived 163Gln allele, which is present in very high frequencies in East Asian populations (>60%), but very low frequencies in European and African populations, has been recently associated with lighter skin in an East Asian sample [30]. Polymorphisms in the *TYRP1* gene have been associated with hair and iris

color in European populations [57], and a non-synonymous mutation that is found only in Oceania was recently associated with blond hair in Melanesians [58]. Frudakis et al. [59] reported association of haplotypes in the *PMEL (SILV)* gene with iris color. The *BCN2* gene has been associated with skin pigmentation and freckling in European populations [60,61]. Variants in the *LYST* gene have been associated with eye color in a Dutch sample [53]. Finally, polymorphisms in the genes *EGFR* and *OPRM1* have been recently associated with skin pigmentation in admixed samples from the New World [49]. We provide more detailed information about each gene, its relevance in the pigmentation pathway, and a description of previous natural selection scans or association studies with pigmentary phenotypes, if relevant, as Additional file 3.

One of the methods employed in our study (LSBL) was designed to highlight genomic regions showing extreme differentiation in one population, with respect to other groups. In this study, we were primarily interested in identifying genomic regions that differentiate East Asian populations with respect to Europeans and Africans. In particular, we would like to find regions in which positive selection may have driven the reduction of melanin levels specifically in East Asia. In order to explore this in more detail, we used the program Haploview [62] to compare the haplotype structure of the candidate pigmentation genes showing large LSBL values in East Asians (*EDAR, OCA2, EGFR* and *DCT*) with the haplotype structure observed in European populations, which are also characterized by reduced melanin content. For the top LSBL windows found for each of these genomic regions, we identified overlapping common markers (frequency higher than 1%) between the East Asian and European 1000 Genomes reference samples, and used the program Haploview to generate the haplotype block structure in each population, using the default algorithm [63]. As expected, we observed very large differences in haplotype frequencies in these regions between the East Asian and European populations. These haplotype differences range between 59% (*DCT*) and 90% (*EDAR*) and these contrasting patterns indicate that positive selection may have favored specific haplotypes in East Asian populations. Figure 1 shows the haplotype structure observed for the *OCA2* gene in East Asian and European populations. The largest differences in frequency are observed for East Asian haplotype blocks 4 and 5, which span slightly more than 10 kilobases, from position 28,187,772 to 28,199,863 on chromosome 15. Interestingly, the non-synonymous variant that has been associated with skin pigmentation in East Asians, rs1800414 [23,29] is located within this region (genomic position 28,197,037 in Genome Built 37.3). Other studies have also reported distinct signatures of positive selection and different haplotype distributions for OCA2 in Europe and East Asia [12,14,24,64,65]. The

haplotype structure of the genes *EDAR, DCT,* and *EGFR* in European and East Asian populations is provided as Additional file 4. These analyses confirm that some genes relevant for pigmentation biology show extreme haplotype differentiation between European and East Asian populations, and suggest that a careful analysis of haplotype variation, in combination with a detailed annotation of the variants present in the relevant windows, may help to identify the genetic variants responsible for the selective sweeps in East Asians.

In summary, we have carried out a genome-wide analysis of selection signatures in East Asian populations, focusing on genes that are relevant for pigmentation biology. This analysis allowed us to identify a number of genes that show extremely unusual patterns of genetic variation in East Asia. It is in principle possible that some of these findings are false positives and are not due to the action of recent positive selection in East Asian populations. However, most of these genes are outliers for different tests and/or different populations (CHB, CHS, JPT), and have been described in previous scans for positive selection, providing strong support to the hypothesis that recent selective sweeps left a signature in these regions. It is important to note that, even if selective sweeps are responsible for these unusual patterns of variation, it is possible that the selective factors involved did not have any effect on melanin levels in East Asian populations. Many of these genes are expressed widely and have a broad range of functions, and these selective sweeps may be related to phenotypes other than pigmentation. For example, certain mutations of the Ectodysplasia A receptor gene *(EDAR),* which is an extreme outlier based on three different types of test, are associated with pigmentary phenotypes in mice (http://www.informatics.jax.org/), and for this reason *EDAR* is a pigmentation candidate gene. However, this gene is also important in the development of hair, teeth, and other ectodermal derivatives, and mutations in this gene have been associated with several traits in humans, including hypohidrotic ectodermal dysplasia [66] shovel-shaped incisors [67] and hair thickness [68]. In a recent study [69], researchers generated a knock-in mouse to test the phenotypic consequences of the EDARV370A (370A) polymorphism, which has been associated with hair thickness and incisor shoveling in East Asian populations. The researchers found that 370A homozygous mice had thicker hair than 370V homozygous mice, similarly to the patterns observed in human populations. Importantly, they also observed that the 370A mice had smaller mammary fat pads and increased eccrine gland numbers. An association study in individuals of Han descent showed that the 370A allele was associated with shoveling of the upper incisors and also eccrine gland density. These findings suggest that the dramatic
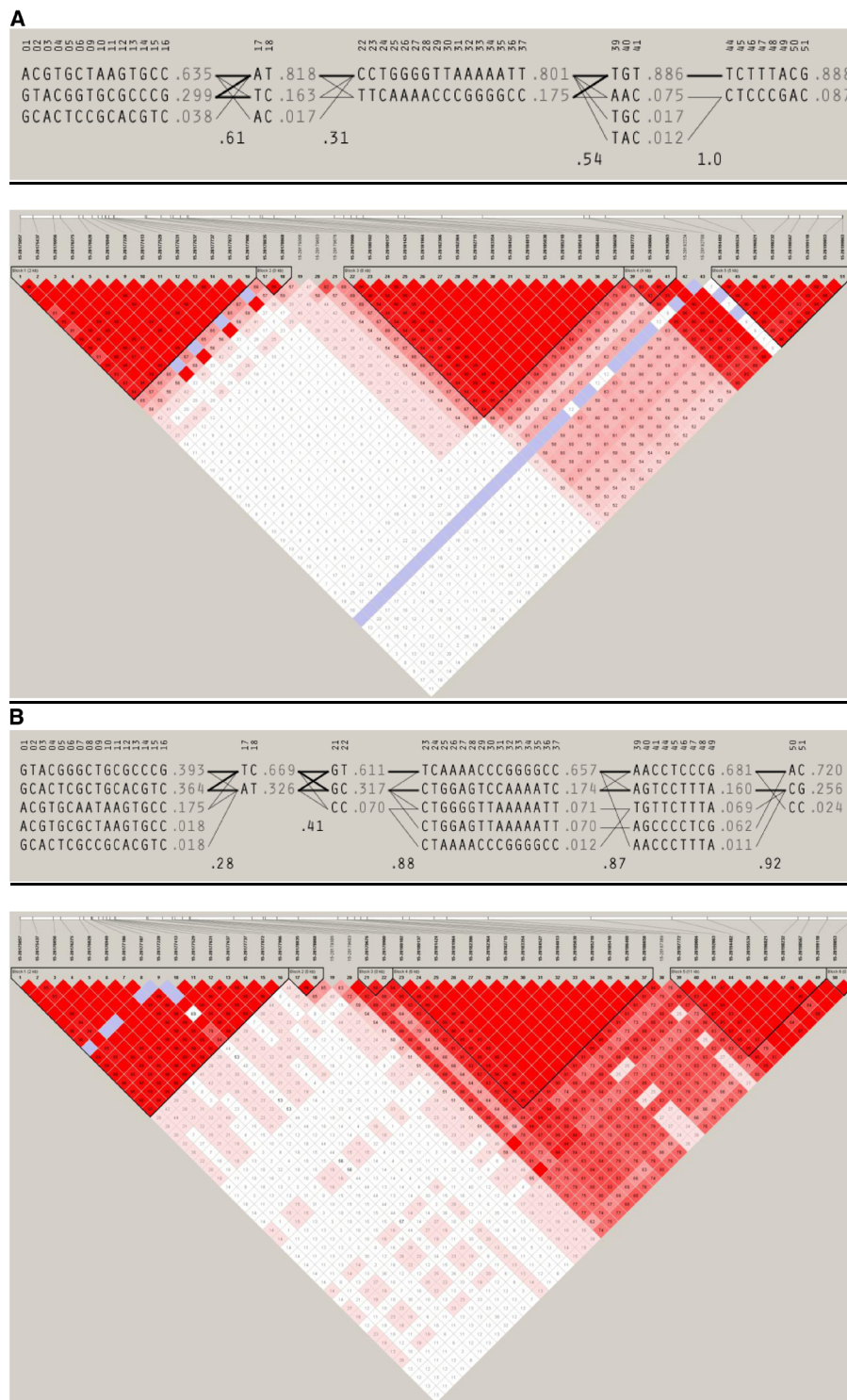
**Figure 1 Haplotype block structure of OCA2 gene (window chr15:28,175,001-28,200,000) A. In East Asian 1000 Genomes samples.
B. In European 1000 Genomes samples.**

increase in the frequency of the 370A allele in East Asia could have been driven by selection favoring more efficient evapo-transpiration, although the authors also mentioned the possibility that reduced mammary fat pad size could also have been adaptive, or, given the clear pleiotropic effects of the 370A mutation, that selection

acted on multiple traits. This fascinating example highlights the challenges encountered when trying to explain the ultimate selective factors responsible for some of the selective sweeps observed in human populations, and emphasizes the importance of association studies, functional studies, and studies in animal models to complement genome-wide scans of selection signatures.

Conclusions

Our study has identified a list of genes that could potentially explain the reduction of melanin levels that took place in East Asia after the out-of-Africa migration of anatomically modern humans. The application of recently developed methods, such as the Composite of Multiple Signals (CMS) test [42] or Approximate Bayesian Computation (ABC) tests [70], and a more extensive annotation of the polymorphisms present within and around these genes, may be useful to narrow down the genic regions that were the target of positive selection, and to distinguish between selection that has acted on newly arisen mutations or standing variation. However, it will be necessary to carry out association studies in samples for which quantitative data on pigmentary traits are available, and functional studies in melanocytes to confirm the implication of these genes in normal pigmentation variation.

Additional files

Additional file 1: Table S1. List of the top 0.1% windows for all the statistics used in this study.

Additional file 2: Table S2. Basic annotation of the genes overlapping the top 0.1% windows identified in this study. The annotations were obtained with the DAVID database. The table reports the gene, the gene name, chromosome location, statistical tests for which the gene is an outlier, other references that have reported signatures of selection in the relevant genes, disease class for which genetic associations have been reported, OMIM_disease, KEGG pathway and GOTERM Biological Pathway.

Additional file 3. Description of the pigmentation genes identified in the scan for signatures of selection in East Asia.

Additional file 4. Haplotype block structure of the genes DCT, EDAR and EFGR in the East Asian and European 1000 Genomes samples.

Abbreviations

HGDP: Human Genome Diversity Project; LD: Linkage Disequilibrium; SNP: Single Nucleotide Polymorphism; JPT: Japanese; CHB: Han Chinese from Beijing; CHS: Southern Han Chinese; LSBL: Locus Specific Branch Length; iHS: Integrated Haplotype Score; SFS: Site Frequency Spectrum; DAVID: Database for Annotation, Visualization and Integrated Discovery; CMS: Composite of Multiple Signals; ABC: Approximate Bayesian Computation.

Competing interests

The authors have declared that no competing interests exist.

Authors' contributions

JLH participated in acquisition of data, contributed to the analysis and interpretation of data, and wrote the first draft of the manuscript, TS, RG, AR and JMA contributed to the analysis and interpretation of data, ME contributed to the acquisition of data, EJP was responsible for study conception and design, contributed to the analysis and interpretation of

data and participated in the preparation of the final version of the manuscript. All authors read and approved the final version of the manuscript.

Author details

[1]Department of Anthropology, University of Toronto at Mississauga, Mississauga, Ontario, Canada. [2]Department of Genome Sciences, University of Washington, Seattle, Washington, USA. [3]Department of Mathematical and Computational Sciences, University of Toronto at Mississauga, Mississauga, Ontario, Canada.

References

1. Henn BM, Cavalli-Sforza LL, Feldman MW: **The great human expansion.** *Proc Natl Acad Sci USA* 2012, **109**:17758–17764.
2. Jablonski NG, Chaplin G: **Human skin pigmentation as an adaptation to UV radiation.** *Proc Natl Acad Sci* 2010, **107**(Suppl 2):8962–8968.
3. Scherer D, Kumar R: **Genetics of pigmentation in skin cancer.** *Mutat Res* 2010, **705**:141–153.
4. Jablonski NG, Chaplin G: **The evolution of human skin coloration.** *J Hum Evol* 2000, **39**:57–106.
5. Parra EJ: **Human pigmentation variation: evolution, genetic basis, and implications for public health.** *AM J Phys Anthropol Suppl* 2007, **45**:85–105.
6. Juzeniene A, Setlow R, Porojnicu, Steindal AH, Moan J: **Development of different human skin colors: a review of highlighting photobiological and photobiophysical aspects.** *Photochem Photobiol B Biol* 2009, **96**:93–100.
7. Elias PM, Menon G, Wetzel BK, Williams JW: **Barrier requirements as the evolutionary driver of epidermal pigmentation in humans.** *Am J Hum Biol* 2010, **22**:526–537.
8. Jablonski NG, Chaplin G: **Human skin pigmentation, migration and disease susceptibility.** *Phil Trans R Soc B* 2012, **367**:785–792.
9. Izagirre N, Garcia I, Junquera, de la Rua C, Alonso S: **A scan for signatures of positive selection in candidate loci for skin pigmentation in humans.** *Mol Biol Evol* 2006, **23**:1697–1706.
10. Nakayama K, Soemantri A, Jin F, Dashnyam B, Ohtsuka R, Duanchang P, Isa MN, Settheetham-Ishida W, Harihara S, Ishida T: **Identification of novel functional variants of the melanocortin 1 receptor gene originated from Asians.** *Hum Genet* 2006, **119**:322–330.
11. Tang K, Thornton KR, Stoneking M: **A New approach for using genome scans to detect recent positive selection in the human genome.** *PLOS Bio* 2007, **5**:1587–1602.
12. McEvoy B, Beleza S, Shriver MD: **The genetic architecture of normal variation in human pigmentation: an evolutionary perspective and model.** *Hum Mol Genet* 2006, **15**(suppl 2):176–181.
13. Myles S, Somel M, Tang K, Kelso J, Stoneking M: **Identifying genes underlying skin pigmentation differences among human populations.** *Hum Genet* 2007, **120**:613–621.
14. Lao O, De Gruijter JM, Van Duijn K, Navarro A, Kayser M: **Signatures of positive selection in genes associated with human skin pigmentation as revealed from analyses of single nucleotide polymorphisms.** *Ann Hum Genet* 2007, **71**:354–369.
15. Sabeti PC, Varilly P, Fry B, Lohmueller J, Elizabeth Hostetter E, Cotsapas C, Xie X, Byrne EH, McCarroll SA, Gaudet R, Schaffner SF, Lander ES, The International HapMap Consortium: **Genome-wide detection and characterization of positive selection in human populations.** *Nature* 2007, **449**:913–918.
16. Alonso S, Izagirre N, Smith-Zubiaga I, Gardeazabal J, Díaz-Ramón JL, Díaz-Pérez JL, Zelenika D, Boyano MD, Smit N, de la Rúa C: **Complex signatures of selection for the melanogenic loci TYR, TYRP1 and DCT in humans.** *BMC Evol Biol* 2008, **8**:1471–2148.
17. Pickrell JK, Coop G, Novembre J, Kudaravalli S, Li JZ, Absher D, Srinivasan BS, Barsh GS, Myers RM, Feldman MW, Pritchard JK: **Signals of recent positive selection in a worldwide sample of human populations.** *Genome Res* 2009, **19**:826–837.

18. Coop G, Pickrell JK, Novembre J, Kudaravalli S, Li J, et al: **The role of geography in human adaptation.** *PLoS Genet* 2009, **5**:e1000500.

19. Hancock AM, Witonsky DB, Alkorta-Aranburu G, Beall CM, Gebremedhin A, Sukernik R, Utermann G, Pritchard JK, Coop G, Di Rienzo A: **Adaptations to climate-mediated selective pressures in humans.** *PLoS Genet* 2011, **7**:1–16.

20. Chen H, Patterson N, Reich D: **Population differentiation as a test for selective sweeps.** *Genome Res* 2010, **20**:393–402.

21. Williamson SH, Hubisz MJ, Clark AG, Payseur BA, Bustamante CD, Nielsen R: **Localizing recent adaptive evolution in the human genome.** *PLoS Genet* 2007, **3**:e90.

22. Norton HL, Kittles RA, Parra E, McKeigue P, Mao X, Cheng K, Canfield VA, Bradley DG, McEvoy B, Shriver MD: **Genetic evidence for convergent evolution of light skin in European and East Asians.** *Mol Bio Evol* 2007, **24**:710–722.

23. Edwards M, Bigham A, Tan J, Li S, Gozdzik A, Ross K, Jin L, Parra EJ: **Association of the *OCA2* Polymorphism His615Arg with Melanin Content in East Asian Populations: Further Evidence of Convergent Evolution of Skin Pigmentation.** *PLoS Genet* 2010, **6**:e1000897.

24. Donnelly MP, Paschou P, Grigorenko E, Gurwitz D, Barta C, Lu RB, Zhukova OV, Kim JJ, Siniscalco M, New M, Li H, Kajuna SL, Manolopoulos VG, Speed WC, Pakstis AJ, Kidd JR, Kidd KK: **A global view of the OCA2-HERC2 region and pigmentation.** *Hum Gen* 2012, **13**:683–696.

25. Beleza S, Alonso SM, McEvoy, Alves I, Martinho C, Cameron E, Shriver MD, Parra EJ, Rocha J: **The timing of pigmentation lightening in europeans.** *Mol Evol Bio* 2013, **1**:24–35.

26. 1000 genomes project consortium: **An integrated map of genetic variation from 1092 human genomes.** *Nature* 2012, **491**:56–65.

27. Rees JL, Harding RM: **Understanding the evolution of human pigmentation: recent contributions from population genetics.** *J Invest Dermatol* 2012, **132**:846–853.

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

29. Abe Y, Tamiya G, Nakamura T, Hozumi Y, Suzuki T: **Association of melanogenesis genes with skin color variation among Japanese females.** *J Dermatol Sci* 2013, **69**:167–172.

30. Yamaguchi K, Watanabe C, Kawaguchi A, Sato T, Naka I, Shindo M, Moromizato K, Aoki K, Ishida H, Kimura R: **Association of melanocortin 1 receptor gene (MC1R) polymorphisms with skin reflectance and freckles in Japanese.** *J Hum Genet* 2012, **57**:700–708.

31. Ang KC, Ngu MS, Reid KP, Teh MS, Aida ZS, Koh DX, Berg A, Oppenheimer S, Salleh H, Clyde MM, Md-Zain BM, Canfield VA, Cheng KC: **Skin color variation in Orang Asli tribes of Peninsular Malaysia.** *PLoS One* 2012, **7**:e42752.

32. Tajima F: **Statistical method for testing the neutral mutation hypothesis by DNA polymorphism.** *Genetics* 1989, **123**:585–595.

33. Fu YX, Li WH: **Statistical tests of neutrality of mutations.** *Genetics* 1993, **133**:693–709.

34. Fu X: **Statistical tests of neutrality of mutations against population growth, hitch-hiking, and background selection.** *Genetics* 1997, **147**:915–925.

35. Fay JC, Wu CI: **Hitchhiking under positive Darwinian selection.** *Genetics* 2000, **155**:1405–1413.

36. Shriver MD, Kennedy GC, Parra EJ, Lawson HA, Sonpar V, Huang, Akey JM, Jones KW: **The genomic distribution of population substructure in four populations using 8,525 autosomal SNPs.** *Hum Genomics* 2004, **4**:274–286.

37. Weir BS, Cockerham CC: **Estimating F-statistics for the analysis of population structure.** *Evolution Int J Or Evol* 1984, **38**:1358–1370.

38. Voight BF, Kudaravalli S, Wen X, Pritchard JK: **A map of recent positive selection in the human genome.** *PLoS Biol* 2006, **4**:e72.

39. Gautier M, Vitalis R: **rehh: An R package to detect footprints of selection in genome-wide SNP data from haplotype structure.** *Bioinformatics* 2012, **28**:1176–1177.

40. Dennis G, Sherman BT, Hosack DA, Yang J, Baseler MW, Lane HC, Lempicki RA: **DAVID: Database for Annotation, Visualization, and Integrated Discovery.** *Genome Biol* 2003, **4**:P3. Epub.

41. Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, Palma A, Mikkelsen S, Altshuler D, Lander ES: **Positive natural selection in the human lineage.** *Science* 2006, **312**:614–620.

42. Grossman SR, Andersen KG, Shlyakhter I, Tabrizi S, Winnicki S, Yen A, Park DJ, Griesemer D, Karlsson EK, Wong SH, Cabili M, Adegbola RA, Bamezai RNK, Hill AVS, Vannberg FO, Rinn JL, 1000 Genomes Project, Lander ES, Schaffner SF, Sabeti PC: **Identifying recent adaptations in large-scale genomic data.** *Cell* 2013, **4**:703–713.

43. The International HapMap Consortium: **A second generation human haplotype map of over 3.1 million SNPs.** *Nature* 2007, **449**:851–861.

44. Kelley JL, Madeoy J, Calhoun JC, Swanson W, Akey JM: **Genomic signatures of positive selection in humans and the limits of outlier approaches.** *Genome Res* 2006, **16**:980–989.

45. Kimura R, Fujimoto A, Tokunaga K, Ohashi J: **A practical genome scan for population-specific strong selective sweeps that have reached fixation.** *PLoS One* 2007, **2**:e286.

46. Duffy DL, Montgomery GM, Chen W, Zhao ZZ, Le L, James MR, Hayward NK, Martin NG, Sturm RA: **A Three–Single-Nucleotide Polymorphism Haplotype in Intron 1 of *OCA2* Explains Most Human Eye-Color Variation.** *Am J Hum Genet* 2007, **80**:241–252.

47. Zhong M, Lange K, Papp JC, Fan R: **A powerful score test to detect positive selection in genome-wide scans.** *Europ J Hum Genet* 2010, **18**:1148–1159.

48. Akey JM, Zhang G, Zhang K, Jin L, Shriver MD: **Interrogating a High-Density SNP Map for Signatures of Natural Selection.** *Genome Res* 2002, **12**:1805–1814.

49. Quillen EE, Bauchet M, Bigham AW, Delgado-Burbano ME, Faust FX, Klimentidis YC, Mao X, Stoneking M, Shriver MD: **OPRM1 and EGFR contribute to skin pigmentation differences between Indigenous Americans and Europeans.** *Hum Genet* 2011, **13**:1073–1080.

50. Sulem P, Gudbjartsson DF, Stacey SN, Helgason A, Rafnar T, Magnusson KP, Manolescu A, Karason A, Palsson A, Thorleifsson G, Jakobsdottir M, Steinberg S, Palsson S, Jonasson F, Sigurgeirsson B, Thorisdottir K, Ragnarsson R, Benediktsdottir KR, Aben KK, Kiemeney LA, Olafsson JH, Gulcher J, Kong A, Thorsteinsdottir U, Stefansson K: **Genetic determinants of hair, eye and skin pigmentation in Europeans.** *Nat Genet* 2007, **39**:1443–1452.

51. Kayser M, Liu F, Janssens AC, Rivadeneira F, Lao O, van Duijn K, Vermeulen M, Arp P, Jhamai MM, van Ijcken WF, Den Dunnen JT, Heath S, Zelenika D, Despriet DD, Klaver CC, Vingerling JR, De Jong PT, Hofman A, Aulchenko YS, Uitterlinden AG, Oostra BA, Van Duijn CM: **Three genome-wide association studies and a linkage analysis identify *HERC2* as a human iris color gene.** *Am J Hum Genet* 2008, **82**:411–423.

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

53. Liu F, Wollstein A, Hysi PG, Ankra-Badu GA, Spector TD, Park D, Zhu G, Larsson M, Duffy DL, Montgomery GW, Mackey DA, Walsh S, Lao O, Hofman A, Rivadeneira F, Vingerling JR, Uitterlinden AG, Martin NG, Hammond CJ, Kayser M: **Digital quantification of human eye color highlights genetic association of three new loci.** *PLoS Genet* 2010, **6**:e1000934.

54. Cook AL, Chen W, Thurber AE, Smit DJ, Smith AG, Bladen TG, Brown DL, Duffy DL, Pastorino L, Bianchi-Scarra G, Leonard JH, Stow JL, Sturm RA: **Analysis of cultured human melanocytes based on polymorphisms within the SLC45A2/MATP, SLC24A5/NCKX5, and OCA2/P loci.** *J Invest Dermatol* 2009, **129**:392–405.

55. Visser M, Kayser M, Palstra RJ: **HERC2 rs12913832 modulates human pigmentation by attenuating chromatin-loop formation between a long-range enhancer and the OCA2 promoter.** *Genome Res* 2012, **22**:446–455.

56. Duffy D, Box N, Chen W, Palmer JS, Montgomery GW, James MR, Hayward NK, Martin NG, Sturm RA: **Interactive effects of MC1R and OCA2 on melanoma risk phenotypes.** *Hum Mol Genet* 2004, **13**:447–461.

57. Sulem P, Gudbjartsson DF, Stacey SN, Helgason A, Rafnar T, Jakobsdottir M, Steinberg S, Gudjonsson SA, Palsson A, Thorleifsson G, Pálsson S, Sigurgeirsson B, Thorisdottir K, Ragnarsson R, Benediktsdottir KR, Aben KK, Vermeulen SH, Goldstein AM, Tucker MA, Kiemeney LA, Olafsson JH, Gulcher J, Kong A, Thorsteinsdottir U, Stefansson K: **Two newly identified genetic determinants of pigmentation in Europeans.** *Nat Genet* 2008, **40**:835–837.

58. Kenny EE, Timpson NJ, Sikora M, Yee MC, Moreno-Estrada A, Eng C, Huntsman S, Burchard EG, Stoneking M, Bustamante CD, Myles S: **Melanesian blond hair is caused by an amino acid change in TYRP1.** *Science* 2012, **336**:554.

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

60. Jacobs LC, Wollstein A, Lao O, Hofman A, Klaver CC, Uitterlinden AG, Nijsten T, Kayser M, Liu F: **Comprehensive candidate gene study highlights UGT1A and BNC2 as new genes determining continuous skin color variation in Europeans.** *Hum Genet* 2013, **132**:147–158.

61. Eriksson N, Macpherson JM, Tung JY, Hon LS, Naughton B, Saxonov S, Avey L, Wojcicki A, Pe'er I, Mountain J: **Web-Based Participant-Driven Studies Yield Novel Genetic Associations for Common Traits.** *PLoS Genet* 2010, **6**:e1000993.

62. Barrett JC, Fry B, Maller J, Daly MJ: **Haploview: analysis and visualization of LD and haplotype maps.** *Bioinformatics* 2005, **21**:263–265.

63. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D: **The structure of haplotype blocks in the human genome.** *Science* 2002, **296**:2225–2229.

64. Yuasa HJ, Takubo M, Takahashi A, Hasegawa T, Noma H, Suzuki T: **Evolution of vertebrate indoleamine 2,3-dioxygenases.** *J Mol Evol* 2007, **65**:705–714.

65. Anno S, Abe T, Yamamoto T: **Interactions between SNP alleles at multiple loci contribute to skin color differences between Caucasoid and Mongoloid subjects.** *Int J Biol Sci* 2008, **4**:81–86.

66. Cluzeau C, Hadj-Rabia S, Jambou M, Mansour S, Guigue P, Masmoudi S, Bal E, Chassaing N, Vincent MC, Viot G, Clauss F, Maniere MC, Toupenay S, Le Merrer M, Lyonnet S, Cormier-Daire V, Amiel J, Faivre L, de Prost Y, Munnich A, Bonnefont JP, Bodemer C, Smahi A: **Only four genes (EDA1, EDAR, EDARADD, and WNT10A) account for 90% of hypohidrotic/anhidrotic ectodermal dysplasia cases.** *Hum Mutat* 2011, **32**:70–72.

67. Kimura R, Yamaguchi T, Takeda M, Kondo O, Toma T, Hanejk K, Hanihara T, Matsukusa H, Kawamura S, Maki K, Osawa M, Ishida H, Oota H: **A common variation in EDAR is a genetic determinant of shovel-shaped incisors.** *Am J Hum Genet* 2009, **85**:528–535.

68. Fujimoto A, Kimura R, Ohashi J, Omi K, Yuliwulandari R, Batubara L, Mustofa MS, Samakkarn U, Settheetham-Ishida W, Ishida T: **A scan for genetic determinants of human hair morphology: EDAR is associated with Asian hair thickness.** *Hum Mol Genet* 2008, **17**:835–843.

69. Kamberov YG, Wang S, Tan J, Gerbault P, Wark A, Tan L, Yang Y, Li S, Tang K, Chen H, Powell A, Itan Y, Fuller D, Lohmueller J, Mao J, Schachar A, Paymer M, Hostetter E, Byrne E, Burnett M, McMahon AP, Thomas MG, Lieberman DE, Jin L, Tabin CJ, Morgan BA, Sabeti PC: **Modeling recent human evolution in mice by expression of a selected EDAR variant.** *Cell* 2013, **152**:691–702.

70. Peter BM, Huerta-Sanchez E, Nielsen R: **Distinguishing between selective sweeps from standing variation and from a de novo mutation.** *PLoS Genet* 2012, **8**:e1003011.