Population Diversity and Adaptive Evolution in Keratinization Genes: Impact of Environment in Shaping Skin Phenotypes

Pramod Gautam,[†,1] Amit Chaurasia,[†,1] Aniket Bhattacharya,[1,2] Ritika Grover,[2,3] Indian Genome Variation Consortium[1] Mitali Mukerji,*[1,2] and Vivek T. Natarajan*[2,3]

[1]Genomics and Molecular Medicine, Council for Scientific and Industrial Research-Institute of Genomics and Integrative Biology, Delhi, India

[2]Academy of Scientific and Innovative Research, Delhi, India

[3]Systems Biology Group, Council for Scientific and Industrial Research-Institute of Genomics and Integrative Biology, Delhi, India

†These authors contributed equally to this work.

***Corresponding authors:** E-mail: mitali@igib.res.in, tn.vivek@igib.res.in.

Associate editor: Joshua Akey

Abstract

Several studies have demonstrated the role of climatic factors in shaping skin phenotypes, particularly pigmentation. Keratinization is another well-designed feature of human skin, which is involved in modulating transepidermal water loss (TEWL). Although this physiological process is closely linked to climate, presently it is not clear whether genetic diversity is observed in keratinization and whether this process also responds to the environmental pressure. To address this, we adopted a multipronged approach, which involved analysis of 1) copy number variations in diverse Indian and HapMap populations from varied geographical regions; 2) genetic association with geoclimatic parameters in 61 populations of dbCLINE database in a set of 549 genes from four processes namely keratinization, pigmentation, epidermal differentiation, and housekeeping functions; 3) sequence divergence in 4,316 orthologous promoters and corresponding exonic regions of human and chimpanzee with macaque as outgroup, and 4) protein sequence divergence (Ka/Ks) across nine vertebrate classes, which differ in their extent of TEWL. Our analyses demonstrate that keratinization and epidermal differentiation genes are under accelerated evolution in the human lineage, relative to pigmentation and housekeeping genes. We show that this entire pathway may have been driven by environmental selection pressure through concordant functional polymorphisms across several genes involved in skin keratinization. Remarkably, this underappreciated function of skin may be a crucial determinant of adaptation to diverse environmental pressures across world populations.

Key words: **keratinization, comparative genomics, noncoding DNA, copy-number variation, adaptive variation, climatic factors, population.**

Introduction

Skin provides the first line of defense against environmental stress and is the most exposed organ system to external perturbations. It is thus conceivable that climatic factors shape skin physiology and the resultant phenotypes. Allied epithelia like that of the lung and gut are involved in exchange of gases and nutrients, respectively, and the barrier function is ancillary at these interfaces. In sharp contrast, skin is optimally designed for protection from ultraviolet (UV) radiations and to act as a barrier for transepidermal water loss (TEWL), via pigmentation and keratinization processes, respectively. Melanin-based skin pigmentation system protects molecules from UV damage in the skin. Across populations, the extent of skin pigmentation inversely correlates with the incident flux of UV rays (Off et al. 2005; Steindal et al. 2006; Han et al. 2007; Jablonski and Chaplin 2010). Consequently, skin pigmentation is adaptive and is evident in the spectrum of skin color phenotypes in populations across the latitude (Jablonski and Chaplin 2010). The other skin-specific process keratinization is mediated by several structural and associated regulatory proteins to form the barrier that restricts the entry and exit of molecules and shields the body from environmental changes in humidity (Ny and Egelrud 2004; Katsuta et al. 2009). It is therefore intuitive that similar environmental forces would have shaped skin keratinization, and consequent adaptations would result in a diversity of this phenotype.

A large number of genes linked to skin pigmentation have been identified through population-based genetic studies vis-a-vis environmental adaptation. Model systems and inborn pigmentation disorders have further highlighted the role of specific genes (Jin et al. 2012; Sturm and Duffy 2012; Baxter and Pavan 2013). Many of these loci exhibit variability across continental populations and contribute to the diversity of skin pigmentation phenotypes. Among the keratinization genes, allelic variations in KRT77, presumed to be involved in thermoregulation via sweating response in skin, have been linked to the flux of solar radiations (Hancock et al. 2011). In addition, genes such as ABCA12 and ATP2A2 involved in keratinization show parallel divergence and are comparable to the divergence in TYR, a gene involved in skin pigmentation (Tennessen and Akey 2011). In the evolutionary context, although it is generally accepted that the process of keratinization is different among the vertebrate classes, genome-scale diversity and the influence of environmental parameters remain largely unexplored.

Comparative genomics provides an excellent opportunity to address mechanistic determinants of this diversity. In chimpanzee, the closest relative of humans, pigmentation is absent in the epidermal skin (Jablonski and Chaplin 2000). Additionally, the human lineage has witnessed the loss of functional body hair. This presents dual physiological challenge at the skin surface, UV damage, and thermoregulation coupled with excessive TEWL. This characteristic difference in physiology enables the assessment of genetic signatures that govern adaptive changes in skin physiology, as the human and chimpanzee genomes are almost identical. Earlier comparative genomic studies have identified a locus of epidermal differentiation complex at 1q21 (of the human genome) as the most rapidly diverging gene cluster (Baxter and Pavan 2013). In addition, 17q21 and 21q22 encode genes that contain several keratins and are also rapidly evolving in the chimpanzee–human clade. Recently, transgenomic capture and sequencing studies have suggested positive selection of keratinization genes to be operational in the primate exomes (George et al. 2011). This could be a key adaptive feature for the restoration of skin barrier compromised due to the loss of hair. In this context, comparative genomic strategy provides valuable support to large-scale genomic data from human populations and could be utilized to address the role of evolutionary forces in dictating selection of phenotypic traits.

In an earlier genome-wide analysis of large copy number variations (CNVs), we have observed extensive diversity in keratinization genes across different Indian populations (Gautam et al. 2012). We postulated that variability in genes governing skin keratinization would be a result of selection imposed by the climatic factors. To address this possibility, we adopted a multipronged approach using two unbiased and two focused analyses. Interestingly, keratinization also appears as one of the most enriched processes when we compare noncoding divergence in the 5-kb promoter-proximal sequences in the orthologous genes between chimpanzee and human. Also it is substantially contributed by genomic repeat sequences, which implies novel mechanisms of gene regulation in this lineage. To examine the influence of climate on skin functions, we studied the association of single-nucleotide polymorphisms (SNPs) in keratinization genes with the environmental parameters enlisted in dbCLINE and found that approximately three-fourth of the genes contain at least one SNP having significant association with a climatic variable. We also observe that keratinization as a process is under accelerated evolution contributed by many structural genes linked to the core process of keratinization and regulators of epidermal differentiation. Throughout our analyses, we have compared the diversity in keratinization with pigmentation, as both of these skin-related processes are evidently under similar selection pressure. Our multipronged analyses from both unbiased and focused approaches reveal that keratinization-related processes show variation in the population, and environmental parameters contribute to the genetic diversity of skin keratinization. In addition to providing a basis to skin phenotypic diversity, our study has direct implications in the incidence of skin disorders such as psoriasis, atopic dermatitis, and vitiligo, whose prevalence is known to significantly vary across geographic locations and seasons (Enamandram and Kimball 2013; Parisi et al. 2013).

Results

Keratinization Is Highly Diverse among Populations

In an earlier work, we had mapped large CNVs in a group of unrelated healthy subjects across 26 diverse Indian populations having multiple ethnic backgrounds and residing in varied geoclimatic zones (Gautam et al. 2012). The identification of CNVs was carried out using different analyses methods following stringent criteria for selection of SNPs for defining CNV regions (CNVRs). Further, we had validated the CNVs using different platforms such as sequenom. In our analysis, we had considered genes that are encompassed in large CNVRs ($>$100 kb) deletion or duplication in 26 diverse Indian populations represented in the Indian Genome Variation Consortium (IGV) panel (Brahmachari et al. 2008). We had employed cn.FARMs, and CNVs were called with ten contiguous probes, which resulted in FDR-controlled CNVR predictions (Gautam et al. 2012). Additionally, using two independent CNV calling algorithms (GTC console and SVS7), we could illustrate a substantial overlap defining the CNVR (supplementary fig. S1, Supplementary Material online). GO enrichment analysis was carried out on a pooled set of genes with CNVRs from each individual population using DAVID. The complete genome was provided as background since we had used data from a genome-wide array (21,095 RefSeq genes from UCSC hg18 build). In the functional annotation clustering (FAC) module of DAVID, we used classification stringency as “high” and Fisher’s exact P-value cutoff $<$ 0.001 to identify functionally enriched clusters (supplementary table S1, Supplementary Material online). The plots (fig. 1*a* and *b*) depict all the GO-enriched terms in different populations. As is evident, keratinization is found to be enriched in many populations, and it remains significant even after FDR cutoff (0.05). Any systematic bias in probe representation could be ruled out as the average probe count per kb gene length was not very different: Epidermal differentiation (0.7), keratinization (0.8), and pigmentation (0.7) (supplementary fig. S1, Supplementary Material online). Despite similar representation in the array, pigmentation did not feature among the enriched classes, whereas keratinization and epidermal differentiation were significantly enriched in multiple populations. The extent of diversity in keratinization was similar to that of the olfactory processes, which are known to be highly diversified across populations (supplementary fig. S2, Supplementary Material online) (Hasin et al. 2008).

Following this lead, we extended our study to explore the extent of CNV in keratinization and epidermal differentiation in the HapMap populations. Because our earlier study was carried out on a low-density array (Affymetrix 50K), we

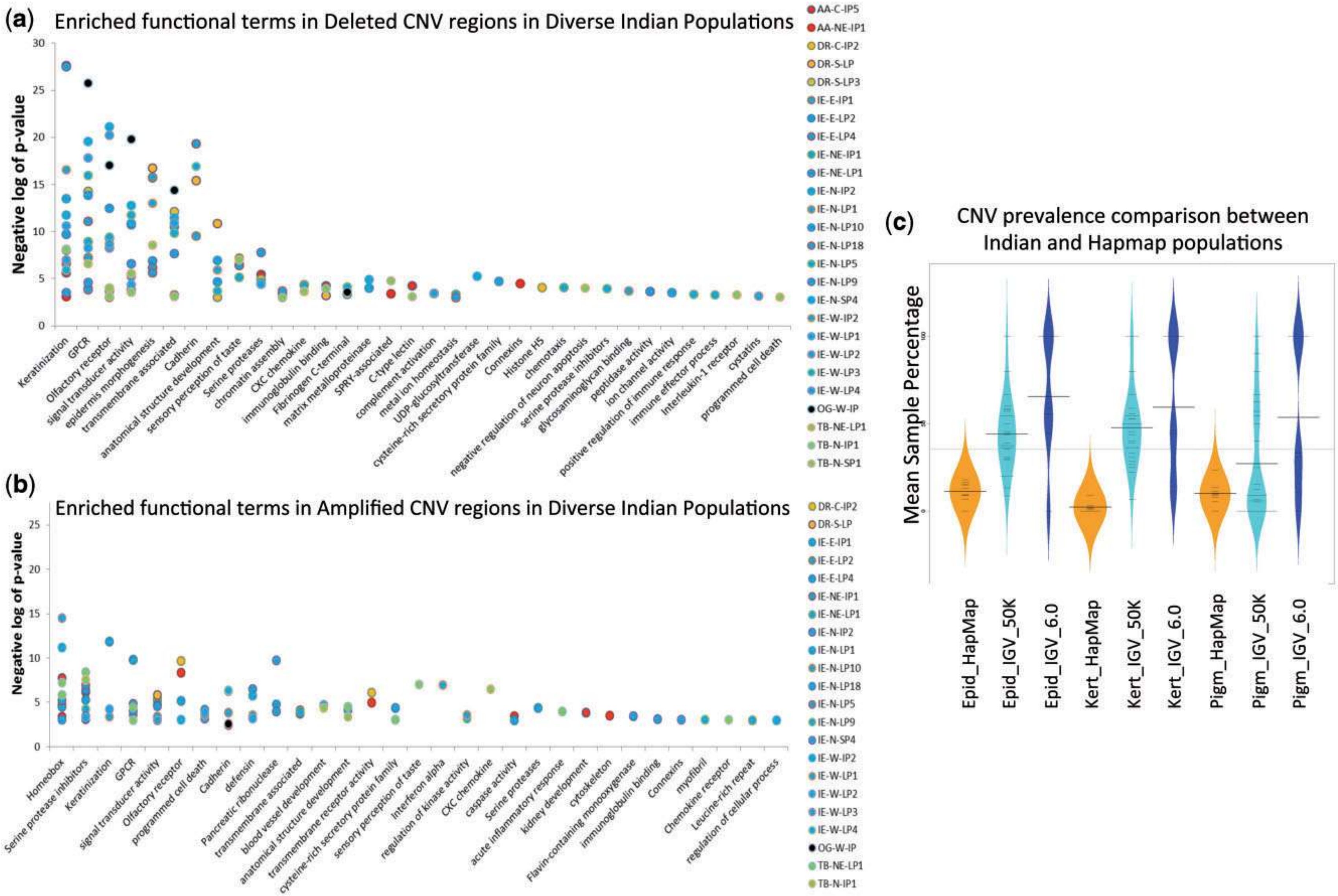


Fig. 1. Diversity of keratinization and epidermal differentiation in Indian and HapMap populations. Functional enrichment of genes containing CNVs in 26 diverse Indian populations is shown for deletion (*a*) and amplification (*b*) data set. *x* axis depicts the FAC terms enriched significantly in different populations is represented as scatter plot. *y* axis represents the negative log of the Fisher's exact *P* values of enrichment obtained from FAC in DAVID. Only populations where enrichment scores were obtained have been plotted. Populations belonging to different ethnicities are color coded differently. Processes related to keratinization/epidermis morphogenesis emerge as one of the most enriched FAC terms across many populations. (*c*) The bean plot depicts the comparison of CNV prevalence in sample percentage between Indian and HapMap populations. CNVs in genes of three classes namely keratinization, pigmentation, and epidermal differentiation are represented. Indians compared with HapMap populations have higher CNVs. CNVs identified from both Affymetrix 50 K and 6.0 arrays in the Indian populations are shown and compared against HapMap CNVs genotyped on Affymetrix 6.0 array.

genotyped a subset of our samples on Affymetrix 6.0 in seven populations, the same array used for genotyping the HapMap populations. Surprisingly, the overall sample proportion is higher among the Indian compared with individual HapMap populations across all the three categories: Epidermal differentiation, keratinization, and pigmentation (fig. 1*c*). Nearly, all the CNVs reported in HapMap were represented in the Indian population. This amounted to 30% of the total CNVs present in the Indian population owing to higher prevalence of CNVs. The details of the IGV and HapMap populations are provided in supplementary table S1, Supplementary Material online. The fidelity of this comparison is substantiated by the fact that it was carried out on the same genotyping platform Affymetrix 6.0 (supplementary fig. S3, Supplementary Material online). It is quite possible that diversity in the geoclimatic conditions along with the ethnic diversity in India have contributed to the CNV differences in selected gene sets with respect to that of the world populations.

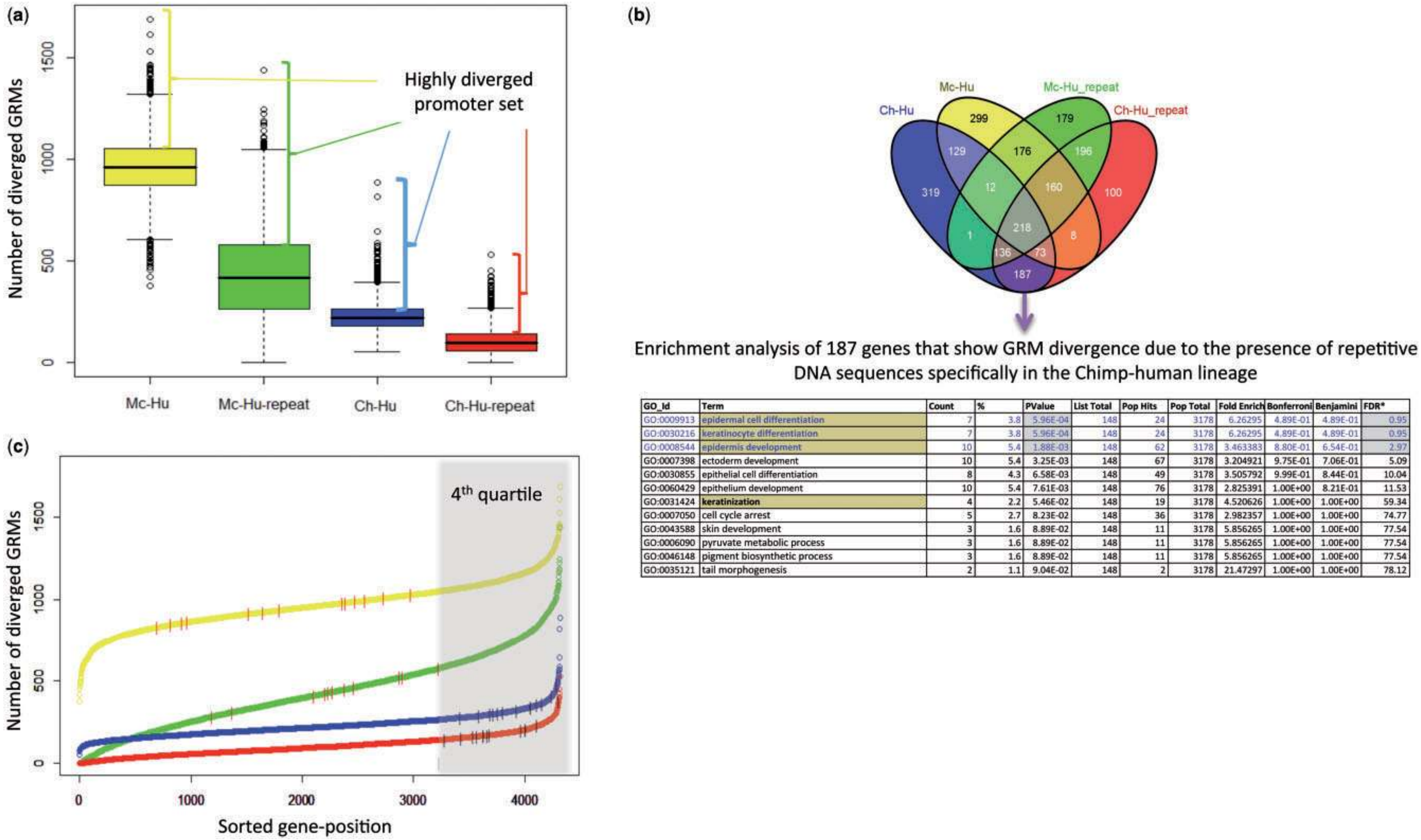
Divergence of Keratinization in the Chimpanzee–Human Lineage

Loss of hair and gain of epidermal pigmentation mark the transition from chimpanzee to humans and provide a suitable opportunity to address the role of selection in skin-related processes. The coding sequence, the entire 5-kb promoter-proximal region, and the same after removing the nonrepetitive (unique) sequences were compared in 4,316 genes for which orthologs could be identified across human, chimpanzee, and macaque. The details of this analysis are provided in Materials and Methods section (also see supplementary text S1, Supplementary Material online). Divergence in the coding region was computed using substitution-based Jukes–Cantor (JC)-divergence measure (Chen et al. 2001). This measure, however, does not essentially capture the diversity of noncoding regions due to the presence of indels (insertions and deletions). Hence, we studied divergence in the gene regulatory motifs (GRMs) defined by the binding sites of transcription factors in the promoter regions (identified using TRANSFAC [http://www.gene-regulation.com/pub/databases.html, last accessed December 20, 2014]). We compared the orthologous pairs of genes between human and chimpanzee, taking macaque as an outgroup, to essentially capture divergence events specific to the chimpanzee–human lineage. The contribution of repeats to this divergence was also studied. Enrichment analysis was carried out on the genes (after applying an FDR cutoff <5%), which were present in the fourth quartile of divergence in all the comparisons (supplementary table S2, Supplementary Material online).

On the basis of the observation of divergence distribution pattern (see supplementary text S1, Supplementary Material online), we used a box plot to divide GRM-divergence values in four quartiles, the first being lowest and the fourth being highest divergence values between the two species (fig. 2*a*). A total of 269 genes that were highly diverged because of repeats along with overall high divergence at their 5-kb promoter-proximal region between human and chimpanzee were identified (fig. 2*b*). FAC was carried out for 187 genes that harbored highly diverged GRMs between human and chimpanzee and did not feature in the macaque–human comparison (fig. 2*c*). Similar to our observations in the CNV analysis, we could observe high divergence in olfaction-related process in both the coding and the promoter regions, corroborating previous observations (Gilad et al. 2005; Hasin et al. 2008). Interestingly, keratinization and epidermal differentiation processes were also found to be enriched in both the coding and the regulatory region. This remained significant (P value < 0.05) even after applying FDR correction (cutoff < 0.05) in the coding regions (supplementary fig. S4, Supplementary Material online). In sharp contrast, pigmentation process was not enriched with similar scores. Surprisingly, keratinization does not feature with significant scores in the human–chimpanzee comparison of the promoter GRMs (P value < 0.02, FDR 31.6%) but appears as one of the most enriched processes upon inclusion of genomic repeats (P value < 0.0005, FDR 0.95%) (supplementary fig. S5, Supplementary Material online). This suggests a functional involvement of repeats in generating keratinization diversity. It is interesting that sensory perception despite being a significant process in the pair-wise comparison (P value < 0.0018, FDR 3.3%) is not contributed by repeats when macaque is used as the outgroup. Our study suggests that keratinization process is under selection in the human lineage, and different polymorphic DNA elements contribute to the diversity in the coding and the regulatory regions of these genes.

Mapping all keratinization genes revealed 57 highly divergent promoters in the chimpanzee–human lineage (due to the differential presence of GRMs within repetitive DNA). Ten of these were exclusively represented in the fourth quartile and were unique to the chimpanzee–human lineage (fig. 2*b*). Notably, these diverged genes were distributed across several loci in the genome including the previously reported 1q21 position (The Chimpanzee Sequencing and Analysis Consortium) suggesting that keratinization process as a whole is under selection (supplementary fig. S6, Supplementary Material online).

As CpG islands are known to have a relatively higher rate of mutation (Subramanian and Kumar 2006; Parker-Katiraee et al. 2007; Farcas et al. 2009), we wondered whether the observed GRM divergence was because of the differential presence of CpG methylation sites in the repetitive and unique DNA. Surprisingly, CpG methylation sites within GRMs in the repetitive region was found to be much lower compared with that of CpG sites within GRMs across the entire promoter (repetitive + unique regions). Upon removal of CpG stretches from our analyses, there was a marginal reduction in the number of genes present in the fourth quartile, and the enrichment analysis output was comparable to the set without CpG removal (supplementary table S2, Supplementary Material online). Hence, we conclude that the GRM divergence in the human–chimpanzee comparison is majorly influenced by repeats. Our observations thus suggest that a concerted diversification in keratinization process mediated by GRMs in repetitive DNA could contribute to



Enrichment analysis of 187 genes that show GRM divergence due to the presence of repetitive DNA sequences specifically in the Chimp-human lineage

GO_Id	Term	Count	%	PValue	List Total	Pop Hits	Pop Total	Fold Enrich	Bonferroni	Benjamini	FDR*
GO:0009913	epidermal cell differentiation	7	3.8	5.96E-04	148	24	3178	6.26295	4.89E-01	4.89E-01	0.95
GO:0030216	keratinocyte differentiation	7	3.8	5.96E-04	148	24	3178	6.26295	4.89E-01	4.89E-01	0.95
GO:0008544	epidermis development	10	5.4	1.88E-03	148	62	3178	3.463383	8.80E-01	6.54E-01	2.97
GO:0007398	ectoderm development	10	5.4	3.25E-03	148	67	3178	3.204921	9.75E-01	7.06E-01	5.09
GO:0030855	epithelial cell differentiation	8	4.3	6.58E-03	148	49	3178	3.505792	9.99E-01	8.44E-01	10.04
GO:0060429	epithelium development	10	5.4	7.61E-03	148	76	3178	2.825391	1.00E+00	8.21E-01	11.53
GO:0031424	**keratinization**	4	2.2	5.46E-02	148	19	3178	4.520626	1.00E+00	1.00E+00	59.34
GO:0007050	cell cycle arrest	5	2.7	8.23E-02	148	36	3178	2.982357	1.00E+00	1.00E+00	74.77
GO:0043588	skin development	3	1.6	8.89E-02	148	11	3178	5.856265	1.00E+00	1.00E+00	77.54
GO:0006090	pyruvate metabolic process	3	1.6	8.89E-02	148	11	3178	5.856265	1.00E+00	1.00E+00	77.54
GO:0046148	pigment biosynthetic process	3	1.6	8.89E-02	148	11	3178	5.856265	1.00E+00	1.00E+00	77.54
GO:0035121	tail morphogenesis	2	1.1	9.04E-02	148	2	3178	21.47297	1.00E+00	1.00E+00	78.12

FIG. 2. Genome-wide promoter and coding region divergence analysis between chimpanzee and human with macaque as an out-group. (*a*) A total of 4,316 orthologous 5-kb promoter-proximal regions were compared for GRM divergences between human and chimpanzee, taking macaque as an outgroup. Distribution of GRMs in human chimpanzee and human macaque promoters in 5-kb promoter-proximal regions and those mapping to only repeat regions are represented in box plots. Divergence contributed exclusively by repeats in this region was compared between chimpanzee–human and macaque–human (represented by the red and green boxes, respectively). The fourth quartile of each of these comparison sets of GRM divergence has been considered as the highly diverged set of promoters. The *y* axis represents the number of diverged GRMs. (*b*) The Venn diagram represents the overlap of the promoters from the fourth quartile of each comparison. The colors of each set represent the corresponding comparison in the box plot. We could identify 187 highly diverged set of promoters, which were unique to human chimpanzee lineage and from the repeat region. Table represents the gene-ontology analysis of these 187 genes for enrichment analysis. Interestingly, genes associated with development of epidermal and keratinization events were found to be enriched. (*c*) Comparative visualization of these enriched promoters from keratinization and epidermal differentiation process has been overlaid on each pair-wise comparison of GRM divergence where the promoters have been arranged based on their order of increasing divergence. The red vertical bars represent the relative placement of these highly diverged promoters in each of the comparisons.

phenotypic differences between chimpanzee and human skin.

Skin Keratinization Is under Environmental Selection

Divergence in several keratinization and epidermal differentiation genes that affect skin phenotypes raises the question whether these observed variations are a consequence of drift or natural selection. As skin is the organ most exposed to external perturbations, keratinization process could be under selection from the environmental factors. Earlier studies (Hancock et al. 2011) have elegantly demonstrated the genome-wide association of SNPs represented in the dbCLINE data set (http://genapps2.uchicago.edu:8081/dbCLINE/main.jsp, last accessed March 25, 2014.) with climatic variables across similar ecoclines from diverse continental populations. These had been identified after regressing out several confounding factors. Specifically in the context of skin pigmentation, SNPs in several genomic loci were reported with high ranks. Thus, this was an ideal data set to examine the effect of climate variables on keratinization genes. We have analyzed the association of SNPs in our curated set of four processes (comprising keratinization, epidermal differentiation, pigmentation, and housekeeping genes) with 15 environmental parameters reported in dbCLINE. Details of the gene sets for this analysis are provided in supplementary tables S3 and S4, Supplementary Material online.

To identify a pathway-specific signal and prevent few high scoring SNPs from skewing the analysis, we considered all those genes which contain at least one SNP with transformed rank (TR) < 0.05 as reported by Hancock et al. (2011) (supplementary table S5, Supplementary Material online). In polygenic traits like pigmentation and keratinization, several independent loci are likely to contribute to the selection of phenotypes. As the association of human skin pigmentation with climatic variables is well appreciated, we first validated this hypothesis in this set of genes. Further, the earlier study also indicated that the environmental parameters are not completely independent of each other and hence several SNPs associate with more than one climate or ecoregion variable simultaneously. Manhattan plot representing the association of SNPs in each category with various environmental parameters reveals significant SNPs spanning the entire genome for each of the three processes (fig. 3*a*). This included the two pigmentation SNPs in *SLC45A2* and *OCA2* that were earlier reported to associate strongly with solar radiation (Hancock et al. 2011). Additionally, several independent loci were found to contain multiple significant SNPs above the threshold (-log-transformed TF 0.05) across several pigmentation genes and correlated climatic parameters. This provides an indication that climate-mediated selection may operate across the entire process.

Having verified this, we set out to address the association of environmental variables with the other three processes. We observed that pigmentation demonstrated maximal association, and housekeeping genes were poorly associated with all the variables. The spider plot (fig. 3*b*) represents the proportion of genes associated with each of these variables. Both epidermal differentiation and keratinization process associated with all the parameters (supplementary fig. S7, Supplementary Material online). Maximal association could be observed with winter precipitation rate (28% of keratinization and 33% of epidermal differentiation genes) and winter relative humidity (28% and 31%, respectively). It is noteworthy that despite lower representation per gene, epidermal differentiation demonstrates substantial association with several parameters (supplementary table S6, Supplementary Material online). Interestingly, the keratinization genes showed a lower percentage of association with all variables compared with epidermal differentiation. Figure 3*a* compares these processes with constituent SNPs and the enhanced association of SNPs in epidermal differentiation is visible in a number of genes spread throughout the genome. Strikingly, the epidermal differentiation had more SNPs with moderately high scores than pigmentation. In contrast, several high-ranking SNPs contributed by few loci such as *SLC45A2* and *OCA2* for pigmentation. As epidermal differentiation governs the process of keratinization, it is likely that the climate-mediated selection operative in this process would also indirectly affect keratinization. We hypothesize that keratinization could be under selection due to environmental humidity. Our observations strongly support the determining role of environment in dictating the selection of skin keratinization phenotype. It is tempting to speculate that the environmental humidity would be the determining factor central to the associations. As many of the clinal variables are interconnected and are also related to solar process, it is difficult to disambiguate the specific role of humidity from other factors in dictating the selection and needs to be carefully addressed in future studies.

Keratinization of Skin across Vertebrates Is under Accelerated Evolution

To further dissect the role of environmental pressure in dictating genome evolution, we studied keratinization across nine different vertebrate species. These reside in diverse environments and have inherently different exposure to the selection pressure from the immediate environment. For the phylogenetic analyses, we have considered nine species taking at least one representative from each of the five major classes of vertebrates. Besides the three primates (human, chimpanzee, and rhesus monkey), we have also included house mouse (since rodents have branched off from an ancestral line, which is phylogenetically close to that of primates) and bottlenose dolphin (because we wanted to check if secondary adaptation to an aquatic mode of life has altered genomic signatures compared with that of other terrestrial mammals). In direct contrast with the other organisms, bottlenose dolphin is unique because they do not have to deal with TEWL, and hence, we anticipated changes in the genomic regions that affect the process of keratinization. Apart from these organisms, we have also included zebrafish, African tree frog, Anole lizard, and chicken. Notwithstanding the discordance in their absolute numbers, we have taken all the genes from each category to avoid introducing any bias (89 genes for keratinization, 208 for

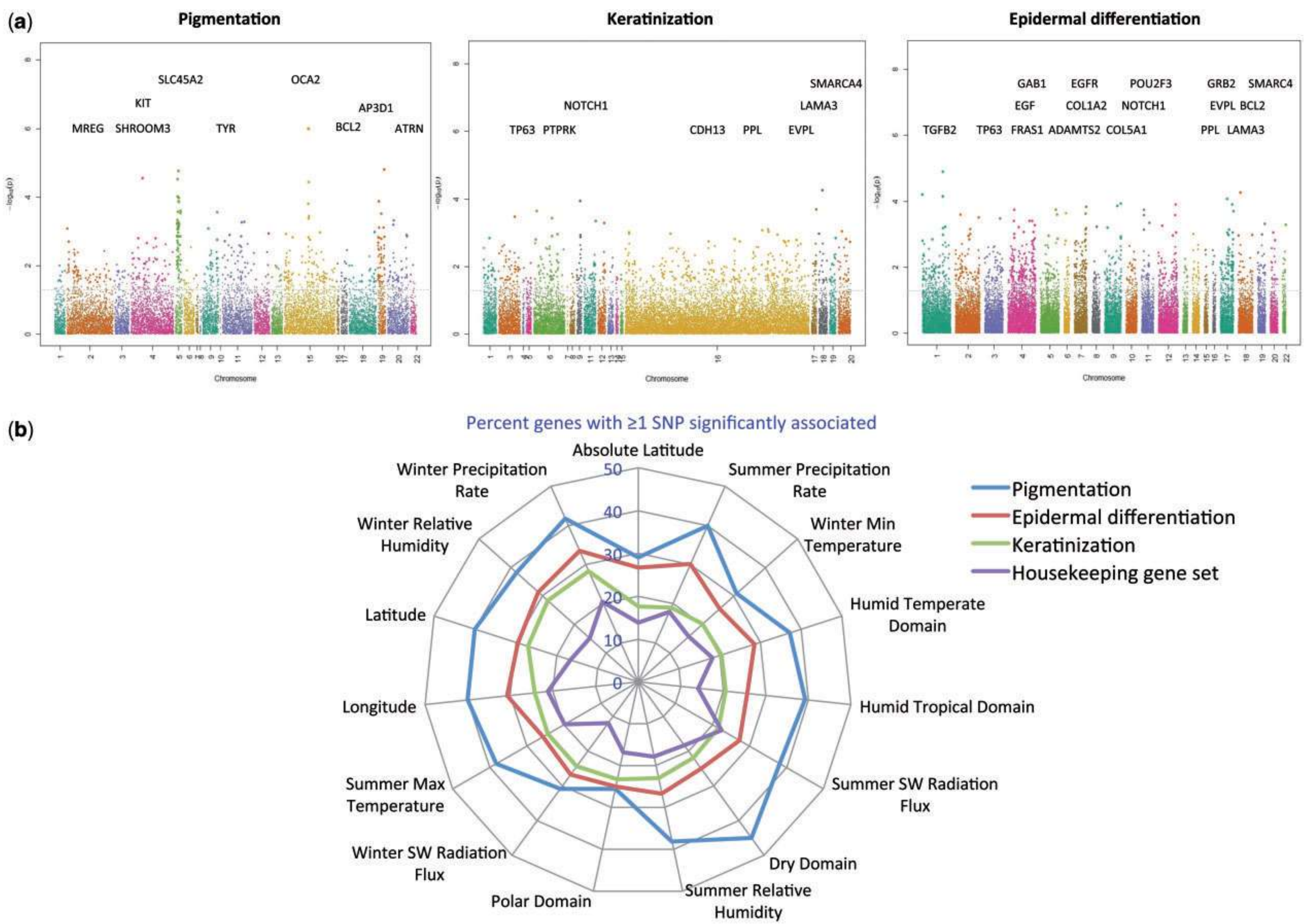


Fig. 3. Pattern of association of the genes related to skin physiology with climate variables. (*a*) The Manhattan plots shows all the SNPs associated with environmental variables in dbCLINE database (each SNP's association with 15 independent variables plotted as one plot). The three plots are for three gene classes of keratinization, pigmentation, and epidermal differentiation. The *y* axis is negative log of TRs (empirical *P* value) and *x* axis is all autosomes. The horizontal cutoff lines represents significance threshold of 0.05. Genes with more than 30 SNPs are labeled. (*b*) TR statistics from dbCLINE database was used for each of the category of genes. The radial axis depicts the percentage of genes, which show significant association with the environmental variables (labeled on the periphery). Green line represents keratinization, red line is for epidermal differentiation, light blue is for pigmentation, and violet line represents a set of housekeeping genes used as control.

epidermal differentiation, and 68 for pigmentation, supplementary table S3, Supplementary Material online). Orthologous mRNA and protein sequences of the four classes of genes (keratinization, epidermal differentiation, pigmentation, and housekeeping genes) were retrieved. Protein sequence diversity of the orthologs is represented as a heatmap (fig. 4). For some entries, orthologs were not found, and for some others, the genes were highly divergent at the sequence level and could not be traced by protein Basic Local Alignment Search Tool (BLAST) analysis (represented in white). Some of the entries were eliminated because they did not meet the quality cutoff criteria (query cover $< 30\%$) (represented in blue). Pigmentation proteins emerge as much more conserved in their sequence, and all the proteins under study were found in all nine organisms. Surprisingly, the level of diversity in the sequence was comparable to that of the housekeeping proteins, which show the highest conservation across the groups. However, genes involved in the process of keratinization are highly diverged across the vertebrate classes, especially among the nonmammals. The entire *LCE* and *SPRR* clusters are missing in most of the nonmammalian vertebrates. Epidermal differentiation-related proteins exhibit a great deal of variation across organisms, yet number of unidentifiable orthologs is less. It is noteworthy that most proteins from epidermal differentiation-related processes, pigmentation, and housekeeping show conservation across primates, whereas keratinization proteins are divergent with several proteins (like *LOR*, *IVL*, *LCE3E*, *LCE5A*, etc.) having marked changes.

Toward quantitating the diversity for assessing the directionality of gene change, Ka/Ks substitution ratio was taken as the measure of evolution. This parameter is scaled to neutral divergence and provides a convenient scale for comparing the average and distribution of the diversity index across the four processes. Ka/Ks indices were calculated for each of the nine species with respect to human, using the combined information of mRNA and protein sequences (elaborated in Materials and Methods section). Process-specific index was calculated by taking the average of the indices of the individual genes under that process. Average value of Ka/Ks for each comparison across four pathways is represented as a scatter plot (fig. 5*a*). Strikingly, the values are lower for the housekeeping and pigmentation processes when compared with keratinization and epidermal differentiation, clearly indicating that the later processes are under selection.

Only statistically significant Ka/Ks ratio values (*P* value $<=$ 0.05; Fisher's exact test) were selected for further analyses. Distribution of Ka/Ks values was found to be between 0.001 to 2.495 with an average value of 0.143 and a standard deviation of 0.179 for all four categories considered together. To account for the contribution of CpG islands to this, we reanalyzed the Ka/Ks scores for the top hits after masking the CpG sites. Although most of the genes did not contain CpG sites within coding regions, for those that did, the reduction was nominal (supplementary table S4, Supplementary Material online). To identify whether the Ka/Ks distributions were significantly different for each of the skin-related biological processes in comparison to the housekeeping set of genes for the region covered within the box ($0.4 \geq \text{Ka/Ks} < 1.0$) (fig. 5*b*), we carried out nonparametric two-sample Kolmogorov–Smirnov test. We could observe that only keratinization genes were distributed significantly differently with a *P*-value score of 0.002. It is noteworthy that the bottlenose dolphin features at the top for keratinization, with the highest Ka/Ks value of 0.26.

Discussion

Our study suggests a population-wide diversity in keratinization process and provides a compelling evidence for selection by environmental forces that could affect TEWL. The level of variations in keratinization genes mediated by CNVs is rather striking, as the diversity is comparable to processes such as olfaction. Albeit with varying frequencies across populations, there is a substantial overlap of genes under CNV in keratinization and epidermal differentiation across Indian and HapMap populations. This indicates that the mechanism of diversity in skin-related functions transcends populations. Although the human population diversity could have resulted from a drift or neutral selection, focused analysis of skin function genes in dbCLINE data revealed association of multiple genes of keratinization linked with a large number of geoclimatic variables, indicating a role of selection (table 1 and supplementary tables S6 and S7, Supplementary Material online). This was further substantiated by the observation of this being one of the most prominent processes that was highly diverged between chimpanzee and human, both in the promoter and coding regions. Comparative analysis of nine vertebrates revealed higher accelerated evolution of keratinization and epidermal differentiation genes compared with housekeeping genes. All these observations establish that keratinization process is under selection in relation to humidity.

Several interesting observations emerged out of our analysis on comparing the SNP frequencies of keratinization and epidermal differentiation genes with the environmental parameters. One hundred forty-nine out of 208 of epidermal differentiation and 63 out of 89 keratinization genes show association with one or more climatic variables. Among them, around 75% of keratinization genes and around 80% of epidermal differentiation genes are associated with humidity-related parameters. The entire process could be fine tuned by the interplay of selection operational at different levels. Genes with CNV polymorphism that associate with the environmental variables include a set of 14 genes, and all of them are involved in the core process of keratinization. Envoplakin (*EVPL*) and Pariplakin (*PPL*) link the cornified envelope to desmosomes and intermediate filaments and are associated with ten environmental variables (table 1). The key process of crosslinking during the terminal differentiation is mediated by transglutaminases 1 and 3, and they together associate with ten variables. Such association across several steps involved in keratinization strongly suggests a subtle but widespread selection signature for this skin process in different environments. Keratinization is a structural process, and the genes are highly repetitive, therefore it is possible that CNV polymorphisms are prevalent in this class of genes. Additionally, these genes are clustered at certain loci (like

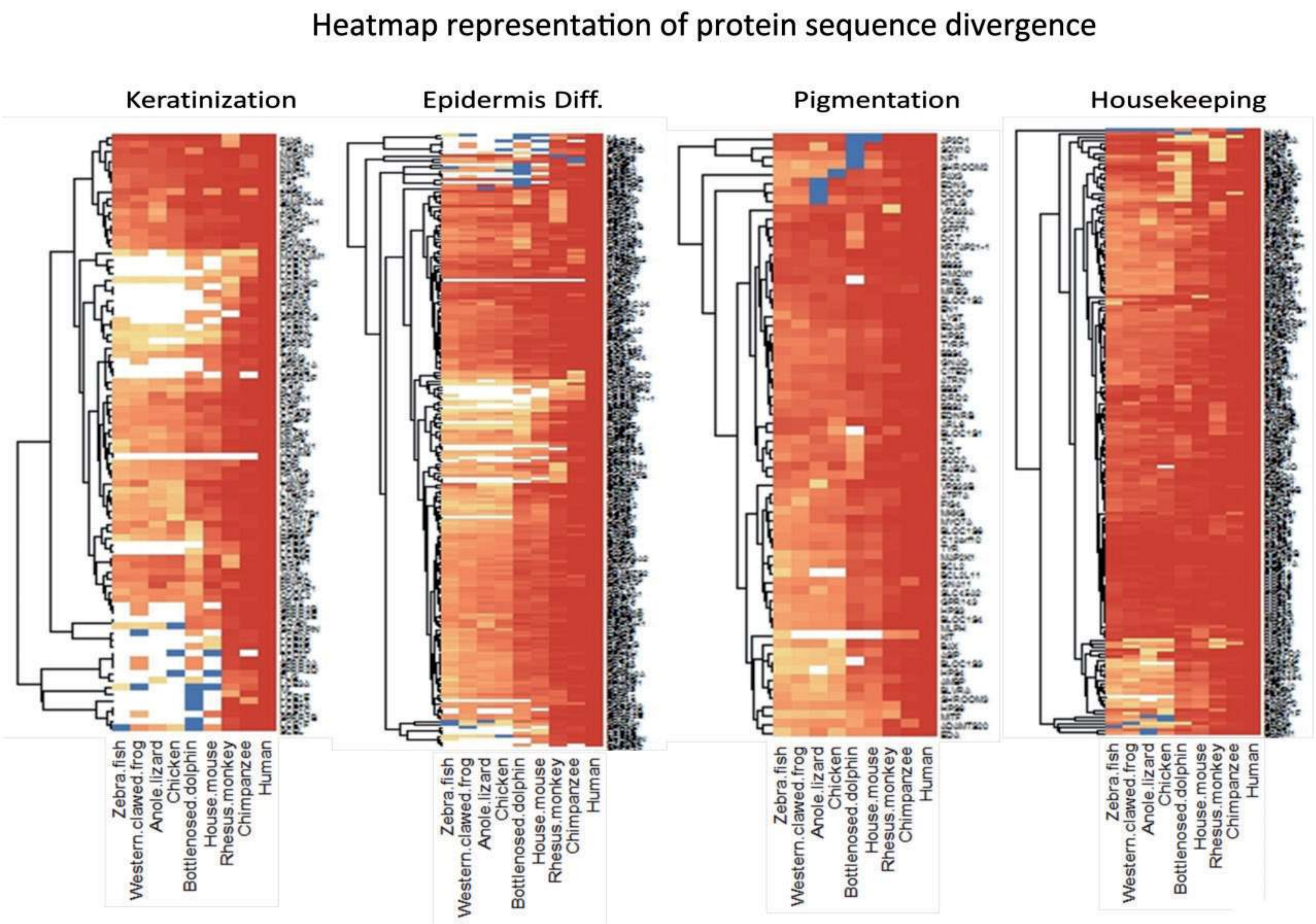


Fig. 4. Heatmap representation of protein BLAST showing sequence divergence. The figure illustrates the extent of protein sequence divergence across three different processes involved in skin physiology compared with a control set of housekeeping proteins. The nine species that have been considered substantially differ in the environmental challenges they face. As is evident, the protein sequences in keratinization show maximum divergence compared with other processes in the skin and the control set of housekeeping proteins, which appear to be the most conserved. The gradient of shades from orange to red indicates the progressively increasing percentage identity of the protein among the different representative organisms. White denotes entries that did not feature in our protein BLAST result although they were reported to exist at gene or/and protein level or were removed due to *E*-value cutoff ($<10^{-4}$). Blue denotes entries that did not pass the cutoff filter (for query cover $\geq 30\%$).

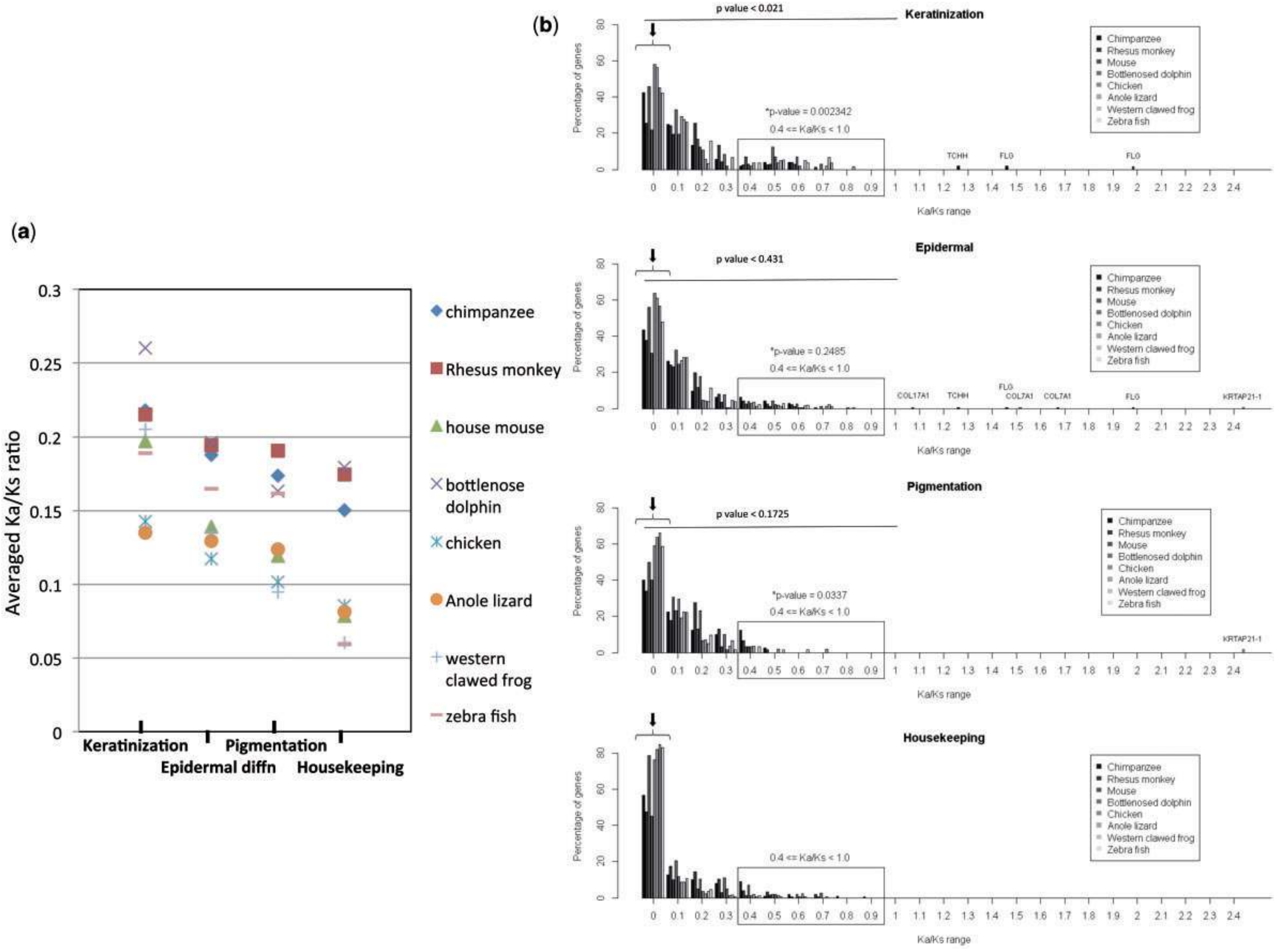


Fig. 5. Divergence of keratinization proteins across major vertebrate classes. (*a*) Average Ka/Ks ratio for all the four processes computed for eight species against the human reference is represented in the scatter plot. For keratinization, bottlenose dolphin has a higher ratio compared with all other species, which may reflect the fact that it has got secondarily adapted to an aquatic mode of life. (*b*) Distribution of Ka/Ks ratio in eight different comparisons. The *x* axis represents the values of Ka/Ks ratio and the *y* axis denotes the number of genes. The region where Ka/Ks values range between 0.4 and 0.9 is boxed for each process, and this region is compared for all three skin-related processes against the housekeeping set. Keratinization is significantly different from housekeeping genes in the selected window (P value < 0.002), whereas epidermal differentiation was not (P value < 0.2), pigmentation was significantly different compared with housekeeping genes; however, the distribution was skewed toward lower Ka/Ks values. Outliers that have > 1, Ka/Ks ratio have been indicated in the plots. The number of proteins that are under highly purifying selective pressure (Ka/Ks = 0) is most for housekeeping and least for keratinization (highlighted by downward pointing arrows).

Table 1. Genes with Significant Associations to At Least Ten Climate Variables.

Sl. No	Gene Class	Gene Name	Chr.	No. of Significantly Associated SNPs	No. of Climate Variables Associated	Name of the Climate Variables
1	Keratinization	CDH13	16	270	15	P, HTemp, D, HTr, AbsLt, Ts, Tw, Lt, Ln, Ps, Pw, RHs, RHw, RFs, RFw
2	Keratinization	PTPRK	6	30	15	P, HTemp, D, HTr, AbsLt, Ts, Tw, Lt, Ln, Ps, Pw, RHs, RHw, RFs, RFw
3	Keratinization epidermis differentiation	TP63	3	45	15	P, HTemp, D, HTr, AbsLt, Ts, Tw, Lt, Ln, Ps, Pw, RHs, RHw, RFs, RFw
4	Keratinization, epidermis differentiation	POU2F3	11	17	13	P, HTemp, D, HTr, AbsLt, Ts, Tw, Lt, Ln, RHs, RHw, RFs, RFw
5	Keratinization, epidermis differentiation	NOTCH1	9	7	11	P, HTemp, D, Ts, Lt, Ps, Pw, RHs, RHw, RFs, RFw
6	Keratinization, epidermis differentiation	EVPL	17	5	10	P, HTr, AbsLt, Tw, Lt, Ps, Pw, RHs, RFs, RFw
7	Keratinization, epidermis differentiation	LAMA3	18	19	10	HTemp, D, HTr, AbsLt, Ts, Lt, Ln, RHs, RHw, RFs
8	Keratinization, epidermis differentiation	PPL	16	9	10	P, HTemp, Tw, Lt, Ln, Ps, Pw, RHs, RFs, RFw
9	Epidermis differentiation	EGFR	7	41	15	P, HTemp, D, HTr, AbsLt, Ts, Tw, Lt, Ln, Ps, Pw, RHs, RHw, RFs, RFw
10	Epidermis differentiation	TP63	3	45	15	P, HTemp, D, HTr, AbsLt, Ts, Tw, Lt, Ln, Ps, Pw, RHs, RHw, RFs, RFw
11	Epidermis differentiation	ADAMTS2	5	24	14	P, D, HTr, AbsLt, Ts, Tw, Lt, Ln, Ps, Pw, RHs, RHw, RFs, RFw
12	Epidermis differentiation	FRAS1	4	75	14	HTemp, D, HTr, AbsLt, Ts, Tw, Lt, Ln, Ps, Pw, RHs, RHw, RFs, RFw
13	Epidermis differentiation	PLCE1	10	23	14	P, HTemp, HTr, AbsLt, Ts, Tw, Lt, Ln, Ps, Pw, RHs, RHw, RFs, RFw
14	Epidermis differentiation	TGFB2	1	15	14	P, HTemp, D, HTr, AbsLt, Ts, Tw, Lt, Ps, Pw, RHs, RHw, RFs, RFw
15	Epidermis differentiation	GAB1	4	17	13	D, HTr, AbsLt, Ts, Tw, Lt, Ps, Pw, RHs, RHw, RFs, RFw
16	Epidermis differentiation	NTF3	12	18	13	HTemp, D, HTr, AbsLt, Ts, Tw, Lt, Ln, Ps, Pw, RHs, RHw, RFs
17	Epidermis differentiation	POU2F3	11	17	13	P, HTemp, D, HTr, AbsLt, Ts, Tw, Lt, Ln, RHs, RHw, RFs, RFw
18	Epidermis differentiation	PPARA	22	17	13	P, HTemp, D, HTr, AbsLt, Ts, Tw, Ln, Ps, Pw, RHs, RHw, RFs
19	Epidermis differentiation	BCL2	18	25	12	HTemp, D, HTr, AbsLt, Ts, Tw, Lt, Ln, Pw, RHs, RFs, RFw
20	Epidermis differentiation	COL5A1	9	23	12	P, HTr, AbsLt, Ts, Lt, Ln, Ps, Pw, RHs, RHw, RFs, RFw
21	Epidermis differentiation	EGF	4	13	12	P, HTemp, D, HTr, AbsLt, Ts, Lt, Ln, RHs, RHw, RFs, RFw
22	Epidermis differentiation	COL5A3	19	8	11	P, HTemp, D, HTr, AbsLt, Tw, Lt, Ps, Pw, RFs, RFw
23	Epidermis differentiation	NOTCH1	9	7	11	P, HTemp, D, Ts, Lt, Ps, Pw, RHs, RHw, RFs, RFw
24	Epidermis differentiation	SPINK5	5	12	11	P, HTemp, D, Ts, Tw, Lt, Ln, RHs, RHw, RFs, RFw
25	Epidermis differentiation	COL1A2	7	17	10	P, HTemp, D, Ts, Lt, Ln, RHs, RHw, RFs, RFw
26	Epidermis differentiation	DGKD	2	12	10	P, HTemp, D, HTr, Ln, Ps, Pw, RHs, RHw, RFs
27	Epidermis differentiation	EVPL	17	5	10	P, HTr, AbsLt, Tw, Lt, Ps, Pw, RHs, RFs, RFw
28	Epidermis differentiation	GRB2	17	11	10	HTr, AbsLt, Ts, Tw, Lt, Ps, Pw, RHs, RFs, RFw

(continued)

Table 1. Continued

Sl. No	Gene Class	Gene Name	Chr.	No. of Significantly Associated SNPs	No. of Climate Variables Associated	Name of the Climate Variables
29	Epidermis differentiation	LAMA3	18	19	10	HTemp, D, HTr, AbsLt, Ts, Lt, Ln, RHs, RHw, RFs
30	Epidermis differentiation	PPL	16	9	10	P, HTemp, Tw, Lt, Ln, Ps, Pw, RHs, RFs, RFw
31	Epidermis differentiation	WNT7A	3	11	10	HTemp, D, HTr, Ts, Lt, Ln, Ps, Pw, RHw, RFs
32	Pigmentation	OCA2	15	51	15	P, HTemp, D, HTr, AbsLt, Ts, Tw, Lt, Ln, Ps, Pw, RHs, RHw, RFs, RFw
33	Pigmentation	SHROOM3	4	38	15	P, HTemp, D, HTr, AbsLt, Ts, Tw, Lt, Ln, Ps, Pw, RHs, RHw, RFs, RFw
34	Pigmentation	ATRN	20	31	12	P, HTemp, D, AbsLt, Ts, Tw, Lt, Ln, Pw, RHs, RHw, RFw
35	Pigmentation	BCL2	18	25	12	HTemp, D, HTr, AbsLt, Ts, Tw, Lt, Ln, Pw, RHs, RFs, RFw
36	Pigmentation	RAB27A	15	11	12	HTemp, D, HTr, AbsLt, Ts, Tw, Lt, Ln, Ps, RHs, RFs, RFw
37	Pigmentation	SLC45A2	5	12	12	HTemp, D, HTr, AbsLt, Ts, Lt, Ps, Pw, RHs, RHw, RFs, RFw
38	Pigmentation	AP3D1	19	11	11	P, HTemp, HTr, AbsLt, Ts, Tw, Lt, Ln, Ps, RHs, RFw
39	Pigmentation	MYO7A	11	15	11	P, D, HTr, Ts, Lt, Ln, Ps, Pw, RHs, RHw, RFs
40	Pigmentation	TYR	11	15	11	P, HTemp, D, HTr,AbsLt, Tw, Lt, Ps, RHs, RFs, RFw
41	Pigmentation	AMBP	9	6	10	Ts, Tw, Lt, Ln, Ps, Pw, RHs, RHw, RFs, RFw
42	Pigmentation	GNA11	19	6	10	P, HTr, Ts, Tw, Lt, Ln, Ps, Pw, RHw, RFw
43	Pigmentation	PAX3	2	14	10	P, HTemp, AbsLt, Ts, Tw, Lt, Ps, Pw, RFs, RFw

NOTE.—Genes with significant association with at least ten climate variables are shown along with the associated variables coded as follows: AbsLt, absolute latitude; Ts, Ave_max_T (summer); Tw, Ave_min_T (winter); P, Bailey100 (polar domain); HTemp, Bailey200 (humid temperate domain); D, Bailey300 (dry domain); HTr, Bailey400 (humid tropical domain); Lt, latitude; Ln, longitude; Ps, Precip_rate (summer); Pw, Precip_rate (winter); RHs, Rel_hum (summer); RHw, Rel_hum (winter); RFs, SW_rad_flux (summer); RFw, SW_rad_flux (winter).

the 1q21 epidermal differentiation complex), further enabling recombination-based generation of CNV polymorphisms. It is interesting that despite this clustering of genes, CNV polymorphisms are also contributed by discrete loci spread throughout the genome thereby highlighting the role of selection in mediating this diversity. Studies with filaggrin (*FLG*) gene support these observations, and the coding repeat polymorphisms in this gene is a strong indicator of population ancestry (Ginger et al. 2005; Nemoto-Hasebe et al. 2009; Brown et al. 2012).

During the course of our study, detailed admixture mapping of modern humans with Neanderthals was reported (Sankararaman et al. 2014; Vernot and Akey 2014). Intriguingly, the parts of the Neanderthal genome that show signatures of adaptive introgression feature a set of keratins (Vernot and Akey 2014). This has led to the hypothesis that Neanderthal alleles that affect skin keratinization may have helped modern humans reside in regions outside Africa (Sankararaman et al. 2014). Therefore, we compared the level of protein sequence divergence for a set of hair and follicular keratins against epidermal keratins in skin among human and other higher primates. The list of keratins was taken from a detailed study by Moll et al. (2008). Although the hair and follicular keratins were found to be highly conserved among the great apes, the epidermal keratins were relatively more diverged, at least at the protein sequence level (supplementary fig. S8 and table S8, Supplementary Material online). This observation further corroborates the role of selection in skin keratinization. Such variability may have very well provided primates with different fitness advantages to adapt to different environmental niches.

Recently analysis of the gorilla genome has indicated that for *EVPL*, the rate of evolution is the highest (Scally et al. 2012). It is plausible that *EVPL* and other keratinization genes might have evolved to confer a specific immunity to abrasions caused by knuckle-walking, a feature common in gorilla. In our analysis, higher divergence of epidermal keratins against more conserved hair and follicular keratins observed among great apes further supports these conjectures (supplementary fig. S8, Supplementary Material online). With the loss of body hair, the skin is likely to be more prone to abrasions, hence the need to evolve altered keratinization could be rationalized. These observations add further credence to this study and suggest that selection of keratinization could be operative at several levels in parallel. Further mechanistic basis of the observed diversity were obtained by comparative genome analysis. The human–chimpanzee divergence provided perfect system to compare near-identical genomes fixed by speciation, yet the skin phenotypes are distinct in the interaction with the environment. The adaptive changes in the keratinization correlate well with the loss of body hair in the hominin lineage, which could have posed a major physiological challenge with respect to TEWL and thermoregulation. Orchestrated genetic changes in keratinization genes provide a tangible solution to ameliorate this stress. This was also corroborated by the observation of accelerated evolution in genes (estimated by the Ka/Ks ratios) of the keratinization process compared with pigmentation and housekeeping genes in humans with respect to eight vertebrate species. These organisms differ in their interaction with immediate environment specifically the TEWL. Strikingly, the bottlenose dolphin showed remarkably higher Ka/Ks for keratinization process (0.26 compared with the average 0.19), whereas for all other three processes, the Ka/Ks values were comparable to the other species. Comparison of GRM divergence and accelerated evolution across human–chimpanzee pair (Ka/Ks > 0.4) revealed no overlap of genes, suggesting that the evolution in coding and regulatory regions are operative in different sets of genes.

Migration of humans across continents exposes them to varied conditions. The observed genetic diversity in keratinization process could confer differences in skin adaptation to changing environments in a manner similar to that of pigmentation. Although the importance of pigmentation in protecting the skin from UV radiations is well appreciated, the protection mediated by altered keratinization in this regard is not clear (Natarajan et al. 2014). It is likely that a thick stratum corneum acts as an effective UV barrier. Alternately, keratinization changes could affect the manifestation of pigmentation in the human skin. This possibility arises from the observation that skin pigmentation is due to the presence of melanosomes in keratinocytes. Hence, keratinization changes may affect the distribution of melanosomes and thereby modulate the appearance of skin color. Although testing this relationship is beyond the scope of this study, this aspect needs to be evaluated carefully both at the molecular and tissue level in future.

In human populations, seasonal variations in climatic conditions predispose individuals to develop skin disorders like psoriasis. It is interesting to note that the prevalence of psoriasis is more in the colder northern region including the polar belt and contrastingly low in the tropic (Chandran and Raychaudhuri 2010). We observe *LCE* cluster genes to be significantly associated with all winter-related climate variables, for example, temperature, precipitation rate, relative humidity, and short wave radiation flux. Similarly, trichohyalin (*TCHH*) is associated mostly with summer-related variables (temperature, relative humidity, and SW radiation flux). In several independent population-based studies, CNVs in three loci consistently associate with psoriasis (Bergboer et al. 2010; Huffmeier et al. 2010; Prans et al. 2013). Two of the dermatological disorders affecting skin keratinization, psoriasis, and atopic dermatitis have strong genetic links. Deletion CNV of the *LCE3C-LCE3B* locus strongly correlates with the incidence of psoriasis (Bergboer et al. 2010; Huffmeier et al. 2010). In atopic dermatitis, CNV in *FLG* gene affects the risk of atopic dermatitis in dose-dependent manner (Brown et al. 2012). Atopic dermatitis is prevalent in northern Europe, and its incidence is lower in Asia and south-east Asian populations (Naldi et al. 2009). Thus, differences in CNV prevalence in keratinization genes across populations suggest that phenotypes and disorders are likely to manifest differently in different ethnicity and geographical conditions. Decreased TEWL and increased skin hydration have been reported in the lesional psoriatic skin (Lee et al. 2012). At the molecular level exposure to simulated changes in environmental humidity of stratum corneum induces enhanced keratinocyte proliferation and inflammatory markers in mice models (Denda et al. 1998). The effect of environment may operate at two distinct levels in the disease manifestation. Although the direct effect would be transient and homeostatic in nature, the role of environment in selecting genetic variants would compound the manifestation and etiopathological investigation of these dermatological disorders. In our analysis, *LCE1D*, *LCE6A*, and *SCEL* featured across all four analyses, suggesting that these structural keratinization components are under immense selection (supplementary fig. S9 and table S7, Supplementary Material online). Our study also provides a list of candidates that could be important in understanding disease prevalence and manifestations linked to dermatopathological conditions.

Given the selection of skin pigmentation by environmental cues, it is not surprising that other aspects of skin functions are also under selection. However, substantiating evidence in this regard and the extent of genetic involvement have so far remained enigmatic. On the basis of our multipronged study, we propose that similar to the evolution of skin pigmentation (Jablonski and Chaplin 2000; Pickrell et al. 2009), genomic alterations in keratinization are an adaptive response and are under selection from the environment. It is likely that individual genotypes are selected for appropriate balance between opposing environmental pressures culminating in the diversity across individuals and populations. In conclusion, our study provides multiple lines of evidence of diversity in keratinization and provides possible mechanistic basis of selection operational in this phenotype. A better understanding of the possible gene-environment links and their downstream effects will have prognostic implications.

Materials and Methods

Design of Strategies for Analysis

To address the influence of environment in shaping the keratinization function of skin, we adopted a multipronged approach involving four independent analyses. The study was synthesized from serendipitous observations converging from two independent unbiased genome-wide analyses. This was further substantiated following two focused comparative analyses. An earlier study by our group on large CNVRs in 26 diverse Indian populations using a low density Affymetrix 50K array had revealed enrichment in a large number of processes that could contribute to phenotypic diversity.

Keratinization process was amongst the top enriched classes. We followed the lead to study and compare their distribution between Indian and HapMap populations and validated using a higher density array (Affymetrix 6.0). The second genome-wide approach was aimed at understanding the contribution from promoter regions in functional divergence between chimpanzee and human. Macaque was considered as an outgroup to trace the event that was specific to the chimpanzee–human lineage. A total of 4,316 orthologous promoters that could be mapped with confidence across all three organisms were considered. Keratinization was observed as one of the most enriched processes. Loss of body hair has been one of the prominent events during speciation in chimpanzee–human lineage. A major adaptive process to regulate TEWL could be mediated through changes in the keratinization process. This lead was further pursued in detail to dissect the functional elements that could contribute to this divergence. The contribution of repeats in divergence was addressed by comparing the divergence between sequences in the presence and absence of repetitive DNA identified using RepeatMasker algorithm (http://www.repeatmasker.org/, last accessed December 20, 2014). Also the divergence in the coding regions from the same genes was compared with that of their noncoding counterparts. We analyzed and curated a set of genes involved in skin-related processes, namely pigmentation, epidermal differentiation, and keratinization, to assess variations and signatures of selection throughout this study (supplementary table S3, Supplementary Material online). As keratinization and epidermal differentiation processes are related, several genes were considered under both categories based on annotation. A set of well-characterized housekeeping genes reported on two studies were taken (Eisenberg and Levanon 2003; Chang et al. 2011) (supplementary table S4, Supplementary Material online).

The two focused approaches involved exploring the role of climatic variables in keratinization and its evolution across different vertebrate species. For the first analysis, we compared the association of SNPs with environmental parameters among the four classes of genes in the dbCLINE database (http://genapps2.uchicago.edu:8081/dbCLINE/main.jsp, last accessed December 20, 2014). This database harbors information on SNPs that are significantly associated with climatic parameters across different continental populations from similar ecoclines. Analysis of the curated set of genes that harbor SNPs with significant TR provided a measure for the influence of environmental parameters on these processes. The second approach involved a comparative analysis of protein level sequence divergence in the four gene sets across nine vertebrates, which differ in their skin/integument physiology owing to their differential interaction with the environment and TEWL. Zebrafish is clearly aquatic and is constantly exposed to water as a part of the environment; western clawed frog being an amphibian would need to balance aquatic and terrestrial niches; anole lizard is completely terrestrial but it thrives in humid tropical forests; chicken has additional integuments in the form of feathers on the skin that is likely to give protection; mouse is a nocturnal furry mammal with burrowing habits; and rhesus monkey, chimpanzee, and humans share genomic similarity although humans differ with respect to the loss of epidermal hair. Additionally, the recently sequenced bottlenose dolphin genome provides an interesting comparison, as this mammal has secondarily adapted to and reacquired an aquatic lifestyle. Ka/Ks ratio that provides an index for assessing the role of selection in governing diversity across the four processes was calculated pair-wise between each animal and human. Details of each analysis are provided below.

Identification of Gene Sets for the Analysis

All the GO terms for keratinization, epidermal differentiation, and pigmentation were fetched from DAVID (http://david.abcc.ncifcrf.gov/, last accessed December 20, 2014.). We first created a list of genes with known function in each of the processes involved: Like for keratinization, we enlisted all the genes that have been reported in published scientific literature to affect keratinization of epidermal skin and cornification of keratinocytes, either directly or via intermediates. This set of manually curated genes was treated as reference.

To finalize the gene lists for further downstream analyses, we performed the following steps: We removed all those categories, which were not related to skin (like eye pigmentation-related GO terms). Next we compared all the genes from each GO term with our reference set. Those that overlapped between the two sets were included. For those genes in a particular GO term which did not overlap with our reference set, we again manually checked for literature support. Although deciding whether to include a gene is the particular category, we essentially checked for three criteria: If mutation in that gene results in a disease phenotype related to that process, if that gene is a structural component of the particular cell type where that process occurs, and if that gene codes for enzymes catalyzing some steps of biochemical reactions involved in the process. With all these done, if 70% of the total genes of a particular GO term could be binned to the process, that GO term was taken as a whole. This was done because two of our approaches downstream were GO term centric. Terms that were removed are highlighted in red in supplementary table S3, Supplementary Material online. Additionally, to give further credence to the list of gene chosen, we checked their expression values from microarray data (done in our laboratory, data unpublished). For genes enlisted under keratinization and epidermal differentiation, we checked their expression in a microarray data on primary human keratinocytes and whole epidermal skin of subjects (average value of 13 different individuals was considered). For genes of pigmentation, we looked up for their expression values and found them highly expressed in primary human melanocytes.

Thus, we arrived at three sets of genes for keratinization, epidermal differentiation, and pigmentation. These sets also have genes, which occur in more than one process since all of them show considerable overlap in their biology. We retained the genes in whichever process it was represented and treated them independently.

As a suitable control set against our skin-related genes, we have taken a consensus of two published studies (Eisenberg and Levanon 2003; Chang et al. 2011) (supplementary table S4, Supplementary Material online), which have quantified gene expression in a wide range of physiological conditions and classified genes as "housekeeping" only if they show a relatively constant expression pattern across multiple tissues and conditions. Because their expression is mostly unperturbed, we anticipated that these genes will also exhibit a nondirectional basal level of association with different climate variables and hence would be an appropriate control in our study.

CNV Detection and Enrichment Analysis

To calculate if there are any differences between Indian and global populations in terms of CNV representation for three gene classes under study, we used ten HapMap populations (International HapMap et al. 2007) (CHB, JPT, MEX, YRI, ASW, LWK, CHD, TSI, CEU, and MKK) and 26 Indian populations as reported earlier (Gautam et al. 2012) (refer to supplementary table S1, Supplementary Material online, for details). Furthermore, we have also taken seven Indian populations, which are genotyped on Affymetrix 6.0 array, similar to the one used for HapMap populations. Six of these populations (IE-N-LP5, OG-W-IP, IE-W-LP4, IE-N-LP9, IE-N-LP1, and IE-W-LP3) are also present among 26 Indian populations in our previous population panel along with one more North Indian (Population ID—IND) population included in the data.

Further curation and analyses were performed as detailed below. For all the HapMap populations, we downloaded the raw intensity data (.CEL files) of SNP array (Affymetrix Genome Wide Human SNP Array 6.0). We detected the CNV in HapMap as well Indian data using Affymetrix Genotyping Console 4.1.3. For CNV calling, we used a set of 270 HapMap subjects as reference set for both the population sets. We followed QC criteria of minimum of five contiguous probes per CNV segment and interprobe distance less than 10 kb. Finally, CNVRs generated for all the populations were annotated using RefSeq gene list from UCSC hg18 build (http://genome.ucsc.edu/, last accessed December 20, 2014).

For each population, we calculated the number of samples reporting genes overlapping with CNVs (deletion and amplification) in each gene class. The sample count thus obtained was divided by the total number of samples in a given populations to get sample frequency. This was plotted for worldwide and Indian populations. The significance of difference between HapMap and the Indian population for each category was calculated using Student's *t*-test.

Results of FAC using DAVID Bioinformatics Resources were extracted from the earlier work (Gautam et al. 2012). There was significant (Fisher exact P value < 0.001) enrichment in processes like keratinization/epidermis morphogenesis. The FAC involves module-centric approach to find the functionally related clusters of genes in large gene list (supplementary table S1, Supplementary Material online).

Human–Chimpanzee Divergence Study

Collection of Orthologous Sequences

We retrieved the orthologous sequences for the 5-kb promoter-proximal and the coding regions from human, chimpanzee, and macaque. The human transcripts were fetched from the RefSeq database through UCSC table browser (http://genome.ucsc.edu/, last accessed December 20, 2014) (Karolchik et al. 2003). On the basis of common gene annotation between human and chimpanzee in RefSeq, we identified a total of 19,199 promoter sequences to search for orthologs. The chimpanzee and the macaque genomes were also downloaded to perform Blast-Like Alignment Tool (BLAT) against the promoter sequences for finding the best match in these genomes. BLAT gave multiple hits based on the position of matches in both the genomes. We filtered these hits to identify the best match, applying several cutoffs to minimize false positives: We excluded promoters which had multiple hits, mapped to Y chromosome, had less than 80% nucleotide match, a total alignment span of less than 4,000 and greater than 6,000 bases, and those alignments starting after the tenth base and terminating before the 4,989th base (supplementary text S1, Supplementary Material online). This yielded 8,788 and 14,332 orthologous promoters for macaque and chimpanzee, respectively; 7,314 being common between the two sets. Additionally, these promoters were also quality checked for poorly sequenced regions to increase the confidence of the data. Finally, a set of 4,316 orthologous promoters among human, chimpanzee, and macaque was obtained (illustrated with the help of a flowchart in supplementary text S1, Supplementary Material online).

Calculation of JC Divergence in the Coding Region between Human and Chimpanzee

Exon coordinates for all the transcripts of 4,316 genes were retrieved from UCSC (Karolchik et al. 2003) (http://genome.ucsc.edu/, last accessed December 20, 2014) table browser in BED format. These files were then submitted to online GALAXY (Goecks et al. 2010) tool (http://galaxy.psu.edu/, last accessed December 20, 2014) to extract pair-wise MAF alignment data between human and chimpanzee for the given coordinates. We then joined pair-wise sequences with each other for a given transcript using excel utilities. Subsequent to this, the exon sequence for each species (human and chimpanzee) were fetched separately followed by alignment with each other using stretcher tool from the EMBOSS (Rice et al. 2000) package (http://emboss.sourceforge.net/, last accessed December 20, 2014). A stand-alone code written in C++ was used to calculate substitutions between the aligned sequences, followed by the calculation of JC divergence using the formula:

JC divergence (for a pair of orthologous sequences) = number of substitutions in alignment/total number of bases in the reference sequence

JC divergence in the coding sequence region for each of the transcripts for all 4,316 genes was calculated and partitioned into 4 quartiles. The highly diverged gene set in the fourth quartile of the box plot was then used to query for enriched biological process within this set using DAVID (Huang da et al. 2009) online tool (http://david.abcc.ncifcrf.gov/, last accessed December 20, 2014), and the original set of 4,316 genes was considered as background.

Identification of GRMs

A stand-alone global alignment tool stretcher, from EMBOSS (Rice et al. 2000) package (http://emboss.sourceforge.net/, last accessed December 20, 2014), was utilized with its default parameters to align the orthologous promoter sequences between human–chimpanzee, human–macaque, and chimpanzee–macaque. This alignment data were used to annotate the positions of the aligned bases and to calculate the exact number of GRMs gained or lost.

We identified 202 putative GRMs for the promoter sequence with vertebrate_non_ redundant_minSUM profile (along with high quality matrices) with the help of P-match tool from TRANSFAC database (Matys et al. 2006) version 12.1 (http://www.gene-regulation.com/cgi-bin/pub/programs/pmatch/bin/p-match.cgi, last accessed December 20, 2014). The data were parsed based on the length of GRM and its matrix score (GRM length $\leq$ 8, matrix score $\geq$ 0.9 and GRM length $>$ 8, matrix score $\geq$ 0.8 only were selected) to minimize false positives. We assumed that shorter sequences would have higher chances of false positive matches in the genome.

RepeatMasker (http://www.repeatmasker.org/, last accessed December 20, 2014) was used to retrieve the coordinates of repetitive sequences in the promoters.

Measurement of GRM Divergence between the Two Species

With the help of alignment and the GRM positions for an orthologous promoter, it was possible to annotate the novel/unique and conserved GRM between the two species. Codes were written in C++ to fetch the position-specific novel GRM in both species for the orthologous sequences. GRM divergence for each pair was calculated as the number of unique GRMs, which differed between the two species. Unique GRM in each species was mapped along with the repeat positions to count for GRM divergence harbored by repetitive elements between human, chimpanzee and macaque using C++ program.

Identification of Highly Diverged Gene Set and Biological Processes Enrichment Analysis

Taking 4,316 genes in the background to avoid any bias, we subjected this set to biological process enrichment analysis using DAVID bioinformatics resources tool version 6.7 (http://david.abcc.ncifcrf.gov/, last accessed December 20, 2014) (Huang da et al. 2009). Data of GRM gain in the promoter between human and chimpanzee for two biological processes (viz. keratinocyte differentiation and epidermal cell differentiation) harbored by repetitive sequences were obtained from the 202 GRMs from 187 genes. We also calculated the Euclidian distance for each GRM for all the genes in an enriched biological process category, and outliers were treated as highly diverged GRMs for that particular category.

Methodology for Calculating the GRM Divergence in Repetitive Region After Masking CpG Sites

Data for CpG sites for human (hg18) and chimpanzee (panTro2) were obtained from the UCSC table browser. The coordinates of CpG sites and the 5Kb promoter sequence from human and chimpanzee were validated extensively before mapping the GRMs coordinates to the CpG coordinates. A C++ program was written and executed to map the unique GRMs to human and chimpanzee promoters (GRM divergence in total) over the CpG coordinates. CpG-mapped GRMs were also overlaid with the GRMs mapped within repetitive region to find out GRMs which were contributed by repetitive region and were located within CpG islands. We further used this information to calculate GRM divergence contributed by repetitive region which were not contributed by CpG sites for each gene (4,316 genes). Genes present within the fourth quartile (highly diverged set) was overlapped with fourth quartile gene set (obtained with reference to total GRM divergence between chimpanzee–human) and fourth quartile gene set (obtained with reference to GRM divergence contributed by repetitive region between human–macaque). A set of 262 genes was obtained overlapping between total GRM-diverged genes and repetitive region (without CpG) for chimpanzee–human but unique to genes with GRM divergence contributed by repetitive region in macaque–human. This gene set was then analyzed for biological process enrichment wherein we find very minor changes in the biological processes when compared with the previous analysis where no CpG masking was done.

Exploring the Association of Genes in Different Classes with Environmental Variables

We used the climate and allelic frequency data for 60 populations from the study reported earlier (Hancock et al. 2011). This data comprise 938 unrelated subjects from Human Genome Diversity panel along with four HapMap populations from Phase 3 and four populations genotyped by them. We downloaded the latest data (dbCLINE ver 4.0) for climate and ecoregion variables along with allele frequencies (with Bayes factors and TR statistics) from dbCLINE database (http://genapps2.uchicago.edu:8081/dbCLINE/main.jsp, last accessed December 20, 2014). To check if there are any discernible correlations with gene in the different classes with environmental data, we used the TR statistic for each SNP–environmental pair. The value of TR ranges between 0 and 1 and can be treated as empirical *P* value. Our analysis data set included three classes of genes (keratinization, epidermal differentiation, and pigmentation). As a control, we used a set of 184 housekeeping genes (details in supplementary tables S3–S5, Supplementary Material online).

We selected all the SNPs within and 2 kb upstream and downstream of the genes from UCSC hg18 build (http://genome.ucsc.edu/, last accessed December 20, 2014) (refer to supplementary table S5, Supplementary Material online). From all possible SNP–environmental variable pairs, we selected those that showed TR statistic $<$ 0.05 (which is indicative of significant association with climate/eco-region variables). Then to compare among four gene classes, we calculated the percentage of the genes, which have SNPs exhibiting significant association with the environmental variable.

The representation of SNPs (average number of SNP/kb) across the categories was observed to be slightly higher for keratinization and epidermal differentiation when compared with pigmentation and housekeeping genes (5.77 and 5.26 vs. 3.35 and 3.16, respectively). On comparing the SNP representation per gene, keratinization and pigmentation were well represented (245 and 215, respectively) compared with epidermal differentiation and housekeeping genes (158 and 55, respectively) (supplementary table S6, Supplementary Material online). To identify a pathway-specific signal and prevent few high scoring SNPs to skew the analysis, we considered genes with at least one SNP with TR < 0.05 as reported in Hancock et al. (2011). This would also reduce the representation bias across genes and processes.

We displayed the SNP association results in Manhattan plots (fig. 3*a* and supplementary fig. S7, Supplementary Material online). For which all the SNP–climate variable pairs are displayed in Manhattan plot for all the variables. This was done for all the three gene classes of epidermis differentiation, keratinization, and pigmentation. A cutoff of 0.05 is applied and depicted as horizontal line. All plotting was done in R Programming language (http://CRAN.R-project.org/, last accessed December 20, 2014).

Phylogenetic Analysis

For all genes, their HGNC symbols were converted to corresponding NP IDs (representing all the protein isoforms that a particular gene is reported to code for). Only well-annotated protein isoforms were retained; all others including predicted proteins (represented by XP IDs) were eliminated. Then we selected only human-specific NP IDs, thereby removing all protein entries which have come from some other species. For genes with multiple protein isoforms, we chose only that NP ID which represents the longest protein isoform (downloaded from NCBI).

Next, we performed protein sequence BLAST for each of human protein homologs carried out against the database of all nonredundant protein sequences in the NCBI repository for rest of the eight selected species. The BLAST results were downloaded as text files from dropdown menu itself and then best isoform match was identified as the first reported match against each species. The values of percentage identity (calculated based on given alignment criteria), as well as query cover and *E* value obtained from the BLAST result were extracted from the text files with the help of excel utilities. The data obtained had gone through extensive manual quality checkups. A cutoff of 30% was kept for query cover (at least 30% of the protein should have been queried for it to be included; those entries that did not clear this cutoff were denoted as blue boxes in fig. 4). *E* values of $<10^{-4}$ were only considered; rest were discarded. There were some NP IDs for which protein identity values could not be found in the BLAST results for certain species (White boxes are used to denote both of these in fig. 4). The tabulated data have been provided in supplementary table S8, Supplementary Material online. Upon procuring the protein identity percentage values for all the proteins for all nine species, heat map was constructed using R package (http://www.r-project.org/, last accessed December 20, 2014).

Ka/Ks Estimation Method

Orthologous protein sequence IDs for all eight species (except human) were identified against each human protein NP ID using the data obtained from the best hits of BLAST results. The sequence IDs of mRNAs (NM IDs) encoding these proteins were also identified for each of these NP IDs from the NCBI database. Protein and mRNA sequences (corresponding to these NP and NM IDs, respectively) were then fetched from NCBI database. ParaAT (a Parallel Alignment and back-Translation tool) (Zhang et al. 2012) was used to construct protein-coding DNA alignment for a pair of orthologous genes (between human and each of the eight species) using their respective protein sequence information (http://code.google.com/p/paraat/wiki/ParaAT, last accessed December 20, 2014). The protein sequence information was necessary to match each triplet codon with their respective amino acid. Input files for ParaAT were prepared with the help of excel utilities, text editors, and command-line scripts from linux terminal. Aligned coding DNA for each gene between a pair of the species, that is, human and each of the eight other species (obtained as a result of output from ParaAT tool) was used as input for the KaKs_Calculator program (Zhang et al. 2006) (http://code.google.com/p/kaks-calculator/wiki/KaKs_Calculator, last accessed December 20, 2014) to calculate Ka, Ks, and Ka/Ks values along with their significance score. We selected model averaging (MA) parameter as a maximum likelihood method for calculating the Ka/Ks values. Maximum likelihood method takes into account the sequence evolutionary features such as ratio of transition/transversion rate and nucleotide frequency to incorporate all these features into a codon-based model. In the case of MA, it takes an average of substitution rates for given seven models to calculate Ka/Ks ratio.

Ka/Ks Analyses

To compare Ka/Ks distribution (calculated between human with each of the eight other species) among four biological processes, we first binned the Ka/Ks values in the range of 0–2.5 with an interval of 0.1 (data can be found in supplementary table S8, Supplementary Material online). The number of genes falling within each bin were calculated based on their Ka/Ks values for each species compared followed by their calculation in percentage. These percentage values were then plotted (for each of the compared species) for all the four biological process using R-software (http://CRAN.R-project.org/, last accessed December 20, 2014).

Study of the Effect of CpG Masking upon Ka/Ks Ratio (Coding Divergence) for Selected Candidate Genes

Twelve genes having high Ka/Ks ratio (between human and any other species under study as the case might be) from the keratinization gene set were prioritized for this study.

Aligned human mRNA for the pair of compared species was then subjected to BLAT against human hg18 database to get the coordinates on the human genome. These coordinates were then overlapped with the CpG sites of the same genome build to identify and mask these from the alignment, so as to create one without having these sites (with some manual curation). The aligned mRNA sequences were used to calculate the Ka/Ks ratio using KaKs_Calculator program (http://code.google.com/p/kaks-calculator/wiki/KaKs_Calculator, last accessed December 20, 2014).

Members of IGV

Samir K. Brahmachari, Partha P. Majumder, Mitali Mukerji, Saman Habib, Debasis Dash, Kunal Ray, Samira Bahl, Lalji Singh, Abhay Sharma, Susanta Roychoudhury, G.R. Chandak, K. Thangaraj, D. Parmar, Shantanu Sengupta, Dwaipayan Bharadwaj, Srikanta K. Rath, Jagmohan Singh, Ganga Nath Jha, Komal Virdi, V.R. Rao, Swapnil Sinha, Ashok Singh, Amit K. Mitra, Shrawan K. Mishra, Qadar Pasha, Sridhar Sivasubbu, Rajesh Pandey, Aradhita Baral, Prashant K. Singh, Amitabh Sharma, Jitender Kumar, Tsering Stobdan, Yasha Bhasin, Chitra Chauhan, Ashiq Hussain, Elyanambi Sundaramoorthy, S.P. Singh, Arun Bandyopadhyay, Krishanu Dasgupta, A.K. Reddy, Charles J Spurgeon, M. Mohd Idris, Vinay Khanna, Alok Dhawan, Mohini Anand, R. Shankar, R.S. Bharti, Madhu Singh, Arvind P. Singh, Anwar J. Khan, Parag P. Shah, A.B. Pant, Rupinder Kaur, Kamlesh K. Bisht, Ashok Kumar, Victor Rajamanickam, Eugene Wilson, Antony Thangadurai, Pankaj K. Jha, Mahua Maulik, Neelam Makhija, Abdur Rahim, Sangeeta Sharma, Rupali Chopra, Pooja Rana, M. Chidambaram, Arindam Maitra, Ruchi Chawla, Suruchika Soni, Preeti Khurana, Mohd. Nadeem Khan, Sushanta Das Sutar, Amit Tuteja, K. Narayansamy, Rachna Shukla, Swami Prakash, Swapna Mahurkar, K Radha Mani, J. Hemavathi, Seema Bhaskar, Pankaj Khanna, G.S. Ramalakshmi, Shalini Mani Tripathi, Nikita Thakur, Balaram Ghosh, Ritushree Kukreti, Taruna Madan, Ranjana Verma, G. Sudheer, Anubha Mahajan, Sreenivas Chavali, Rubina Tabassum, Sandeep Grover, Meenal Gupta, Jyotsna Batra, Amrendra Kumar, Abdoulazim Nejatizadeh, Mudit Vaid, Swapan K. Das, Shilpy Sharma, Mamta Sharma, Rajshekhar Chatterjee, Jinny A. Paul, Pragya Srivastava, Charu Rajput, Uma Mittal, Mridula Singh, Manoj Hariharan, Sumantra Das, Keya Chaudhuri, Mainak Sengupta, Moulinath Acharya, Ashima Bhattacharyya, Atreyee Saha, Arindam Biswas, Moumita Chaki, Arnab Gupta, Saibal Mukherjee, Suddhasil Mookherjee, Ishita Chattopadhyay, Taraswi Banerjee, Meenakshi Chakravorty, Chaitali Misra, Gourish Monadal, Shiladitya Sengupta, Dipanjana Dutta De, Swati Bajaj, Ishani Deb, Arunava Banerjee, Rajdeep Chowdhury, Debalina Banerjee, Deepak Kumar, Sumit Ranjan Das, Shrish Tiwari, Anshu Bharadwaj, Sangeeta Khanna, Ikhlak Ahmed, Sumera Parveen, Nivedita Singh, Dipayan Dasgupta, Siddharth Singh Bisht, Rashmi Rajput, Biswaroop Ghosh, Naveen Kumar, Amit Chaurasia, James K. Abraham, Amit Sinha, Vinod Scaria, Tav Pritesh Sethi, Amit K. Mandal, Arijit Mukhopadhyay.

Supplementary Material

Supplementary text S1, tables S1–S13, and figures S1–S9 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This work was supported by the Council for Scientific and Industrial Research (grant numbers MLP6201, MLP901, BSC0302, CMM0017). Fellowships to A.B. (JRF-CSIR), A.C. (SRF-ICMR) and P.G. and R.G. (SRF-CSIR) are acknowledged. The authors thank Drs Rajesh S. Gokhale, Abhay Sharma, and Ravi Shankar for their valuable and critical comments.

References

Baxter LL, Pavan WJ. 2013. The etiology and molecular genetics of human pigmentation disorders. *Wiley Interdiscip Rev Dev Biol.* 2:379–392.

Bergboer JG, Zeeuwen PL, Irvine AD, Weidinger S, Giardina E, Novelli G, Den Heijer M, Rodriguez E, Illig T, Riveira-Munoz E, et al. 2010. Deletion of late cornified envelope 3B and 3C genes is not associated with atopic dermatitis. *J Invest Dermatol.* 130:2057–2061.

Brahmachari SK, Majumder PP, Mukerji M, Habib S, Dash D, Ray K, Bahl S, Singh L, Sharma A, Roychoudhury S, et al. 2008. Genetic landscape of the people of India: a canvas for disease gene exploration. *J Genet.* 87(1):3–20.

Brown SJ, Kroboth K, Sandilands A, Campbell LE, Pohler E, Kezic S, Cordell HJ, McLean WH, Irvine AD. 2012. Intragenic copy number variation within filaggrin contributes to the risk of atopic dermatitis with a dose-dependent effect. *J Invest Dermatol.* 132:98–104.

Chandran V, Raychaudhuri SP. 2010. Geoepidemiology and environmental factors of psoriasis and psoriatic arthritis. *J Autoimmun.* 34:J314–J321.

Chang CW, Cheng WC, Chen CR, Shu WY, Tsai ML, Huang CL, Hsu IC. 2011. Identification of human housekeeping genes and tissue-selective genes by microarray meta-analysis. *PLoS One* 6:e22859.

Chen FC, Vallender EJ, Wang H, Tzeng CS, Li WH. 2001. Genomic divergence between human and chimpanzee estimated from large-scale alignments of genomic sequences. *J Hered.* 92:481–489.

Denda M, Sato J, Tsuchiya T, Elias PM, Feingold KR. 1998. Low humidity stimulates epidermal DNA synthesis and amplifies the hyperproliferative response to barrier disruption: implication for seasonal exacerbations of inflammatory dermatoses. *J Invest Dermatol.* 111:873–878.

Eisenberg E, Levanon EY. 2003. Human housekeeping genes are compact. *Trends Genet.* 19:362–365.

Enamandram M, Kimball AB. 2013. Psoriasis epidemiology: the interplay of genes and the environment. *J Invest Dermatol.* 133:287–289.

Farcas R, Schneider E, Frauenknecht K, Kondova I, Bontrop R, Bohl J, Navarro B, Metzler M, Zischler H, Zechner U, et al. 2009. Differences in DNA methylation patterns and expression of the CCRK gene in human and nonhuman primate cortices. *Mol Biol Evol.* 26(9):1379–1389.

Gautam P, Jha P, Kumar D, Tyagi S, Varma B, Dash D, Mukhopadhyay A, Mukerji M. 2012. Spectrum of large copy number variations in 26 diverse Indian populations: potential involvement in phenotypic diversity. *Hum Genet.* 131:131–143.

George RD, McVicker G, Diederich R, Ng SB, MacKenzie AP, Swanson WJ, Shendure J, Thomas JH. 2011. Trans genomic capture and sequencing of primate exomes reveals new targets of positive selection. *Genome Res.* 21:1686–1694.

Gilad Y, Man O, Glusman G. 2005. A comparison of human and chimpanzee olfactory receptor gene repertoires. *Genome Res.* 15:224–230.

Ginger RS, Blachford S, Rowland J, Rowson M, Harding CR. 2005. Filaggrin repeat number polymorphism is associated with a dry skin phenotype. *Arch Dermatol Res.* 297:235–241.

Goecks J, Nekrutenko A, Taylor J. 2010. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* 11:R86.

Han J, Colditz GA, Hunter DJ. 2007. Polymorphisms in the MTHFR and VDR genes and skin cancer risk. *Carcinogenesis* 28:390–397.

Hancock AM, Witonsky DB, Alkorta-Aranburu G, Beall CM, Gebremedhin A, Sukernik R, Utermann G, Pritchard JK, Coop G, Di Rienzo A. 2011. Adaptations to climate-mediated selective pressures in humans. *PLoS Genet.* 7:e1001375.

Hasin Y, Olender T, Khen M, Gonzaga-Jauregui C, Kim PM, Urban AE, Snyder M, Gerstein MB, Lancet D, Korbel JO. 2008. High-resolution copy-number variation map reflects human olfactory receptor diversity and evolution. *PLoS Genet.* 4:e1000249.

Huang da W, Sherman BT, Lempicki RA. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc.* 4:44–57.

Huffmeier U, Bergboer JG, Becker T, Armour JA, Traupe H, Estivill X, Riveira-Munoz E, Mossner R, Reich K, Kurrat W, et al. 2010. Replication of LCE3C-LCE3B CNV as a risk factor for psoriasis and analysis of interaction with other genetic risk factors. *J Invest Dermatol.* 130:979–984.

International HapMap C, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, et al. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449:851–861.

Jablonski NG, Chaplin G. 2000. The evolution of human skin coloration. *J Hum Evol.* 39:57–106.

Jablonski NG, Chaplin G. 2010. Colloquium paper: human skin pigmentation as an adaptation to UV radiation. *Proc Natl Acad Sci U S A.* 107(Suppl 2), 8962–8968.

Jin Y, Birlea SA, Fain PR, Ferrara TM, Ben S, Riccardi SL, Cole JB, Gowan K, Holland PJ, Bennett DC, et al. 2012. Genome-wide association analyses identify 13 new susceptibility loci for generalized vitiligo. *Nat Genet.* 44:676–680.

Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu YT, Roskin KM, Schwartz M, Sugnet CW, Thomas DJ, et al. 2003. The UCSC Genome Browser Database. *Nucleic Acids Res.* 31:51–54.

Katsuta Y, Iida T, Hasegawa K, Inomata S, Denda M. 2009. Function of oleic acid on epidermal barrier and calcium influx into keratinocytes is associated with N-methyl D-aspartate-type glutamate receptors. *Br J Dermatol.* 160:69–74.

Lee Y, Je YJ, Lee SS, Li ZJ, Choi DK, Kwon YB, Sohn KC, Im M, Seo YJ, Lee JH. 2012. Changes in transepidermal water loss and skin hydration according to expression of aquaporin-3 in psoriasis. *Ann Dermatol.* 24:168–174.

Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, et al. 2006. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.* 34:D108–D110.

Moll R, Divo M, Langbein L. 2008. The human keratins: biology and pathology. *Histochem Cell Biol.* 129:705–733.

Naldi L, Parazzini F, Gallus S, GISED Study Centres. 2009. Prevalence of atopic dermatitis in Italian schoolchildren: factors affecting its variation. *Acta Derm Venereol.* 89:122–125.

Natarajan VT, Ganju P, Ramkumar A, Grover R, Gokhale RS. 2014. Multifaceted pathways protect human skin from UV radiation. *Nat Chem Biol.* 10(7):542–551.

Nemoto-Hasebe I, Akiyama M, Nomura T, Sandilands A, McLean WH, Shimizu H. 2009. FLG mutation p.Lys4021X in the C-terminal imperfect filaggrin repeat in Japanese patients with atopic eczema. *Br J Dermatol.* 161:1387–1390.

Ny A, Egelrud T. 2004. Epidermal hyperproliferation and decreased skin barrier function in mice overexpressing stratum corneum chymotryptic enzyme. *Acta Derm Venereol.* 84:18–22.

Off MK, Steindal AE, Porojnicu AC, Juzeniene A, Vorobey A, Johnsson A, Moan J. 2005. Ultraviolet photodegradation of folic acid. *J Photochem Photobiol B.* 80:47–55.

Parisi R, Symmons DP, Griffiths CE, Ashcroft DM. 2013. Global epidemiology of psoriasis: a systematic review of incidence and prevalence. *J Invest Dermatol.* 133:377–385.

Parker-Katiraee L, Carson AR, Yamada T, Arnaud P, Feil R, Abu-Amero SN, Moore GE, Kaneda M, Perry GH, Stone AC, et al. 2007. Identification of the imprinted KLF14 transcription factor undergoing human-specific accelerated evolution. *PLoS Genet.* 3(5):e65.

Pickrell JK, Coop G, Novembre J, Kudaravalli S, Li JZ, Absher D, Srinivasan BS, Barsh GS, Myers RM, Feldman MW, et al. 2009. Signals of recent positive selection in a worldwide sample of human populations. *Genome Res.* 19:826–837.

Prans E, Kingo K, Traks T, Silm H, Vasar E, Koks S. 2013. Copy number variations in IL22 gene are associated with *Psoriasis vulgaris*. *Hum Immunol.* 74:792–795.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* 16:276–277.

Sankararaman S, Mallick S, Dannemann M, Prüfer K, Kelso J, Päbo S, Patterson N, Reich D. 2014. The genomic landscape of Neanderthal ancestry in present-day humans. *Nature* 507:354–357.

Scally A, Dutheil JY, Hillier LW, Jordan GE, Goodhead I, Herrero J, Hobolth A, Lappalainen T, Mailund T, Marques-Bonet T, et al. 2012. Insights into hominid evolution from the gorilla genome sequence. *Nature* 483:169–175.

Steindal AH, Juzeniene A, Johnsson A, Moan J. 2006. Photodegradation of 5-methyltetrahydrofolate: biophysical aspects. *Photochem Photobiol.* 82:1651–1655.

Sturm RA, Duffy DL. 2012. Human pigmentation genes under environmental selection. *Genome Biol.* 13:248.

Subramanian S, Kumar S. 2006. Higher intensity of purifying selection on > 90% of the human genes revealed by the intrinsic replacement mutation rates. *Mol Biol Evol.* 23(12):2283–2287.

Tennessen JA, Akey JM. 2011. Parallel adaptive divergence among geographically diverse human populations. *PLoS Genet.* 7:e1002127.

Vernot B, Akey JM. 2014. Resurrecting surviving Neandertal lineages from modern human genomes. *Science* 343:1017–1021.

Zhang Z, Li J, Zhao XQ, Wang J, Wong GK, Yu J. 2006. KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. *Genomics Proteomics Bioinformatics.* 4:259–263.

Zhang Z, Xiao J, Wu J, Zhang H, Liu G, Wang X, Dai L. 2012. ParaAT: a parallel tool for constructing multiple protein-coding DNA alignments. *Biochem Biophys Res Commun.* 419:779–781.