Recent Positive Selection in Genes of the Mammalian Epidermal Differentiation Complex Locus

Zane A. Goodwin and Cristina de Guzman Strong*

Division of Dermatology, Department of Internal Medicine, Center for Pharmacogenomics and Center for the Study of Itch, Washington University School of Medicine, St. Louis, MO, USA

The epidermal differentiation complex (EDC) is the most rapidly evolving locus in the human genome compared to that of the chimpanzee. Yet the EDC genes that are undergoing positive selection across mammals and in humans are not known. We sought to identify the positively selected genetic variants and determine the evolutionary events of the EDC using mammalian-wide and clade-specific branch- and branch-site likelihood ratio tests and a genetic algorithm (GA) branch test. Significant non-synonymous substitutions were found in filaggrin, SPRR4, LELP1, and S100A2 genes across 14 mammals. By contrast, we identified recent positive selection in SPRR4 in primates. Additionally, the GA branch test discovered lineage-specific evolution for distinct EDC genes occurring in each of the nodes in the 14-mammal phylogenetic tree. Multiple instances of positive selection for FLG, TCHHL1, SPRR4, LELP1, and S100A2 were noted among the primate branch nodes. Branch-site likelihood ratio tests further revealed positive selection in specific sites in SPRR4, LELP1, filaggrin, and repetin across 14 mammals. However, in addition to continuous evolution of SPRR4, site-specific positive selection was also found in S100A11, KPRP, SPRR1A, S100A7L2, and S100A3 in primates and filaggrin, filaggrin2, and S100A8 in great apes. Very recent human positive selection was identified in the filaggrin2 L41 site that was present in Neanderthal. Together, our results identifying recent positive selection in distinct EDC genes reveal an underappreciated evolution of epidermal skin barrier function in primates and humans.

Keywords: positive selection, evolution, skin, epidermis, barrier, epidermal differentiation complex

INTRODUCTION

The evolution of modern humans (Homo sapiens sapiens) is driven by ongoing adaptations to local environments and niches (Jeong and Di Rienzo, 2014). Insights into the biological pathways and genes underlying human evolution have been identified from early pairwise comparisons between the coding sequences of human and its closest primate relative, the chimpanzee (Pan troglodytes) (Clark et al., 2003). Additional studies discovered positive selection for gene variants involved in sensory perception, amino acid catabolism, host defense/immunity, reproduction, hair follicle development, and skin pigmentation in human evolution (Nielsen et al., 2005; Voight et al., 2006; Sabeti et al., 2007; Kosiol et al., 2008). The inclusion of additional mammalian genomes provided higher resolution into specific biological function, including innate complement immunity and taste perception (Bakewell et al., 2007; Kosiol et al., 2008).
A direct comparison to the complete genome of the chimpanzee enabled a closer investigation of more recent evolution in the human genome (Chimpanzee Sequencing and Analysis Consortium, 2005). From this study, the epidermal differentiation complex (EDC) locus was identified as the most rapidly evolving locus in the human genome among other loci involved in immunity, perception, and epithelia (Chimpanzee Sequencing and Analysis Consortium, 2005).

The EDC exhibited the highest proportion of amino acid substitutions ($K_\alpha/K_\varepsilon > 1$ where $K_\alpha$ and $K_\varepsilon$ denotes nucleotide substitutions that affect codon and intron/intragenic nucleotide changes, respectively). This finding provided evidence for positive selection in the EDC. The EDC on human 1q21 spans approximately 1.6 Mb and comprises 65 genes representing four gene families including the Filaggrin (FLG)-like or SFTP (S100 fused type protein), Late Cornified Envelope (LCE), Small Proline Repeat-Rich (SPRR) and S100-domain (S100) genes (Mischke et al., 1996; de Guzman Strong et al., 2010). The expression of key genes in the EDC [including filaggrin (FLG), loricrin (LOR), involucrin (IVL), SPRRs, LCEs, and S100A7, A10, A11] is a hallmark feature of terminally differentiated epidermal cells (or keratinocytes) that comprise the mature, stratified layers of the interfollicular epidermis at the skin surface and found between hair follicles (Candi et al., 2005). EDC proteins including FLG, LOR, IVL, and many SPRRs and LCEs are covalently cross-linked in the formation of the cornified envelope that surrounds the keratinocyte as a single structural unit of the epidermal barrier (Candi et al., 2005). The linearity and synteny of the EDC in both eutherian and metatherian mammals have greatly facilitated a more accurate identification of orthologous EDC genes in other primates and mammals (Mischke et al., 1996; Hardman et al., 1999; Cabral et al., 2001b; Jackson et al., 2005; de Guzman Strong et al., 2010; Henry et al., 2012; Jiang et al., 2014; Strasser et al., 2014).

Early comparative genome-wide scans were successful in identifying olfaction and spermatogenesis in human evolving traits (Clark et al., 2003; Nielsen et al., 2005; Kosiol et al., 2008). However, the EDC had not been implicated despite the inclusion of only a small subset of $100$ annotated genes from the EDC at the time of their analyses. These genomic studies included a range of 3–6 species whose genomes had been sequenced at the time. Furthermore, despite the discovery of the human EDC with the highest amino acid substitutions compared to the chimpanzee, we still do not understand which individual genes and their variants that are under positive selection in the human genome (Chimpanzee Sequencing and Analysis Consortium, 2005). Moreover, we also have a poorer understanding of the evolutionary history of the mammalian EDC that is expressed in the interfollicular epidermis. The inclusion of additional, high-quality mammalian genomes for our study will enable more improved comparative genomic analyses to determine such genes.

Here, we sought to more comprehensively identify the genes that underlie the rapid evolution of the human EDC. We aim to gain a better understanding of the evolution of both mammalian and human interfollicular epidermis. Knowledge of evolving EDC genes is critical toward developing hypotheses that will be tested for skin barrier function in mammals and humans. Ultimately, the knowledge gained from these comparative genomics studies and downstream functional analyses will motivate parallel studies in other tissue types and advance our understanding of mammalian and human evolution.

To identify positively selected EDC genes, we used robust statistical measures from manually curated annotations of EDC genes obtained from a comprehensive set of nine publicly available primate genome builds (Lindblad-Toh et al., 2011) as well as the inclusion of dog, opossum, rat, and mouse genomes to the human genome, totaling 14 genomes. Specifically, we aimed to identify both the genes and the single nucleotide changes responsible for the high non-synonymous substitution ratio observed across the entire EDC and to estimate when each gene underwent positive selection during mammalian and primate evolution.

Our results among the studied genes collectively identified significant mammalian-wide positive selection in Filaggrin (FLG), SPRR4, late cornified envelope-like proline-rich 1 (LELP1), and S100A2 in the branch likelihood ratio tests (B-LRTs). Clade-specific B-LRT analyses further identified more recent positive selection in SPRR4 in primates. Using genetic algorithm (GA)-branch tests, we pinpoint multiple instances of positive selection for FLG, TCHHL1, SPRR4, LELP1, and S100A2 across many nodes of primate origin and discover lineage-specific evolution for distinct EDC genes. We also identified site-specific positive selection in SPRR4, LELP1, FLG, and RPTN when testing across all 14 mammals using branch site-specific likelihood ratio tests (BS-LRTs). Clade-specific BS-LRTs highlighted recent evolution in S100A11, KPRP, SPRR1A, S100A7L2, and S100A3 (primates) and FLG and FLG2 (great apes). Finally, we determine even more recent site-specific positive selection for L41 in FLG2 in humans.

Thus, our study provides a deeper molecular understanding of the evolution of human and primate skin for epidermal barrier function.

**MATERIALS AND METHODS**

**EDC Ortholog Analyses**

Orthologs for each of the human EDC reference genes were downloaded from NCBI and ENSEMBL v. 81 ID (Smedley et al., 2015) using eutils and BioMart, respectively (Yates et al., 2016). For those orthologs which could not be retrieved from the aforementioned method, we manually obtained and annotated the ortholog using either best-hit reciprocal BLAST or the BLAT feature to the respective genome [UCSC genome browser (Speer et al., 2016)]. Direct ortholog comparisons exclude potential biases that can stem from gene flow or recombination when estimating selection (Anisimova et al., 2003; Arenas and Posada, 2010, 2014). Where possible, the longest isoform or ORF [predicted using Geneious v8.1 (Kearse et al., 2012)] for each gene was used for the multiple alignment using MUSCLE (Edgar, 2004). After alignment, codon substitutions were converted to rates whereby $d_N$ = non-synonymous and $d_S$ = synonymous substitutions (also known as $K_\alpha$ and $K_\varepsilon$ or $K$N and $K_S$). Positive selection was determined by the significance
of the calculated \( dN/dS \) likelihood ratios using phylogenetic analysis with maximum likelihood (PAML) (Yang, 1998, 2007), specifically branch and branch-site likelihood ratio tests (B-LRT and BS-LRT, respectively). For each B-LRT and BS-LRT that was tested on each gene, the \( dN/dS \) ratios were tested under the assumptions of two pair-wise statistical models (M1a–M2a and M7–M8). M1a models genetic drift by constraining \( dN/dS \) values \(<1\) in comparison to M2a’s assumption of \( dN/dS > 1\) for positive selection. In the more sensitive pairwise model that is beta distributed, M7 models genetic drift \((0 \leq dN/dS \leq 1)\) vs. the M8 positive selection model whereby \( dN/dS > 1\) (Anisimova et al., 2001; Wong et al., 2004). For each model, raw \( dN, dS\) and \( dN/dS\) ratios for all sites and all lineages were estimated using the Nei-Gojobori Method (Nei and Gojobori, 1986). Both B-LRTs and BS-LRTs were performed, with B-LRT testing the entire length of a tested gene and BS-LRT for individual sites for a specific gene. With respect to the mammalian phylogeny of each genome, each B-LRT or BS-LRT was tested across the entire 14-mammal tree (according to Murphy et al., 2001) in accordance with the general consensus in the vertebrate phylogeny community and in two foreground vs. background clade-specific tests (also Figure 1). By definition, as the M7 and M8 model is designed only to test for selection across all branches in a phylogenetic tree, clade-specific LRTs are not possible for the M7–M8 comparison. Each model (M) also computes likelihood values that can be tested for significance using a chi-squared test \((df = 2\) for the M1a–M2a and M7–M8 comparisons). \( p\)-values were adjusted to account for both tests across multiple genes and across four different lineages (23 genes in four lineages = 92 hypotheses) using the p.adjust package in R with method = “fdr” according to Benjamini and Hochberg’s method to control for the false discovery rate (Anisimova and Yang, 2007; Bakewell et al., 2007). A \( p\)-value cutoff \((\alpha)\) of 0.01 corresponds to a FDR of 0.05 for this dataset, hence \( \alpha = 0.01\) is the significance cutoff for the B-LRT. The posterior probabilities (Posterior Prob.) of positive selection for the BS-LRT were calculated using the Bayes’ Empirical Bayes method in PAML for the M1a–M2a and the M7–M8 comparisons as previously described (Yang et al., 2005).

Positive selection for internal branches of the phylogenetic tree were determined by \( dN-dS \) and estimated using the GA-branch method (Branch-SiteREL) in HyPhy (Pond and Frost, 2005; Pond et al., 2005). A universal genetic code was assumed and the same phylogenetic tree that was used for the PAML tests as previously described above. We allowed the method to automatically decide on model complexity among branches. Both \( dN\) and \( dS\) were allowed to vary along branch-site combinations. Internal branch-specific \( dN-dS\) totals were extracted from the data lines labeled “baselineTree” in the “mglocal.fit” output files produced by HyPhy’s GA-branch method.

Validation of EDC Variants in Neanderthal and Denisovan Genomes

Alignments of reads from the Denisovan and Altai Neandertal genome sequencing projects were downloaded from http://cdn.eva.mpg.de/denisova/alignments/ and http://cdn.eva.mpg.de/neandertal/altai/AltaiNeandertal/bam/, respectively. Read count

FIGURE 1 | Positive selection of FLG, SPRR4, LELP1, and S100A2 in 14 mammals and SPRR4 in primates. (A) B-LRT conducted for all 14 mammals, primates, and the HGC clades. The following genes were identified for the respective B-LRT: (B) FLG, SPRR4, LELP1, and S100A2 (M7–M8 Test) and (C) SPRR4 (primate B-LRT [M1a–M2a test]). Dotted line 1% significance cutoff corresponds to a FDR of 0.05 for this dataset as determined by the Benjamini and Hochberg method. Genes with \( p = 0\) were assigned –log10 \( p\)-values of 0.
TABLE 1 | Branch likelihood ratio test (B-LRT) results across 14 mammals for each EDC gene that exhibited site-s–pecific proportions with \( dN/dS \) Ratios > 1 for the M7–M8 comparison.

| Branch likelihood ratio test (B-LRT) | Gene | \( dN \) | \( dS \) | \( dN/dS \) | Proportion | \( dN/dS \) | LRT \( \chi^2 \) statistic | \( p \) |
|-------------------------------------|------|---------|---------|---------|-----------|---------|-----------------|------|
| **M7 vs. M8**                       |      |         |         |         |           |         |                 |      |
| THXL1                               | 0.27 | 0.52    | 0.52    | 0.25    | 1.39      | 2.22    | \( 3.29 \times 10^{-1} \) |      |
| RPN1                                | 0.20 | 0.71    | 0.30    | 0.21    | 2.90      | 48.74   | 0.00            |      |
| FLG                                 | 0.48 | 1.04    | 0.46    | 0.23    | 1.82      | 19.25   | \( 6.60 \times 10^{-5} \) |      |
| FLG2                                | 0.08 | 0.59    | 0.12    | 0.12    | 3.47      | 4.08    | \( 1.30 \times 10^{-1} \) |      |
| KPRP                                | 0.16 | 0.49    | 0.30    | 0.01    | 5.01      | 6.79    | \( 3.35 \times 10^{-2} \) |      |
| IVL                                 | 0.41 | 0.98    | 0.38    | 0.46    | 1.31      | 5.27    | \( 7.17 \times 10^{-2} \) |      |
| SPRR4                               | 0.17 | 0.84    | 0.21    | 0.28    | 5.30      | 20.56   | \( 3.43 \times 10^{-5} \) |      |
| SPRR1A                              | 0.08 | 0.51    | 0.16    | 0.36    | 1.09      | 0.61    | \( 7.38 \times 10^{-1} \) |      |
| SPRR3                               | 0.26 | 0.91    | 0.23    | 0.38    | 1.24      | 6.84    | \( 3.27 \times 10^{-2} \) |      |
| LELP1                               | 0.11 | 0.93    | 0.16    | 0.24    | 3.19      | 13.86   | \( 9.78 \times 10^{-4} \) |      |
| S100A9                              | 0.34 | 0.79    | 0.42    | 0.07    | 2.38      | 7.64    | \( 2.29 \times 10^{-2} \) |      |
| S100A6                              | 0.05 | 0.30    | 0.11    | 0.03    | 1.23      | 3.14    | \( 2.98 \times 10^{-1} \) |      |
| S100A5                              | 0.04 | 0.53    | 0.12    | 1.0E-05 | 2.87      | \(-1.03 \times 10^{-3}\) | 0.00 |      |
| S100A4                              | 0.03 | 0.39    | 0.06    | 0.03    | 1.59      | 8.74    | \( 1.27 \times 10^{-2} \) |      |
| S100A2                              | 0.10 | 0.45    | 0.23    | 0.06    | 1.95      | 13.56   | \( 1.14 \times 10^{-3} \) |      |
| S100A1                              | 0.01 | 0.39    | 0.02    | 1.0E-05 | 6.11      | \(-1.88 \times 10^{-3}\) | 0.00 |      |

EDC genes are sorted by genomic start position (hg38). Significant genes (\( p < 0.01 \), based on FDR [false discovery rate]) in bold.

**RESULTS**

Evidence of Positive Selection for *FLG*, *SPRR4*, *LELP1*, and *S100A2* across Mammals and *SPRR4* in Primates

We sought to accurately determine which of the EDC genes had undergone positive selection in primate and human lineages in the context of mammalian phylogeny. To do this, we utilized comparative alignments among 14 mammalian species genomes. We included the genomes from human (*Homo sapiens*), nine primate species (chimp [*Pan troglodytes*], gorilla [*Gorilla gorilla*], Sumatran orangutan [*Pongo abelii*], macaque [*Macaca mulatta*], baboon [*Papio anubis*], marmoset [*Callithrix jacchus*], mouse lemur [*Microcebus murinus*], galago [*Otolemur garnettii*], tarsier [*Tarsius syrichta*], and four phylogenetically distant mammalian species (rat [*Rattus norvegicus*], mouse [*Mus musculus*], dog [*Canis lupus familiaris*], and opossum [*Monodelphis domestica*]) (Figure 1A) (Anisimova et al., 2001). The nine primate species met our selection criteria for the most complete EDC orthology to the human reference among the 15 publicly available primate orthologous gene sets, the 14-mammal B-LRT in the more sensitive M1a–M2a model comparison were identified (Supplementary Tables S1 and S2). Across the whole tree and for all sites in the entire alignment in both the M1a–M2a and M7–M8 comparisons, all genes demonstrated \( dN/dS < 1 \) consistent with purifying (negative) selection (Supplementary Tables S1 and S2). However, when considering individual sites exhibiting \( dN/dS > 1 \) for each of the orthologous gene sets, the 14-mammal B-LRT in the M1a–M2a comparison identified 12 genes but was not significant (\( dN/dS \) ratio range, 1.42–24.89; 1–28% of sites) (Table 1). By contrast, a total of 16 genes for the 14-mammal B-LRT in the more sensitive M7–M8 model comparison were identified (\( dN/dS \) ratio range, 1.09–6.11; 0.001–25% of sites) (Table 1). Of the 16 genes, four genes in the M7–M8 test were significant (*FLG*, *SPRR4*, *LELP1*, and *S100A2*; Figure 1B). Together, our 14-mammal B-LRT results identified positive selection for *FLG*, *SPRR4*, *LELP1*, and *S100A2* genes across the mammalian lineage.

Clade-specific B-LRT analyses using only M1a and M2a enabled us to determine more recent occurrences of positive selection.
TABLE 2 | Branch likelihood ratio test results in the primate foreground clade-specific test for each EDC gene with \( d_{N/dS} \) Ratios \( > 1 \) using the M1a–M2a comparison.

| Likelihood ratio test | Gene  | % Foreground sites with \( d_{N/dS} > 1 \) | Background Clade \( (d_{N/dS}) \) | Foreground Clade \( (d_{N/dS}) \) | LRT \( \chi^2 \) Statistic | LRT \( p \)-value |
|-----------------------|-------|----------------------------------------|-------------------------------|--------------------------|------------------|------------------|
| Primate test          | RPTN  | 0.1                                    | 0.25                          | 1.44                     | 2.65             | \( 2.65 \times 10^{-1} \) |
|                       | FLG2  | 0.15                                   | 0.12                          | 2.54                     | 1.69             | \( 4.30 \times 10^{-1} \) |
|                       | KPRP  | 0.22                                   | 0.12                          | 1.15                     | 0.47             | \( 7.90 \times 10^{-1} \) |
|                       | SPRR4 | 0.16                                   | 0.04                          | 4.88                     | 21.64            | \( 2.00 \times 10^{-6} \) |
|                       | SPRR1A| 0.02                                   | 0.04                          | 4.93                     | 3.40             | \( 1.83 \times 10^{-1} \) |
|                       | S100A9| 0.04                                   | 0.18                          | 2.67                     | 1.16             | \( 5.59 \times 10^{-1} \) |

Significant genes \( (p < 0.01, \text{based on FDR [false discovery rate]}) \) in bold.

selection unique to either primates or great ape clades (each tested as a foreground) vs. background (respective remaining clade) (Supplementary Tables S3 and S4). In the primate-specific B-LRT, we identified eight EDC genes exhibiting regions with \( d_{N/dS} \) ratios \( > 1 \) compared to the background clade (Table 2). This was in contrast to many EDC genes exhibiting \( d_{N/dS} < 1 \) associated with purifying selection (Supplementary Table S3). Of the six genes exhibiting \( d_{N/dS} > 1 \) in the primate-specific B-LRT, only SPRR4 was significant \( (p = 2.00 \times 10^{-5}, d_{N/dS} = 4.88) \) (Figure 1C, Table 2).

When testing for more recent positive selection in the EDC using the great ape HGC as the foreground in the B-LRT (Supplementary Table S4), our analysis identified seven genes exhibiting \( d_{N/dS} > 1 \) with \( d_{N/dS} > 1 \) in the primate-specific B-LRT, only SPRR4 was significant \( (p < 0.01) \). Nevertheless, the observations of the significances of positive selection in SPRR4 in both the 14-mammal and primate-specific B-LRTs highlight the ongoing evolution of SPRR4 that extends to primates.

We next sought to further discover where positive selection in the EDC was occurring in our 14-mammal tree. Using the GA-branch method, we identified branch-specific occurrences of positive selection in the EDC in 12 ancestral nodes (Figure 2; Supplementary Table S5). At least two genes were found to have undergone positive selection in each of the 12 nodes. All genes were also identified as positively selected in at least one node except for S100A4. Both FLG and TCHHL1 were each identified in six different nodes and thus represent the top two genes that underwent positive selection in multiple nodes. Positive selection for FLG was found in primate nodes (Nodes 3, 5, 6, 11, and 10) but also in rodents (Node 12). Identification of TCHHL1 for positive selection was found in primate nodes (Nodes 2, 5, 8, 9, and 10) and in rodents as well (Node 12). As TCHHL1 was not identified in the B-LRT (observed \( d_{N/dS} = 1 \)), the observations of positive selection for TCHHL1 in the GA-branch test assert an evolution that is more lineage-specific and parallel. The GA-branch test also revealed that LELP1, S100A2, and SPRR4 (found in B-LRT) underwent multiple rounds of positive selection in the primate lineage. Furthermore, SPRR4 exhibited \( d_{N-dS} > 0 \) in the primate clade (Nodes 7 and 10), supporting its identification by the M1a–M2a B-LRT. Overall, our GA-branch data further support positive selection for FLG, LELP1, S100A2, and SPRR4 that occurred in multiple nodes in the primate lineage. As well, our results additionally identify evolution for distinct EDC genes in a lineage-specific manner.

A Majority of Site-Specific Positive Selection Occurred in Conserved Protein Domains of EDC genes

Using the branch-site likelihood ratio test (BS-LRT), we next sought to identify the individual codon substitutions that explain the positive selection in EDC genes. The BS-LRT was performed on 14-mammal alignments for each EDC gene across all species (M1a–M2a and M7–M8) as well as in the pairwise comparisons of the primate and HGC foreground to their respective background clades (with M1a–M2a models’ comparison) as previously described (Figure 3). The posterior probability of positive selection (that is, the probability that a given site is in a class of sites with \( d_{N/dS} > 1 \) [Posterior Prob.], see Materials and Methods) acting on specific codons was subsequently determined for each of the three BS-LRTs. We further determined the locations for the positively selected sites identified by the BS-LRT to gain insight into the functional impact.

The 14-mammal BS-LRT in the M1a–M2a comparison identified evidence for positive selection in codon 60 in the conserved cornifin domain of SPRR4 (60, human reference position; \( d_{N/dS} = 3.42; \text{Posterior Prob} = 0.97 \) (Table 3). Positive selection was also found in the same codon 60 site in SPRR4 in the 14-mammal BS-LRT (M7–M8 comparison) as well as in sites for LELP1, FLG, and RPTN. Four sites were found in LELP1 \( (d_{N/dS} \text{ range, 3.33–3.43}) \) and were located within LELP1’s conserved cysteine and proline-rich domains. Six sites were found in FLG \( (1.51–1.53) \), and 17 sites were all found within the conserved glutamine rich protein domain of RPTN \( (2.48–2.56) \) (Posterior Prob ≥ 0.95). Six sites (including codon 60 and 5 additional) were determined in SPRR4 \( (d_{N/dS} \text{ range, 5.30–5.35}) \). The codon 60 site was identified in both the M7–M8 and the more stringent M1a–M2a comparison indicating the significance of codon 60 in SPRR4’s cornifin domain across the mammalian lineage.

When testing for recent positive selection in the primate clade, we observed site-specific positive selection in six genes \( (S100A11 \text{ [four sites], KPRP [three sites], SPRR4 [five sites], SPRR1A [one site], S100A7L2 [21 sites], and S100A3 [two sites]}) \) (all, Posterior Prob ≥ 0.95) (Table 4). We next considered the functional impact of these significant substitutions with respect...
KPRP sites did not map within conserved protein domains. However, three out of four sites (codons 69, 72, 73, and 78) in S100A11 and one out of the two sites (codon 83) in S100A3 both mapped within the S100 EF-hand domains of the respective proteins. As well, 15 out of the 21 positively selected sites in S100A7L2 also mapped within an S100 EF-hand domain (Table 4). Each EF hand is comprised of two alpha helices separated by a calcium binding domain that...
impacts calcium signaling for S100 proteins (Santamaria-Kisiel et al., 2006). Observance of the positively selected substitutions occurring within S100 EF hands suggests modulations of either calcium binding or downstream binding of target proteins upon calcium-binding-induced conformational changes to the S100 protein. The P32 site and all five sites (codons 26, 39, 43, 70, and 78) were located within the cornifin domains of SPRR1A and SPRR4, respectively. Although the codon 60 site in SPRR4 in its cornifin domain showed significant variation in $dN/dS$ across 14 mammals, this same site was not detected in the primate clade test (M1a–M2a) indicating that different sites in SPRR4 were under positive selection in primates vs. mammals. The cornifin domain found in SPRR proteins is crosslinked by transglutaminase to form the scaffold of the cornified envelope, a structural unit for the epidermal skin barrier (Marvin et al., 1992). The substitutions in the cornifin domains of SPRR1A and SPRR4 suggest modulation of cornified envelope scaffold with an anticipated effect on skin barrier function. Together, we find a majority of the positively selected substitutions within conserved protein domains further highlighting significant evolutionary changes in S100A11, S100A3, S100A7L2, SPRR1A, and SPRR4.

Site-specific positive selection in the more recent great ape HGC clade was identified in a new set of genes; FLG [one site], FLG2 [five sites], and S100A8 [one site] (Posterior Prob ≥ 0.95) (Table 4). Two of the five sites in FLG2 (codons 31 and 41) were found within the S100 EF-hand domain. We address additional significance of codon 41 below. By contrast, the sites in FLG and S100A8 did not overlap any known conserved domains for these genes but does not necessarily preclude the functional impact of these sites. Together with the 14-mammal and primate specific BS-LRTs, the data suggests site-specific positive selection in SPRR4 across all 14 mammals as well as episodic positive selection for KPRP, S100A11, S100A3, S100A7L2, and SPRR1A in primates, and very recent positive selection for FLG, FLG2, and S100A8 in the great ape clade.

The BS-LRT Identifies Human-Specific Positive Selection in FLG2 and Is Found in Neanderthal But Not Denisova

We next sought to identify evidence of positive selection specific to humans. To do this, we examined our BS-LRT results for evidence of human and site-specific positive selection. Human-specific substitutions were only observed in FLG2 in the HGC test using the M1a–M2a comparison [L41, (F)TT$\rightarrow$(L)CTT] (L41, Posterior Prob. = 0.994, respectively)
TABLE 3 | Branch site-specific likelihood ratio test (BS-LRT) results across 14 mammals for each EDC gene for site-specific positive selection (Posterior Prob ≥ 0.95) in either M1a–M2a or M7–M8 comparisons.

| Branch site likelihood ratio test (BS-LRT) | Gene   | Reference amino acid (human) | Human position | Amino acid variation | Alignment position | Posterior probability | Site dN/dS | Site dN/dS SD | Affected domain |
|-----------------------------------------|--------|------------------------------|----------------|---------------------|--------------------|----------------------|------------|--------------|----------------|
| M1a vs. M2a                             | SPRR4  | I                            | 60             | I/V/P/A/T/C         | 87                 | 0.972                | 3.42       | 2.04         | Cornifin       |
| M7 vs. M8                               | SPRR4  | G                            | 58             | G/K/G/S/Q           | 85                 | 0.994                | 5.32       | 1.83         |                |
|                                         |        | I                            | 60             | I/V/P/A/T/C         | 87                 | 1.00                 | 5.35       | 1.81         |                |
|                                         |        | I                            | 61             | I/E/N/P             | 88                 | 0.991                | 5.31       | 1.84         |                |
|                                         |        | C                            | 68             | C/V                 | 94                 | 0.989                | 5.3        | 1.85         |                |
|                                         |        | Q                            | 72             | Q/C/A               | 124                | 0.99                 | 5.3        | 1.84         |                |
|                                         |        | A                            | 73             | A/S/Q/D/P           | 125                | 0.997                | 5.34       | 1.81         |                |
|                                         |        | LELP1                        | K              | K/P/Q/N             | 63                 | 0.989                | 3.39       | 1.35         | Cys/Pro-region |
|                                         |        |                              | S              | S/P/G/M/P/L/F       | 75                 | 0.999                | 3.43       | 1.35         | Cys/Pro-region |
|                                         |        |                              | K              | K/C/P               | 123                | 0.992                | 3.41       | 1.36         | Cys/Pro-region |
|                                         |        |                              | K              | K/P/T/S/K           | 135                | 0.973                | 3.33       | 1.35         | Cys/Pro-region |
|                                         |        | FLG                          | S              | S/Q/H/V/A/T/K       | 171                | 0.971                | 1.53       | 0.23         |                |
|                                         |        |                              | S              | S/L/A/Q/R           | 199                | 0.961                | 1.52       | 0.24         |                |
|                                         |        |                              | N              | H/R/S/T/Q/K/N       | 2636               | 0.966                | 1.52       | 0.23         |                |
|                                         |        |                              | F              | V/S/T/Q/G/A/S/I     | 2641               | 0.960                | 1.52       | 0.24         |                |
|                                         |        |                              | Q              | E/Q/S/T/P/          | 2648               | 0.952                | 1.51       | 0.25         |                |
|                                         |        |                              | R              | R/H/Q/A/S           | 3556               | 0.955                | 1.51       | 0.25         |                |
|                                         |        |                              | R              | R/Q/T               | 221                | 0.965                | 2.5        | 0.38         |                |
|                                         |        |                              | H              | L/Q/C/F/S           | 305                | 0.998                | 2.56       | 0.25         |                |
|                                         |        |                              | H              | H/R/Y/C/A           | 322                | 0.991                | 2.55       | 0.28         |                |
|                                         |        |                              | P              | P/M/T/S             | 864                | 0.964                | 2.5        | 0.39         |                |
|                                         |        |                              | M              | M/T/P/K/S           | 878                | 0.966                | 2.5        | 0.38         |                |
|                                         |        |                              | K              | K/S/R/T             | 945                | 0.955                | 2.48       | 0.41         |                |
|                                         |        |                              | T              | Q/T/K/R             | 954                | 0.993                | 2.55       | 0.27         |                |
|                                         |        |                              | L              | L/S/P/A             | 969                | 0.951                | 2.48       | 0.43         |                |
|                                         |        |                              | W              | W/G/R/S             | 1029               | 0.956                | 2.49       | 0.42         |                |
|                                         |        |                              | S              | S/L/K               | 1031               | 0.976                | 2.52       | 0.35         |                |
|                                         |        |                              | W              | W/Y/G/O/Q           | 1068               | 0.974                | 2.52       | 0.35         |                |
|                                         |        |                              | H              | H/R/P/V/C           | 1090               | 0.966                | 2.5        | 0.38         |                |
|                                         |        |                              | C              | C/Y/N/W/Q           | 1097               | 0.971                | 2.51       | 0.36         |                |
|                                         |        |                              | R              | R/H/Q/H/S           | 1146               | 0.954                | 2.48       | 0.41         |                |
|                                         |        |                              | R              | R/D/Q/N/E           | 1190               | 0.965                | 2.5        | 0.38         |                |
|                                         |        |                              | T              | T/S/N/R/Q           | 1195               | 0.978                | 2.52       | 0.34         |                |
|                                         |        |                              | E              | E/G/K/S/N           | 1199               | 0.960                | 2.49       | 0.40         |                |

(Figure 3; Table 4). Further investigation revealed a common SNP (rs3818831, 121T > C transition, with C as the major allele in modern humans) underlying the L41 substitution with an observed major allele frequency of 0.63 (Sherry et al., 2001). L41 occurs close to a cluster of calcium binding sites in the S100 EF-hand domain of FLG2, suggesting a possible role for this variant in a FLG2-mediated response to calcium.

We next addressed the origins of the human-specific substitutions in FLG2. We determined whether the variants in support of the human-specific residues identified by the BS-LRT arose in the ancient hominids, the Denisova and Neanderthals. The common ancestor of Neanderthals and Denisova diverged from modern humans 700,000–800,000 years ago, while the Denisova lineage diverged from Neanderthals approximately 600,000 years ago (Green et al., 2010; Meyer et al., 2012). Together, the Neanderthals and Denisova represent two separate ancient human lineages with evidence of detectable gene flow events with each other and with modern humans (Green et al., 2010). Closer inspection of the L41 substitution in FLG2 indicates that the modern allele (Derived Allele: L41, Ancestral Allele: F41) occurred following the split between humans and non-human primates (Figure 4). The variant underlying L41 was observed in Neanderthal but not Denisova orthologous sequences (Neanderthal coverage: Total = 48 reads, 19 T [Ancestral], 29 C [Derived]; Denisova coverage: Total = 41 reads, 41 T [Ancestral], 0 C [Derived], Figures 4A,B). Thus, the alignments indicate that L41 appeared early in human evolution but has not yet reached fixation in modern humans.
TABLE 4 | Branch site-specific likelihood ratio test results in the primate and HGC foreground clade-specific tests for each EDC gene with site-specific positive selection (Posterior Prob ≥ 0.95) using the M1a–M2a comparison.

| Branch site likelihood ratio test (BS-LRT) | Gene | Reference amino acid (human) | Human position | Amino acid variation | Alignment position | Posterior probability | Affected domain |
|------------------------------------------|------|-------------------------------|----------------|---------------------|---------------------|----------------------|------------------|
| M1a vs. M2a (primate clade)              | S100A11 | T | 69 | T/K/L/I | 78 | 0.99 | S100 EF-Hand |
|                                          | KPRP | C | 113 | C/R/G/E | 204 | 0.967 | S100 EF-Hand |
|                                          | S100A7L2 | P | 32 | P/T/I | 97 | 0.963 | Cornifin |
|                                          | S100A3 | C | 83 | C/A/V | 103 | 0.952 | S100 EF-Hand |
| M1a vs. M2a (HGC clade)                  | FLG | E | 3451 | K/Q/E/D/R | 3690 | 0.977 | FLG |
|                                          | FLG2 | G | 31 | G/D/K/N/R | 31 | 0.954 | S100 EF-Hand |
|                                          | FLG | L | 41 | F | 41 | 0.994 | S100 EF-Hand |
|                                          | FLG | V | 44 | I/V | 44 | 0.974 | FLG |
|                                          | FLG | N | 47 | N/E/R | 47 | 0.998 | FLG |
|                                          | FLG | D | 70 | D/N | 70 | 0.994 | FLG |
|                                          | S100A8 | D | 22 | D/E | 61 | 0.987 | |

Human-specific site are indicated in bold face.

DISCUSSION

Our results identify the genes and their variants in the EDC locus that are undergoing positive selection across mammalian phylogeny and specific to primates and human. Using both B-LRT and BS-LRT and GA-branch tests, we find significance for more non-synonymous vs. synonymous changes in FLG, SPRR4, LELP1, and S100A2 across 14 mammals. Using foreground clade-specific analyses to determine more recent episodes of positive selection, we further identify positive selection in
SPRR4 that was specific to the primate lineage. GA-branch tests implicated all tested EDC genes except for S100A4 to have undergone positive selection in at least one mammalian node. Furthermore, the GA-branch test further resolved lineage-specific evolution across the mammalian nodes and highlighted FLG and TCHHL1 as the top two genes to evolve across multiple mammalian lineages. Positive selections for SPRR4, LELP1, and S100A2 among many mammalian nodes in the GA-branch test further supports the B-LRT finding. Using the BS-LRT, we also determined site-specific positive selection in SPRR4, LELP1, FLG, and RPTN across mammalian phylogeny. Recent evolution at specific sites in primates were also found in S100A11, KPRP, SPRR4, SPRR1A, S100A7L2, and S100A3 and FLG, FLG2, and S100A8 in great apes. More recent positive selection was identified in a human-specific FLG2 variant (L41) in modern humans that was also found in Neanderthal. Together, our study finds positive selection in a diverse set of key EDC genes thus highlighting recent evolution of epidermal skin barrier function in mammalian and human skins.

Our focused study identifying positively selected genes in the EDC in mammals, primates, and human contributes to our understanding of mammalian evolution. Previous genome-wide scans in search of genes undergoing positive selection specifically in humans have implicated several EDC genes in their analyses (Clark and Kosiol papers). Using genome-wide comparisons of human-chimpanzee-mouse genes, Clark et al. (2003) investigated members of the S100 cluster but did not detect significance in their likelihood ratio test. Kosiol et al. (2008) improved on the analysis to identify positively selected genes by using a deeper phylogenetic data set, consisting of three primates (Human, Chimpanzee, Macaque) and four non-primates (mouse, rat, dog, and opossum). They performed likelihood ratio tests on the entire tree, and then on the primate branch to identify episodes of positive selection. In their data set, they identified SPRR3, LOR, and SPRR4 ($p = 6.64 \times 10^{-3}$, $6.44 \times 10^{-3}$, and $3.32 \times 10^{-3}$, respectively) as being under positive selection.

SPRR4 and LELP1 belong to the SPRR gene family. SPRR (or small proline rich region) proteins are expressed in the
terminally differentiated upper layers of cornified epithelia and at low levels in the cervix and the esophagus (Cabral et al., 2001b). SPRR proteins function as substrates for transglutaminase that crosslinks many EDC proteins together during the formation of the keratinocyte's cornified envelope for the skin barrier. LELP1 is both expressed in the epidermis and although their exact functions remain unknown, genetic LELP1 variation was associated with high IgE levels in humans (Sharma et al., 2007). Site-specific positive selection in LELP1 sites in the conserved cysteine and proline rich domain highlight the evolving biochemical properties of this novel SPRR protein. Our identification of positive selection in SPRR4 evolution further validates the protein diversification of these genes that belong to the group 1 SPRR cluster (Cabral et al., 2001b). SPRR4 is highly expressed in the human stratum corneum upon exposure to UV radiation and further supports an adaptive role for SPRR4 (Cabral et al., 2001a,b; Henry et al., 2012). The BS-LRT enabled us to further determine the molecular evolution of SPRR4 with positive selection of codon 60 site in the conserved cornifin domain across mammals in contrast to non-cornifin domain in codons 26, 39, 43, 70, and 78 that were selected in the primate lineage. Together, our genomic findings pinpoint the occurrences of these specific SPRR genes in the mammalian lineage with recent selection for biochemical sites outside the cornifin domain suggesting ongoing molecular evolution for SPRR4.

Positive selection across mammalian phylogeny was also found in S100A2, a member of the S100 family. Many of the S100 proteins encoded in the EDC are associated with calcium signal transduction and are expressed in the granular layer of the epidermis (Eckert et al., 2004; Glaser et al., 2005). Although antimicrobial activity has not been demonstrated for S100A2, the paralogy to S100A9 of known AMP activity (Brandtzaeg et al., 1995; Clark et al., 2016) suggests that S100A2 may also exhibit AMP activity as well.

Finally, we also observed positive selection in FLG across our 14-mammal study and human-specific site selection for L41 in FLG2. Both genes belong to the FLG-like or SFTP fused domain family whose members possess both a fused S100 domain and an EF hand domain (Wu et al., 2009). FLG is a key structural protein in the epidermis that aggregates with keratin filaments (Sandilands et al., 2009; Brown and McLean, 2012). Initially a profilaggrin precursor, FLG is post-translationally cleaved to single filaggrin monomers that metabolically contribute to the natural moisturizing factor of the skin. Like FLG, FLG2 is also expressed in the differentiated granular layer of the epidermis and is proteolytically degraded (Hsu et al., 2011). The observance of the FLG2 L41 substitution in ancient hominids suggests that an additional episode of positive selection led to changes in the epidermal barrier integrity in modern humans and has not reached fixation in modern humans. We speculate that the phenotype associated with this positive selection may have been a fitness advantage in the context of dry arid environments during human migration in Eastern and Southern Africa (Carrier et al., 1984). To extrapolate on filaggrin evolution, interestingly, loss-of-function (LOF) mutations for FLG are strong risk factors for atopic dermatitis, a common inflammatory skin disease (Sandilands et al., 2006; Brown et al., 2008; Irvine et al., 2011). The allele frequencies for FLG LOF of European descent (specifically, Irish) are common, approximately 10% (Sandilands et al., 2006, 2007). Moreover, while LOF mutations in FLG are widely replicated across many populations for AD susceptibility, the LOF mutations are inherently unique to each ethnically distinct population that has been studied (in other words, no two FLG LOF mutations are alike). Together, the observations suggest recent and independently parallel emergences of potentially positively selected mutations that also converge on AD risk. Similarly, in the absence or rarity of FLG LOF mutations in AD patients of African descent, stop gain mutations in FLG2 instead have been found (Margolis et al., 2014). These observations suggest more recent and parallel selective pressures acting on the evolution for FLG and FLG2 even in the context of disease susceptibility in modern humans. Together, positive selection of the FLG epidermal genes highlight mammalian macroevolution and perhaps even more recent human microevolution at the environmental interface for which further biological investigations are warranted.

Together, our results reveal, for the first time, the genetic underpinnings that highlight recent episodic positive selection in epidermal barrier function in mammalian, primate, and modern human skin evolution. This is in contrast to previous studies in search of observable differences in pigmentation and hair follicle density across human populations (Carrier et al., 1984; Elias et al., 2009; Jablonski and Chaplin, 2010). Major changes in human skin have also been reported and associated with habitation of xeric environments characterized by dryness, high temperatures, and high levels of UV-B exposure (Elias et al., 2009). By contrast, genomic scans discovered variations in EDAR and EDA2R for human hair follicle variation and SLC24A5 and SLC45A2 in pigmentation in human skin evolution (Sabeti et al., 2007) and for which the EDAR V370A variant was further functionally determined to affect hair thickness as well as a higher density of sweat glands (Kamberov et al., 2013). As much as we have found compelling evidence for positive selection in the skin barrier in modern, ancient humans, and primates, it is likely that the evolution of EDC genes is underestimated. The shared homology within the paralogous members of the FLG-like (SFTP), SPRR, LCE, and S100 families in the EDC, including gene fusion in FLG-like genes, provides further evidence of the evolution and innovation of the EDC arising via gene duplication and repeat expansion but has been difficult to glean from short sequencing reads and downstream alignments (Cabral et al., 2001b; Strasser et al., 2014). Current sequencing strategies also could have contributed to the lack of fully assessing the evolution of members of the LCE (Late Cornified Envelope) gene family (one exon, average 350 bp coding length) that were tested but did not reach significance in our analyses. Furthermore, structural variation including copy number variation, gene duplication, and tandem repeat expansions have been known to contribute to both primate and human evolution (Teumer and Green, 1989; Dumas et al., 2007; Vanhoutteghem et al., 2008; Conrad et al., 2010). It is clear that we need more comparative genomic analyses to better understand the historical events that have shaped skin barrier.
function and the selection for these genetic variants. In doing so, we will be better equipped to interpret contemporary variation as it pertains to modern disease. Nevertheless, future experiments will address the functional impact for the variants underlying positive selection in the EDC.

**AUTHOR CONTRIBUTIONS**

ZG performed all the experiments. ZG and CdGS did the analyses and wrote the paper.

**FUNDING**

The work cited in this publication was supported in part by FUNDING and wrote the paper.

**REFERENCES**

Anisimova, M., Bielawski, J. P., and Yang, Z. (2001). Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Mol. Biol. Evol.* 18, 1585–1592. doi: 10.1093/oxfordjournals.molbev.a003945

Anisimova, M., Nielsen, R., and Yang, Z. (2003). Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. *Genetics* 164, 1229–1236.

Anisimova, M., and Yang, Z. (2007). Multiple hypothesis testing to detect lineages under positive selection that affects only a few sites. *Mol. Biol. Evol.* 24, 1219–1228. doi: 10.1093/molbev/msm042

Arenas, M., and Posada, D. (2010). Coalescent simulation of intracodon recombination. *Genetics* 184, 429–437. doi: 10.1534/genetics.109.109736

Arenas, M., and Posada, D. (2014). “The influence of recombination on the estimation of selection from coding sequence alignments,” in *Natural Selection: Methods and Applications*, ed. M. A. Fares (Boca Raton, FL: CRC Press), 112–125.

Bakewell, M. A., Shi, P., and Zhang, J. (2007). More genes underwent positive selection in chimpanzee evolution than in human evolution. *Proc. Natl. Acad. Sci. U.S.A.* 104, 7489–7494. doi: 10.1073/pnas.0701705104

Brandtzæg, P., Gabrielsen, T. O., Dale, I., Müller, F., Steinbakk, M., and Fagerhol, M. K. (1995). The leucocyte protein L1 (calprotectin): a putative nonspecific defence factor at epithelial surfaces. *Adv. Exp. Med. Biol.* 1219–1228. doi: 10.1093/molbev/msm042

Brown, S. J., and McLean, W. H. I. (2012). One remarkable molecule: filaggrin. *J. Invest. Dermatol.* 132, 751–762. doi: 10.1038/jid.2011.393

Brown, S. J., Sandilands, A., Zhao, Y., Liao, H., Retlon, C. L., Meggitt, S. J., et al. (2008). Prevalent and low-frequency null mutations in the filaggrin gene are associated with early-onset and persistent atopic eczema. *J. Invest. Dermatol.* 128, 1591–1594. doi: 10.1038/sj.jid.5701206

Cabral, A., Sayin, A., de Winter, S., Fischer, D. F., Pavel, S., and Bachendorf, C. (2001a). SPRR4, a novel cornified envelope precursor: UV-dependent epidermal expression and selective incorporation into fragile envelopes. *J. Cell Sci.* 114, 3837–3843.

Cabral, A., Voskamp, P., Cleton-Jansen, A. M., South, A., Nizetic, D., and Bachendorf, C. (2001b). Structural organization and regulation of the small proline-rich family of cornified envelope precursors suggest a role in adaptive barrier function. *J. Biol. Chem.* 276, 19231–19237. doi: 10.1074/jbc.M100336200

Candi, E., Schmidt, R., and Melino, G. (2005). The cornified envelope: a model of cell death in the skin. *Nat. Rev. Mol. Cell Biol.* 6, 328–340. doi: 10.1038/nrm1619

Carrier, D. R., Kapoor, A. K., Kimura, T., and Nickels, M. K. (1984). The energetic paradox of human running and hominid evolution [and comments and reply]. *Curr. Anthropol.* 25, 483–495. doi: 10.1086/203165

Chimpanzee Sequencing and Analysis Consortium (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437, 69–87. doi: 10.1038/nature04072

Clark, A. G., Glanowski, S., Nielsen, R., Thomas, P. D., Kejarival, A., Todd, M. A., et al. (2003). Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science* 302, 1960–1963. doi: 10.1126/science.1088821

Clark, H. L., Jhingran, A., Sun, Y., Varechon, C., de Jesus Carrion, S., Skaar, E. P., et al. (2016). Zinc and manganese chelation by neutrophil S100A8/A9 (Calprotectin) limits extracellular aspergillus fumigatus hyphal growth and corneal infection. *J. Immunol.* 196, 336–344. doi: 10.4049/jimmunol.1502037

Conrad, D. P., Pinto, D., Redon, R., Feuk, L., Gokcen, O., Zhang, Y., et al. (2010). Origins and functional impact of copy number variation in the human genome. *Nature* 464, 704–712. doi: 10.1038/nature08516

Dumas, L., Kim, Y. H., Karimpour-Fard, A., Cox, M., Hopkins, J., Pollack, J. R., et al. (2007). Gene copy number variation spanning 60 million years of human and primate evolution. *Genome Res.* 17, 1266–1277. doi: 10.1101/gr.6557307

Eckert, R. L., Broome, A.-M., Ruse, M., Robinson, N., Ryan, D., and Lee, K. (2004). S100 Proteins in the Epidermis. *J. Invest. Dermatol.* 123, 23–33. doi: 10.1111/j.0022-202X.2004.22719.x

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797. doi: 10.1093/nar/gkh340

Elias, P. M., Menon, G., Wetzel, B. K., and Williams, J. J. W. (2009). Evidence that stress to the epidermal barrier influenced the development of pigmentation in humans. *Pigment Cell Melanoma Res.* 22, 420–434. doi: 10.1111/j.1755-148X.2009.00588.x

Glasner, R., Harder, J., Lange, H., Bartels, J., Bartels, J., Christophs, E., et al. (2005). Antimicrobial psoriasin (S100A7) protects human skin from *Escherichia coli* infection. *Nat. Immunol.* 6, 57–64. doi: 10.1038/ni1142

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., et al. (2010). A draft sequence of the Neandertal genome. *Science* 328, 710–722. doi: 10.1126/science.1180801

Hardman, M. J., Moore, L., Ferguson, M. W., and Byrne, C. (1999). Barrier formation in the human fetus is patterned. *J. Invest. Dermatol.* 113, 1106–1113. doi: 10.1046/j.1523-1747.1999.00800.x

Henry, J., Toulza, E., Hsu, C.-Y., Pellerin, L., Balica, S., Mazereeuw-Hautier, J., et al. (2012). Update on the epidermal differentiation complex. *Front. Biosci.* (Landmark Ed) 17:1517–1532. doi: 10.2741/4001

Hsu, C.-Y., Henry, J., Raymond, A.-A., Mèchin, M.-C., Pendaries, V., Nassar, D., et al. (2011). Deimination of human filaggrin-2 promotes its proteolysis by calpain 1. *J. Biol. Chem.* 286, 23222–23233. doi: 10.1074/jbc.M110.197400

Irvine, A. D., McLean, W. H. I., and Leung, D. Y. M. (2011). Filaggrin mutations and primate evolution. *Genome Res.* 21, 1106–1113. doi: 10.1101/gr.161976.110

Jablonski, N. G., and 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. doi: 10.1073/pnas.0914628107

**ACKNOWLEDGMENT**

We thank Gerald Dorn and John Edwards for critical reading of the manuscript.

**SUPPLEMENTARY MATERIAL**

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fgene.2016.00227/full#supplementary-material
Jackson, B., Tilli, C. M. L. J., Hardman, M. J., Avilion, A. A., MacLeod, M. C., Ashcroft, G. S., et al. (2005). Late cornified envelope family in differentiating epidermis—response to calcium and ultraviolet irradiation. J. Invest. Dermatol. 124, 1062–1070. doi: 10.1111/j.0022-202X.2005.23699.x

Jeong, C., and Di Rienzo, A. (2014). Adaptations to local environments in modern human populations. Curr. Opin. Genet. Dev. 29, 1–8. doi: 10.1016/j.gde.2014.06.011

Jiang, Y., Xie, M., Chen, W., Talbot, R., Maddox, J. F., Faraut, T., et al. (2014). The human genome at 10X coverage. Cell 159, 969–987. doi: 10.1016/j.cell.2014.12.012

Kamberov, Y. G., Wang, S., Tan, J., Gerbault, P., Wark, A., Tan, L., et al. (2013). Modeling recent human evolution in mice by expression of a selected EDAR variant. Cell 152, 691–702. doi: 10.1016/j.cell.2013.01.016

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., et al. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28, 1647–1649. doi: 10.1093/bioinformatics/bts199

Kosiol, C., Vinar, T., da Fonseca, R. R., Hubisz, M. J., Bustamante, C. D., Nielsen, R., et al. (2008). Patterns of positive selection in six Mammalian genomes. PLoS Genet. 4:e1000141. doi: 10.1371/journal.pgen.1000141

Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M. F., Parker, B. J., Washietl, S., et al. (2011). A high-resolution map of human evolutionary constraint using 29 mammals. Nature 478, 476–482. doi: 10.1038/nature10530

Margolis, D. J., Gupta, J., Apt, A. J., Ganguly, T., Hoffman, O., Papadopoulos, M., et al. (2014). Filaggrin-2 variation is associated with more persistent atopic dermatitis in African American subjects. J. Allergy Clin. Immunol. 133, 1–6. doi: 10.1016/j.jaci.2013.09.015

Marvel, K. W., George, M. D., Fujimoto, W., Saunders, N. A., Bernacki, S. H., and Jetten, A. M. (1992). Corinnin, a cross-linked envelope precursor in keratinocytes that is down-regulated by retinoids. Proc. Natl. Acad. Sci. U.S.A. 89, 11026–11030. doi: 10.1073/pnas.89.22.11026

Meyer, M., Kircher, M., Gansauge, M.-T., Li, H., Racimo, F., Mallick, S., and Lunter, G. (2014). A map of recent human evolutionary constraint using 141 genomes. Nature 518, 425–431. doi: 10.1038/nature14170

Murphy, W. J., Eizirik, E., O’Brien, S. J., Madsen, O., Scally, M., Douady, C. J., et al. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28, 1647–1649. doi: 10.1093/bioinformatics/bts199

Smedley, D., Haider, S., Durinck, S., Pandini, L., Provero, P., Allen, J., et al. (2015). The BioMart community portal: an innovative alternative to large, centralized data repositories. Nucleic Acids Res. 43, W589–W598. doi: 10.1093/nar/gkv350

Strasser, B., Mizt, V., Hermann, M., Rice, R. H., Eigenheer, R. A., Alibardi, L., et al. (2014). Evolutionary origin and diversification of epidermal barrier proteins in anamniote. Mol. Biol. Evol. 12, 2394–2405. doi: 10.1093/molbev/msu251

Teumer, J., and Green, H. (2009). Divergent evolution of part of the involucrin gene in the hominoids: unique intragenic duplications in the gorilla and human. Proc. Natl. Acad. Sci. U.S.A. 86, 1283–1286. doi: 10.1073/pnas.86.4.1283

Vanhuyttem, A., Dijian, P., and Green, H. (2008). Ancient origin of the gene encoding involucrin, a precursor of the cross-linked envelope of epidermis and related epithelia. Proc. Natl. Acad. Sci. U.S.A. 105, 15481–15486. doi: 10.1073/pnas.0807643105

Voight, B. F., Kudaravalli, S., Wen, X., and Pritchard, J. K. (2006). A scan for positively selected genes in the genomes of humans and chimpanzees. Nature 449, 913–918. doi: 10.1038/nature06250

Wong, W. S. W., Yang, Z., Goldman, N., and Nielsen, R. (2004). Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. Genetics 168, 1041–1051. doi: 10.1034/j.1600-0528.2004.005227

Yang, Z. (1998). Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol. Biol. Evol. 15, 568–573. doi: 10.1093/oxfordjournals.molbev.a025957

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24, 1586–1591. doi: 10.1093/molbev/msm088

Yang, Z., Wong, W. S. W., and Nielsen, R. (2005). Bayes empirical bayes inference of amino acid sites under positive selection. Mol. Biol. Evol. 22, 1107–1118. doi: 10.1093/molbev/msi097

Yates, A., Akanni, W., Amode, M. R., Barrett, D., Billis, K., Carvalho-Silva, D., et al. (2016). Ensembl 2016. Nucleic Acids Res. 44, D710–D716. doi: 10.1093/nar/gkv1157

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2017 Goodwin and de Guzman Strong. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.