Regulatory Network Decoded from Epigenomes of Surface Ectoderm-Derived Cell Types

Correspondence: Jeffrey B. Cheng, MD, PhD, University of California, San Francisco, Department of Dermatology, 1701 Divisadero Street, 4th Floor, San Francisco, CA 94143, chengjb@derm.ucsf.edu.

Author Contributions
R.F.L., J.F.C., T.W., and J.B.C. designed the study; T.M., B.B., and T.D.T. supervised sample collection and processing; P.G., M.S., and J.B.C. processed cell samples; P.J.F., M.A.M, M.H., J.F.C, and J.B.C. designed and supervised library production and sequencing assays; R.F.L., B.Z., and T.W. performed data analysis; M.B. and D.L. contributed computational tools for data processing and quality control; R.F.L., J.F.C., T.W., and J.B.C. wrote the manuscript.

Competing financial interests
The authors declare no competing financial interests.

Accession codes
Accession codes for keratinocyte skin01 MeDIP-seq, MRE-seq, mRNA-seq, miRNA-seq, H3K4me3 ChIP-seq, H3K4me1 ChIP-seq, and input ChIP-seq datasets have been deposited in the Gene Expression Omnibus (GEO) database under the accession codes GSM707022, GSM707018, GSM751278, GSM817253, GSM669589, GSM669591, and GSM817242 respectively. Accession codes for keratinocyte skin02 MeDIP-seq, MRE-seq, mRNA-seq, miRNA-seq, H3K4me3 ChIP-seq, H3K4me1 ChIP-seq, and input ChIP-seq datasets have been deposited in the GEO database under the accession codes GSM941726, GSM941723, GSM941745, GSM1127113, GSM941735, GSM941736, and GSM941742 respectively. Accession codes for keratinocyte skin03 MeDIP-seq, MRE-seq, mRNA-seq, miRNA-seq, WGBS, H3K4me3 ChIP-seq, H3K4me1 ChIP-seq, H3K27ac ChIP-seq, and input ChIP-seq datasets have been deposited in the GEO database under the accession codes GSM958180, GSM958169, GSM958177, GSM1127111, GSM1127056/GSM1127058, GSM958155, GSM958161, GSM958156, and GSM958167 respectively. Accession codes for fibroblast skin01 MeDIP-seq, MRE-seq, mRNA-seq, miRNA-seq, H3K4me3 ChIP-seq, H3K4me1 ChIP-seq, and input ChIP-seq datasets have been deposited in the GEO database under the accession codes GSM707021, GSM707017, GSM751277, GSM817252, GSM817235, GSM817234, and GSM817246 respectively. Accession codes for fibroblast skin02 MeDIP-seq, MRE-seq, mRNA-seq, H3K4me3 ChIP-seq, H3K4me1 ChIP-seq, and input ChIP-seq datasets have been deposited in the GEO database under the accession codes GSM941725, GSM941722, GSM941744, GSM941718, GSM941717, and GSM817247 respectively. Accession codes for fibroblast skin03 MeDIP-seq, MRE-seq, mRNA-seq, miRNA-seq, H3K4me3 ChIP-seq, H3K4me1 ChIP-seq, H3K27ac ChIP-seq, and input ChIP-seq datasets have been deposited in the GEO database under the accession codes GSM958182, GSM958171, GSM958178, GSM1127116, GSM958158, GSM958164, GSM958163, and GSM958168 respectively. Accession codes for melanocyte skin01 MeDIP-seq, MRE-seq, mRNA-seq, miRNA-seq, H3K4me3 ChIP-seq, H3K4me1 ChIP-seq, and input ChIP-seq datasets have been deposited in the GEO database under the accession codes GSM707020, GSM707016, GSM751276, GSM817251, GSM941719, GSM941728, and GSM941740 respectively. Accession codes for melanocyte skin02 MeDIP-seq, MRE-seq, mRNA-seq, H3K4me3 ChIP-seq, H3K4me1 ChIP-seq, and input ChIP-seq datasets have been deposited in the GEO database under the accession codes GSM941727, GSM941724, GSM941743, GSM941731, GSM941730, and GSM941741 respectively. Accession codes for melanocyte skin03 MeDIP-seq, MRE-seq, mRNA-seq, H3K4me3 ChIP-seq, H3K4me1 ChIP-seq, H3K27ac ChIP-seq, and input ChIP-seq datasets have been deposited in the GEO database under the accession codes GSM958181, GSM958170, GSM958174, GSM958151, GSM958152, GSM958157, and GSM958166 respectively. Accession codes for breast luminal epithelia RM071 MeDIP-seq and MRE-seq datasets have been deposited in the GEO database under the accession codes GSM1517154 and GSM613826 respectively. Accession codes for breast luminal epithelia RM080 mRNA-seq, H3k4me1 ChIP-seq, and input ChIP-seq datasets have been deposited in the GEO database under the accession codes GSM669620, GSM669595, and GSM959124 respectively. Accession codes for breast myoepithelia RM071 MeDIP-seq and MRE-seq datasets have been deposited in the GEO database under the accession codes GSM1517153 and GSM613908 respectively. Accession codes for breast myoepithelia RM080 H3K4me3 ChIP-seq, H3k4me1 ChIP-seq and input ChIP-seq datasets have been deposited in the GEO database under the accession codes GSM693277, GSM613885, and GSM613897 respectively. Accession codes for the Fetal Brain Germinal Matrix HuFGM02 WGBS dataset have been deposited in the GEO database under the accession code GSM941747. Accession codes for PBMC TC015 H3K4me3 ChIP-seq, H3k4me1 ChIP-seq, and input ChIP-seq datasets have been deposited in the GEO database under the accession codes GSM613811, GSM613814, and GSM613816 respectively. Accession codes for CD8 Naïve TC001 H3K4me3 ChIP-seq, H3k4me1 ChIP-seq, and input ChIP-seq datasets have been deposited in the GEO database under the accession codes GSM1127126, GSM1127143, and GSM1127151 respectively.

Rebecca F. Lowdon[1], **Bo Zhang**[1], **Misha Bilenky**[3], **Thea Mauro**[6], **Daofeng Li**[1], **Philippe Gascard**[4], **Mahvash Sigaroudinia**[4], **Peggy J. Farnham**[5], **Boris C. Bastian**[6], **Thea D. Tlsty**[4], **Marco A. Marra**[3], **Martin Hirst**[3,7], **Joseph F. Costello**[2], **Ting Wang**[1], and **Jeffrey B. Cheng**[6]

[1]Dept. of Genetics, Center for Genome Sciences and Systems Biology, Washington University in St. Louis, MO 63108

[2]Dept. of Neurological Surgery, Helen Diller Family Comprehensive Cancer Center, UC-San Francisco, CA 94143

[3]Canada's Michael Smith Genome Sciences Centre, BC Cancer Agency, Vancouver, British Columbia, Canada V5Z 1L

[4]Department of Pathology, Center for Translational Research in the Molecular Genetics of Cancer, UCSF 94143

[5]Department of Biochemistry & Molecular Biology, Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, CA

[6]Department of Dermatology, University of California San Francisco, CA 94143

[7]Centre for High-Throughput Biology, Department of Microbiology & Immunology, University of British Columbia, Vancouver, British Columbia, Canada V6T 1Z4

Abstract

Developmental history shapes the epigenome and biological function of differentiated cells. Epigenomic patterns have been broadly attributed to the three embryonic germ layers. Here we investigate how developmental origin influences epigenomes. We compare key epigenomes of cell types derived from surface ectoderm (SE), including keratinocytes and breast luminal and myoepithelial cells, against neural crest-derived melanocytes and mesoderm-derived dermal fibroblasts to identify SE differentially methylated regions (SE-DMRs). DNA methylomes of neonatal keratinocytes share many more DMRs with adult breast luminal and myoepithelial cells than with melanocytes and fibroblasts from the same neonatal skin. This suggests that SE origin contributes to DNA methylation patterning, while shared skin tissue environment has limited effect on epidermal keratinocytes. Hypomethylated SE-DMRs are in proximity to genes with SE relevant functions. They are also enriched for enhancer- and promoter-associated histone modifications in SE-derived cells, and for binding motifs of transcription factors important in keratinocyte and mammary gland biology. Thus, epigenomic analysis of cell types with common developmental origin reveals an epigenetic signature that underlies a shared gene regulatory network.

Introduction

While epigenetic mechanisms are crucial in establishing and maintaining cell identity, the role of developmental origin and tissue microenvironment in shaping the epigenome is just beginning to be unraveled. Marked epigenomic transitions occur upon directed embryonic stem cell differentiation into the three major embryonic lineages[1,2] and over the course of development[3]. Differentiated cells and tissues have specific DNA hypomethylation signatures, particularly at enhancers[4,5]; however, a subset of hypomethylated enhancers are

actually dormant in adult tissues and active only in corresponding fetal tissues suggesting that a DNA methylation memory of fetal origin may be retained in adult cells[6]. Similarly, DNase I-hypersensitive patterns in differentiated cells can reflect embryonic lineage and mark a subset of embryonic enhancers[7]. Tissue microenvironment influences cell identity and morphogenesis[8] and consequently, may affect epigenomes. Accordingly, perturbation of tissue microenvironment is associated with epigenomic alteration[9,10]. These studies suggest that embryonic origin and tissue environment may influence normal cellular epigenomic states and that differentiated cell epigenomes can be utilized to infer epigenomic patterns of precursor embryonic cell populations.

To investigate how developmental origin and tissue environment contribute to cell type-specific epigenetic patterns, we utilize skin as a model system. The three most prevalent skin cell types are each derived from a different embryonic origin (keratinocytes from surface ectoderm, fibroblasts from mesoderm, and melanocytes from neural crest[11]), but exist within a shared tissue environment (Figure 1). We generate DNA methylation and histone modification profiles for these three skin cell types and compared their epigenomes among the skin cell types and against breast, blood, and brain tissue epigenomes. The three skin cell types share few regions with common DNA methylation and histone modification states, that were not also present in the other tissue samples. Surface-ectoderm derived skin keratinocytes and breast cells however, share many common differentially DNA methylated regions (SE-DMRs). SE-DMRs are enriched for enhancer- and promoter-associated histone modifications in SE-derived cell types and for binding motifs of relevant transcription factors. Reconstruction of the gene regulatory network connecting these transcription factors and putative target genes with nearby SE-DMRs demarcated epigenetic and regulatory events associated with structural components and signaling pathways in SE-derived cell types. Thus, for surface ectoderm-derived cells, their shared developmental origin influences their epigenomes to a greater extent than tissue environment. Furthermore, a shared gene regulatory network emerged from the SE-DMR signature.

Results

Skin cell type-specific differentially methylated regions

Fibroblasts, melanocytes, and keratinocytes were individually isolated from each of three neonatal human foreskins and expanded as short-term primary cultures. From these samples, we generated nine high-resolution epigenomes encompassing key histone modifications (H3K4me1, H3K4me3, and H3K27ac) and DNA methylation, along with mRNA and miRNA expression profiles (Supplementary Data 1 and 2). The effects of aging and environmental exposure were minimized by utilizing neonatal samples. Since each set of three cell types shares a common genome, the effect of genetic variation on epigenetic variability was also minimized.

We identified 12,892 regions encompassing 193,202 CpGs with a DNA methylation status unique to one of the three skin cell types and consistent across all three individuals (Methods, Figure 2a, Supplementary Fig. 1–3, 4a, Supplementary Notes 1–3, Supplementary Table 1). The majority of these skin cell type-specific DMRs were hypomethylated (Figure 2a), suggesting potential cell type-specific regulatory activity at these regions[4,12,13]. 40–

46% of the DMRs were intergenic and 5–9% were associated with RefSeq annotated gene promoters (Supplementary Fig. 5); non-CGI promoters were enriched among cell type-specific DMRs (Supplementary Note 4; Supplementary Table 2). 80–91% of hypomethylated cell type-specific DMRs overlapped with regulatory element-associated histone modifications in the same cell type (Figure 2b). Accordingly, hypomethylation of cell type-specific DMRs at gene promoters correlated with increased gene expression relative to the other two cell types where the DMR was hypermethylated (Figure 2c, Supplementary Tables 3–5). Gene Ontology (GO) analysis using the GREAT[14] tool on hypomethylated cell type-specific DMRs showed strong enrichment for biological processes relevant to each cell type (e.g. extracellular matrix organization for fibroblasts (*P*-value=9.05E-45) and pigmentation for melanocytes (*P*-value=2.43E-06); Figure 2d; Supplementary Data 3). These data suggest skin cell type-specific DMRs occur primarily at distal enhancers and regulate genes relevant to each cell type.

Skin tissue-specific epigenomic features

We next examined whether the common tissue environment of the three skin cell types would impose an identifiable skin tissue epigenetic signature. For comparison, we generated complete epigenomes and transcriptomes for a panel of non-skin cell types and tissues (including brain tissue and breast and blood cell types) and identified DMRs shared by all three skin cell types relative to other tissues (Supplementary Fig. 4b, Supplementary Note 5). Of the 28,776 total DMRs detected, only 8 regions shared the same methylation status in skin cell types and the opposite methylation status in all other samples (Figure 3a, b). Hierarchical clustering based on methylation levels at the 28,776 DMRs reveals that while samples of the same cell type cluster together, the three skin cell types do not (Figure 3c). These results suggested that skin cell type methylomes did not share many differences compared to breast, brain, and blood cell methylomes and that skin tissue lacks a specific and substantive DNA methylation signature.

To determine whether skin tissue also lacks a shared histone modification signature, we identified cell type-specific chromatin states from H3K4me1, H3K4me3, and H3k27ac ChIP-seq data for each skin cell type, as well as for breast, brain and blood samples. Among the 259,297 enhancer-associated H3K4me1 peaks and 55,859 promoter-associated H3K4me3 peaks identified in the above samples, only 997 H3K4me1 and 57 H3K4me3 peaks are present in all three skin cell types and absent in the other samples (Supplementary Fig. 6,7). Only 100 of the 997 exclusively skin-specific H3K4me1 peaks overlapped with H3K27ac peaks (a combination which marks active enhancers) in all three skin cell types (Figure 3d). While GO enrichment analysis for cell type-specific histone modification patterns showed enrichment for relevant terms, analysis for the few exclusively skin tissue shared histone modification peaks did not reveal any relevant enrichment (Supplementary Data 4). The minimal DNA methylation and histone modification commonalities that separate skin cell types from other tissues and the lack of functional enrichment for these common shared regions strongly suggest that the shared skin tissue environment does not significantly influence its constituent cell type epigenomes at this developmental stage.

Developmental origin influences epigenomes

In the absence of a strong skin tissue-specific epigenetic signature, we hypothesized that developmental origin is a major determinant of skin cell type epigenetic patterns. We explored this hypothesis by focusing on skin keratinocytes and breast epithelial cells, which are both derived from surface ectoderm[15]. Consistent with their shared developmental origin, neonatal skin keratinocytes clustered with adult breast epithelial cell types based on DNA methylation values at the DMRs previously identified in skin and non-skin cell pairwise comparisons (Figure 3c). To specifically define the DNA methylation signature of surface ectoderm-derived cell types, we identified DMRs for each of the surface ectodermal cell types in a pairwise manner compared to neonatal skin melanocytes and fibroblasts, which are derived from other embryonic germ layers (Supplementary Fig. 4c). There were 1,392 DMRs with the same methylation state exclusively in keratinocyte, breast myoepithelial, and breast luminal epithelial cells relative to the two other cell types, which we inferred to be surface ectoderm-specific DMRs (SE-DMRs) (Methods, Figure 4a). Therefore, common developmental origin influences surface ectoderm-derived cell epigenomes to a greater extent than does the shared skin tissue environment.

We examined whether SE-DMRs, like cell type-specific DMRs, possessed regulatory potential. The majority (97%) of surface ectoderm DMRs (SE-DMRs) were hypomethylated with 12% located in gene promoters and 40% within intergenic regions (Supplementary Fig. 8a). Hypomethylated SE-DMRs were enriched for promoter- and enhancer-associated histone modifications in both keratinocytes and breast myoepithelial cells, and for DNase I-hypersensitive sites in keratinocytes (Figure 4b, Supplementary Fig. 8b). Hypomethylated SE-DMRs were also enriched for transcription factor binding motifs including TFAP2 and KLF4 (Figure 4c); transcription factors that bind to these two motifs function in keratinocyte and mammary epithelium development, differentiation, and/or maintenance of cell fate[16–20]. Genes associated with hypomethylated SE-DMRs were enriched for functions relevant to the biology of these cell types, such as "epidermis development" (*P*-value=4.35e-15) and "mammary gland epithelium development" (*P*-value=2.10e-9) (Figure 4d, Supplementary Data 5). DNA hypomethylation status of genes with hypomethylated SE-DMRs in their promoter regions correlated with increased expression in SE-derived cells relative to non-SE cells (Figure 4e, Supplementary Table 6). These annotations suggested that the majority of surface ectoderm-DMRs were at distal enhancer or gene promoter elements and regulate genes important for keratinocyte and mammary gland development. More generally, these results offer a new and deeper level of interrogating the origin and function of adult epigenomes, adding significantly to the recent attribution of epigenome signatures to germinal layers[1,2].

Epigenome-derived surface ectoderm regulatory network

Given their regulatory element signatures, overlap with DNase I-hypersensitive sites, and enrichment for relevant transcription factor binding site (TFBS) motifs, we hypothesized that hypomethylated SE-DMRs may be regulatory elements that coordinate expression of genes essential for function of surface ectoderm-derived cells. To test this, we sought to connect these putative regulatory elements to genes in a surface ectoderm gene network. We associated DMRs with nearby putative target genes and queried databases of TF-target

genes and gene-gene interactions to construct regulatory relationships among these genes (Methods). The result is a highly connected network with a statistically significant number of connections (1458 edges, 278 nodes; *P*-value=1.25e-4; Methods; Supplementary Data 6, Supplementary Table 7), whose distribution follows a power law (R^2=0.89; Supplementary Fig. 9).

Strikingly, the transcription factors near the top of the inferred SE network were those whose motifs were enriched in the hypomethylated SE-DMRs (Figure 4c). This observation, along with the network connectivity data, suggested that TFAP2a, TFAP2c, and KLF4 may regulate many of the downstream genes in this network. To identify biological processes associated with each set of hypomethylated DMRs containing either TFAP2 or KLF4 TFBSs, we performed GREAT analysis[14]. The network was characterized by two partially overlapping major branches (summarized data in Figure 5a, Supplementary Table 8). The first branch included the transcription factors TFAP2a and TFAP2c and connected to genes associated with surface ectoderm relevant GO terms, e.g. "hemidesmosome assembly" which is a structural complex critical for epithelial cells[21] and Notch signaling which functions in mammary cell fate committment[22] and keratinocyte homeostasis[23] (Figure 5b). The second branch was characterized by KLF4 and associated with mammary gland development and Wnt signaling which influences both breast and keratinocyte cell fate decisions[24,25] (Figure 5c). Thus, we observed a highly structured set of connections between regulatory elements and putative target genes that underlie and integrate signaling pathways vital for both keratinocyte and mammary gland epithelial cell function.

Surface ectoderm hypomethylated DMRs were located near the TSS of six genes that encode hemidesmosome/epidermal basement membrane zone components, five of which contain the TFAP2 TFBS motif (Figure 5e). These genes were highly expressed in all surface ectodermal cell types (Figure 5d). Mutations occur in any one of five of these genes in various forms of the inherited epidermolysis bullosa blistering skin diseases [26,27]. These findings suggest SE-DMRs may coordinately regulate a suite of genes that encode for components of a key structural complex in surface ectoderm-derived cells, that when perturbed leads to a clinically relevant phenotype.

Hypomethylated SE-DMRs containing TFAP2 motifs were also identified near the transcription start site of two genes, *IRF6* and *Stratifin*, that are highly expressed in surface ectoderm-derived cells (Figure 6a–d). *IRF6* is a transcription factor, known to be regulated by TFAP2a[28], that coordinates keratinocyte and breast epithelium proliferation and differentiation[29,30]. *Stratifin* is a member of the 14-3-3 protein family which functions as an adaptor protein and binds to phosphorylated proteins mediating diverse cellular processes, such as cell cycle control, apoptosis, and keratinocyte differentiation[31]. Stratifin promoter DNA hypermethylation and expression downregulation is found in both breast and skin cancers[32]. Mutations in *IRF6* or *SFN* lead to similar phenotypes with limb and craniofacial developmental abnormalities and an impaired skin barrier due to defective keratinocyte differentiation[33,34].

A KLF4 motif containing hypomethylated SE-DMR was noted near the mir-200c/141 locus. These two microRNAs promote epithelial cell fate and mir-200c/141 expression is often lost

in breast cancers[35]. Our findings of mir-200c/141 surface ectoderm-specific expression and DNA hypomethylation (Figure 6e, f) are consistent with previously demonstrated epigenetic regulation of this locus[36]. Thus, SE-DMRs may modulate key genes that regulate proliferation, differentiation, and epithelial cell fate maintenance in surface ectoderm-derived cells.

Developmental dynamics of SE regulatory elements

To explore the developmental dynamics of DNA methylation at SE-DMRs, we obtained whole genome bisulfite sequencing data for samples representing early stages in surface ectoderm development: H1 embryonic stem cells (ESCs) and ESCs differentiated to represent an early ectoderm developmental stage[2]. A majority of hypomethylated SE-DMRs were methylated in both early developmental stages, but hypomethylated in keratinocytes and mammary gland epithelia (Methods, Figure 7a). The few exceptions are transcription factors that are upstream in the regulatory hierarchy. For example, the DMR near the *TFAP2a* promoter was demethylated in ES cells, whereas the DMR in *KLF4* was methylated in ES cells but demethylated in early surface ectoderm differentiated cells. Both genes are most highly expressed in keratinocytes (Figure 7b–e). The remaining hypomethylated SE-DMRs, many of which putatively regulate genes that are TFAP2a, TFAP2c, or KLF4 targets in the network analysis, were lowly methylated in differentiated cells. Accordingly, expression of these genes was generally increased in keratinocytes relative to H1 ESCs (Figure 7f). Additionally, hypomethylated SE-DMRs were highly methylated in fetal brain tissue, which is predominantly neuroectoderm-derived, concordant with their specific assignment to surface ectoderm-derived cells rather than embryonic ectoderm as a whole (Supplementary Fig. 10).

Discussion

Analysis of an increasingly diverse collection of epigenomes has revealed tissue- and cell type-specific regulatory elements important for cell fate and development[13,37–40]. However, the developmental origins of these epigenomic features have been less explored. Studies utilizing in vitro ESC differentiation systems have uncovered early developmental DNA methylation dynamics that are believed to occur with specification of the embryonic germ layers[1,2]. There is a growing realization that this developmental lineage-specific information is maintained in differentiated cells, as DNA methylation and DNase I hypersensitive site profiles of cell types and tissues cluster by their embryonic germ layer of origin[6,7]. The persistence of a subset of DNA hypomethylated enhancers, which are active in early development but quiescent in adulthood, also suggests a developmental memory is encoded in the epigenome of differentiated cells[6].

Here we present our analysis of the epigenomic features of human skin cell types and their origins. In our experimental design, we used three different skin cell types from the same individual, and identified DNA methylation signatures which are consistent for three individuals across each cell type, minimizing variables that confound many other study designs including genetic background, age, and external environmental exposures. Consistent with findings in other cell types, we found many skin cell type-specific DMRs at

distal enhancers, enriched for association with cell type-relevant genes, and correlated with expression at hypomethylated promoters. Thus we demonstrated that, as expected, the cell types within skin tissue possess many regions with cell type-specific epigenomic patterns.

Next we assessed whether the shared environment within skin tissue imparts common epigenomic features upon its constituent cell types to create a skin tissue-specific signature. To investigate this question, we developed an approach to identify "shared differences" between epigenomes. This approach prioritized specificity and minimized the influence of variation between biological replicates. Thus, shared epigenomic signatures should be robust to sources of variation and attributable to the common biological factor of the grouped samples, for example, the shared tissue environment of skin cell types. Utilizing this approach on the skin cell type epigenomes revealed few shared regions compared to epigenomes of other tissues and cell types, suggesting that skin tissue environment had little uniform impact on the epigenomes of its constituent cell types.

Since tissue environment had minimal effect on skin cell type epigenomes, we hypothesized that developmental origin may influence differentiated cell epigenomes and confer features specific to their shared origin. We compared the DNA methylomes of surface ectoderm-derived cells, epidermal keratinocytes and breast luminal and myoepithelial cells, to methylomes of non-SE-derived cells to identify "shared differences." We found that SE-derived cell types share many DMRs when compared to non-SE derived cells and that these DMRs possess regulatory potential. This suggests that the common developmental origin of these surface ectoderm-derived cells impacts their epigenomes, and that this influence is greater than that of tissue environment on keratinocyte methylomes.

To gain better insight into the SE-DMR signature, which we defined indirectly through neonatal and adult cell epigenomes, we identified target genes putatively regulated by SE-DMRs and then connected these genes based on known interactions[41] (Methods). The resulting SE network predicted both upstream regulators and co-regulated suites of genes. Transcription factors predicted to bind to SE-DMRs (Figure 4c) were encoded by genes with the highest number of connections in the network (Supplementary Data 6, Supplementary Fig. 9). The presence of SE-DMRs containing TFAP2 TFBSs near the transcription start site of hemidesmosome genes suggests their co-regulation by TFAP2 family transcription factors. Additionally, TFAP2 TFBS-containing SE-DMRs are found near the TSSs of the cell cycle regulators *IRF6* and *SFN*. Given the genetic interaction of these two genes in epidermal development[30], a TFAP2 motif-binding transcription factor may coordinately co-regulate their expression in SE-derived cells. These examples of predicted regulatory relationships illustrate the significant value afforded by incorporating epigenetically-defined regulatory elements into gene networks.

A more direct approach to define epigenomic features that arise from a developmental origin would involve isolation and profiling of human embryonic tissues and their derivatives at various time points along a single developmental lineage and comparing their epigenomes and transcriptomes. As this type of experiment is not possible for ethical reasons, we selected cell types arising from a major germ layer derivative, surface ectoderm, to infer for the first time a DNA methylation signature derived from this inaccessible human embryonic

cell population. Our approach builds upon previous studies that utilized induced differentiation of ESCs to elucidate DNA methylation patterns of the three main embryonic germ layers[1,2]. Our SE-specific signature findings substantially extend the general concept that epigenomes of differentiated cell types cluster by their embryonic origin[6,37]. We demonstrate that a gene network regulating shared biological processes and functional components can be decoded from DNA methylation profiles of cell types specifically chosen for their common embryonic origin. Thus, analysis of differentiated cell types with shared developmental origin may be widely applicable for inference of regulatory epigenomic states derived from other inaccessible precursor human cell populations.

Methods

Cell Type and Tissue Isolation

Fibroblasts, keratinocytes, and melanocytes were isolated from neonatal foreskins obtained from circumcision using standard techniques[42]. Briefly, epidermis was mechanically separated from dermis after overnight incubation at 4 degrees Celsius with dispase solution. The epidermal sheet was incubated with trypsin for 15 minutes at 37 degrees Celsius. The disassociated cells were then incubated in selective growth media. Keratinocytes were grown in keratinocyte growth media (Medium 154CF supplemented with 0.07 mM $CaCl_2$ and Human Keratinocyte Growth Supplement (Life Technologies)). Melanocytes were grown in melanocyte growth media ((Medium 254 with Human Melanocyte Growth Supplement (Life Technologies)). Fibroblasts were extracted from the dermis by mincing and digesting with collagenase. The cell suspension was plated in Medium 106 supplemented with Low Serum Growth Supplement (Life Technologies). All skin cell types were harvested after two passages by snap freezing in liquid Nitrogen.

A pure population of keratinocytes was verified by examination of cell morphology and immunofluoresence staining for keratinocyte markers (cytokeratin (acidic), clone AE1, Life Technologies, 18-0153) and lack of staining for melanocyte markers (HMB45+Mart-1+Tyrosinase cocktail, Biocare Medical, CM165 or Mel-5, Covance, Sig-38150). A pure population of melanocytes was verified by examination of cell morphology and immunofluoresence staining for melanocyte markers and lack of staining for keratinocyte markers. A pure population of fibroblasts was verified by examination of cell morphology and positive staining for vimentin (Sigma, V6630) and lack of staining for keratinocyte and melanocyte markers.

Breast, blood, and fetal brain samples were isolated as previously described[4]. Briefly, for blood cell types, peripheral blood mononuclear cells (PBMCs) were isolated from buffy coat using Histopaque 1077 separation medium (Sigma-Aldrich) according to the manufacturer's protocol. CD4 naïve, CD4 memory, and CD8 naïve cells were isolated from PBMCs using the following isolation kits: EasySep Human Naive CD4+ T Cell Enrichment Kit, EasySep Human Memory CD4+ T Cell Enrichment Kit, and Custom Human Naive CD8+ T Cell Enrichment Kit (Stemcell Technologies). Pure populations of PBMCs and T cell subsets were confirmed by staining with the following antibodies (anti-CD3 TRI-COLOR (Invitrogen), anti-CD4 PE (BD Biosciences), anti-CD8 FITC (BD Biosciences), anti-CD4

TRI-COLOR (Invitrogen), anti-CD45RO PE (Invitrogen), anti-CD45RA FITC (BD Biosciences), and anti-CD8 TRI-COLOR (Invitrogen)) and FACS analysis.

Briefly, for breast cell types, breast tissue from disease-free premenopausal women was obtained from reduction mammoplasty samples under UCSF CHR protocol #10-01563. Tissue was mechanically and enzymatically dissociated with collagenase and hyaluronidase. Cell suspensions were serially filtered through 150-um and 40-um nylon mesh to obtain epithelial cell enriched clusters (breast cell organoids). To obtain single cell suspensions, organoids were further digested with trypsin and dispase and filtered with a 40-um cell strainer followed by incubation for 60–90 minutes in MEGM medium (Lonza). The resulting cells were stained and sorted by FACS to isolated purified breast myoepithelial and luminal epithelial cells. For positive selection, a PE-Cy7 labelled anti-CD10 antibody (for myoepithelial cells, BD Biosciences, 341092) and a FITC labelled anti-CD227/MUC1 antibody (for luminal epithelial cells, BD Biosciences, 559774) were used. For negative selection of hematopoietic, endothelial, and leukocyte cells, cells were stained with the following antibodies respectively: anti-CD2, -CD3, CD16, CD64 (BD Biosciences, 555325, 555338, 555405, and 555526); CD31 (Invitrogen, MHCD3115); and CD45, CD140b (BioLegend, 304003 and 323604).

Briefly, for fetal brain samples, brain tissue was obtained post-mortem from fetuses whose death was attributed to environmental/placental etiology, under Partner's Healthcare/ Brigham and Women's Hospital IRB Protocol #2010P001144. Fetal brain germinal matrix was grossly dissected.

Genomic DNA isolation

Cells were lysed in extraction buffer (50 mM Tris (pH 8.0), 1 mM EDTA (pH 8.0), 0.5 % SDS, and 1 mg/ml proteinase K) at 55 degrees Celsius for 12–16 hours. The lysed cells were incubated with 40 ug/ml of RNase A for 1 hour at 37 degrees Celsius to remove RNA. DNA was purified by two rounds of phenol/chloroform/isoamyl alcohol extractions and then two rounds of chloroform extractions. DNA was precipitated with 1/10 volume of 3 M sodium acetate (pH 5.2) and 2.5 volumes of ethanol, washed in 70% ethanol, and resuspended in TE.

Methylation-sensitive restriction enzyme (MRE)-seq

MRE-seq was performed as in Maunakea, et al.[2] with modifications as detailed below. Five parallel restriction enzyme digestions ((HpaII, Bsh1236I, SsiI(AciI) and Hin6I (Fermentas) and HpyCH4IV (NEB)) were performed, each using 1 ug of DNA per digest for each of the skin cell type samples. Five units of enzyme were initially incubated with DNA for 3 hours and then an additional five units of enzyme was added to the digestion for a total of 6 hours of digestion time. DNA was purified by phenol/chloroform/isoamyl alcohol extraction, followed by chloroform extraction using phase lock gels. Digested DNA from the different reactions was combined and precipitated with 1/10 volume of 3 M sodium acetate (pH 5.2) and 2.5 volumes of ethanol. The purified DNA was size selected and purified (50–300 bp) by gel electrophoresis and Qiagen MinElute extraction. Library construction was performed as per the Illumina Genomic DNA Sample Prep Kit protocol with the following

modifications. During the end repair reaction, T4 DNA polymerase and T4 PNK were excluded and 1 uL of 1:5 diluted Klenow DNA polymerase was utilized. For the adapter ligation reaction, 1 uL of 1:10 diluted PE adapter oligo mix was utilized. 10 uL from the 30 uL of purified adapter ligated DNA was utilized for the PCR enrichment reaction with PCR PE Primers 1.0 and 2.0. PCR products were size selected and purified (170–420 bp) by gel electrophoresis and Qiagen Qiaquick extraction. DNA libraries were checked for quality by Nanodrop (Thermo Scientific) and Agilent DNA Bioanalyzer (Agilent).

Reads were aligned to hg19 using BWA, and pre-processed using methylQA (an unpublished C program; available at http://methylqa.sourceforge.net/). MRE reads were normalized to account for differing enzyme efficiencies, and methylation values were determined by counting reads with CpGs at fragment ends[43]. To enable comparison between MRE-seq data from blood, brain, and breast samples which utilized three restriction enzymes and skin cell types which utilized five restriction enzymes, skin cell type MRE reads that resulted from the use of additional restriction enzymes (Bsh1236I and HpyCH4IV) were removed. Detailed library construction protocols for MRE-seq, MeDIP-seq, ChIP-seq, RNA-seq, and miRNA-seq are publicly available at the NIH Roadmap Epigenomics project website http://www.roadmapepigenomics.org/protocols/type/experimental/

Methylated DNA immunoprecipitation (MeDIP)-seq

MeDIP-seq was performed as in Maunakea et al.[2]. 5 ug of genomic DNA was sonicated to a fragment size of ~100–400 bp using a Bioruptor sonicator (Diagenode). End-repair, addition of 3′ A bases, and PE adapter ligation with 2 ug of sonicated DNA was performed as per the Illumina Genomic DNA Sample Prep Kit protocol. Adapter-ligated DNA fragments were size selected to 166–366 bp and purified by gel electrophoresis. DNA was heat denatured and then immunoprecipitated with 5-Methylcytidine antibody (Eurogentec) (1 ug of antibody per 1 ug of DNA) in 500 uL of immunoprecipitation buffer (10 uM sodium phosphate, pH 7.0, 140 mM sodium chloride, and 0.05% Triton X-100) overnight at 4 degrees Celsius. Antibody/DNA complexes were isolated by addition of 1 uL of rabbit anti-mouse IgG secondary antibody (2.4 mg/ml, Jackson Immunoresearch) and 100 uL protein A/G agarose beads (Pierce Biotechnology) for 2 hours at 4 degrees C. Beads were washed six times with immunoprecipitation buffer and then DNA was eluted in TE buffer with 0.25% SDS and 0.25 mg/ml of proteinase K for 2 hours at 50 degrees Celsius. DNA was then purified with the Qiagen Qiaquick kit and eluted in 30 uL EB buffer. 10 ul of DNA was utilized for a PCR enrichment reaction with PCR PE Primers 1.0 and 2.0. PCR products were size selected (220–420 bp) and purified by gel electrophoresis. Methylated DNA enrichment was confirmed by PCR on known methylated (SNRPN and MAGEA1 promoters) and unmethylated (a CpG-less sequence on chromosome 15 and GADPH promoter) sequences. DNA libraries were checked for quality by Nanodrop (Thermo Scientific) and Agilent DNA Bioanalyzer (Agilent). Reads were aligned to hg19 using BWA, and preprocessed using methlyQA.

methylCRF

Genome-wide DNA methylation value predictions were made using a conditional random field model that integrates MRE and MeDIP sequencing data for a given sample. The program was run using default parameters[44], and can be downloaded from http://methylcrf.wustl.edu/. In Figure 7, methylCRF predicted values were averaged for each DMR.

Differential DNA Methylated Region Analysis

The M&M statistical model[4] which integrates MeDIP-seq and MRE-seq data to identify differentially methylated regions between two samples was implemented with a window size of 500 bp and a q-value (FDR corrected p-value) cutoff = 1e-5. Scripts utilized for pair-wise comparison are shown in Supplementary Note 2. Adjacent 500 bp DMRs were merged into a single DMR for further analysis unless otherwise noted. The specific pairwise comparisons performed to generate each DMR set are summarized in Supplementary Fig. 4. Additional details and discussion of the DMR calling strategy and false discovery rate for M&M analyses are in Supplementary Notes 1 and 3. Comprehensive lists of identified skin cell type specific DMRs are available online (http://epigenome.wustl.edu/SE).

Whole Genome Bisulfite Sequencing

1–5 ug of Qubit quantified genomic DNA was utilized for library construction. Unmethylated Lambda DNA (Promega) was added to genomic DNA for a 0.1% final concentration. DNA was fragmented to ~300 bp using Covaris E series shearing. End-repair, addition of 3′ A bases, and adapter ligation was performed as per the Illumina PE Genomic DNA Sample Prep Kit protocol except methylated cytosine PE adapters were used. After each of the previous steps, DNA was purified using Ampure XP beads (Agencourt). Bisulfite conversion of purified adapter ligated DNA was performed using the Epitect bisulfite kit (Qiagen) according to manufacturer's instructions. The DNA was amplified by PCR enrichment using Kapa HiFi Hot Start Uracil+Ready (Kapa Biosystems) for 5 cycles with PCR PE primers 1.0 and 2.0. PCR products were purified with the Qiagen Minelute kit and size selected with PAGE gel purification. DNA libraries were checked for quantity by Qubit (Life Technologies) and quality by Agilent DNA Bioanalyzer (Agilent). Libraries were sequenced using paired-end 100 nt sequencing chemistry on an Illumina HiSeq2000 following manufacturer's protocols (Illumina).

Raw WGBS sequences were examined for quality, sample swap and reagent contamination using custom in house scripts. Sequence reads were directionally aligned to the human genome (GRCh37-lite) using Bismark[5] v. 0.7.6) running Bowtie[45] (v. 0.12.5) allowing up to two mismatches in the 50 bp seed region (using -n 2 -l 50 parameters). Methylation status for each aligned CpG was calculated using Bismark Methylation Extractor (v. 0.7.10) at a minimum of 5x coverage per site in a strand-specific manner (run-time parameters: -p, no_overlap, --comprehensive, --bedGraph, --counts). Overlapping methylation calls from read_1 and read_2 were scored once.

All WGBS data was processed using custom scripts to obtain CpG methylation values. CpG methylation values were filtered such that only CpGs with 10x coverage were subsequently

averaged for each DMR in each sample. Lowly methylated regions were called as DMRs for which the average CpG methylation values were 0.3. Averaged values were plotted as in Figure 4a using the R package *pheatmaps*.

ChIP-seq

Standard operating procedures for ChIP-seq library construction are available at http://www.roadmapepigenomics.org/protocols/type/experimental/. ChIP-seq library construction involves the following protocols in order: 1) Crosslinking of frozen cell pellet, 2) DNA sonication using Sonic Dismembrator 550, and 3) SLX-PET protocol for Illumina sample prep. Antibodies used in this study were subjected to rigorous quality assessment to meet Reference Epigenome Mapping Quality Standards (http://www.roadmapepigenomics.org/protocols) including western blot of whole cell extracts, 384 peptide dot blot (Active Motif MODified Histone Peptide Array) and ChIP-seq using control cell pellets (HL60). Antibody vendor, catalog number and lot are provided along with ChIP-seq library construction details as part of the metadata associated with all ChIP-seq datasets and available through GEO and the NCBI epigenomics portals (e.g. http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM669589). Final library distributions were calculated using an Agilent Bioanalyzer and quantified by fluorometric quantification (Qubit, Life Technologies). Libraries were sequenced using single-end 76 nt sequencing chemistry on an Illumina GAiix or HiSeq2000 following manufacturer's protocols (Illumina) as either single or multiplexed libraries using custom index adapters added during library construction.

Sequencing reads were aligned to NCBI GRCh37-lite reference using BWA 0.6.2-r126 with default parameters. MethylQA (an unpublished C program; available at http://methylqa.sourceforge.net/) was used to directionally extend aligned reads to the average insert size of DNA fragments (150 bp) and to generate a bigWig file for downstream visualization. Reads with BWA mapping quality scores < 10 were discarded and reads that aligned to the same genomic coordinate were counted only once.

Differential ChIP-seq enrichment analysis

Mapped read density was generated from aligned sequencing reads using methlyQA. Read density overlapping DMRs and their 5 kb upstream/downstream regions were extracted at 50 bp resolution as RPKM values.

The default parameters were used to apply MACS2[46] to histone modification ChIP-seq data for the identification of peaks at a 1% false discovery rate. A DMR was defined as enriched for histone signal when at least 60% of the DMR overlapped with histone peaks. Skin cell type-specific histone peaks were identified using the following two criteria: 1) peaks were identified in at least two of three biological replicates of a skin cell type and 2) peaks were not identified in any of the other two skin cell types or other tissue types (brain, breast, and blood). Skin tissue-specific histone peaks were identified using the following three criteria: 1) peaks were identified in at least two of three biological replicates of a skin cell type, 2) peaks were identified in all three skin cell types, and 3) peaks were not identified in any other tissue type (brain, breast, or blood).

Genomic features

CpG islands, gene bodies, and RefSeq gene annotations (including 5′ and 3′ UTRs, exons, and introns) were downloaded from the UCSC Genome Browser. Promoters were defined as the 3.5 kb surrounding the TSS (−3 kb/+500 bp) of all RefSeq genes. Intergenic regions were defined as all regions outside RefSeq gene bodies and promoters.

Gene ontology enrichment analysis

Gene Ontology (GO) analyses for biological processes were performed using the GREAT package[14]. Gene regulatory domains were defined by default as the regions spanning 5 kb upstream and 1 kb downstream of the TSS (regardless of other nearby genes). Gene regulatory domains were extended in both directions to the nearest gene's basal domain but no more than a maximum extension in one direction. Only categories that were below a false discovery rate of 0.05 were reported.

Transcription Factor Binding Site Identification Enrichment

Genome sequences were obtained for hypomethylated SE-DMRs from the hg19 human genome assembly. Motif finding analysis was performed using the FIMO tool from the MEME suite and default vertebrate databases[47,48], with a q-value (FDR-corrected p-value) cutoff of 0.04 Motif enrichment was calculated as the number of motif instances found in the test data compared to the number found genome wide (for hg19), normalized for length.

$$E(motif) = \frac{\frac{n_{DMRs}}{820000}}{\frac{N_{hg19}}{3200000000}} \quad (1)$$

where n_{DMRs} = number of a given motif found in the hypomethylated SE-DMRs and N_{hg19} = number of a given motif found in hg19. 820000 = number of base pairs in hypomethylated SE-DMRs; 3200000000 = number of base pairs in the human genome.

Regulatory network construction

Regulatory networks were constructed in the following steps. First, genes (nodes) were identified as putative targets of regulatory (hypomethylated) SE-DMRs either by their association with DMRs that fell in the promoter region of RefSeq genes (−3 kb/+500 bp TSS) or by association as distal intergenic regulatory elements. Intergenic DMRs were associated with genes whose TSS fell in a window of +/− 35 kb (This window size is chosen based on literature assessing the average distance of enhancer-promoter associations[38]). The collection of these promoter- and distal enhancer-associated gene lists were then filtered for a gene expression level 1 RPKM in any of the surface ectoderm-derived cell types.

To obtain interactions between genes in this list, the gene list was used as nodes in the UCSC Interaction Browser[41]. The Interaction Browser queries known databases for connections (links) between a given set of genes (nodes). Four pathway collections (GEA_CLR TF-targets network; UCSC_Superpathway; UCSC_Superpathway_collapsed; CHEA transcription factors) were used to query for interactions between the given genes. For the SE-DMR network, KLF4 was added to the gene list because its motif was enriched

in hypomethylated SE-DMRs (Figure 2c) and because it is known to be important for keratinocyte differentiation[49]. *Klf4* does have two hypomethylated SE-DMRs in its second exon, suggesting it is regulated, but the exonic location of the *Klf4* DMRs excluded it from the stringent method for identifying putatively regulated genes, above. Similarly, TFAP2C was added to the gene list because it is known to be important in keratinocyte differentiation[16] and its motif (shared with TFAP2A) was enriched in our motif analysis (Figure 2c). For the network overview presented in Figure 3a, the transcription factor p63 was added at the top of the network as it integrates both network branches, is a known regulator of the ZNF750–KLF4 transcriptional cascade[50], and interacts genetically with TFAP2a/c[51,52]; however, p63 and its edges are not included in the data (Supplementary Data 6) or network structural analysis (Supplementary Table 7, Supplementary Fig. 9).

We applied the same method for generating links between a set of 374 random genes to obtain an expected distribution of links given the number of genes in the test network. This resulted in a distribution as described in Supplementary Table 7 with a mean of 958 and variance of 136.5. By a *t*-test, the number of links in the SE network is statistically significant (*P*-value=1.245e-4). To assess the scale-free properties of the SE network, we calculated the number of edges assigned to each node and plotted this distribution in Supplementary Fig. 9.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

References

1. Xie W, et al. Epigenomic analysis of multilineage differentiation of human embryonic stem cells. Cell. 2013; 153:1134–1148. [PubMed: 23664764]

2. Gifford CA, et al. Transcriptional and epigenetic dynamics during specification of human embryonic stem cells. Cell. 2013; 153:1149–1163. [PubMed: 23664763]

3. Lister R, et al. Global epigenomic reconfiguration during mammalian brain development. Science. 2013; 341:1237905. [PubMed: 23828890]

4. Zhang B, et al. Functional DNA methylation differences between tissues, cell types, and across individuals discovered using the M&M algorithm. Genome Res. 2013; 23:1522–1540. [PubMed: 23804400]

5. Schlesinger F, et al. De novo DNA demethylation and noncoding transcription define active intergenic regulatory elements. Genome Res. 2013; 23:1601–1614. [PubMed: 23811145]

6. Hon GC, et al. Epigenetic memory at embryonic enhancers identified in DNA methylation maps from adult mouse tissues. Nat Genet. 2013; 45:1198–1206. [PubMed: 23995138]

7. Stergachis AB, et al. Developmental fate and cellular maturity encoded in human regulatory DNA landscapes. Cell. 2013; 154:888–903. [PubMed: 23953118]

8. Nelson CM, Bissell MJ. Of extracellular matrix, scaffolds, and signaling: tissue architecture regulates development, homeostasis, and cancer. Annu Rev Cell Dev Biol. 2006; 22:287–309. [PubMed: 16824016]

9. Dumont N, et al. Sustained induction of epithelial to mesenchymal transition activates DNA methylation of genes silenced in basal-like breast cancers. Proc Natl Acad Sci. 2008; 105:14867–14872. [PubMed: 18806226]

10. DesRochers TM, et al. The 3D tissue microenvironment modulates DNA methylation and E-cadherin expression in squamous cell carcinoma. Epigenetics Off J DNA Methylation Soc. 2012; 7:34–46.

11. James, WD. Andrews' Diseases of the skin: clinical dermatology. Saunders Elsevier; 2011.

12. Stadler MB, et al. DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature. 2011; 480:490–495. [PubMed: 22170606]

13. Xie M, et al. DNA hypomethylation within specific transposable element families associates with tissue-specific enhancer landscape. Nat Genet. 2013; 45:836–841. [PubMed: 23708189]

14. McLean CY, et al. GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol. 2010; 28:495–501. [PubMed: 20436461]

15. Sell, S. Stem cells handbook. 2013. at <http://dx.doi.org/10.1007/978-1-4614-7696-2>

16. Wang X, Pasolli HA, Williams T, Fuchs E. AP-2 factors act in concert with Notch to orchestrate terminal differentiation in skin epidermis. J Cell Biol. 2008; 183:37–48. [PubMed: 18824566]

17. Dai X, Segre JA. Transcriptional control of epidermal specification and differentiation. Curr Opin Genet Dev. 2004; 14:485–491. [PubMed: 15380238]

18. Yori JL, Johnson E, Zhou G, Jain MK, Keri RA. Kruppel-like factor 4 inhibits epithelial-to-mesenchymal transition through regulation of E-cadherin gene expression. J Biol Chem. 2010; 285:16854–16863. [PubMed: 20356845]

19. Zhang J, Brewer S, Huang J, Williams T. Overexpression of transcription factor AP-2α suppresses mammary gland growth and morphogenesis. Dev Biol. 2003; 256:128–146.

20. Jäger R, Werling U, Rimpf S, Jacob A, Schorle H. Transcription factor AP-2gamma stimulates proliferation and apoptosis and impairs differentiation in a transgenic model. Mol Cancer Res MCR. 2003; 1:921–929. [PubMed: 14573793]

21. Green KJ, Jones JC. Desmosomes and hemidesmosomes: structure and function of molecular components. FASEB J Off Publ Fed Am Soc Exp Biol. 1996; 10:871–881.

22. Bouras T, et al. Notch signaling regulates mammary stem cell function and luminal cell-fate commitment. Cell Stem Cell. 2008; 3:429–441. [PubMed: 18940734]

23. Okuyama R, Tagami H, Aiba S. Notch signaling: its role in epidermal homeostasis and in the pathogenesis of skin diseases. J Dermatol Sci. 2008; 49:187–194. [PubMed: 17624739]

24. Slavik MA, Allen-Hoffmann BL, Liu BY, Alexander CM. Wnt signaling induces differentiation of progenitor cells in organotypic keratinocyte cultures. BMC Dev Biol. 2007; 7:9. [PubMed: 17306035]

25. Alexander CM, Goel S, Fakhraldeen SA, Kim S. Wnt signaling in mammary glands: plastic cell fates and combinatorial signaling. Cold Spring Harb Perspect Biol. 2012; 4

26. Fine JD, et al. The classification of inherited epidermolysis bullosa (EB): Report of the Third International Consensus Meeting on Diagnosis and Classification of EB. J Am Acad Dermatol. 2008; 58:931–950. [PubMed: 18374450]

27. Karamatic Crew V, et al. CD151, the first member of the tetraspanin (TM4) superfamily detected on erythrocytes, is essential for the correct assembly of human basement membranes in kidney and skin. Blood. 2004; 104:2217–2223. [PubMed: 15265795]

28. Rahimov F, et al. Disruption of an AP-2alpha binding site in an IRF6 enhancer is associated with cleft lip. Nat Genet. 2008; 40:1341–1347. [PubMed: 18836445]

29. Bailey CM, Hendrix MJC. IRF6 in development and disease: a mediator of quiescence and differentiation. Cell Cycle Georget Tex. 2008; 7:1925–1930.

30. Richardson RJ, et al. Irf6 is a key determinant of the keratinocyte proliferation-differentiation switch. Nat Genet. 2006; 38:1329–1334. [PubMed: 17041603]

31. Medina A, Ghaffari A, Kilani RT, Ghahary A. The role of stratifin in fibroblast–keratinocyte interaction. Mol Cell Biochem. 2007; 305:255–264. [PubMed: 17646930]

32. Lodygin D, Hermeking H. Epigenetic silencing of 14-3-3sigma in cancer. Semin Cancer Biol. 2006; 16:214–224. [PubMed: 16698281]

33. Herron BJ, et al. A mutation in stratifin is responsible for the repeated epilation (Er) phenotype in mice. Nat Genet. 2005; 37:1210–1212. [PubMed: 16200063]

34. Ingraham CR, et al. Abnormal skin, limb and craniofacial morphogenesis in mice deficient for interferon regulatory factor 6 (Irf6). Nat Genet. 2006; 38:1335–1340. [PubMed: 17041601]

35. Gregory PA, et al. The miR-200 family and miR-205 regulate epithelial to mesenchymal transition by targeting ZEB1 and SIP1. Nat Cell Biol. 2008; 10:593–601. [PubMed: 18376396]

36. Neves R, et al. Role of DNA methylation in miR-200c/141 cluster silencing in invasive breast cancer cells. BMC Res Notes. 2010; 3:219. [PubMed: 20682048]

37. Ziller MJ, et al. Charting a dynamic DNA methylation landscape of the human genome. Nature. 2013; 500:477–481. [PubMed: 23925113]

38. Shen Y, et al. A map of the cis-regulatory sequences in the mouse genome. Nature. 2012; 488:116–120. [PubMed: 22763441]

39. Dunham I, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489:57–74. [PubMed: 22955616]

40. Varley KE, et al. Dynamic DNA methylation across diverse human cell lines and tissues. Genome Res. 2013; 23:555–567. [PubMed: 23325432]

41. Wong CK, et al. The UCSC Interaction Browser: multidimensional data views in pathway context. Nucleic Acids Res. 2013; 41:W218–224. [PubMed: 23748957]

42. Normand J, Karasek MA. A method for the isolation and serial propagation of keratinocytes, endothelial cells, and fibroblasts from a single punch biopsy of human skin. In Vitro Cell Dev Biol Anim. 1995; 31:447–455. [PubMed: 8589888]

43. Maunakea AK, et al. Conserved role of intragenic DNA methylation in regulating alternative promoters. Nature. 2010; 466:253–257. [PubMed: 20613842]

44. Stevens M, et al. Estimating absolute methylation levels at single-CpG resolution from methylation enrichment and restriction enzyme sequencing methods. Genome Res. 2013; 23:1541–1553. [PubMed: 23804401]

45. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009; 10:R25. [PubMed: 19261174]

46. Zhang Y, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008; 9:R137. [PubMed: 18798982]

47. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinforma Oxf Engl. 2011; 27:1017–1018.

48. Mathelier A, et al. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. Nucleic Acids Res. 201310.1093/nar/gkt997

49. Koster MI, Roop DR. Mechanisms Regulating Epithelial Stratification. Annu Rev Cell Dev Biol. 2007; 23:93–113. [PubMed: 17489688]

50. Sen GL, et al. ZNF750 Is a p63 Target Gene that Induces KLF4 to Drive Terminal Epidermal Differentiation. Dev Cell. 2012; 22:669–677. [PubMed: 22364861]

51. Koster MI, Kim S, Huang J, Williams T, Roop DR. TAp63α induces AP-2γ as an early event in epidermal morphogenesis. Dev Biol. 2006; 289:253–261. [PubMed: 16324689]

52. Antonini D, et al. An Autoregulatory Loop Directs the Tissue-Specific Expression of p63 through a Long-Range Evolutionarily Conserved Enhancer. Mol Cell Biol. 2006; 26:3308–3318. [PubMed: 16581802]
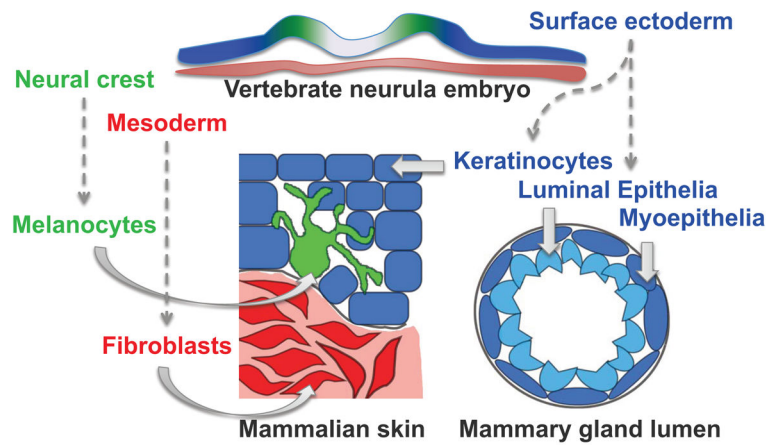
Figure 1. Developmental origins of samples

Developmental origins of skin and breast cell types utilized in this study. Embryonic surface ectoderm from the vertebrate neurula stage embryo (blue) gives rise to keratinocytes in the skin and cells of the mammary gland lumen. Embryonic neural crest cells (green) will produce melanocytes that intercalate with epidermal keratinocytes, and skin fibroblasts are derived from embryonic mesoderm (red).
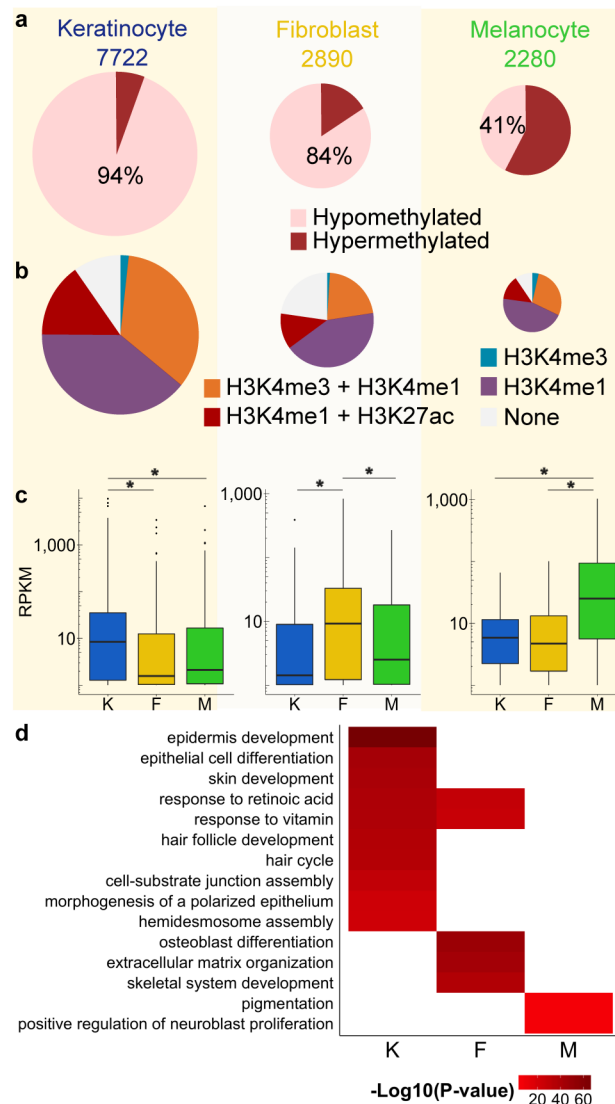
Figure 2. Identification and characterization of skin cell type-specific DMRs

(a) Hypomethylation and hypermethylation percentages for each set of skin cell type-specific DMRs defined by comparison against the other two skin cell types. The total number for each set of cell type-specific DMRs is listed above the pie chart. DMRs are 500bp windows.

(b) Histone modification patterns at skin cell type-specific hypomethylated DMRs.

(c) Skin cell type RNA expression levels for genes with hypomethylated cell type-specific DMRs in their promoter regions. Each panel depicts expression values for a set of cell type-specific DMR-associated genes. Plotted values are RNA-seq RPKM values over exons, averaged (mean) over three biological replicates. For each boxplot, the middle line indicates the median value, top and bottom box edges are the third and first quartile boundaries respectively. The upper whisker is the highest data value within 1.5 times the interquartile range; the lower whisker indicates the lowest value within 1.5 times the interquartile range. The interquartile range is the distance between the first and third quartiles. Points indicate

data beyond whiskers. Logarithmic scale transformations were applied before boxplot statistics were computed. RPKM distributions for a given set of cell type-specific DMR-associated genes in the specified cell type compared to other cell types were statistically significant (Wilcoxon ranked test, paired, * indicates *P*-value < 0.003, Keratinocyte-DMRs $n = 602$, Fibroblast-DMRs $n = 108$, Melanocyte-DMRs $n = 74$; K = keratinocytes, F = fibroblasts, M = melanocytes; Supplementary Tables 3–5).

(d) Heat map depicting selected gene ontology terms enriched for keratinocyte, fibroblast, and melanocyte hypomethylated cell type-specific DMRs. K = keratinocytes, F = fibroblasts, M = melanocytes. Color intensity represents the negative log10 transformed p-value of enrichment of a given cell type-specific DMR set for association with the listed gene ontology term. Full datasets are in Supplementary Data 3.
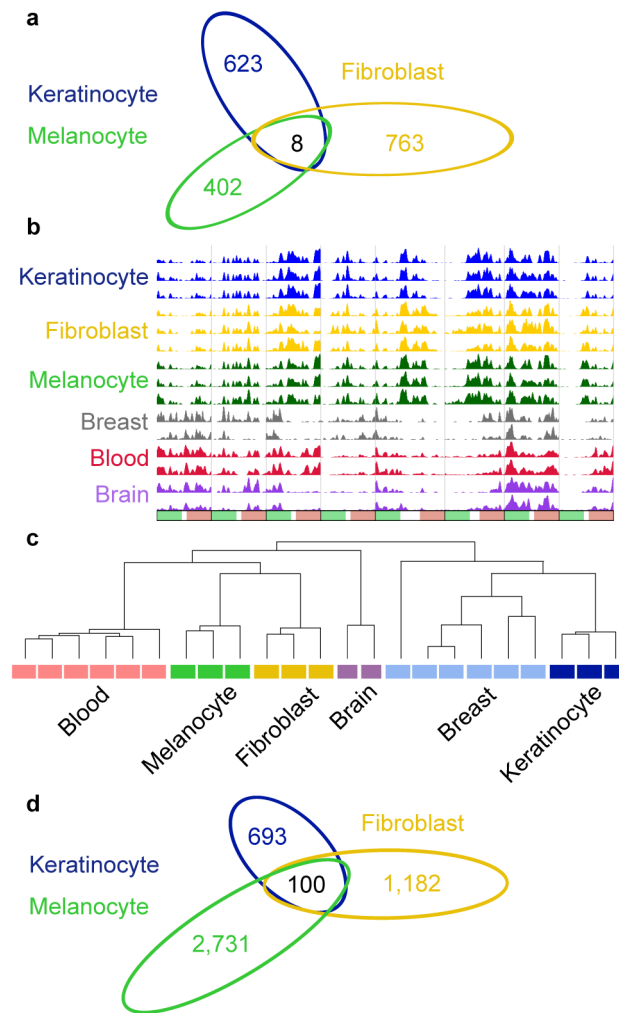
Figure 3. Skin-tissue level epigenomic features

(a) Venn diagram showing number of DMRs for each of the skin cell types compared to non-skin samples (brain, breast, and blood). 8 DMRs (overlap region) share the same methylation status in the three skin cell types and have the opposite methylation status in all non-skin samples.

(b) WashU Epigenome Browser screenshot of the 8 DMRs where the three skin cell types share the same methylation status and all non-skin cell types have the opposite methylation status. Each column represents a 500bp window +/− 2.5 kb except for two columns which represent multiple contiguous 500bp windows +/− 2.5 kb. Each row is a MeDIP-seq track for the indicated cell type. Three replicates for each skin cell type and two replicates for each non-skin sample are depicted.

(c) Clustering dendrogram based on average DNA methylation levels (predicted by methylCRF[44]) at 28,776 DMRs found between skin and brain tissue, breast, and blood cell types.

(d) Venn diagram showing number of H3K4me1 peaks for each skin cell type that are absent in all non-skin samples (brain, breast, and blood), which also have overlapping H3K27ac signal. The intersection represents the 100 overlapping regions where H3K4me1 and
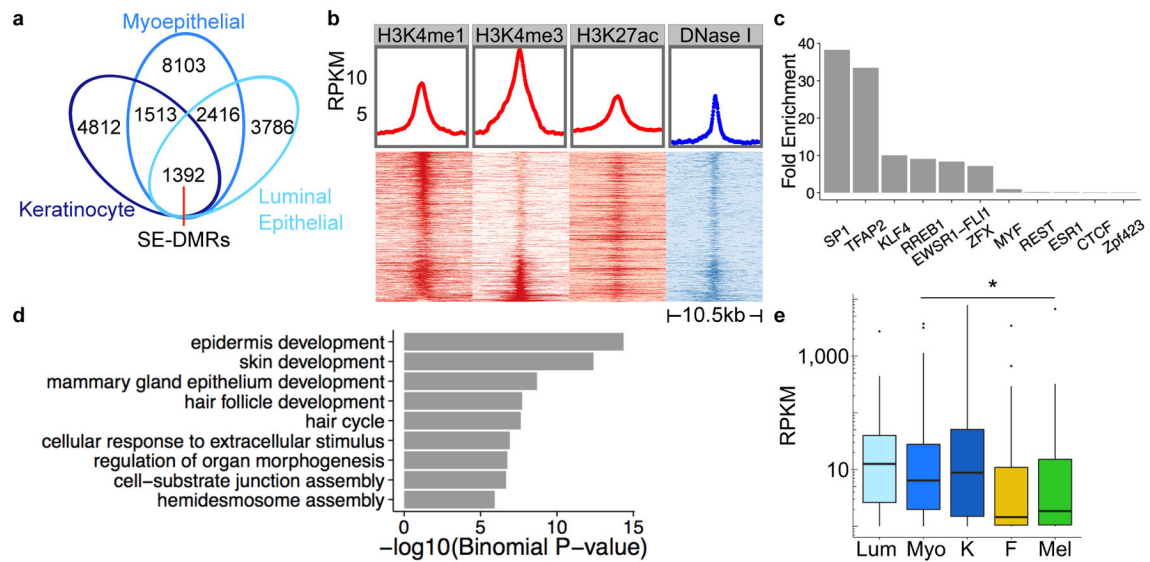
H3K27ac peaks are present in all three skin cell types and H3K4me1 peaks are absent in all non-skin samples.

Figure 4. Identification and characterization of surface ectoderm-DMRs

(a) Venn diagram showing surface ectoderm-specific DMRs, defined as the overlap of keratinocyte, breast myoepithelial, and luminal epithelial cell DMRs.

(b) Enrichment of H3K4me1, H3K4me3, H3K27ac, and DNAse I-hypersensitivity at SE-DMRs. Each heat map column represents histone modification ChIP-seq or DNAse-seq signal at 500bp SE-DMRs +/− 5 kb. Each heat map row represents a single hypomethylated SE-DMR, ordered by decreasing H3K4me1 signal, then increasing H3K4me3 signal.

(c) Bar plot of enrichment values for top ten enriched TFBS motifs determined by motif scanning of hypomethylated SE-DMRs using FIMO[47] (Methods). Enrichment based on hg19 genome background.

(d) Selected gene ontology terms enriched for hypomethylated surface ectoderm-DMRs. *P*-value of enrichment calculated by GREAT[14]. Full list of enriched GO terms is in Supplementary Data 5.

(e) Box plots showing RNA expression levels for genes with hypomethylated SE-DMRs in promoter regions. Skin cell type RNA-seq RPKM values over exons are averages (mean) of three biological replicates; luminal epithelial and myoepithelial values are a single biological replicate. The middle line indicates the median value, top and bottom box edges are the third and first quartile boundaries respectively. The upper whisker is the highest data value within 1.5 times the interquartile range; the lower whisker indicates the lowest value within 1.5 times the interquartile range. The interquartile range is the distance between the first and third quartiles. Points indicate data beyond whiskers. Logarithmic scale transformation was applied before boxplot statistics were computed. RPKM distributions for SE cell type expression levels vs. non-SE cell type expression levels are statistically significant (Wilcoxon-ranked test, paired, * indicates *P*-value < 0.02; *n* = 150 genes; Lum = breast luminal epithelial cells, Myo = breast myoepithelial cells, K= keratinocytes, F = fibroblasts, M = melanocytes; Supplementary Table 6).
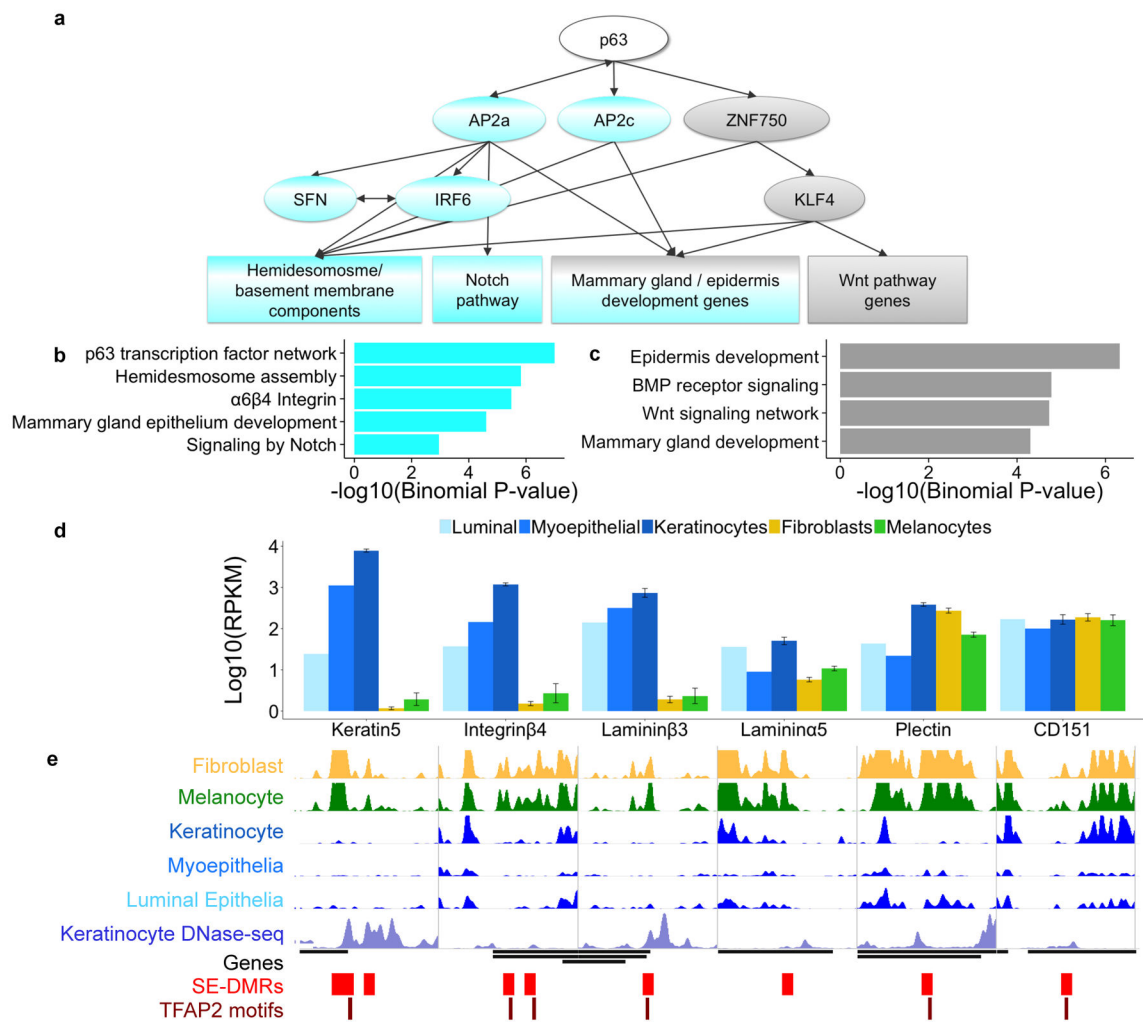
Figure 5. Surface ectoderm-DMRs are regulatory elements in a gene network

(a) Summary of the TF-target gene regulatory network derived from SE-DMR analyses. The categories at the bottom of the panel represent enriched biological processes or pathways for genes associated with DMRs containing TFAP2 or KLF4 motifs. TFAP2 associated TFs/ pathways highlighted in blue; KLF4 associated pathways in gray.

(b) Functional enrichment for TFAP2 motif containing hypomethylated SE-DMRs.

(c) Functional enrichment for KLF4 motif containing hypomethylated SE-DMRs.

(d) RNA expression values for SE-DMR associated hemidesmosome/basement membrane genes for SE and non-SE cell types. Skin cell type values are averages (mean) of three biological replicates. Error bars are standard error of the mean (s.e.m.).

(e) WashU Epigenome Browser screenshot of hemidesmosome/basement membrane genes. MeDIP-seq tracks depicted in green, yellow, and blue; all track y-axes heights are 60 RPKM. DNase-seq track is shown in light blue. Genes depicted as black lines. SE-DMRs depicted as red boxes and TFAP2 motifs as maroon lines.
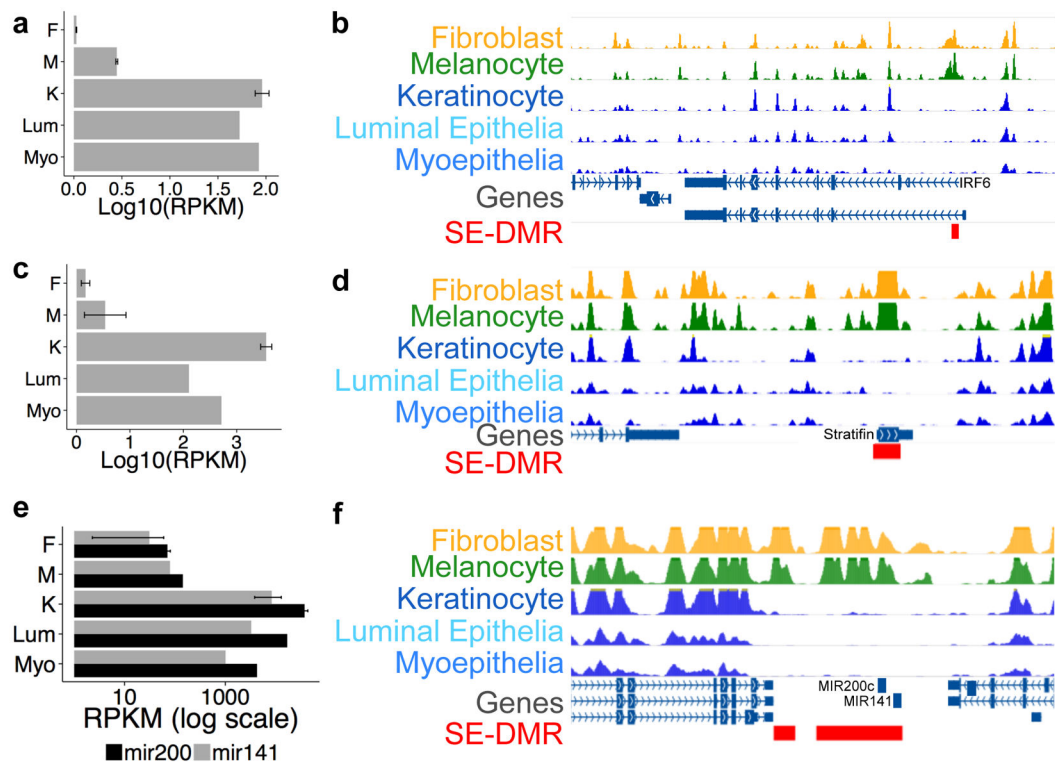
Figure 6. RNA expression levels and browser screenshots of selected loci with SE-DMRs

(a) Expression values for *IRF6* in each cell type as listed on the left. X-axis is expression in RPKM (log10 scale) for each cell type. Skin cell type values are averages (mean) of three biological replicates (error bars are s.e.m.); luminal epithelial and myoepithelial values are a single biological replicate.

(b) Browser screenshot of *IRF6* locus and surrounding genomic region. MeDIP-seq tracks are shown for the indicated cell types; all track y-axes heights are 60 RPKM. Red box = hypomethylated SE-DMR near the *IRF6* promoter.

(c) Expression for Stratifin (*SFN*) as in (a).

(d) Browser screenshot of *SFN* locus. Tracks as in (b). Red box = hypomethylated SE-DMR at *SFN* promoter.

(e) Expression values for mir-200c and mir-141 in each cell type as listed to the left. X-axis is reads per million (RPM, log scale). Keratinocyte value is the average (mean) of three biological replicates; fibroblast value is the mean of two biological replicates, (error bars are s.e.m.); melanocyte, luminal epithelial, and myoepithelial values are a single biological replicate.

(f) Browser screenshot of mir-200c/mir-141 locus and surrounding genomic region. Tracks as in (b). Red boxes = hypomethylated SE-DMRs including and adjacent to both miRNA loci.
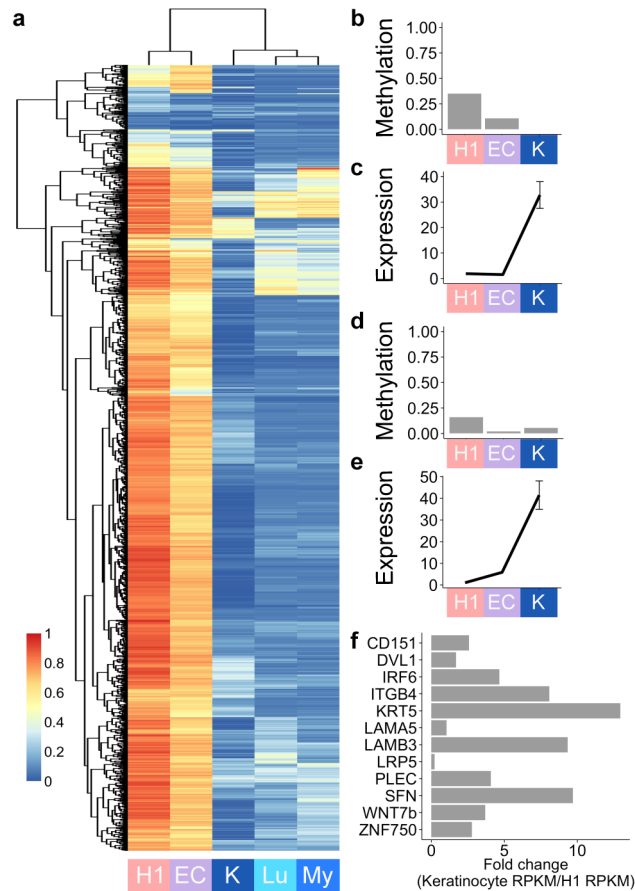
Figure 7. DNA methylation dynamics of SE-DMRs across samples from different developmental stages

(a) Heatmap and clustering dendrogram based on average CpG DNA methylation values of hypomethylated SE-DMRs for different developmental samples. Each row represents one of 1307 DMRs for which there are CpGs with   10x coverage in WGBS data. Methylation values for H1 ESCs, ectoderm differentiated ESCs ("EC"), and keratinocyte ("K") are from WGBS; breast luminal ("Lu") and myoepithelial ("My") values are the average of single CpG methylCRF predictions in each DMR. MethylCRF predictions are based on MeDIP-seq and MRE-seq data for these samples (Methods). A value of "1" is fully methylated; "0" is completely unmethylated.

(b) *KLF4* gene body SE-DMR average CpG DNA methylation levels across developmental stages.

(c) *KLF4* RNA expression across developmental stages. Values are RPKM over coding exons; error bars for keratinocytes are s.e.m., *n* = 3. Sample abbreviations as in (a).

(d) *TFAP2A* promoter SE-DMR average CpG DNA methylation levels across developmental stages.

(e) *TFAP2A* RNA expression across developmental stages. Values are RPKM over coding exons; error bars for keratinocytes are s.e.m., *n* = 3. Sample abbreviations as in (a).

(f) RNA expression levels in keratinocytes relative to H1 ESCs for selected genes with hypomethylated SE-DMRs in their promoters. These SE-DMRs, like the majority of

hypomethylated SE-DMRs, were methylated in H1 and ectoderm-differentiated ESCs but lowly methylated in differentiated SE cell types. Increased expression relative to an earlier developmental sample suggests these DMRs are transcriptional regulatory regions for their associated genes.