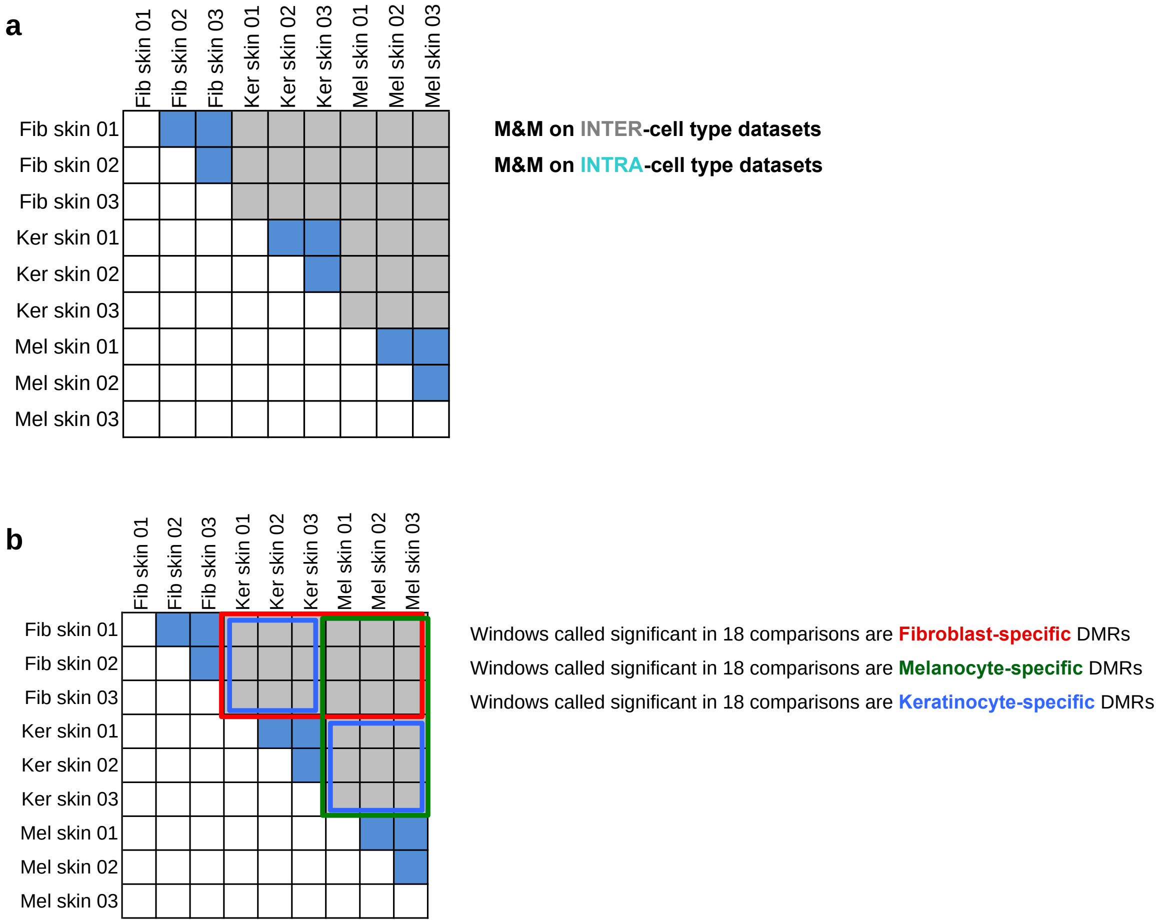
# Supplementary Information

## Supplementary Figures



Windows called significant in 18 comparisons are **Fibroblast-specific** DMRs

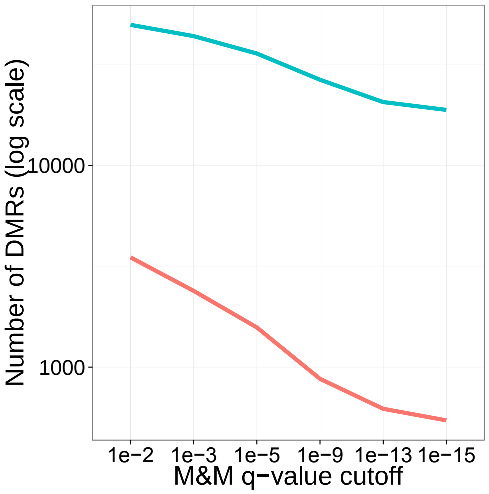
Windows called significant in 18 comparisons are **Melanocyte-specific** DMRs

Windows called significant in 18 comparisons are **Keratinocyte-specific** DMRs

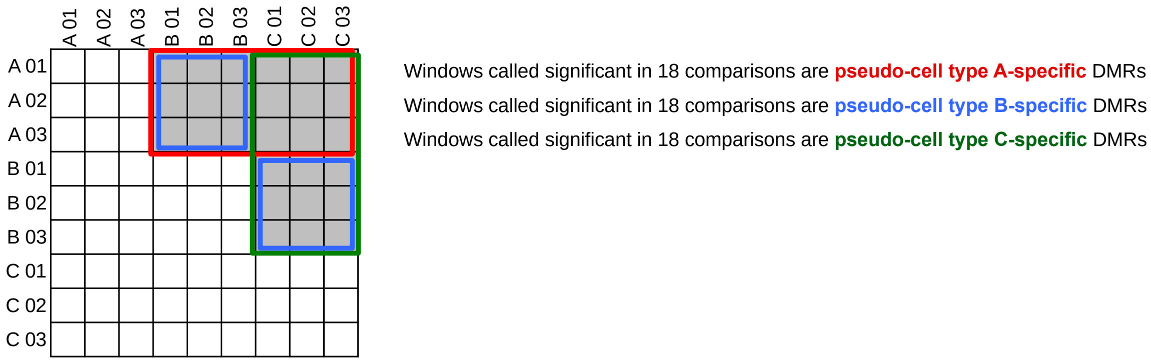
**Supplementary Figure 1. Skin cell type-specific DMR calling strategy.**

(a) Illustration of M&M skin cell type pairwise comparisons.

(b) Illustration of intersection strategy for calling skin cell type-specific DMRs. Each gray cell represents one comparison by M&M. DMRs called in the same direction in each of the indicated comparisons (cells within red, green, or blue outlines) were collected as a given cell type-specific DMR set (fibroblast, melanocyte, or keratinocyte, respectively).



**Supplementary Figure 2. Number of DMRs across M&M q-values.** Red line = intra-Fibroblast DMRs (Fibroblast 02 vs Fibroblast 03); blue line = inter-cell type DMRs (Fibroblast 03 vs Keratinocyte 03).

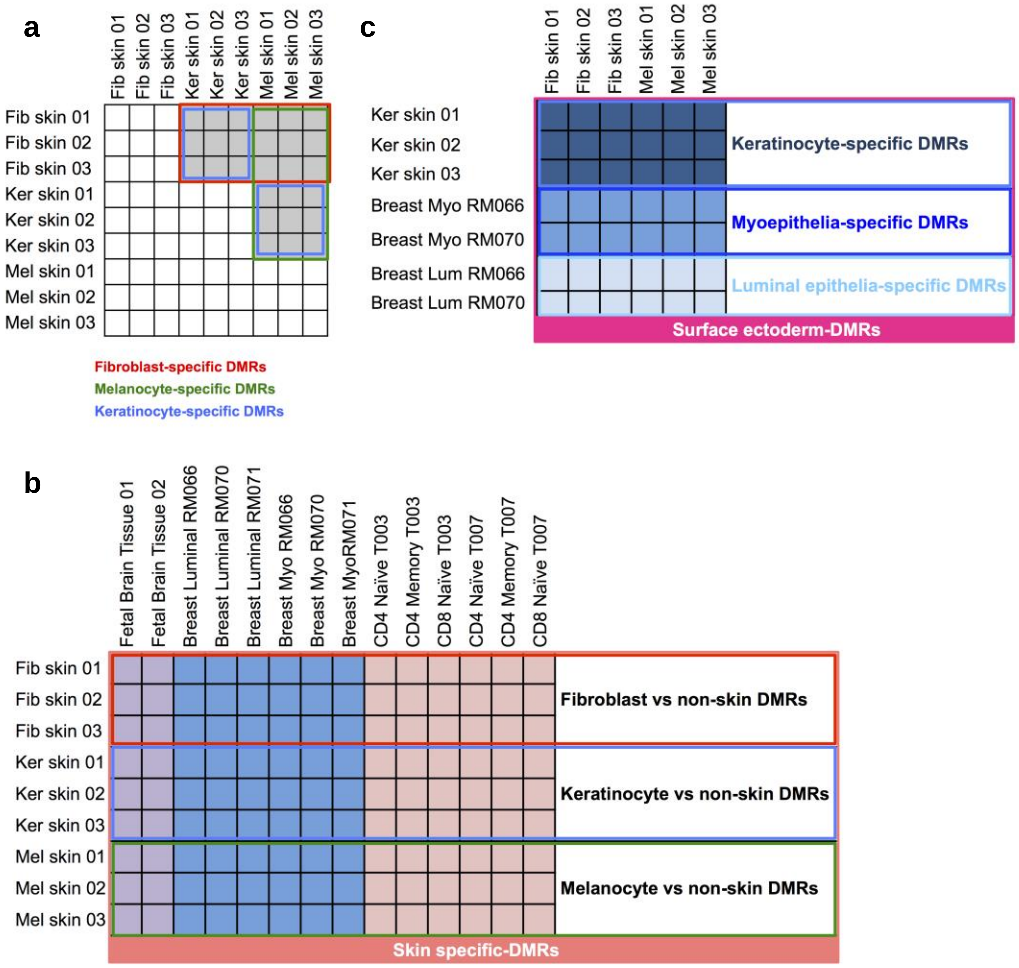


Windows called significant in 18 comparisons are **pseudo-cell type A-specific** DMRs

Windows called significant in 18 comparisons are **pseudo-cell type B-specific** DMRs

Windows called significant in 18 comparisons are **pseudo-cell type C-specific** DMRs

**Supplementary Figure 3. Illustration of intersection strategy for identifying pseudo-cell type-specific DMRs.** Each gray cell represents one comparison by M&M. DMRs called in the same direction in each of the indicated comparisons (cells within red, blue, or green outlines) were collected as a given pseudo-cell type-specific DMR set (A, B, or C, respectively).

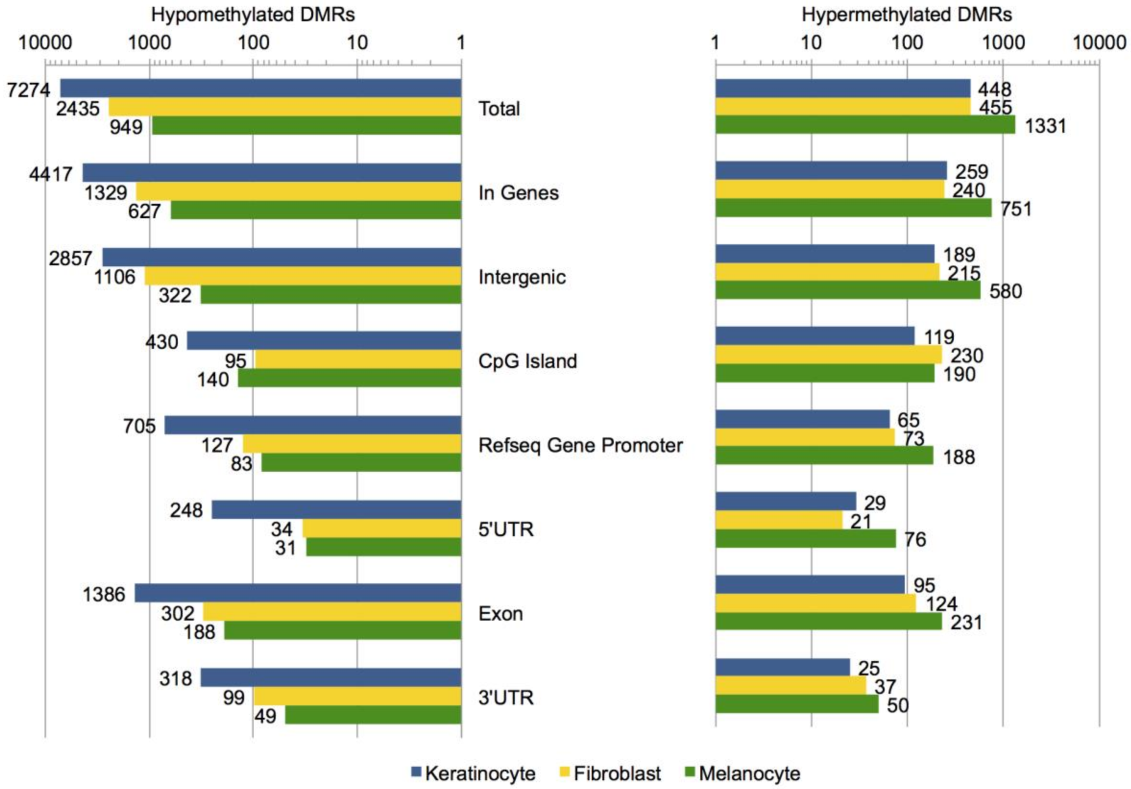


**Supplementary Figure 4: Matrices depicting sample comparisons used to identify differentially DNA methylated regions.**

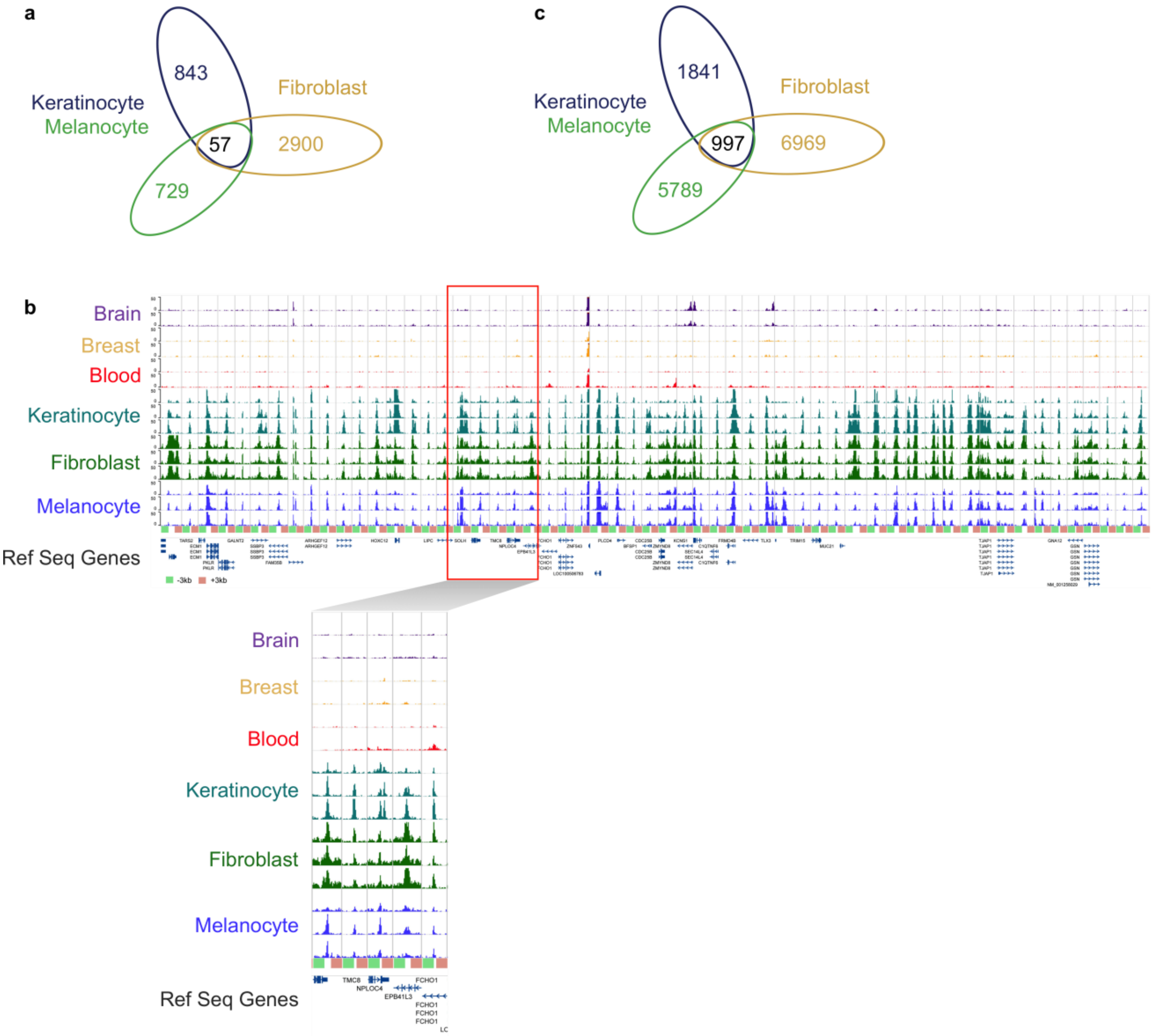
**(**a) Matrix depicting pairwise methylome comparisons used to determine skin cell type-specific DMR sets. Each gray cell represents one comparison by M&M (Methods). DMRs called in the same direction in each of the indicated comparisons (cells within red, green, or blue outlines) were collected as a given cell type-specific DMR set (fibroblast, melanocyte, or keratinocyte, respectively).

(b) Matrix depicting pairwise methylome comparisons used to determine skin tissue-specific DMRs. Each cell represents one M&M pairwise comparison. DMRs called in the same direction in all depicted pairwise comparisons (i.e. for each of the 3 skin cell types compared to non-skin cell types) were called “skin tissue-specific DMRs” (of which there were only 8; Figure 3a).

(c) Matrix depicting pairwise methylome comparisons used to determine surface ectoderm-specific DMRs. Each cell represents one M&M pairwise comparison. DMRs called in the same direction in all depicted pairwise comparisons (i.e. for each of the 3 surface ectoderm cell types) were collected as the surface ectoderm-DMR set.



**Supplementary Figure 5: Genomic annotation of skin cell type-specific DMRs.** Hypomethylated and hypermethylated DMRs plotted independently. DMRs are 500 bp windows. Cell types indicated by bar color. Genomic annotations described in Methods.

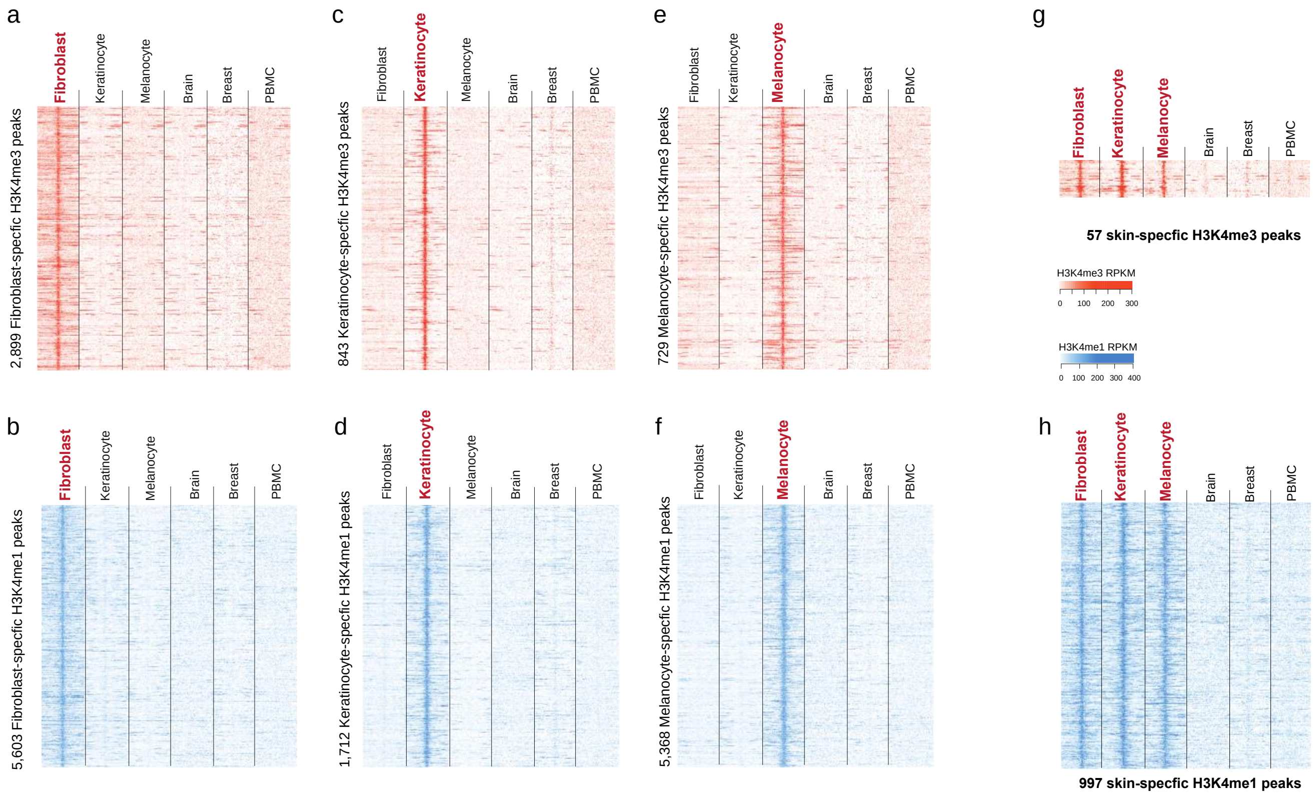


**Supplementary Figure 6: Shared histone modification patterns for skin cell types.**

(a) Venn diagram showing number of H3k4me3 peaks present in each skin cell type that are also absent in all non-skin samples (brain, breast, and blood). There are 57 overlap regions where H3K4me3 peaks are present in all three skin cell types and absent in all non-skin samples. A total of 55,859 H3K4me3 peaks were detected in all samples.

(b) WashU Epigenome Browser screenshot of the 57 regions where H3K4me3 is present in all three skin cell types and absent in all non-skin cell types. Each row is a ChIP-seq track for the indicated cell type. Three replicates for each skin cell type and two replicates for each non-skin sample are depicted. Each column represents one of the 57 different genomic regions +/- 3 kb. Bottom panel is a close-up of the red-boxed region in top panel.

(c) Venn diagram showing number of H3k4me1 peaks for each skin cell type that are absent in all non-skin samples (brain, breast, and blood). There are 997 overlap regions where H3K4me1 peaks are present in all three skin cell types and absent in all non-skin samples. A total of 259,297 H3K4me1 peaks were detected in all samples.



**Supplementary Figure 7: Heat maps of ChIP-seq signal around skin cell type-specific and tissue-specific histone modification peaks.**

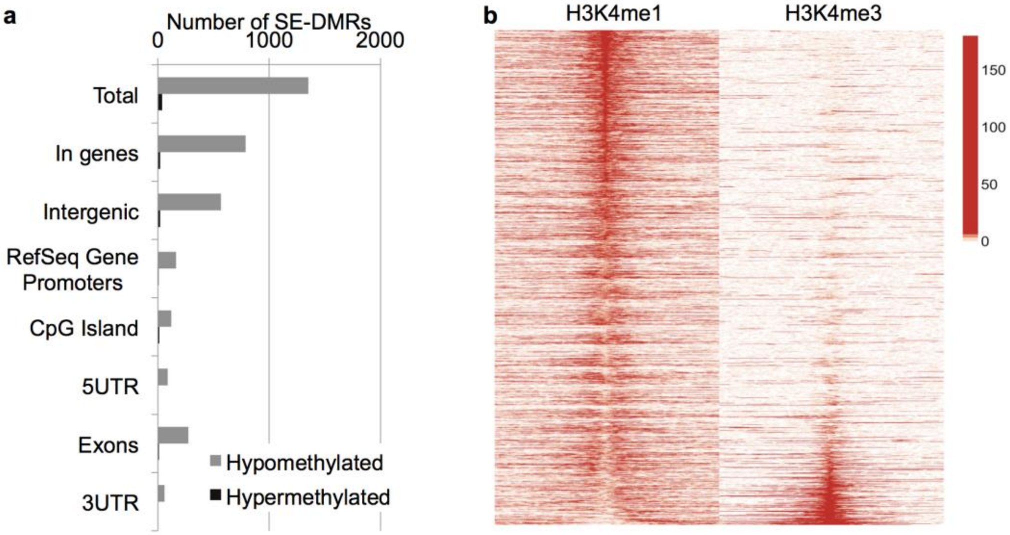
(a) ChIP-seq signal for fibroblast-specific H3K4me3 peaks. Each heat map row represents a 10kb region centered on a fibroblast-specific H3K4me3 peak divided into 200 windows, read density (RPKM) was calculated for each window. Each heat map column represents ChIP-seq signal for the labelled cell type. Breast = breast myoepithelial cell, Brain = fetal brain tissue, and PBMC = peripheral blood mononuclear cells.

(b) Similar to (a), but for fibroblast-specific H3K4me1 peaks.

(c-d) ChIP-seq signal for keratinocyte-specific H3K4me3 peaks (c) and H3K4me1 peaks (d).

(e-d) ChIP-seq signal for melanocyte-specific H3K4me3 peaks (e) and H3K4me1 peaks (f).

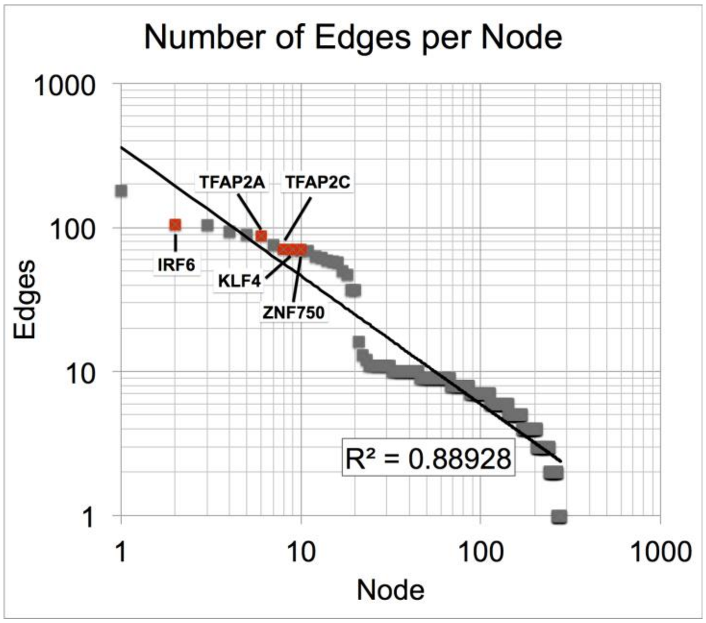
(g-h) ChIP-seq signal for skin tissue-specific H3K4me3 peaks (g) and H3K4me1 peaks (h).



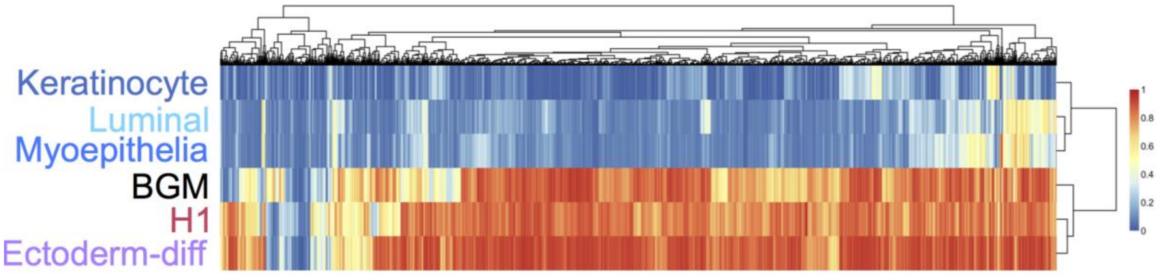
**Supplementary Figure 8: Additional SE-DMR characterization.**

(a) Genomic annotation of SE-DMRs. Hypomethylated and hypermethylated DMRs (1392 total) plotted independently. Genomic annotations described in Methods.

(b) Breast myoepithelial cell histone modification ChIP-seq signal at SE-DMRs. Each row represents a 500 bp DMR +/- 5kb (as in Figure 4b). DMRs are sorted in descending order of H3K4me1 signal, then increasing H3K4me3 signal. Values plotted are RPKM normalized to input.



**Supplementary Figure 9: Distribution of edges per node in the SE network.** Each node is plotted on the x-axis in order of its degree (total number of edges), the number of edges is the y-axis. The distribution fits a power law (black line) with $R^2$=0.88928. Gray and red boxes are individual nodes (genes). Genes of interest are highlighted in red and labeled. Genes with the highest degree are transcription factors at the top of the SE network (as in Figure 5a).



**Supplementary Figure 10: Heatmap and clustering dendrogram based on average CpG methylation values for hypomethylated SE-DMRs**. Each column represents one of 1307 DMRs for which there are CpGs with ≥ 10x coverage. Keratinocyte, brain germinal matrix (BGM), H1 ESC, and ectoderm-differentiated ESC values from WGBS; breast luminal and myoepithelial 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.

## Supplementary Tables

**Supplementary Table 1.** False discovery rate for calling DMRs across M&M q-values.

| q-value | 1.00E-02 | 1.00E-03 | 1.00E-05 | 1.00E-09 | 1.00E-13 | 1.00E-15 |
|---|---|---|---|---|---|---|
| FDR | 0.071 | 0.055 | 0.044 | 0.033 | 0.030 | 0.029 |

**Supplementary Table 2.** Numbers of CGI and non-CGI promoters in all skin cell type-specific DMRs. $\chi^2$ test p-value < 2.2e-16.

|  | # CGI promoters | # non-CGI promoters |
|---|---|---|
| Cell type DMRs | 267 | 974 |
| Genome-wide | 16638 | 9691 |

**Supplementary Table 3: Wilcoxon test for keratinocyte-specific expression analysis.** Wilcoxon ranked test, paired, *P*-values for RPKM distributions for Ref Seq genes with keratinocyte-specific hypomethylated DMRs at promoters.

|  | Fibroblast expression (average, *n*=3) | Melanocyte expression (average, n=3) |
|---|---|---|
| Keratinocyte expression (average, n=3) | 2.20E-16 | 2.39E-13 |

**Supplementary Table 4: Wilcoxon test for fibroblast-specific expression analysis.** Wilcoxon ranked test, paired, *P*-values for RPKM distributions for RefSeq genes with fibroblast-specific hypomethylated DMRs at promoters.

|  | Keratinocyte expression (average, n=3) | Melanocyte expression (average, n=3) |
|---|---|---|
| Fibroblast expression (average, n=3) | 2.59E-09 | 3.14E-03 |

**Supplementary Table 5: Wilcoxon test for melanocyte-specific expression analysis.** Wilcoxon ranked test, paired, *P*-values for RPKM distributions for RefSeq genes with melanocyte-specific hypomethylated DMRs at promoters.

| | Keratinocyte expression (average, n=3) | Fibroblast expression (average, n=3) |
|---|---|---|
| Melanocyte expression (average, n=3) | 3.65E-09 | 3.53E-09 |

**Supplementary Table 6: Wilcoxon test for surface ectoderm-specific expression analysis.** Wilcoxon ranked test, paired, *P*-values for RPKM distributions for RefSeq genes with surface ectoderm-specific hypomethylated DMRs at promoters.

| | Melanocyte expression (average, *n*=3) | Fibroblast expression (average, *n*=3) |
|---|---|---|
| Keratinocyte expression (average, *n*=3) | 1.00E-05 | 1.31E-09 |
| Luminal epithelia expression | 1.07E-04 | 1.18E-07 |
| Myoepithelia expression | 1.51E-02 | 2.25E-04 |

**Supplementary Table 7: Statistics for network analysis.** Control datasets and statistics for SE network.

| Random dataset filenames | one | two | three | four | five | six | seven | eight | nine | ten |
|---|---|---|---|---|---|---|---|---|---|---|
| Number of random genes into the Interaction Browser | 374 | 374 | 374 | 374 | 374 | 374 | 374 | 374 | 374 | 374 |
| Number not found in database | 13 | 16 | 13 | 11 | 8 | 7 | 8 | 11 | 15 | 15 |
| Number of edges | 810 | 841 | 996 | 1068 | 995 | 890 | 1015 | 1193 | 1034 | 738 |

| | |
|---|---|
| Mean (# edges in random datasets) | 958 |
| Standard Deviation (random datasets) | 136.54 |
| Number of edges in SE network | 1458 |
| P-value (t-test, upper-tail) | 1.25E-04 |

**Supplementary Table 8: TFBS motif-containing DMRs.** Number of hypomethylated SE-DMRs (1353 total) which contain TFAP2 and/or KLF4 binding site motifs.

| | Number of SE-DMRs |
|---|---|
| Contains TFAP2 motif only | 283 |
| Contains KLF4 motif only | 273 |
| Contains both TFAP2 and KLF4 motifs | 283 |
| No TFAP2 or KLF4 motifs | 514 |

## Supplementary Notes

### Supplementary Note 1: Skin cell type-specific DMR calling strategy

The specific skin cell type pairwise comparisons processed by M&M are as shown in Supplementary Fig. 1a. Each of the 3 skin cell type datasets from 3 different individuals is compared against every other skin cell type dataset, for a total of 36 pairwise comparisons. Pairwise comparisons between two different cell types are inter-cell type comparisons (27 total, gray boxes), while M&M comparisons between two of the same cell type datasets are intra-cell type comparisons (9 comparisons, 3 per cell type, blue boxes).

To maximize the specificity of our DMR prediction, we took advantage of the presence of three biological replicates for each cell type, and required that the same DMR call was reproduced in all analogous pair-wise comparisons. Therefore, to call cell type-specific DMRs, we took the intersection of all comparisons involving the three replicates of a given cell type, and required that a 500bp window be called significantly differentially methylated (in the same direction) by our M&M statistic in each of 18 pairwise comparisons. Our intersection strategy is illustrated in Supplementary Fig. 1b.

### Supplementary Note 2: M&M command line and output description

R scripts used to generate pairwise comparisons using the methylMnM R package (http://epigenome.wustl.edu/MnM/). The **compare.pv.R** script contains the functions to perform the actual pairwise comparison that generates p-values for each 500bp window (**MnM.test()** function). The **qv.DMR.R** script calculates q-values for every window (**MnM.qvalue()**) and selects significant windows based on a user-given q-value threshold (**MnM.selectDMR()**).

### compare.pv.R

```
library(methylMnM)
cpgbin <- 'num500_cpgbin.bed'
mrecpgbin <- 'num500_Five_mre_cpg.bed'
medip.list = read.table('skin_medip.list')
mre.list = read.table('skin_mre.list')
c_s <- NULL

for  (i in 1:length(medip.list[,1])) {
        medipfile1 <- paste('num500_',medip.list[i,1],sep='')
        mrefile1 <- paste('num500_',mre.list[i,1],sep='')
        name <- paste(medip.list[i,1])
        first <-
paste(strsplit(name,"_")[[1]][1],strsplit(name,"_")[[1]][2],sep="")

        for (j in (i+1):length(medip.list[,1])) {
                medipfile2 <- paste('num500_',medip.list[j,1],sep='')
                mrefile2 <- paste('num500_',mre.list[j,1],sep='')
                name <- paste(medip.list[j,1])
                second <-
paste(strsplit(name,"_")[[1]][1],strsplit(name,"_")[[1]][2],sep="")
                dataset <- c(medipfile1, medipfile2, mrefile1, mrefile2)
                w_f <- paste('pv_',first,"_",second,".bed",sep="")
                r_f <- paste('pv_',first,"_",second,".report",sep="")
                MnM.test(file.dataset=dataset, chrstring=c_s,
file.cpgbin=cpgbin, file.mrecpgbin=mrecpgbin, writefile=w_f, reportfile=r_f,
mreratio=3/7, method='XXYY', psd=2, mkadded=1, a=1e-20, cut=100, top=500)
        }
}
```

### qv.DMR.R

```
library(methylMnM)
qv.list = read.table('skin_qv_files.list')

for  (i in 1:length(qv.list[,1])) {
        name <- paste(qv.list[i,1])
        qval_f <- paste('qv_',name,sep='')
        r_f <- paste('qv_',strsplit(name,".bed")[[1]][1],".report",sep="")
        MnM.qvalue(pval_f, writefile=qval_f, reportfile=r_f)
        frames <- read.table(qval_f, header=TRUE, sep="\t", as.is=TRUE)
        DMR <- MnM.selectDMR(frames=frames, up=2, down=1/2, q.value=1e-5,
cutoff="q-value", quant=0.9)
        fname <- paste(qval_f,sep="")
        sname <-strsplit(fname,"pv")[[1]][2]
        writeDMRfile <- paste('DMR_q1e-5',sname,sep="")
        write.table(DMR, writeDMRfile, sep="\t", quote=FALSE, row.names=FALSE)
}
```

The output of M&M pairwise comparisons were p-value and q-value measurements for the likelihood that the methylation levels of the two samples were different for each 500bp window across the genome. Note that q-value is the false discovery rate analogue of the p-value. The genome-wide false discovery rate (FDR) was controlled using the previously described Group Benjamini-Hochberg method[1]. We then chose a q-value cutoff to call differentially methylated regions. All of our analyses used a q-value cutoff of 1e-5.

## Supplementary Note 3: Estimation of M&M and cell type-specific DMR FDR

To estimate the false discovery rate of DMRs called by M&M, we chose a pairwise comparison as a test case: Fibroblast skin 03 vs Keratinocyte skin 03. These results were compared to the M&M results from a within-cell type comparison: Fibroblast skin 02 vs Fibroblast skin 03, which are biological replicates (i.e. the same cell type from two different newborn males). For these pairwise comparisons, we examined the number of DMRs called at varying q-value cutoffs. As seen in Supplementary Fig. 2, the number of DMRs in both cases decreased with decreasing q-value cutoff. As expected, the numbers of DMRs found between biological replicates is very small. Thus, our pairwise DMR false discovery rate is very low (Supplementary Table 1). We used M&M q-values of 1e-5 throughout, which by this analysis had a FDR of 0.044. FDR calculations using other within-cell type comparisons yielded similar results.

To assay the false discovery rate of our skin cell type-specific DMR calling strategy, we performed a permutation experiment to empirically estimate this value. In this experiment, we randomly shuffled our datasets by labeling them as three “pseudo” cell types (A, B, and C) with three replicates each (01, 02, and 03). Because we have already performed all possible pair-wise comparisons using the M&M algorithm, we called pseudo-cell type specific-DMRs by the same criteria as in Supplementary Note 1, i.e. that a window must be called differentially methylated 18/18 times to be a DMR in any pseudo-cell type at a q-value cutoff of 1e-5. The strategy is illustrated in

Supplementary Fig. 3. We repeated this process of shuffling, assigning pseudo-cell type names, and finding DMRs 10 times. Each time the analysis of the pseudo-cell types returned zero windows called as pseudo-cell type-specific DMRs. Thus, our cell type-specific DMRs are very far from the random expectation for these data.

## Supplementary Note 4: Analysis of CpG Islands in cell type-specific DMRs

It is known that approximately 70% of all gene promoters are associated with a CpG island (CGI)[2,3]. We defined a CGI promoter as any promoter that has ≥ 0.05% of a given CGI contained in it and found 16638 RefSeq gene promoters (or 63.2%) were CGI promoters. Then we counted the numbers of CGI promoters and non-CGI promoters in each DMR class and tested the null hypothesis that the percentage of promoters that contain CGIs for each DMR class is similar to the CGI promoter distribution found across the genome. We found that across our DMR sets, the numbers of CGI promoters in DMRs are significantly depleted relative to their genome-wide distribution, while non-CGI promoters are significantly enriched (Supplementary Table 2). In general, the majority of DMRs at promoters were within non-CGI promoters, which is consistent with the concept that non-CGI promoters are involved in tissue and cell type specificity.

## Supplementary Note 5: Skin tissue-specific DMR calling strategy

We sought to identify the unique DNA methylation signature that the skin environment might contribute to its resident cell types. Therefore, we asked what shared regions of the skin fibroblast, keratinocyte, and melanocyte methylomes were differentially methylated compared to cell types of other tissues. To do this, we compared skin cell type methylomes to those of non-skin cell types and tissues (including brain tissue and breast and blood cell types) to identify DMRs in a pairwise manner. 28,776 total DMRs were identified in these pair-wise comparisons. Compared to the non-skin samples, keratinocytes, fibroblasts, and melanocytes each possessed 623, 763, and 402 consensus DMRs respectively.  We then took the intersection of these three DMR sets to identify the shared differences between skin cell types and cell types residing in a different tissue environment (i.e. the same methylation status in all skin cell types and the opposite methylation status in all non-skin cell types). The result was, surprisingly, a very small set of only 8 regions. To be clear, we do expect much of the methylome for the three skin cell types is similar, but the shared methylome signature that is unique to the skin is very small.

Identification of skin tissue-specific DMRs follows the exact same logic as that of cell type-specific DMRs (for which FDR and reproducibility are documented above in Supplementary Note 3). Both M&M and our DMR identification strategy are designed to optimize specificity. We use the same M&M q-value threshold for our tissue-specific analysis as for the cell type-specific analysis.

## Supplementary Methods

### *RNA isolation*

Total RNA was extracted from cells using Trizol reagent (Life Technologies) following the manufacturer’s instructions.

### *RNA-seq*

Standard operating procedures for RNA-seq library construction are available at http://www.roadmapepigenomics.org/protocols/type/experimental/. RNA-seq library construction involves the following protocols in order: 1) Purification of polyA+ mRNA and mRNA(-) Flow-Through Total RNA using MultiMACS 96 separation unit, 2) Strand specific 96 Well cDNA Synthesis, and 3) Strand specific 96-well library construction for Illumina sequencing. Briefly, polyA+ RNA was purified using the MACS mRNA isolation kit (Miltenyi Biotec) from 2-10 ug of total RNA with a RIN>=7 (Agilent Bioanalyzer) as per the manufacturer's instructions. The process included on-column DNase I treatment (Invitrogen). Double stranded cDNA was synthesized from the purified polyA+ RNA using the Superscript II Double-Stranded cDNA Synthesis kit (Invitrogen) and 200 ng of random hexamers. After first strand synthesis, dNTPs were removed using 2 volumes of AMPure XP beads (Beckman Genomics). GeneAmp 12.5mM dNTPs blend (Invitrogen) was used in the second strand synthesis mixture in the presence of 2 ug of Actinomycin D. Double stranded cDNA was purified using 2 volumes of Ampure XP beads, fragmented using Covaris E series shearing (20% duty cycle, Intensity 5, 55 seconds), and used for paired-end sequencing library preparation (Illumina). Prior to library amplification, uridine digestion was performed at 37 degrees Celsius for 30 minutes, followed by a 10 minute incubation at 95 degrees Celsius in Qiagen Elution buffer (10mM Tris-Cl, pH 8.5) with 5 units of Uracil-N-Glycosylase

(UNG: AmpErase). The resulting single stranded sequencing library was amplified by PCR (10-13 cycles) to add Illumina P5 and P7 sequences for cluster generation. PCR products were purified on Qiaquick MinElute columns (Qiagen) and assessed and quantified using an Agilent DNA 1000 series II assay and Qubit fluorometer (Invitrogen) respectively. Libraries were sequenced using paired-end 76 nt sequencing chemistry on a cBot and Illumina GAiix or HiSeq2000 following manufacturer's protocols (Illumina).

RNA-seq pair-end reads were aligned to a transcriptome reference consisting of the reference genome extended by the annotated exon-exon junctions17. To generate a transcriptome reference, we used the JAGuaR v 1.7.6 pipeline (http://www.bcgsc.ca/platform/bioinfo/software/jaguar) which is specifically developed to allow for a single read to span multiple exons. Reads aligned to a custom transcriptome reference (build from NCBI GRCh37-lite reference and Ensembl v65 (GenCode v10) annotations) are then “repositioned” onto genomic coordinates, transforming reads spanning exon-exon junctions into large-gapped alignment. Using repositioned reads, we generated genome wide coverage profiles (wiggled files) using BMA2WIG java program for further analysis and visualization in genome browsers. To generate profiles we included pairs that are marked as duplicated as well as pairs mapped in multiple genomic locations.

A custom RNA-seq QC and analysis pipeline was applied to the generated profiles and a number of QC metrics were calculated to assess the quality of RNA-seq libraries such as intron-exon ratio, intergenic reads fraction, strand specificity, 3’-5’ bias, GC bias, and RPKM discovery rate. To quantify exon and gene expression we calculated modified RPKM metrics[4]. For the normalization factor in RPKM calculations, we used the total number of reads aligned into coding exons and excluded reads from the mitochondrial genome, that fall within genes encoding ribosomal proteins, or that fall into the top 0.5% expressed exons. RPKM for a gene was calculated using total number of reads aligned into all its merged exons normalized by total exonic length. All mRNA-seq analyses used a pseudocount of 1.

### *miRNA-seq*

Standard operating procedures for miRNA-seq library construction are available at http://www.roadmapepigenomics.org/protocols/type/experimental/. miRNA-seq library construction involves the following protocols in order: 1) purification of polyA+ mRNA and mRNA(-) Flow-Through Total RNA using MultiMACS 96 separation unit, 2) strand specific 96 Well cDNA Synthesis, and 3) strand specific 96-well library construction for Illumina sequencing. A more detailed description of miRNA-seq library construction and data processing in Gascard, P. et al. (submitted REMC companion paper).

## Supplementary References

1. Hu, J. X., Zhao, H. & Zhou, H. H. False Discovery Rate Control With Groups. *J. Am. Stat. Assoc.* **105,** 1215–1227 (2010).
2. Saxonov, S., Berg, P. & Brutlag, D. L. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc. Natl. Acad. Sci.* **103,** 1412–1417 (2006).
3. Deaton, A. M. & Bird, A. CpG islands and the regulation of transcription. *Genes Dev.* **25,** 1010–1022 (2011).
4. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5,** 621–628 (2008).