Stem Cell Reports
Resource
OPEN ACCESS

Dynamic Transcriptional and Epigenetic Regulation of Human Epidermal Keratinocyte Differentiation

Alessia Cavazza,[1] Annarita Miccio,[2] Oriana Romano,[3] Luca Petiti,[4] Guidantonio Malagoli Tagliazucchi,[3] Clelia Peano,[4] Marco Severgnini,[4] Ermanno Rizzi,[4] Gianluca De Bellis,[4] Silvio Bicciato,[3] and Fulvio Mavilio[3,5,*]

[1]Center for Functional Cancer Epigenetics, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA 02215, USA
[2]Imagine Institute, 75015 Paris, France
[3]Department of Life Sciences, University of Modena and Reggio Emilia, 41125 Modena, Italy
[4]Institute for Biomedical Technologies, National Research Council, 20132 Milan, Italy
[5]Genethon, 1bis rue de l'Internationale, 91002 Evry, France
*Correspondence: fmavilio@genethon.fr
http://dx.doi.org/10.1016/j.stemcr.2016.03.003

SUMMARY

Human skin is maintained by the differentiation and maturation of interfollicular stem and progenitors cells. We used DeepCAGE, genome-wide profiling of histone modifications and retroviral integration analysis, to map transcripts, promoters, enhancers, and super-enhancers (SEs) in prospectively isolated keratinocytes and transit-amplifying progenitors, and retrospectively defined keratinocyte stem cells. We show that >95% of the active promoters are in common and differentially regulated in progenitors and differentiated keratinocytes, while approximately half of the enhancers and SEs are stage specific and account for most of the epigenetic changes occurring during differentiation. Transcription factor (TF) motif identification and correlation with TF binding site maps allowed the identification of TF circuitries acting on enhancers and SEs during differentiation. Overall, our study provides a broad, genome-wide description of chromatin dynamics and differential enhancer and promoter usage during epithelial differentiation, and describes a novel approach to identify active regulatory elements in rare stem cell populations.

INTRODUCTION

The epidermis is a stratified epithelium differentiating from keratinocyte stem cells (KSCs) contained in the basal layer and in the bulge of hair follicles. Upon division, KSCs produce transit-amplifying (TA) progenitors that generate differentiated keratinocytes and other epithelial skin components. The available information on the molecular events underlying self-renewing and differentiation of KSCs comes from studies on the murine hair follicle (reviewed in Blanpain et al., 2007). Much less is known about human KSCs, which lack robust markers for prospective isolation and are defined only retrospectively by the nature of their progeny in cell culture or transplantation assays. Clonal analysis in vitro has defined three types of clonogenic cells, giving rise to the so-called holoclones, meroclones, and paraclones. Holoclone-forming cells have the highest self-renewing and proliferative capacity, and define in culture the KSCs of the epidermis or the corneal epithelium (Pellegrini et al., 1999; Rochat et al., 1994). Meroclone- and paraclone-forming cells have proportionally less proliferative capacity and terminally differentiate into keratinocytes after 5–15 cell doublings, as expected for TA progenitors (Barrandon and Green, 1987). Few molecular markers are known for KSCs or TA progenitors: they include the p63, BMI1, CEBPs, MYC, and GATA-3 transcription factors (TFs), integrins, Wnt/β-catenin, NOTCH, HH, SGK3, and some bone morphogenetic proteins (Blanpain et al., 2007). In particular, p63 is considered a master regulator of morphogenesis, identity, and regenerative capacity of stratified epithelia (Pellegrini et al., 2001; Yang et al., 1999). Although some of the targets of p63 and other TFs involved in epidermal cell functions are known, little is known about the chromatin dynamics and the differential usage of promoters and enhancers driving the differentiation of human KSCs and TA progenitors.

Specific histone modifications are currently used to define chromatin regions with different regulatory functions. In particular, monomethylation of lysine 4 of histone 3 (H3K4me1) characterizes enhancer regions, whereas its trimethylation (H3K4me3) defines promoters (Ernst et al., 2011; Heintzman et al., 2009). Acetylation of H3K27 defines transcriptionally active enhancers and large clusters of enhancers (super-enhancers [SEs]) involved in the definition of cell and tissue identity (Hnisz et al., 2013). In this study, we aimed to map transcriptional regulatory elements and define their usage during epithelial differentiation. By combining high-throughput identification of Pol-II-transcribed (capped) RNAs defined by Cap Analysis of Gene Expression (DeepCAGE) (Carninci et al., 2006) with genome-wide profiling of histone modifications determined by chromatin immunoprecipitation (ChIP-seq), we mapped active enhancer and SE elements in prospectively isolated TA progenitors and terminally differentiated keratinocytes. For KSCs, which

lack markers for prospective isolation, we exploited the integration characteristics of the Moloney murine leukemia retrovirus (MLV), which integrates in active promoters and enhancers (Biasco et al., 2011; Cattoglio et al., 2010; De Ravin et al., 2014) as a consequence of the direct binding of the viral integrase to the bromodomain and extra-terminal (BET) proteins BRD2, BRD3, and BRD4 that tether the pre-integration complex to acetylated chromatin regions (De Rijck et al., 2013; Gupta et al., 2013; Sharma et al., 2013). By using MLV vector integration clusters as surrogate genetic markers of active regulatory elements, we mapped a collection of putative enhancers and SEs active in bona fide KSCs, retrospectively defined by their capacity to maintain long-term keratinocyte cultures.

RESULTS

DeepCAGE Mapping of Active Promoters in Keratinocyte Progenitors and Differentiated Keratinocytes

To enrich keratinocyte progenitors (KPs) from a keratinocyte mass culture, we panned β1 integrin-positive cells by adherence to collagen-IV-coated plates (Jones and Watt, 1993). Adhering cells were highly enriched in KPs, as determined by a clonogenic assay, and showed significantly increased expression of the progenitor-related markers *TP63* ($p < 0.05$), *LRIG1* ($p < 0.01$), *ITGB1*, *MCSP*, and *DLL1* ($p < 0.001$) by real-time qPCR, while the non-adhering fraction was depleted in colony-forming cells and expressed the differentiation markers *KRT1*, *IVL*, and *LOR* (Figures S1A–S1D). Differentiated keratinocytes (DKs) were obtained by in vitro differentiation in conditions of contact inhibition (Kouwenhoven et al., 2015; Shen et al., 2013), and showed residual colony-forming capacity and high expression of differentiation markers (Figures S1E and S1F).

To define global promoter usage, we used DeepCAGE on RNA extracted from KPs and re-analyzed an epidermal keratinocytes dataset available from ENCODE as a proxy of DKs. We identified a total of 15,283 CAGE promoters, 14,565 expressed in KPs and 15,027 in DKs. Most CAGE promoters mapped to known promoters (20%) or to immediately downstream 5′ UTR regions (48.6%) or gene bodies (Figure 1A). We grouped CAGE promoters in three clusters based on the tag position with respect to transcription start sites (TSSs): promoters in cluster 3 showed a broad profile around TSSs and represented the majority of alternatively used promoters, cluster 2 represented canonical promoters with a sharp localization at TSSs, while cluster 1 exemplified pervasive transcription within genes (Figure 1B).

Epithelial Differentiation Is Characterized by Quantitative Regulation of a Large Set of Common Promoters

Most CAGE promoters (14,309) were active in both cell populations and represented 98.2% and 95.2% of KP and DK promoters, respectively. Only 256 and 718 promoters were strictly stage specific, the majority of which (>60%) represented uncharacterized TSSs or were associated with non-coding transcripts, mainly long non-coding RNAs (lncRNAs). Most of the changes in transcriptome associated with keratinocyte differentiation were therefore defined by quantitative changes in the expression of promoters active in both KPs and DKs. A total of 5,429 promoters were expressed at significantly different levels between KPs and DKs ($p < 0.001$, χ^2 test), with 1,838 promoters upregulated in KPs and 1,712 in DKs at a $\log_2$ fold change (FC) of ≥ 2 (Figure 1C). In KPs differentially expressed TSSs were more abundant in promoters and introns, while in DKs they were more abundant in introns and 3′ UTRs (Figure 1D). qPCR analysis confirmed differential mRNA expression for 40 of the 46 randomly chosen promoters (Figure S2A). We detected alternative transcription initiation in 1,187 protein-coding genes, 455 of which underwent switch between alternative promoters during the KP-to-DK transition (Figure 1E). As an example, *PLEC1*, encoding six isoforms of the keratinocyte adhesion protein plectrin, is transcribed from different promoters predicting KP-specific, DK-specific, and common isoforms (Figures 1F and S2B).

We annotated all CAGE promoters in six classes on the basis of the combinatorial presence of TATA box and CpG islands, i.e., $TATA^+$ or $TATA^-$, and no-CpG (NCPs), low-CpG (LCPs), and high-CpG (HCPs). The majority (~75%) of the promoters fell in the HCP class and were mostly $TATA^-$, a feature associated with housekeeping functions (Carninci et al., 2006; Schug et al., 2005). As expected, the proportion of LCP and NCP promoters progressively increased in the differentially expressed promoters at increasing FC values (Figure 1G).

Differential Promoter Usage Defines Stage-Specific Gene Expression Programs

Genes associated with differentially expressed CAGE promoters encoded known markers of follicular and interfollicular epidermal progenitors (i.e., *SOX9*, *LRIG1*, *BMI1*, *TCF3*, *TCF4*, *TP63*) and differentiating keratinocytes (*IVL*, *FLG*, *KRT1*, and genes belonging to the epidermal differentiation complex (EDC) on chromosome 1q21). To correlate differential promoter usage with gene-expression patterns, we carried out an RNA-sequencing (RNA-seq) analysis in KPs and DKs. The DK dataset showed a good correlation (Spearman's $r > 0.8$) with the RNA-seq data of human epidermal keratinocytes reported in ENCODE, demonstrating the similarity between the two populations and

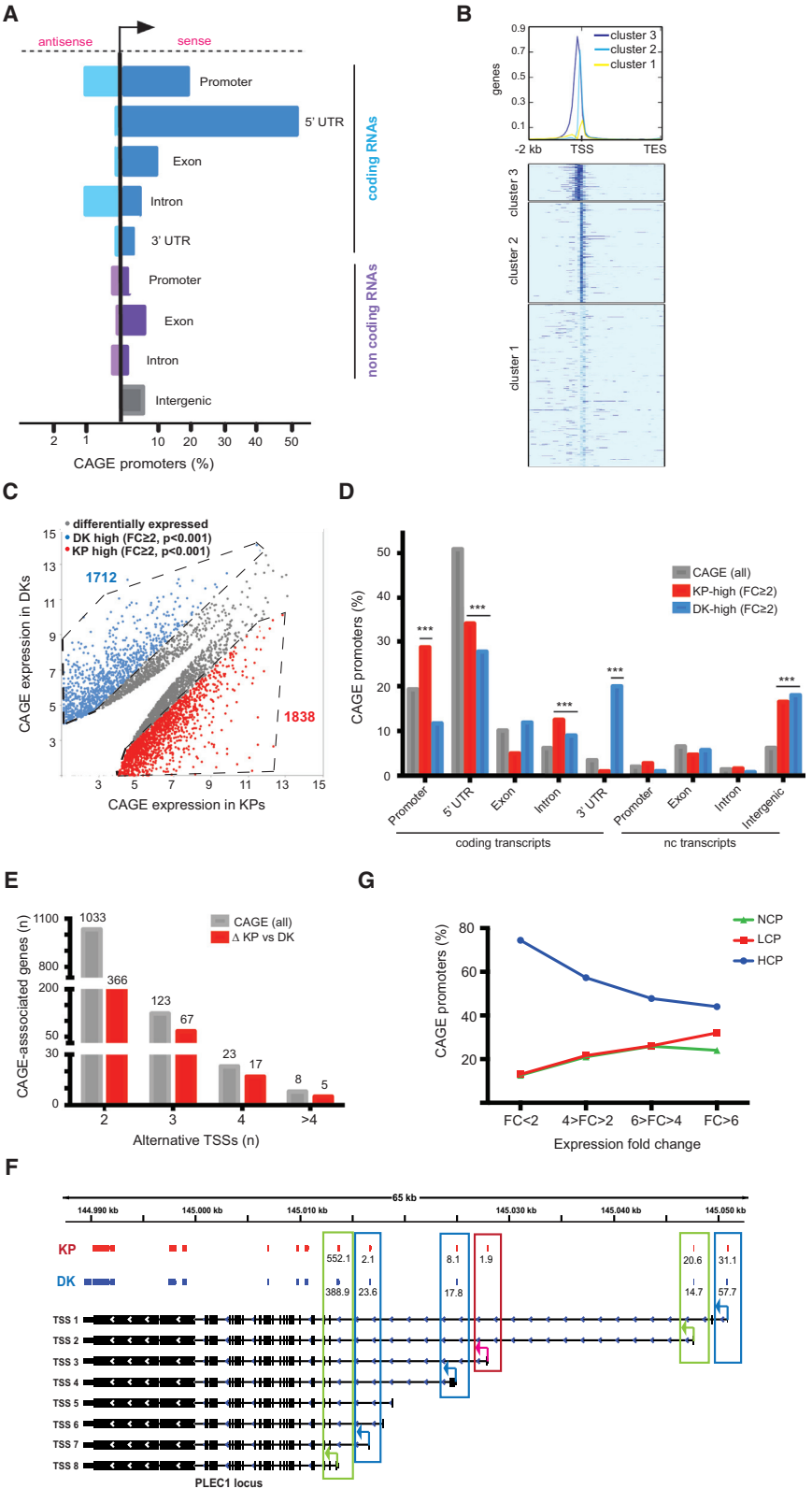


Figure 1. Transcriptome Analyses in Human Epidermal Differentiation by DeepCAGE

(A) Histogram depicting the proportion of CAGE tags aligned to promoters (defined as a 500-bp region upstream of TSS), 5′ UTRs, exons and introns of coding and non-coding transcripts, and CAGE tags mapping to intergenic regions. Bars on the right side of the histograms represent CAGE tags on the same strand as the corresponding annotated transcript, while bars on the left represent tags on the opposite strand.

(B) CAGE tags distribution profile along the region spanning from 2,000 bp upstream of the TSS to the transcription end site (TES) of RefSeq genes. Gene bodies were stretched or shrunk to fit the same 1,000-bp length. CAGE promoters were grouped in three different clusters through k-means clustering, based on their tag distribution along the considered region.

(C) Scatterplot of gene expression profiling of KPs and DKs obtained from three biological replicates. Only genes that are differentially expressed ($p < 0.001$) are represented in the plot. Dashed lines indicate the 2-fold differential expression cut-off to define KP- or DK-high (genes upregulated in KPs or DKs with FC $\geq$ 2, $p < 0.001$) genes. The numbers of differentially expressed genes with FC $\geq$ 2 are indicated.

(D) Percentage of differentially expressed (FC $\geq$ 2) and of all CAGE tags aligned to promoters, 5′ UTRs, exons and introns of coding and non-coding transcripts, and in intergenic regions. The asterisks indicate the statistical significance in the level of enrichment of KP-high or DK-high CAGE tags in each category over all CAGE tags distribution (***$p < 0.001$, $n = 3$).

(E) Numbers of RefSeq genes using multiple TSSs (gray) and of CAGE promoters that are alternatively used in KPs and DKs (red).

(F) Genomic browser screenshot of alternative promoters usage for the *PLEC1* gene. CAGE promoters in KPs and DKs are represented together with their expression levels in transcripts per million (TPM). *PLEC1* TSSs used preferentially by DKs (blue box), KPs (red box), or equally in both cell types (green box) are shown.

(G) Proportion of differentially expressed CAGE promoters falling in the HCP, LCP, and NCP categories with respect to the fold change in expression.

See also Figures S1 and S2.

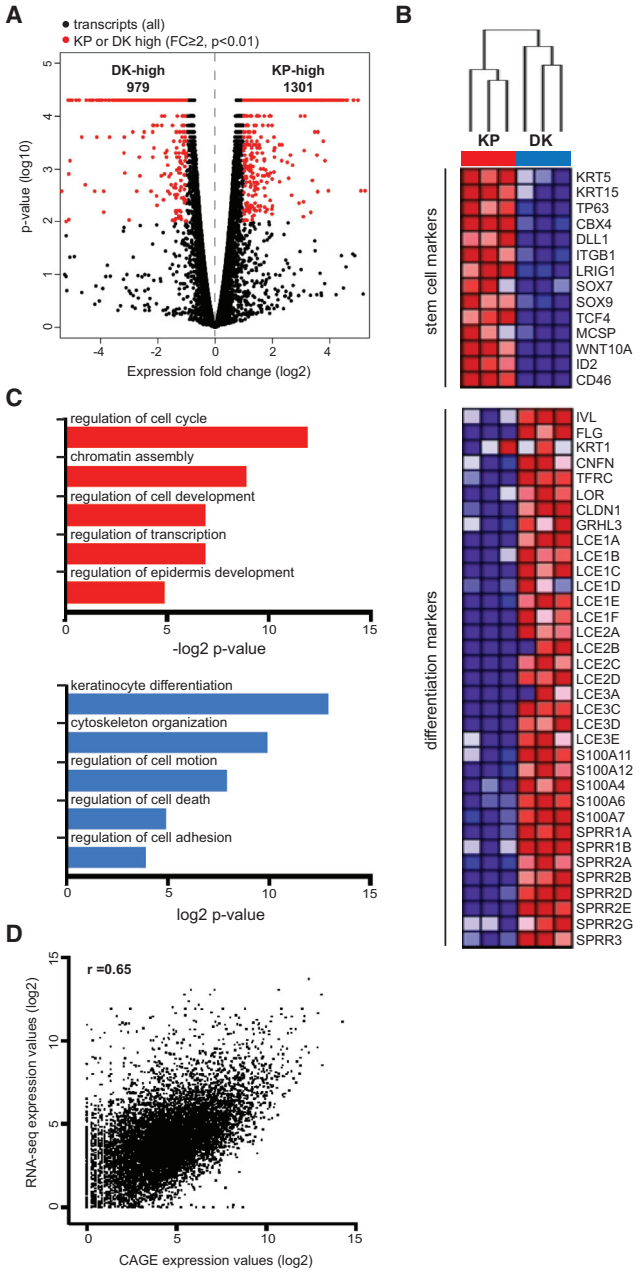


Figure 2. Transcriptome Analyses in Human Epidermal Differentiation by RNA-Seq

(A) Volcano plot of RNA-seq data from three biological replicates of KPs and DKs. Differentially expressed transcripts between KPs and DKs are highlighted in red and numbers are indicated.

(B) Heatmap and clustering of RNA expression profiles of manually selected genes relevant to stem cell or differentiation functions in epidermis.

(C) GO analysis of KP (red) and DK (blue) signature genes.

(D) Pearson's correlation plot of $\log_2$-transformed expression values detected for differentially expressed transcripts/promoters by RNA-seq and DeepCAGE, respectively.

See also Figure S3.

validating the use of the ENCODE CAGE data as a proxy of DKs (Figure S3).

RNA-seq analysis showed a substantially different transcriptome in KPs and DKs, with 2,280 differentially expressed transcripts (FC > 2, p < 0.01) (Figure 2A). Concordantly with CAGE promoter usage, DKs showed activation of well-known differentiation markers and a decreased expression of stem cell-related genes compared to KPs (Figure 2B). A gene ontology (GO) analysis showed statistically significant biases toward regulation of cell proliferation and epithelium morphogenesis in KPs, and epidermal differentiation and regulation of cell motility and apoptosis in DKs (Figure 2C). A correlation of RNA-seq and CAGE expression values for the same genes showed a statistically significant concordance between the two datasets (Pearson's r = 0.6) (Figure 2D).

Dynamic Epigenetic Changes in Active Promoter Regions during Epithelial Differentiation

ChIP-Seq analysis of histone modifications identified 22,813 and 15,440 promoter regions in KPs and DKs respectively, as defined by the H3K4me3$^+$ and H3K4me1$^{-/low}$ signature. The H3K27ac marker identified 8,557 and 11,341 "strong" promoters, respectively, >80% of which overlapped with CAGE promoters of the HCP class (Figures 3A and 3B). HCP and TATA$^+$ promoters showed an H3K4me3$^+$/H3K4me1$^{-/low}$/H3K27ac$^+$ profile, while LCP and NCP elements were barely marked (Figure 3C).

Comparative analysis of ChIP-seq data showed no dramatic changes in chromatin configuration at the promoter level between KPs and DKs (Figure 3D). The 312 KP-specific active promoters were mainly annotated to ncRNAs (85%) and genes such as *RUNX1*, involved in the specification of hair follicle progenitors (Lee et al., 2014), and *CLDN1*, encoding an adhesion molecule (Figures 3D and 3E). Conversely, the 292 DK-specific promoters were annotated to genes in the EDC (*S100* and *SLC* gene clusters) or encoding suprabasal keratins (*KRT6*, *KRT75*), collagens, and the TF *SOX15* (Figures 3D and 3F). Despite the overall modest epigenetic changes, we observed a significant correlation between the intensity of H3K4me3 marking and CAGE expression levels. Promoters highly expressed in KPs were highly enriched in H3K4me3 with respect to DKs, and vice versa (Figure S4A). H3K4me3 levels were significantly increased in 1,363 KP and 458 DK promoters, the majority of which was linked to upregulated CAGE promoters and RNA-seq transcripts in the corresponding cells. In KPs, these regions included many regulators of skin and stem cells homeostasis (Figure S4B).

Analysis of the H3K27ac marker identified 1,119 KP-specific and 567 DK-specific promoters (Figure 3D) and a strong correlation between H3K27ac marking intensity and CAGE expression levels (Figure S4A). H3K4me3 and

H3K27ac intensities were directly correlated in both cell types.

We then looked at the H3K27me3 histone mark, which is associated with repression of gene expression during epidermal lineage transitions (Frye and Benitah, 2012). We found 7,255 promoters marked by H3K27me3 in DKs, the majority of which (72.9%) were driving genes not expressed in KPs or DKs by RNA-seq analysis and related to early embryonic functions such as morphogenesis, cell-fate commitment, neuroectoderm development, and homeobox TFs by GO analysis (Figure S4C). Interestingly, 1,932 promoters marked by H3K27me3 in DKs drove genes expressed in KPs and were downregulated in DKs: they were enriched in transcription regulators and chromatin remodelers, and regulators of cell cycle, ectoderm development, epidermal stem cell biology, and skin homeostasis (Figure S4D). Many of these genes harbored at least one repressed (H3K4me1^{+}/H3K27me3^{+}) enhancer within 100 kb from their repressed promoter (Figure S4D), suggesting epigenetic silencing of entire loci in differentiation.

Keratinocyte Differentiation Is Accompanied by Substantial Changes in Enhancer Usage

We defined enhancers as regions harboring an H3K4me1^{+}/H3K4me13$^{-/low}$ signature at a distance of >2.5 kb from any promoter. Enhancers were considered active when marked by H3K27ac (Ernst et al., 2011; Heintzman et al., 2009; Rada-Iglesias et al., 2011). We mapped 70,011 enhancers in KPs and 84,414 in DKs, located on average 43 kb away from any promoter, 14.5% and 21% of which were marked by H3K27ac. Overall, 1,000 intergenic or intronic H3K27ac^{+} enhancers were actively transcribed, as indicated by overlapping CAGE tags (Figure 3A). These were mainly cell-specific, CpG-poor CAGE clusters (Figure 3B) driving the expression of annotated ncRNAs.

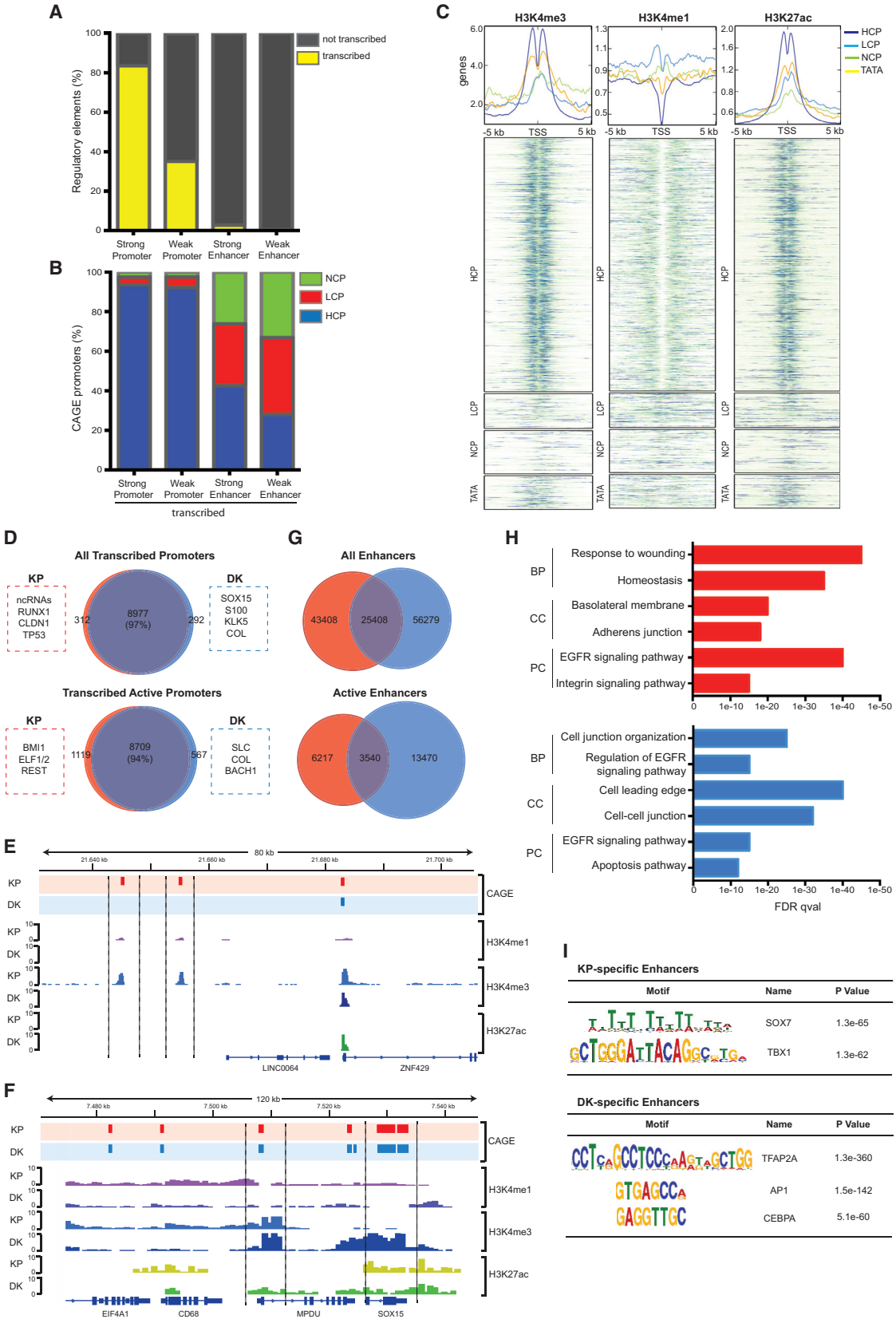
More than 60% of the H3K4me1^{+} regions were uniquely mapped in either KPs or DKs, and among those mapped in both cell types, 24.4% were active (H3K27ac^{+}) exclusively in KPs and 53.3% exclusively in DKs (Figure 3G). Interestingly, the intensity of H3K4me1 and H3K27ac deposition at enhancers correlated with the expression level of the closest CAGE promoter (Figure S4E). Functional annotation of KP- and DK-specific active enhancers using the GREAT tool showed their association with common epidermal cell functions, such as cell-junction organization, integrin and epidermal growth factor receptor pathways, as well as progenitor-specific processes such as homeostasis and wound healing (Figure 3H). The most abundant TF binding motifs present in KP-specific enhancers were those for SOX7 and TBX1, involved in stem cell and mouse hair follicle homeostasis (Chen et al., 2012; Tan et al., 2013), while DK-specific enhancers were highly enriched in binding motifs of differentiation-related TFs such as TFAP2, AP1, and CEBPA (Fuchs, 2009; Lopez et al., 2009; McDade et al., 2012) (Figure 3I).

Super-Enhancers Define Core Transcriptional Regulatory Networks in Epithelial Differentiation

We used H3K27ac ChIP-seq data to identify SEs, i.e., large clusters of enhancers that drive the expression of genes essential for the definition of cell identity (Hnisz et al., 2013; Whyte et al., 2013). We retrieved 953 SEs in KPs and 1,090 in DKs (Figure 4A), a substantial portion of which (56% and 61%, respectively) was unique for each cell type and associated with cell-specific genes. SE-associated genes were expressed at higher levels than genes associated with typical enhancers ($p < 2.2 \times 10^{-16}$) (Figure 4B), and most of them encoded TFs and proteins necessary for key epidermal functions, such as laminins, keratins, cell-adhesion complexes, and components of the TGF, WNT, and SMAD signaling pathways. Among the >200 TF genes associated with SEs, we found fundamental regulators of skin and stem cell biology, such as *TP63*, *SOX9*, *SOX15*, *RUNX1*, *FOXP1*, *TCF4*, *TP53*, *MYC*, *KLF4*, and *TFAP2*. ncRNAs were strongly associated with SEs, including the keratinocyte-specific mir-203.

TF binding motifs for p63 and FOXP1 binding sites were significantly enriched in SEs of both KPs and DKs, while SMAD motifs where specifically enriched in KP-specific SEs and differentiation-related KLF5, AP1 and TFAP2C motifs in DK-specific SEs (Figure 4C). To validate the cues provided by TF-motif discovery, we mapped by ChIP-seq the p63 binding sites in our DK population: virtually all p63 sites overlapped with those previously identified in DKs (Kouwenhoven et al., 2015), validating the use of the latter dataset in our analyses (Figure S5A). Over 80% of the SEs in both KPs and DKs overlapped with at least one p63 binding site, a significantly higher proportion compared to the total enhancer population (35%) (Figure 4D). p63 binding sites were found in the SEs of the *TP63* gene itself (Figure 4E) and in SEs associated with genes encoding TFs enriched in keratinocyte-specific enhancers and SEs, such as *TFAP2A*, *RUNX1*, *SOX9*, *MYC*, *FOXP1*, *SMAD3*, and *KLF5*.

Interestingly, >50% of the SEs bound by p63 in KPs and DKs were cell-specific, indicating that p63 binds and controls cell-specific regulatory regions in both progenitors and differentiated cells. When we integrated genes associated with p63-bound SEs into molecular and transcriptional interaction networks, we observed that KP and DK networks barely overlapped, with only six nodes in common (*SOX9*, *SMAD7*, *LAMC2*, *RAD51B*, *GRHL3*, *EFNB1*) and different hubs. Genes in the KP network are involved in the developmental control of organ and epithelial tissue homeostasis (main hubs: *RUNX2*, *RUNX1*, *CEBPD*, *SOX9*, *JUNB*, *ETS*, *HMGA2*, *STAT6*), while genes in the DK network are involved in signal transduction, cell communication,



(legend on next page)

cell size and apoptosis (main hubs: *TP53*, *CREB1*, *P21*, *YAP1*, *KLF4*, *KLF5*, *HES1*, *SOX9*, *ETS1*) (Figure S5B). When we analyzed all genes driven by p63-bound promoters, enhancers, or SEs, we found 825 KP-specific genes involved in the control of cell cycle and epidermal proliferation, and 591 DK-specific genes encoding lipoproteins and intermediate filament components involved in keratinization and epithelium differentiation (Figure S5C). These data indicate that p63 regulates distinct sets of genes at different differentiation stages through binding of stage-specific regulatory elements.

To gain insight into the combinatorial interactions among TFs operating on SEs, we looked at TF motifs in a 50-bp window around p63 binding sites, and discovered a specific enrichment of TCF4 and SMAD3 motifs in KP-specific SEs and of AP1 in DK-specific SEs. In particular, TCF4 seems to be uniquely enriched next to p63 binding sites in KP-specific SEs. When integrating genes associated with SEs enriched in these specific motifs into interaction networks, we found a tight and specific connection between the TFs and their target SEs in both KPs and DKs, with a significant overlap among target genes associated with SEs containing p63, SMAD3, and TCF4 motifs in progenitors and those containing p63 and AP1 motifs in differentiated cells (Figures S5D and S5E).

Discovering Transcriptional Active Regulatory Elements by Retroviral Integration Site Analysis

We decided to use retroviral integration sites to identify active regulatory elements retrospectively in cells maintaining epithelial cultures in vitro, a bona fide approximation of KSCs. Early-passage (P2) foreskin-derived primary keratinocytes were co-cultured on an NIH3T3-J2 feeder layer to maintain stem cell activity, and transduced with a GFP-expressing MLV vector. Cells were subcultured for six passages (>35 cell doublings) to exhaust the populations of TA progenitors and enrich for the progeny of culture-maintaining KSCs (Figures S6A and S6B). A CFE assay indicated progressive decrease of clonogenic cells and increase of abortive colonies (Figures S6C and S6D). In parallel, we transduced a population of DKs that were collected 72 hr after infection (Figure S6A). Genomic DNA was extracted from the two transduced cell populations, and MLV integration sites mapped genome-wide as previously described (Cattoglio et al., 2010). We mapped 10,819 MLV integration sites in the progeny of KSCs and 9,815 in DKs, and identified 1,478 and 1,326 integration clusters, respectively, as defined by comparison with an adjusted random distribution (Cattoglio et al., 2010).

To validate retroviral scanning as a tool for the identification of regulatory elements, we analyzed the genomic characteristics of the 1,326 clusters mapped in DKs. All clusters overlapped with epigenetically defined active regulatory regions, and in particular 79% with strong enhancers and 19.7% with strong promoters ($p < 10^{-22}$ compared with random sites) (Figures 5A and 5B). Clusters associated with promoters mapped predominantly (82%) in a ±2.5-kb window around TSSs, while those associated with enhancers were in intergenic (50.9%) or intragenic (38.7%) locations >35 kb away from any TSS (Figures 5A and S5E). All clusters showed a strong preference for H3K27ac, conserved non-coding sequences (Figure 5C), and open chromatin regions identified by FAIRE (formaldehyde-assisted isolation of regulatory elements) sequencing and DNase sequencing in keratinocytes (Figure 5D). The average expression level of CAGE promoters in a ±100-kb window from MLV clusters was significantly higher in DKs than in an unrelated control cell population,

Figure 3. Promoter and Enhancer Regions Involved in the Regulation of Keratinocyte Differentiation

(A) Percentage of promoters and enhancers overlapping with sites of active transcription, as detected by DeepCAGE.

(B) Proportion of CAGE-defined TSSs overlapping with transcribed promoters and enhancers falling in the NCP, LCP, and HCP categories.

(C) ChIP-Seq density profiles and heat maps for the H3K4me1, H3K4me3, and H3K27ac histone marks within each CAGE promoter category (HCP, LCP, NCP, and $TATA^{+}$).

(D) Venn diagrams showing genome-wide overlap of transcribed promoter regions between KPs and DKs. Promoters were defined as unique when present in only one dataset (in the "All Transcribed Promoters" category), and " unique active" when mapped in both datasets but active ($H3K27ac^{+}$) only in one of them (in the "Transcribed Active Promoters" category).

(E) Genomic browser screenshot of KP-specific H3K4me3 peaks overlapping with KP-high CAGE promoters that map TSSs of transcripts with unknown function.

(F) Genomic browser screenshot of DK-specific promoters. The promoter of *SOX15* is marked with H3K4me3 uniquely in DKs, while the promoter of *MPDU* shows H3K4me3 signal in both KPs and DKs, but is active only in DKs. Both promoters are marked by DK-high CAGE tags.

(G) Venn diagrams showing genome-wide overlap of enhancers between KPs and DKs. Enhancers were defined as "unique" when present in only one dataset, and "unique active" when mapped in both datasets but active ($H3k27ac^{+}$) only in one of them.

(H) Annotation of KP (red) and DK (blue) active enhancers using GREAT. Gene ontologies are listed by biological process (BP), cellular component (CC), and pathway common (PC) categories. The x axis values correspond to binomial false discovery rate (FDR) (corrected) q values.

(I) Selected TF binding sequence motifs enriched at KP- and DK-unique active enhancers.

See also Figure S4.

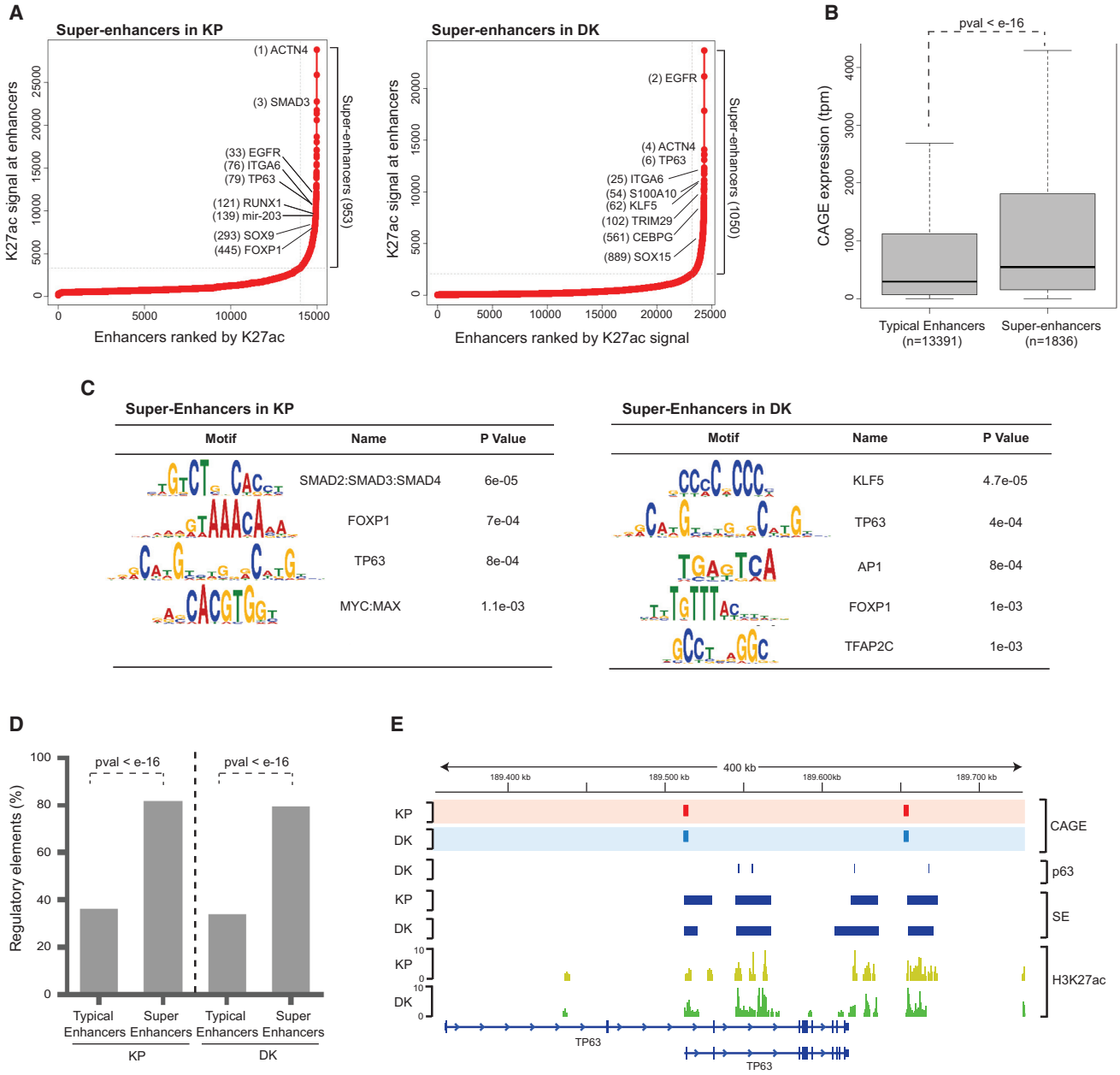


Figure 4. Super-Enhancers Define Specific Regulatory Networks in KPs and DKs

(A) Distribution of H3K27ac ChIP-seq signals across all the H3K27ac-containing enhancers (x axis), where enhancers are ranked by increasing H3K27ac ChIP-seq signal. Super-enhancers are found above the inflection point of the curve. Biologically relevant super-enhancers are highlighted together with their ranks and associated genes.

(B) Expression levels of CAGE promoters associated with typical enhancers and super-enhancers. Boxes show median line and quartiles, whiskers show the minimum and maximum boundary (1.5 times of the interquartile range from the first and third quartile) to define outliers. The p value was calculated using an unpaired Wilcoxon test.

(C) Selected TF binding sequence motifs enriched at enhancers in KPs and DKs.

(D) Percentages of typical enhancers and super-enhancers bound by p63. p Values were calculated using two-sample test for equality of proportion.

(E) Genome browser snapshot of KP- and DK-specific super-enhancers at the *TP63* locus. CAGE tags mark the TSS of the shortest *TP63* isoform and of an unknown transcript next to the 3′ end of the gene. p63 binds to its own super-enhancers, as defined by p63 ChIP-seq. See also Figure S5.

consistent with their enhancer function (Figure 5E). The 30% fraction of transcribed enhancers targeted by MLV integration was significantly more transcribed ($p < 10^{-5}$) than the average population (Figure 5F). Interestingly, 64% of the SEs were hit by at least one MLV integration compared with 7.8% of random sites ($p < 10^{-22}$). Functional annotation performed by GREAT showed a correlation between cluster-associated genes and differentiated cell functions, such as apoptosis, cholesterol biosynthesis, and FAS-, TAp63-, and TP53-linked pathways (Figure 5G).

To further validate the regulatory nature of the regions identified by the MLV clusters, we randomly chose 12 cluster regions and tested their transcriptional activity by a luciferase reporter assay in primary human keratinocytes: 6 of 12 regions scored positive for enhancer function and 2 of 12 for repressor function (Figure S6F).

Retroviral Scanning Uncovers Regulatory Regions Associated with Stem Cell Functions in Retrospectively Defined KSCs

MLV clusters mapped in the progeny of KSCs were intersected with those mapped in DKs to identify common and cell-specific regulatory regions. Less than 15% (195) of the KSC clusters overlapped for at least one base pair with any DK cluster, and <3% (41) overlapped completely, indicating that only a minority of the regulatory regions identified by MLV scanning was shared between the two populations. Only 28% of the remaining 1,283 KSC-specific clusters overlapped with enhancers epigenetically defined in KPs, indicating that MLV scanning identifies a set of potentially stem cell-specific regulatory elements.

KSC-specific clusters were associated with genes with stem cell-related functions, such as *LRIG1*, *ITGB1*, *ITGA6*, *YAP*, *MCSP*, and *WNT10A*. Functional annotation showed a clear correlation with developmentally regulated genes associated with regeneration, wound healing, anchoring and adherence junctions, and ITGB1 and TP63 signaling pathways (Figure 5G). No cluster mapped to the EDC or other genes associated with terminal differentiation functions. qPCR analysis showed that the expression of 9 out of 16 (56.3%) randomly chosen transcripts associated with KSC-specific clusters was higher in KPs than in DKs, indicating that putative stem cell-specific enhancers retain a higher activity in progenitors than in differentiated cells (Figure S6G). Expression of 7 out of 16 transcripts was barely detectable, suggesting that they represent stem cell-specific transcripts downregulated in both KPs and DKs.

KSC-Specific Regulatory Regions Are Characterized by a Unique Combination of Epithelial-Specific TF Binding Sites

A de novo search of TF binding motifs in a ±1-kb interval from KSC- and DK-specific clusters uncovered the same motifs enriched in SEs, and particularly p63 binding sites (Figure 6A). p63 bound 47% of the sequences flanking MLV integration sites in DKs, and up to 73% when considering only SE-associated sites, a significant increase with respect to the 2.4% observed for random control sequences ($p < 10^{-16}$) (Figure 6B). Interestingly, the majority of the genes encoding TFs whose motifs are enriched in SEs and MLV clusters are in turn associated with SEs and MLV clusters (Figure 6C). Some of these genes, like *SMAD3*, *SOX9*, and *RUNX1*, were associated with KSC-specific MLV clusters overlapping KP-specific SEs (Figures 6C and 6D), and therefore identify TFs important for the execution of transcriptional programs in both stem and progenitor cells. Other TFs, such as *TCF4* and *SOX7* (Figures 6C and 6E), were associated with KSC-specific clusters but not KP- or DK-specific SEs, and may thus be involved in the execution of a more stem cell-specific program. These TFs were significantly more expressed in KPs than in DKs, as indicated by CAGE and RNA-seq analysis, and are known to play pivotal roles in the biology of murine hair follicle stem cells (Beck and Blanpain, 2012; Scheitz and Tumbar, 2013).

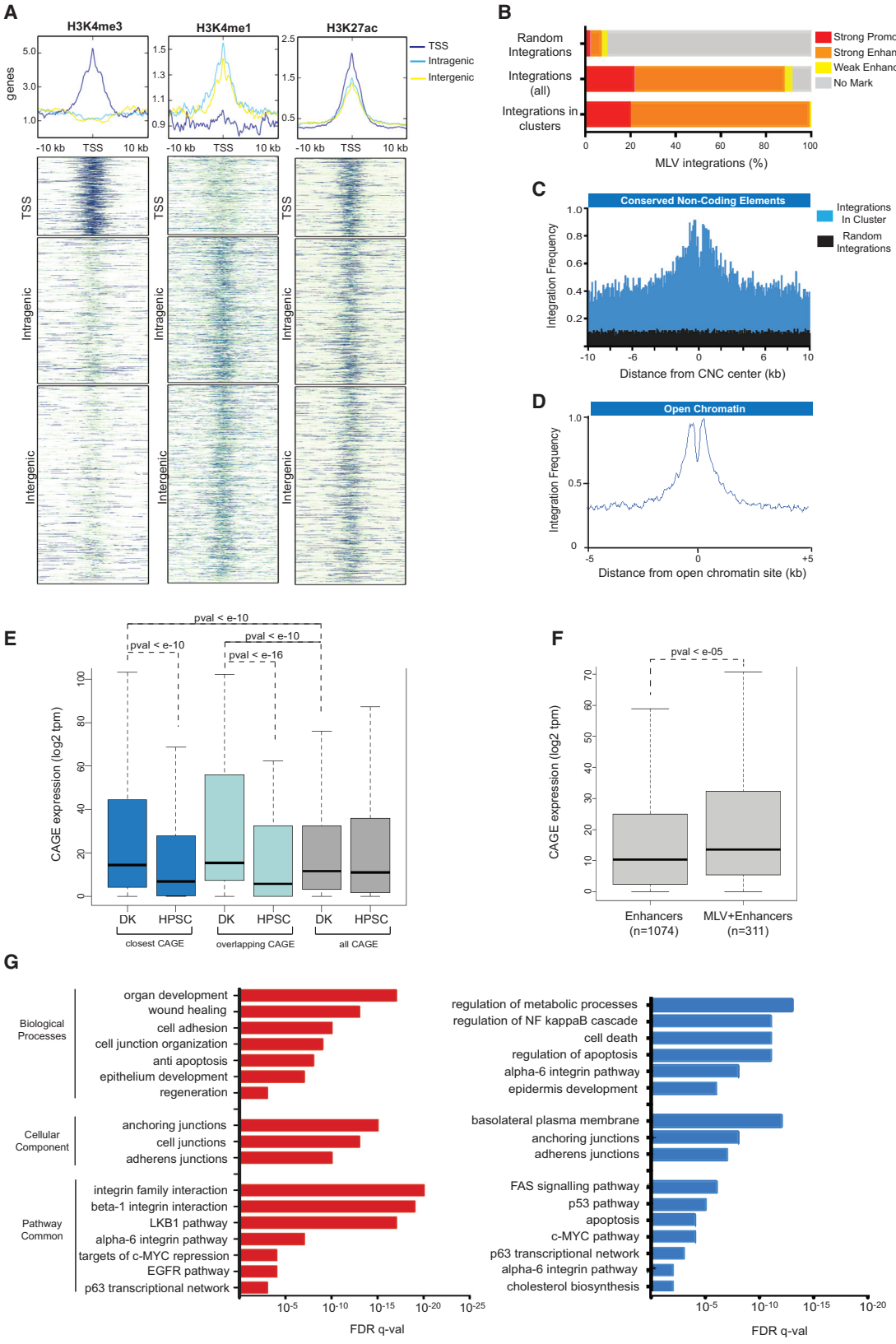
These analyses indicate that the regions uniquely identified by MLV scanning in retrospectively defined KSCs (see list in Table S1) represent bona fide stem cell-specific enhancers.

DISCUSSION

The hierarchy of keratinocyte stem and progenitor cells is defined by cell kinetics parameters, and is an ideal model to study transcriptional and chromatin dynamics driving differentiation of a human somatic stem cell. In this study, we mapped transcripts and transcriptional regulatory elements in prospectively isolated DKs and KPs, and retrospectively defined KSCs. We correlated CAGE promoter maps with epigenetic annotations of active promoters, enhancers, and SEs obtained by ChIP-seq, and integrated this information to discover shared or stage-specific regulatory elements.

Differentiation of Keratinocytes from Progenitors Is Determined by Quantitative Regulation of a Common Set of Promoters

The use of CAGE, ChIP-seq, and RNA-seq allowed the description of two different transcriptomes in KPs and DKs and a robust definition of promoters and their usage. We found that most of the >14,000 mapped promoters are shared between KPs and DKs and differentially expressed, indicating that the substantial transcriptome changes associated with differentiation are determined by quantitative regulation of promoters engaged in both



(legend on next page)

progenitors and differentiated cells rather than by the activation or silencing of stage-specific ones. The few, strictly stage-specific promoters were mostly unannotated or associated with non-coding transcripts and particularly lncRNAs, influential players in the control of lineage commitment and tissue identity. Three-quarters of the shared promoters showed housekeeping characteristics ($TATA^-$ and high-CpG content), while the proportion of the $TATA^+$/low-CpG promoters progressively increased in highly regulated and strictly cell-specific categories. Combining CAGE annotation with histone modification marks showed that the majority of the epigenetically defined "strong" (acetylated) promoters overlapped with CAGE promoters, while only one-fourth of the non-acetylated promoters were actually transcribed. The large overlap in promoter regions between KPs and DKs was found also at the epigenetic level, with just 300 regions specific for progenitors and just as few for differentiated cells. However, the intensity of the promoter-specific histone modifications differed in the two cell types and directly correlated with transcriptional activity. Transcriptional regulation is therefore accompanied by modest, essentially quantitative changes in histone modifications during keratinocyte differentiation, suggesting that the epigenetic landscape around promoters is already established at the progenitor state. Interestingly, silencing and downregulation of a large set of stem/progenitor cell-related genes in KPs and DKs was associated with H3K27 methylation of both promoters and enhancers, suggesting Polycomb-group-mediated repression as a mechanism for negative gene regulation in keratinocyte differentiation. Finally, CAGE analysis identified alternative transcripts in more than 1,100 protein-coding genes. Half of the alternative transcripts showed stage-specific changes in expression level, indicating that switching between alternative protein isoforms is an inherent part of the keratinocyte differentiation program.

Keratinocyte Differentiation Is Accompanied by Dramatic Changes in Enhancer Usage

Strikingly, enhancers were much more regulated than promoters during epithelial differentiation: more than 65% of the acetylated $H3K4me1^+$ regions were strictly stage specific, indicating that enhancers are dramatically redefined during the KP-to-DK transition, and that differential enhancer usage is responsible for the quantitative regulation of promoter activity. Although the role of enhancers has been identified in other differentiation models (Creyghton et al., 2010; Heintzman et al., 2009; Rada-Iglesias et al., 2011), the difference between KPs and DKs is particularly striking given their developmental proximity. Functional annotation of active enhancers showed association with common epithelial pathways in both cells, but also cell-specific pathways such as wound healing in KPs and cell motility and apoptosis in DKs. We identified approximately 1,000 transcribed enhancers in both cell populations, which were mainly cell specific and drove the expression of annotated ncRNAs, consistent with previous reports (Andersson et al., 2014).

Super-Enhancers and TF Regulatory Circuits Play a Major Role in Epithelial Differentiation

SEs are large, highly acetylated clusters of transcriptional enhancers that drive the expression of cell-specific genes defining cell and tissue identity (Hnisz et al., 2013; Whyte et al., 2013). We mapped ~1,000 SEs in both KPs and DKs, the majority of which was cell-specific and associated with cell-specific genes playing key functions in epithelial homeostasis, as already shown in the murine hair follicle (Adam et al., 2015). These included laminins, keratins, cell-adhesion complexes, and components of the TGF, WNT, and SMAD signaling pathways, but also master regulators of skin and stem cell biology, such as p63, FOXP1, MYC, and KLF4, and ncRNAs such as *mir-203*, a suppressor of p63 and a key promoter of keratinocyte differentiation

Figure 5. MLV Integration Clusters Mark Regulatory Regions Associated with Cell-Specific Functions

(A) ChIP-Seq density profiles and heatmaps are shown for the H3K4me1, H3K4me3, and H3K27ac histone marks within each MLV integration category (TSS-proximal, intragenic, and intergenic).

(B) Percentages of randomly generated sites and clustered or unclustered MLV integration sites associated with epigenetically defined regulatory regions.

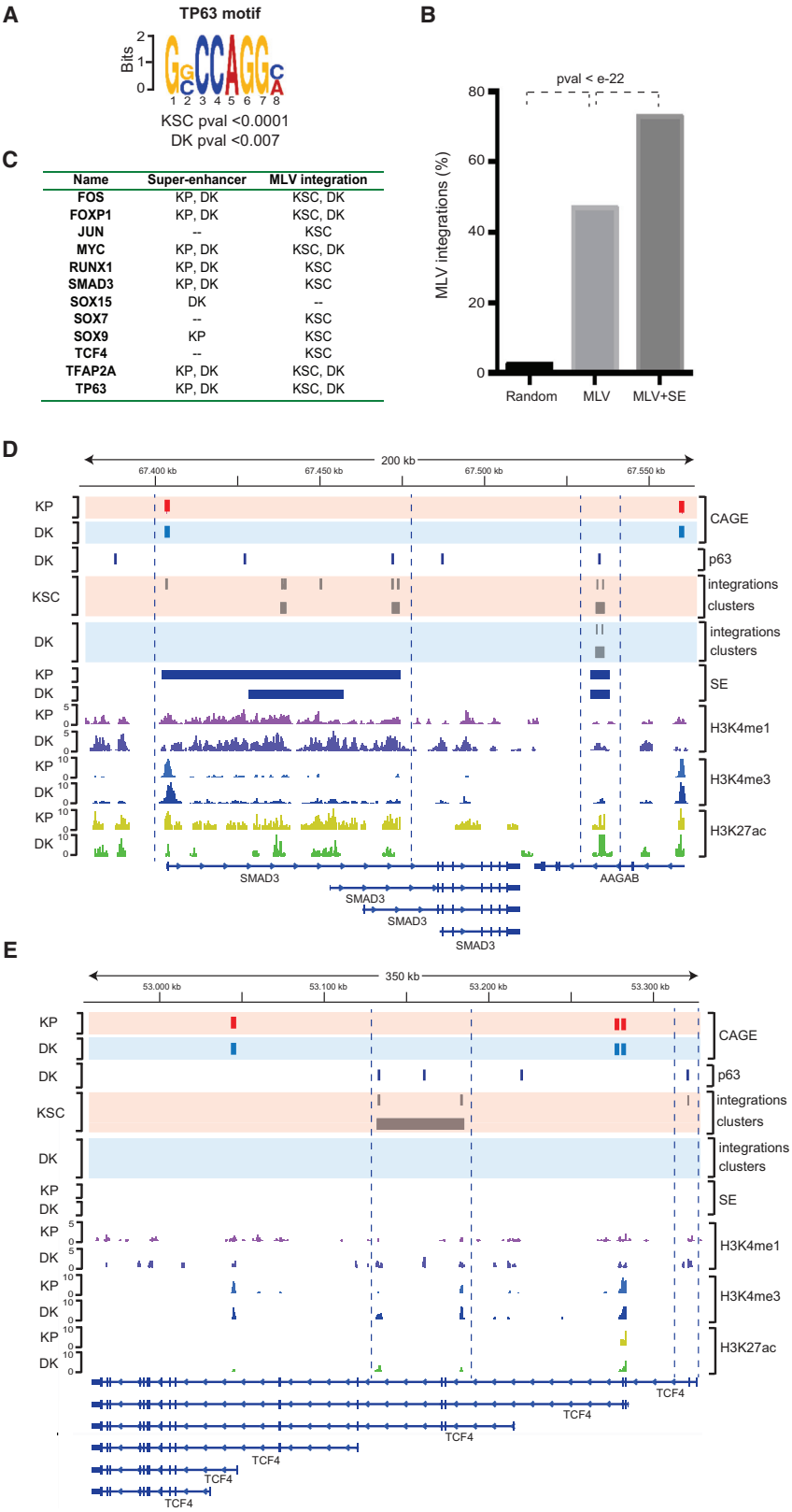
(C and D) Distribution of the distance of MLV integration clusters from the midpoint of (C) conserved non-coding (CNC) elements in a 20-kb window, and of (D) open chromatin regions defined by FAIRE sequencing in a 10-kb window.

(E) Differential expression levels (log_2 of CAGE TPM values) in DKs or hematopoietic stem-progenitor cells (HSPC) of CAGE promoters proximal to, or overlapping, MLV integration clusters in DKs.

(F) Differential expression levels of total transcribed enhancers and transcribed enhancers marked by MLV integration sites. Boxplots in (E) and (F) show median line and quartiles, whiskers show the minimum and maximum boundary (1.5 times of the interquartile range from the first and third quartile) to define outliers. p Values were calculated using an unpaired Wilcoxon test.

(G) Functional annotation of KSC-specific (red) and DK-specific (blue) enhancers identified by MLV clusters using GREAT. Gene ontologies are listed by biological process, cellular component, and pathway common categories. The x axis values correspond to binomial FDR (corrected) q values.

See also Figure S6.



Name	Super-enhancer	MLV integration
FOS	KP, DK	KSC, DK
FOXP1	KP, DK	KSC, DK
JUN	--	KSC
MYC	KP, DK	KSC, DK
RUNX1	KP, DK	KSC
SMAD3	KP, DK	KSC
SOX15	DK	--
SOX7	--	KSC
SOX9	KP	KSC
TCF4	--	KSC
TFAP2A	KP, DK	KSC, DK
TP63	KP, DK	KSC, DK

Figure 6. Retroviral Scanning Identifies Potential Regulatory Networks in Retrospectively Defined KSC

(A) The TP63 motif found overrepresented in a 500-bp window around the center of MLV integration clusters in both KSCs and DKs.

(B) Percentages of randomly generated sites, MLV integration sites, and MLV integration sites mapping in SEs bound by p63 in DKs. p Values were calculated using a two-sample test for equality of proportion.

(C) List of key transcription factors associated with SEs and MLV integration sites in KPs, DKs, and KSCs.

(D) Genome browser snapshot of the *SMAD3* gene locus, harboring KSC-specific MLV integration clusters that overlap with p63 binding sites and with a SE in both KPs and DKs. *SMAD3* transcription (CAGE TPM values) is higher in KPs than in DKs. The expression of a close gene, *AAGAB*, is instead not significantly different between KPs and DKs, and its genomic locus is marked by the same MLV clusters and SEs in both cell types.

(E) Genome browser snapshot of the *TCF4* gene locus. The locus is barely marked by active histone modifications in both KPs and DKs, but harbors KSC-specific MLV clusters that represent putative KSC-specific regulatory regions.

(Yi et al., 2008). SEs in both KPs and DKs were particularly enriched in binding motifs for FOXP1, a regulator of hair follicle quiescence and activation (Leishman et al., 2013), and for the master regulator p63. Actual, ChIP-seq-mapped binding sites for p63 were highly enriched in SEs, validating the predictions provided by TF-motif discovery and indicating the pervasive role of p63 in the control of epithelial SE function. Interestingly, p63 binding sites were enriched in SEs associated with genes encoding p63 itself, FOXP1, and other key TFs binding to keratinocyte-specific enhancers and SEs, such as TFAP2A, RUNX1, SOX9, MYC, SMAD3, AP1, and KLF5. Finally, ChIP-seq and TF-motif analysis indicate that even though p63 is a master regulator throughout keratinocyte differentiation (Kouwenhoven et al., 2015), it regulates distinct sets of genes at each stage through binding of stage-specific promoters, enhancers, and SEs and in combination with stage-specific TFs.

Identification of Regulatory Networks in KSCs by Retrospective MLV Scanning

To identify enhancers and SEs in KSCs, a rare population which lacks robust markers for prospective isolation, we used MLV scanning as a technique for their retrospective identification: primary cultures were transduced by an MLV vector and integration sites were mapped in the progeny of long-term keratinocyte culture-maintaining cells, a characteristic that bona fide defines self-renewing KSCs. The MLV pre-integration complex specifically interacts through its integrase component with proteins (BET) that tether integration to highly acetylated, transcriptionally active regions (De Rijck et al., 2013; Gupta et al., 2013; Sharma et al., 2013). MLV integration clusters can therefore be used as surrogate markers of promoters, enhancers, and SEs, as previously reported in hematopoietic cells (Biasco et al., 2011; Cattoglio et al., 2010; De Ravin et al., 2014). We validated this concept also in DKs by correlating MLV integration clusters with CAGE and ChIP-seq data: MLV clusters were preferentially associated with SEs, probably due to their highly acetylated state, and genes associated with clusters included important regulators of epidermal differentiation and homeostasis such as p63, FOXP1, SOX9, SMAD3, KLF4, GATA3, GRHL3, and TFAP2.

More than 85% of the 1,327 MLV clusters mapped in KSCs were specific to these cells and showed no overlap with regulatory regions defined in KPs or DKs. Many of these KSC-specific regions were associated with genes known to play a role in epidermal stem cell functions, such as *LRIG1*, *ITGB1*, *ITGA6*, *YAP*, *MCSP*, or *WNT10A*, and none was associated with the EDC complex or genes necessary for differentiated cell functions. KSC-specific clusters showed an exceedingly high frequency of p63 binding sites, and binding motifs for other TFs identified also in KP enhancers. Some of these genes, such as *SMAD3*, *SOX9*, and *RUNX1*, and *TP63* itself, were associated with KSC-specific clusters that overlapped to KP-specific SEs, identifying TFs important for the execution of transcriptional programs in both stem and progenitor cells. Other TFs, such as *TCF4* and *SOX7*, known to play pivotal roles in the biology of murine hair follicle stem cells (Beck and Blanpain, 2012; Scheitz and Tumbar, 2013), were associated with KSC-specific clusters but not KP- or DK-specific SEs, and might thus represent TFs involved in the execution of a more stem cell-specific program. In general, the TF circuitries identified by KSC-specific clusters are in close agreement with previous studies in epidermal murine models, which demonstrated the importance of Sox, Ets, and the Wnt and Bmp signaling pathways—to which TCF4 and SMAD3 belong—in ectodermal and epidermal development, and the importance of MYC and GATA3 in keratinocyte differentiation (Fuchs, 2007). Moreover, TFAP2A, RUNX1, and AP1 were shown to cooperate with the epithelial master regulator p63 in the specification of the epidermal fate and differentiation programs (Kouwenhoven et al., 2015; McDade et al., 2012).

The analysis of SEs in progenitors and DKs, and of MLV clusters in stem cells, identify a complex regulatory and auto-regulatory TF network with p63 as the central player, which regulates the specification of the stem and progenitor cell identity and the execution of their differentiation program. In embryonic stem cells, TFs of the pluripotency module form an auto-regulatory loop whereby they cooperatively bind to their promoters and regulate their own expression as well as that of other TFs and ncRNAs, which form a core regulatory circuitry driven in large part by the activity of SEs (Whyte et al., 2013). Our data indicate that SE-mediated auto- and cross-regulatory TF circuitries play a key role in mediating identity and differentiation also in somatic cells, and particularly in the human epithelium.

EXPERIMENTAL PROCEDURES

Cell Culture

Human primary keratinocytes were obtained from foreskin biopsies of healthy donors and expanded on an NIH3T3-J2 cell feeder in FAD medium. KPs were obtained by collagen IV adherence assay. Keratinocyte differentiation was induced by cell-contact inhibition and by exclusion of several growth factors from the medium. See also Supplemental Experimental Procedures.

DeepCAGE

RNA from three different KP selection experiments was isolated using an RNeasy Plus Mini kit (Qiagen) and pooled together. The DeepCAGE library was prepared by DNAFORM at RIKEN Omics Science Center, as described previously (Carninci et al., 2006). Samples were sequenced using the Illumina GA II sequencer,

with an average read length of 36 bases, and tags were extracted and mapped to the hg19 genome. See also Supplemental Experimental Procedures.

Gene-Expression Analysis

Expression profiles were determined by RNA-seq analysis. RNA-seq libraries were prepared from 300 ng of RNA, and 75-bp single-end sequences were obtained on a NextSeq 500 Instrument (Illumina). Sequence tags were mapped to the hg19 genome using TopHat v2.0.6 and transcript levels were calculated using Cufflinks v2.0.2. See also Supplemental Experimental Procedures.

ChIP-Seq

Chromatin was prepared from KPs and DKs and immunoprecipitated with antibodies against H3K4me1, H3K4me3, and H3K27ac, as previously described (Cattoglio et al., 2010). After Illumina sequencing, raw reads were mapped to the hg19 genome using Bowtie (Langmead et al., 2009) and ChIP-seq peaks were called using SICER default parameters (Zang et al., 2009) and using each INPUT data to model the background noise. See also Supplemental Experimental Procedures.

Retroviral Scanning

The MLV-derived retroviral vector expressing GFP under a modified LTR control (MFG.GFP*mod*) was used to transduce KSCs and DKs. Retroviral integration sites were mapped by linker-mediated PCR and Roche/454 pyrosequencing as previously described (Cattoglio et al., 2007). See also Supplemental Experimental Procedures.

ACCESSION NUMBERS

The DeepCAGE, ChIP-seq, and microarray data were deposited in the Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) under accession number GEO: GSE64328. MLV integration sites sequencing data were deposited in the Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra) under accession number SRA: SRP051203.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, six figures, and one table and can be found with this article online at http://dx.doi.org/10.1016/j.stemcr.2016.03.003.

AUTHOR CONTRIBUTIONS

A.C. and F.M. conceived and designed the study; A.C., A.M., and O.R. obtained the data; A.C., L.P., G.M.T., M.S., and E.R. performed computational and statistical analyses under the supervision of C.P., G.D.B., and S.B.; A.C. and F.M. wrote the manuscript, with input from all authors.

ACKNOWLEDGMENTS

This work was supported by grants from the European Research Council (ERC-2010-AdG, GT-SKIN) and the Italian Ministry of Education, University and Research (EPIGEN Epigenomics Flagship Project).

Received: June 15, 2015
Revised: March 5, 2016
Accepted: March 7, 2016
Published: March 31, 2016

REFERENCES

Adam, R.C., Yang, H., Rockowitz, S., Larsen, S.B., Nikolova, M., Oristian, D.S., Polak, L., Kadaja, M., Asare, A., Zheng, D., et al. (2015). Pioneer factors govern super-enhancer dynamics in stem cell plasticity and lineage choice. Nature *521*, 366–370.

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., et al. (2014). An atlas of active enhancers across human cell types and tissues. Nature *507*, 455–461.

Barrandon, Y., and Green, H. (1987). Three clonal types of keratinocyte with different capacities for multiplication. Proc. Natl. Acad. Sci. USA *84*, 2302–2306.

Beck, B., and Blanpain, C. (2012). Mechanisms regulating epidermal stem cells. EMBO J. *31*, 2067–2075.

Biasco, L., Ambrosi, A., Pellin, D., Bartholomae, C., Brigida, I., Roncarolo, M.G., Di Serio, C., von Kalle, C., Schmidt, M., and Aiuti, A. (2011). Integration profile of retroviral vector in gene therapy treated patients is cell-specific according to gene expression and chromatin conformation of target cell. EMBO Mol. Med. *3*, 89–101.

Blanpain, C., Horsley, V., and Fuchs, E. (2007). Epithelial stem cells: turning over new leaves. Cell *128*, 445–458.

Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., Semple, C.A., Taylor, M.S., Engstrom, P.G., Frith, M.C., et al. (2006). Genome-wide analysis of mammalian promoter architecture and evolution. Nat. Genet. *38*, 626–635.

Cattoglio, C., Facchini, G., Sartori, D., Antonelli, A., Miccio, A., Cassani, B., Schmidt, M., von Kalle, C., Howe, S., Thrasher, A.J., et al. (2007). Hot spots of retroviral integration in human CD34+ hematopoietic cells. Blood *110*, 1770–1778.

Cattoglio, C., Pellin, D., Rizzi, E., Maruggi, G., Corti, G., Miselli, F., Sartori, D., Guffanti, A., Di Serio, C., Ambrosi, A., et al. (2010). High-definition mapping of retroviral integration sites identifies active regulatory elements in human multipotent hematopoietic progenitors. Blood *116*, 5507–5517.

Chen, T., Heller, E., Beronja, S., Oshimori, N., Stokes, N., and Fuchs, E. (2012). An RNA interference screen uncovers a new molecule in stem cell self-renewal and long-term regeneration. Nature *485*, 104–108.

Creyghton, M.P., Cheng, A.W., Welstead, G.G., Kooistra, T., Carey, B.W., Steine, E.J., Hanna, J., Lodato, M.A., Frampton, G.M., Sharp, P.A., et al. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc. Natl. Acad. Sci. USA *107*, 21931–21936.

De Ravin, S.S., Su, L., Theobald, N., Choi, U., Macpherson, J.L., Poidinger, M., Symonds, G., Pond, S.M., Ferris, A.L., Hughes, S.H., et al. (2014). Enhancers are major targets for murine leukemia virus vector integration. J. Virol. *88*, 4504–4513.

De Rijck, J., de Kogel, C., Demeulemeester, J., Vets, S., El Ashkar, S., Malani, N., Bushman, F.D., Landuyt, B., Husson, S.J., Busschots, K.,

et al. (2013). The BET family of proteins targets Moloney murine leukemia virus integration near transcription start sites. Cell Rep. *5*, 886–894.

Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., et al. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. Nature *473*, 43–49.

Frye, M., and Benitah, S.A. (2012). Chromatin regulators in mammalian epidermis. Semin. Cell Dev. Biol. *23*, 897–905.

Fuchs, E. (2007). Scratching the surface of skin development. Nature *445*, 834–842.

Fuchs, E. (2009). Finding one's niche in the skin. Cell Stem Cell *4*, 499–502.

Gupta, S.S., Maetzig, T., Maertens, G.N., Sharif, A., Rothe, M., Weidner-Glunde, M., Galla, M., Schambach, A., Cherepanov, P., and Schulz, T.F. (2013). Bromo- and extraterminal domain chromatin regulators serve as cofactors for murine leukemia virus integration. J. Virol. *87*, 12721–12736.

Heintzman, N.D., Hon, G.C., Hawkins, R.D., Kheradpour, P., Stark, A., Harp, L.F., Ye, Z., Lee, L.K., Stuart, R.K., Ching, C.W., et al. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature *459*, 108–112.

Hnisz, D., Abraham, B.J., Lee, T.I., Lau, A., Saint-Andre, V., Sigova, A.A., Hoke, H.A., and Young, R.A. (2013). Super-enhancers in the control of cell identity and disease. Cell *155*, 934–947.

Jones, P.H., and Watt, F.M. (1993). Separation of human epidermal stem cells from transit amplifying cells on the basis of differences in integrin function and expression. Cell *73*, 713–724.

Kouwenhoven, E.N., Oti, M., Niehues, H., van Heeringen, S.J., Schalkwijk, J., Stunnenberg, H.G., van Bokhoven, H., and Zhou, H. (2015). Transcription factor p63 bookmarks and regulates dynamic enhancers during epidermal differentiation. EMBO Rep. *16*, 863–878.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*, R25.

Lee, S.E., Sada, A., Zhang, M., McDermitt, D.J., Lu, S.Y., Kemphues, K.J., and Tumbar, T. (2014). High Runx1 levels promote a reversible, more-differentiated cell state in hair-follicle stem cells during quiescence. Cell Rep. *6*, 499–513.

Leishman, E., Howard, J.M., Garcia, G.E., Miao, Q., Ku, A.T., Dekker, J.D., Tucker, H., and Nguyen, H. (2013). Foxp1 maintains hair follicle stem cell quiescence through regulation of Fgf18. Development *140*, 3809–3818.

Lopez, R.G., Garcia-Silva, S., Moore, S.J., Bereshchenko, O., Martinez-Cruz, A.B., Ermakova, O., Kurz, E., Paramio, J.M., and Nerlov, C. (2009). C/EBPalpha and beta couple interfollicular keratinocyte proliferation arrest to commitment and terminal differentiation. Nat. Cell Biol. *11*, 1181–1190.

McDade, S.S., Henry, A.E., Pivato, G.P., Kozarewa, I., Mitsopoulos, C., Fenwick, K., Assiotis, I., Hakas, J., Zvelebil, M., Orr, N., et al. (2012). Genome-wide analysis of p63 binding sites identifies AP-2 factors as co-regulators of epidermal differentiation. Nucleic Acids Res. *40*, 7190–7206.

Pellegrini, G., Golisano, O., Paterna, P., Lambiase, A., Bonini, S., Rama, P., and De Luca, M. (1999). Location and clonal analysis of stem cells and their differentiated progeny in the human ocular surface. J. Cell Biol. *145*, 769–782.

Pellegrini, G., Dellambra, E., Golisano, O., Martinelli, E., Fantozzi, I., Bondanza, S., Ponzin, D., McKeon, F., and De Luca, M. (2001). p63 identifies keratinocyte stem cells. Proc. Natl. Acad. Sci. USA *98*, 3156–3161.

Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S.A., Flynn, R.A., and Wysocka, J. (2011). A unique chromatin signature uncovers early developmental enhancers in humans. Nature *470*, 279–283.

Rochat, A., Kobayashi, K., and Barrandon, Y. (1994). Location of stem cells of human hair follicles by clonal analysis. Cell *76*, 1063–1073.

Scheitz, C.J., and Tumbar, T. (2013). New insights into the role of Runx1 in epithelial stem cell biology and pathology. J. Cell Biochem. *114*, 985–993.

Schug, J., Schuller, W.P., Kappen, C., Salbaum, J.M., Bucan, M., and Stoeckert, C.J., Jr. (2005). Promoter features related to tissue specificity as measured by Shannon entropy. Genome Biol. *6*, R33.

Sharma, A., Larue, R.C., Plumb, M.R., Malani, N., Male, F., Slaughter, A., Kessl, J.J., Shkriabai, N., Coward, E., Aiyer, S.S., et al. (2013). BET proteins promote efficient murine leukemia virus integration at transcription start sites. Proc. Natl. Acad. Sci. USA *110*, 12036–12041.

Shen, J., van den Bogaard, E.H., Kouwenhoven, E.N., Bykov, V.J., Rinne, T., Zhang, Q., Tjabringa, G.S., Gilissen, C., van Heeringen, S.J., Schalkwijk, J., et al. (2013). APR-246/PRIMA-1(MET) rescues epidermal differentiation in skin keratinocytes derived from EEC syndrome patients with p63 mutations. Proc. Natl. Acad. Sci. USA *110*, 2157–2162.

Tan, D.W., Jensen, K.B., Trotter, M.W., Connelly, J.T., Broad, S., and Watt, F.M. (2013). Single-cell gene expression profiling reveals functional heterogeneity of undifferentiated human epidermal cells. Development *140*, 1433–1444.

Whyte, W.A., Orlando, D.A., Hnisz, D., Abraham, B.J., Lin, C.Y., Kagey, M.H., Rahl, P.B., Lee, T.I., and Young, R.A. (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell *153*, 307–319.

Yang, A., Schweitzer, R., Sun, D., Kaghad, M., Walker, N., Bronson, R.T., Tabin, C., Sharpe, A., Caput, D., Crum, C., et al. (1999). p63 is essential for regenerative proliferation in limb, craniofacial and epithelial development. Nature *398*, 714–718.

Yi, R., Poy, M.N., Stoffel, M., and Fuchs, E. (2008). A skin microRNA promotes differentiation by repressing 'stemness'. Nature *452*, 225–229.

Zang, C., Schones, D.E., Zeng, C., Cui, K., Zhao, K., and Peng, W. (2009). A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. Bioinformatics *25*, 1952–1958.

Stem Cell Reports, Volume *6*

Supplemental Information

Dynamic Transcriptional and Epigenetic Regulation of Human Epidermal Keratinocyte Differentiation

Alessia Cavazza, Annarita Miccio, Oriana Romano, Luca Petiti, Guidantonio Malagoli Tagliazucchi, Clelia Peano, Marco Severgnini, Ermanno Rizzi, Gianluca De Bellis, Silvio Bicciato, and Fulvio Mavilio

SUPPLEMENTAL DATA ITEMS

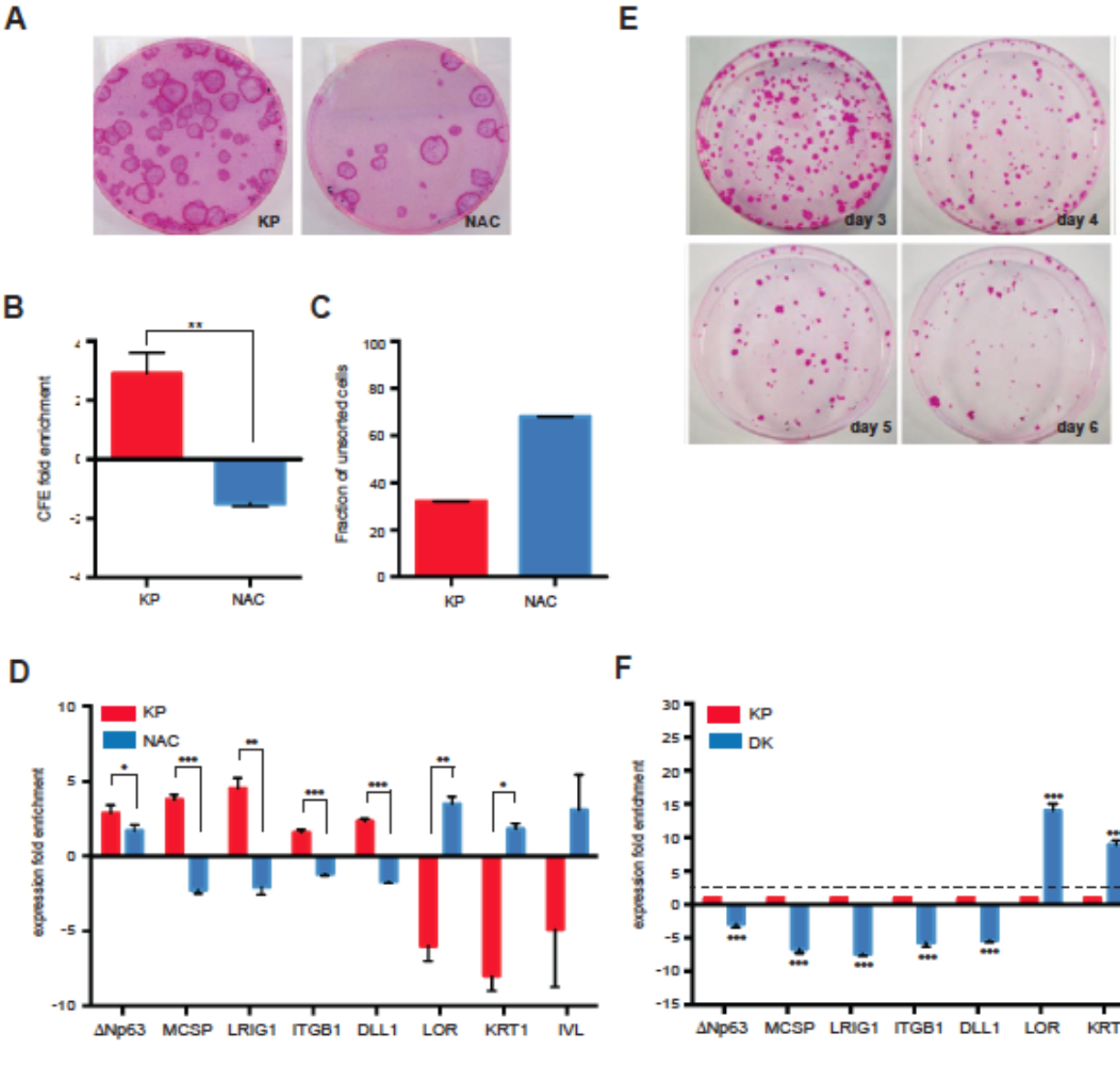
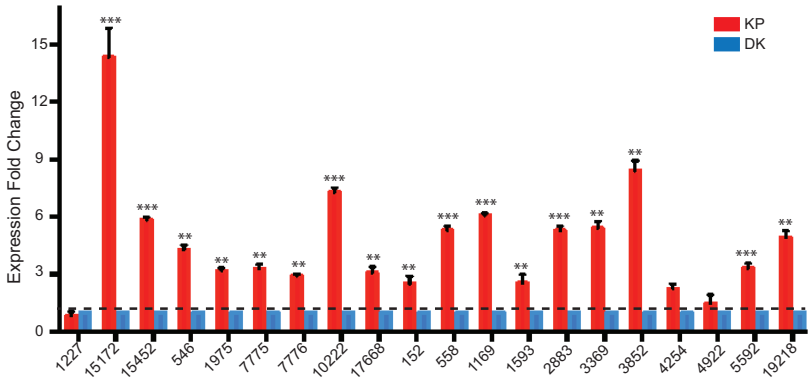
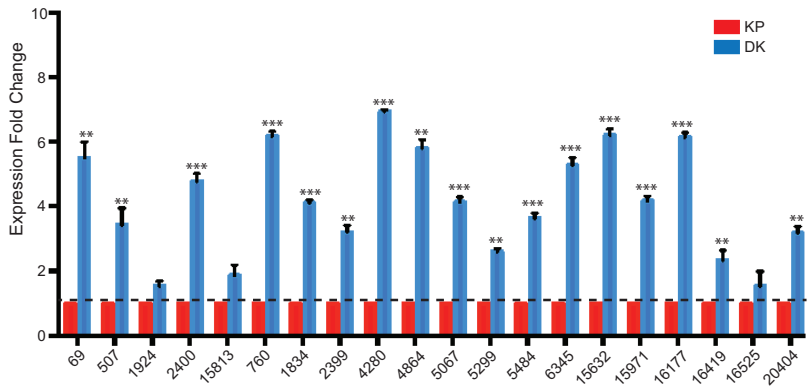


Figure S1. **Related to Figure 1. Prospective isolation of human keratinocyte progenitors and differentiated keratinocytes.** (A and B) Colony Forming Efficiency assay on Collagen IV rapidly-adhering cells (keratinocyte progenitors, KPs) and non-adhering cells (NACs) showed an enrichment in clonogenic cells in KPs over NACs. KPs were enriched on average 2.9-fold in colony-forming cells compared to the original culture, while NACs were 1.5-fold depleted in clonal progenitors. Data are reported as average ± SD, with n=8 (** $p<0.01$, t-test). (C) KPs represented about one third of the unsorted cell culture, in line with the distribution of β1 integrin expression in human epidermis *in vitro* and *in vivo* (Jones and Watt, 1993). Data are reported as average ± SD, with n=8. (D) Confirmation of key KP and differentiating keratinocyte signature genes by RT-PCR. Expression fold changes were normalized to the expression level of unsorted cells. Data are reported as average ± SD, with n=3 (* $p<0.05$; ** $p<0.01$; *** $p<0.001$, t-test). (E) Colony Forming Efficiency assay on differentiated keratinocytes (DKs), showing a decrease in the colony forming ability of the culture from the beginning to the end (day 6) of the differentiation protocol. (F) Confirmation of key KP and DK signature genes by RT-PCR. Fold changes of expression in DKs were normalized to the expression level detected in KPs. Data are reported as average ± SD, with n=3 (* $p<0.05$; ** $p<0.01$; *** $p<0.001$, t-test).

A

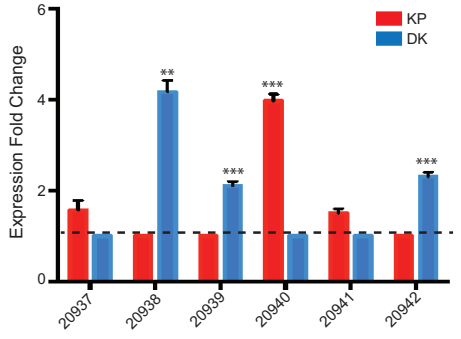


CAGE ID	Transcript	Name	Fold Change
1227	ncRNA	LRIG2 AS	2.1
15172	EST	EST	8.2
15452	ncRNA	ncRNA	4.3
546	ncRNA	ncRNA	2.9
1975	EST	EST	5.8
7775	coding RNA	KRT13	2.1
7776	coding RNA	KRT13	5.2
10222	coding RNA	THBD	6.4
17668	ncRNA	ncRNA	2.2
152	coding RNA	TCF7L2	2.7
558	ncRNA	ncRNA	4.4
1169	coding RNA	BARX2	5
1593	ncRNA	ncRNA	3.2
2883	ncRNA	ncRNA	4.6
3369	coding RNA	NEK2	3.7
3852	coding RNA	ID3	4
4254	ncRNA	ncRNA	4.4
4922	coding RNA	ELF1	3
5592	ncRNA	ncRNA	3.7
19218	ncRNA	ncRNA	4.3

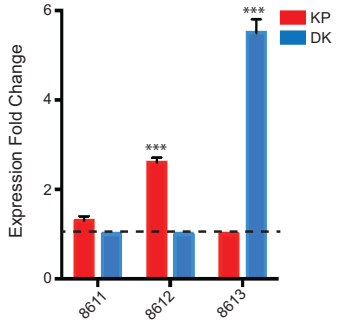


CAGE ID	Transcript	Name	Fold Change
69	coding RNA	NOLC1	-3.6
507	coding RNA	CALML5	-5.1
1924	coding RNA	S100A16	-2.9
2400	ncRNA	ncRNA	-5.5
15813	ncRNA	ncRNA	-4.3
760	ncRNA	ncRNA	-5.2
1834	coding RNA	S100A10	-4.0
2399	ncRNA	ncRNA	-3.3
4280	ncRNA	ncRNA	-5.2
4864	coding RNA	FRY	-5.2
5067	ncRNA	ncRNA	-4.2
5299	coding RNA	CTPS	-2.8
5484	ncRNA	ncRNA	-3.5
6345	ncRNA	ncRNA	-3.7
15632	ncRNA	ncRNA	-4.3
15971	coding RNA	RASSF1	-2.8
16177	ncRNA	ncRNA	-5.2
16419	coding RNA	PLRG1	-2.6
16525	coding RNA	DCTD	-2.2
20404	coding RNA	POR	-3.2

B



CAGE ID	TSS #	Position	Fold Change
20937	8	chr8:145013629	0.2
20938	7	chr8:145016700	-3.3
20939	4	chr8:145025016	-2.2
20940	3	chr8:145027943	3.8
20941	2	chr8:145047678	0.5
20942	1	chr8:145050880	-2.1



CAGE ID	TSS #	Position	Fold Change
8611	1	chr18:21269378	0.3
8612	2	chr18:21269531	2.1
8613	3	chr18:21452769	-4.5

Figure S2. **Related to Figure 1. Gene expression profiling of human keratinocyte progenitors and differentiated keratinocytes with deepCAGE.** (A) Confirmation of the differential expression of randomly chosen CAGE promoters by RT-PCR. Fold changes of expression of KP-high and CAGE promoters were calculated over the expression level detected in DKs; expression fold changes of DK-high CAGE promoters were calculated over the expression level detected in KPs. Data are reported as average ± SD, with n=3 (* $p<0.05$; ** $p<0.01$; *** $p<0.001$, t-test). For each CAGE promoter analyzed, we reported information concerning the transcript type and name, and the fold change of expression detected by deepCAGE. (B) Confirmation of the differential expression of alternatively used TSSs for the PLEC1 and LAMA3 genes by RT-PCR. Expression fold changes were normalized to the expression level detected in the unsorted keratinocyte culture. Data are reported as average ± SD, with n=3 (** $p<0.01$; *** $p<0.001$, t-test). For each CAGE promoter analyzed, we reported information concerning the genomic position of the TSS, and the expression fold change detected by DeepCAGE.

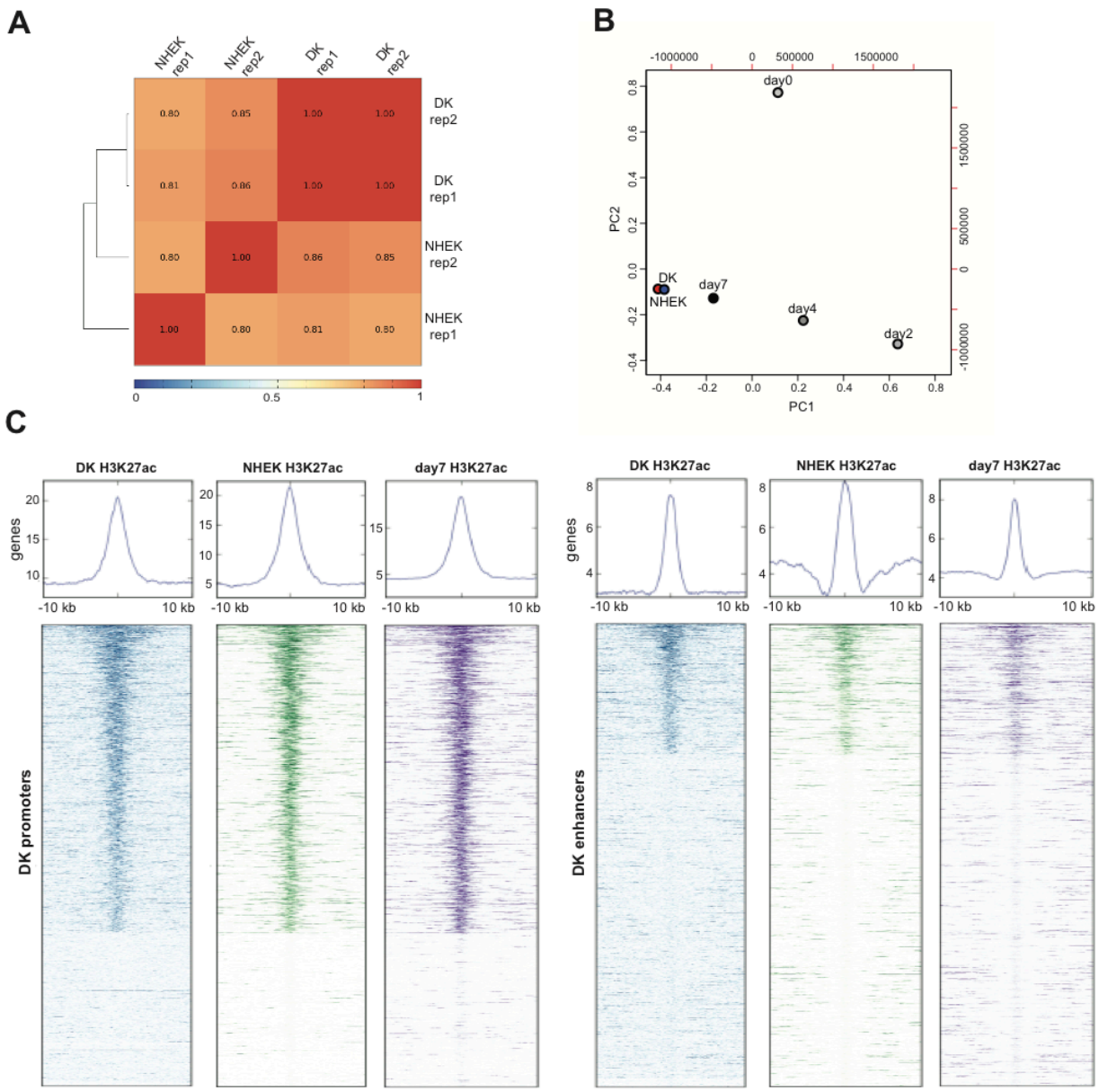
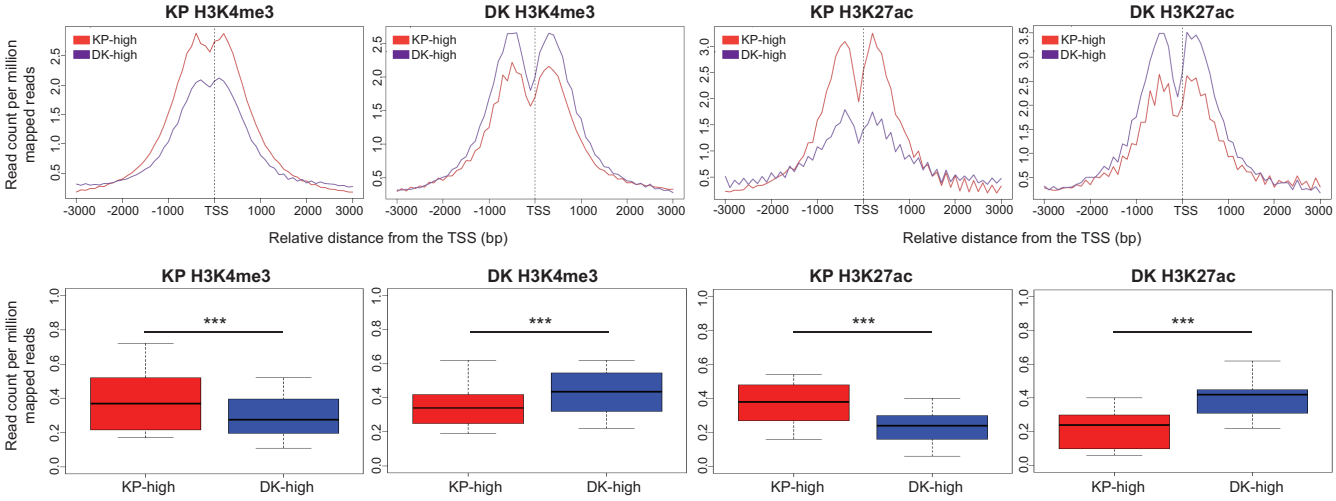


Figure S3. Related to Figure 2. Comparison of genome-wide data of RNA-seq and H3K27ac ChIP-seq. (A) Comparison of RNA-seq gene expression patterns between differentiated keratinocyte (DK) samples from this study and those from the ENCODE project (NHEK). The heatmap shows the Spearman rank correlation coefficients between each pair of samples, indicated by both color and number. (B) Principal component analysis of RNA-seq gene expression data from DK samples from this study, the ENCODE project (NHEK) and from Kouwenhoven et al., 2015 (day0, day2, day4 and day7) (Kouwenhoven et al., 2015). The analysis shows a significant similarity between DK and NHEK transcriptomes, as well as with the RNA-seq data obtained from the latest stage of keratinocyte differentiation (day7) in Kouwenhoven et al., 2015. (C) ChIP-seq density profiles and heatmaps of H3K27ac signals from DKs, NHEK and differentiated keratinocytes at day7 (Kouwenhoven et al., 2015) in a 20-kb window around the center of promoters and enhancers defined in DKs. Promoters and enhancers are ranked by increasing H3K27ac ChIP-seq signal in DKs. As shown by the heatmaps, H3K27ac profiles among the three datasets are mostly overlapping, with strongly marked promoters and enhancers in DKs being marked also in NHEK and differentiated keratinocytes at day7.

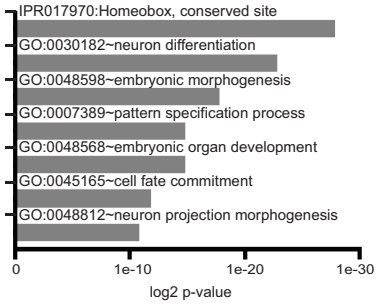
A



B

	Category	Gene Symbol
Increased H3K4me3 in KPs	Transcription and chromatin	Tcf3, Tcf4, Tcf7, Ascl2, Nfatc1, Klf5, Tbx1, Grhl1, Atf3, Bcl11b, Pitx2, Prdm2, Ovol1, Cebpb, Ezh2, Suz12, Bmi1, Dnmt3a, Jmjd8, Cbx2
	Signaling	Fgfr2, Bmp6, Fst, Rac1, Fzd6, Wnt10b, Smad3, Lrig1
	Cell cycle	Bmi1, Cdkn1b
	Cell adhesion and cytoskeleton	Cd34, Prom1, Col2a1, Itgb7, Itgb1, Itga6, Lamc3
Increased H3K4me3 in DKs	Transcription and chromatin	Dlx3, Msx2, Grhl3, Hoxc10, Hoxc13, Runx1, Hes2, Prdm1
	Signaling	Dkk1, Wnt5a, Yap1, Tead2, Lrig2
	Epidermal differentiation complex	Slc7a8, Slc44a2, Slc39a10, Slc2a9, Slc6a9, Slc6a15 S100a2, S100a16, S100a13
	Cell adhesion and cytoskeleton	Krt6a, Col4a2, Col7a1, Col5a2, Lamb3, Klk5, Klk8, Plek2

C



D

	Category	Gene Symbol
H3K27me3 marked in DKs	Transcription and chromatin	Nfatc1, Tbx1, Barx2, Bcl11b, Foxp1, Klf4, Sox7, Sox9, Ascl2, Hopx, Id4, Irx1,Cebpa, Pitx2, FoxA1, Cbx4, Cbx2, Ezh1, Ring1b, Bmi1
	Signaling	Dll1, Fgfr3, Bmp7, Wnt3, Wnt4, Wnt5a, Wnt10a, Wnt6, Wnt9a
	Cell adhesion and cytoskeleton	Col4a6, Col1a1, Col12a1, Itga6, Lamb1, Krt4, Gjb2

E

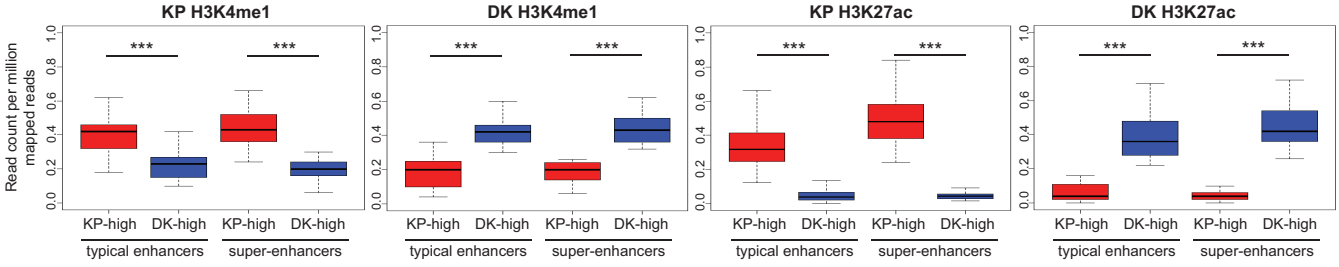


Figure S4. Related to Figure 3. Regulation of promoter and enhancer regions in progenitors and differentiated keratinocytes. (A) Correlation between histone modification intensities and CAGE promoter expression levels. Distribution of H3K4me3 and H3K27ac peaks around CAGE TSSs (top panels) and the corresponding box-whisker plots (bottom panels). A significant correlation between H3K4me3 intensity and CAGE promoter expression levels is observed. KP-high CAGE promoters are highly enriched in H3K4me3 in KPs, compared to DK-high promoters. Similarly, DK-high CAGE promoters show significantly higher levels of H3K4me3 in DKs. An even stronger correlation is seen when considering H3K27ac mark around both KP- and DK-high CAGE promoters. Statistical significance was determined by Wilcoxon test with Bonferroni correction of p-value (***p<0,001). (B) Selected list of genes with a reported functional role in skin that show an increased level of H3K4me3 at their promoters, in either KPs or DKs. (C) Gene Ontology analysis of promoters marked by H3K27me3 in DKs and transcriptionally inactive in both KPs and DKs. (D) Selected list of genes with a reported functional role in skin marked by H3K27me3 at the promoters and transcriptionally silent in DKs. Underlined genes are known markers of human or murine hair follicle and interfollicular stem cells. Genes in red are also flanked by a H3K27me3-marked enhancer. (E) H3K4me1 and H3K27ac intensity of typical and super-enhancers close to CAGE promoters (±100 kb). In KPs H3K4me1 and H3K27ac signals of typical and super-enhancers is higher around CAGE promoters highly active in KPs (KP-high) compared to the H3K4me1 and H3K27ac intensities around DK-high CAGE promoters. Similar results are obtained in DKs. Statistical significance was determined by Wilcoxon test with Bonferroni correction of p-value (***p<0,001).

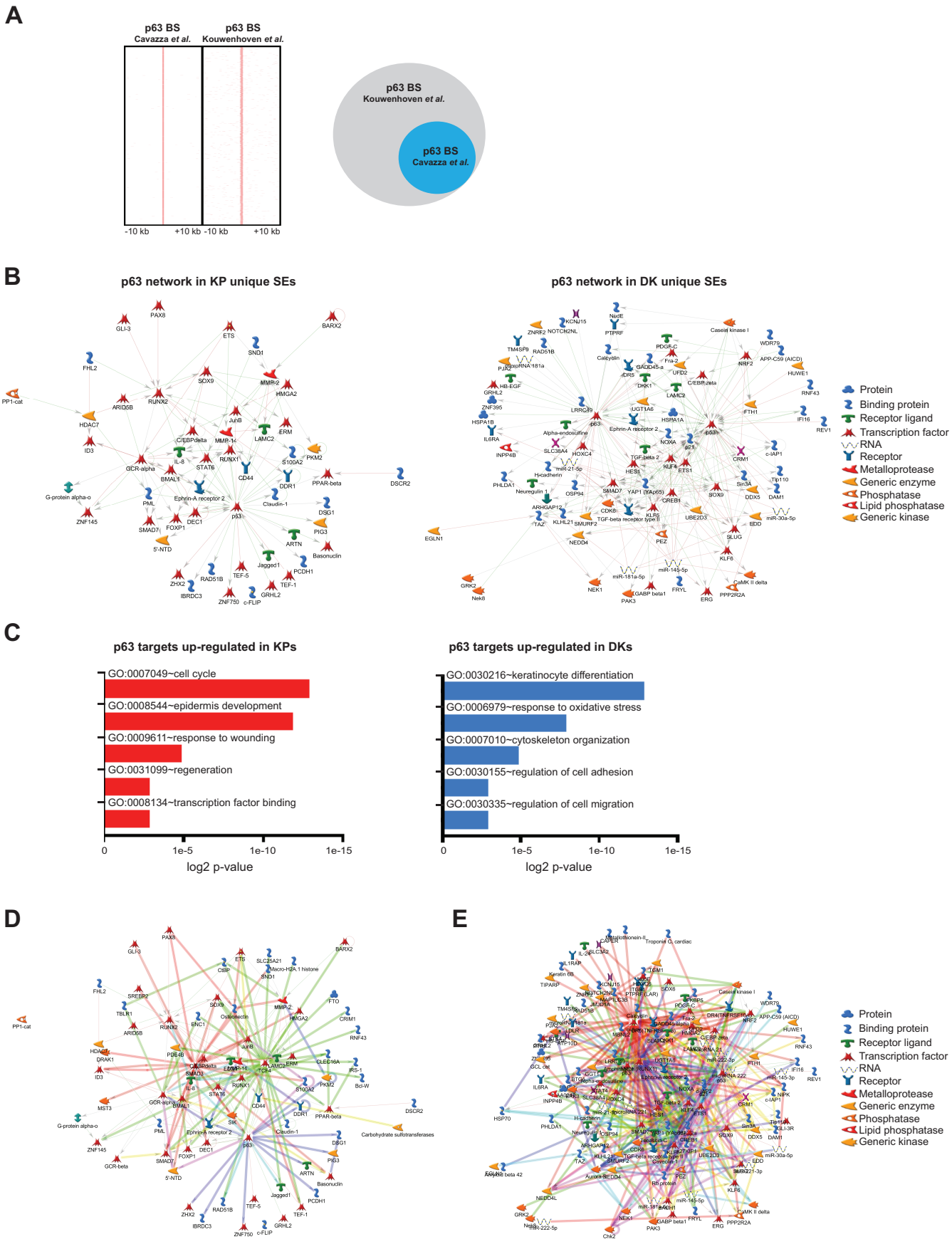
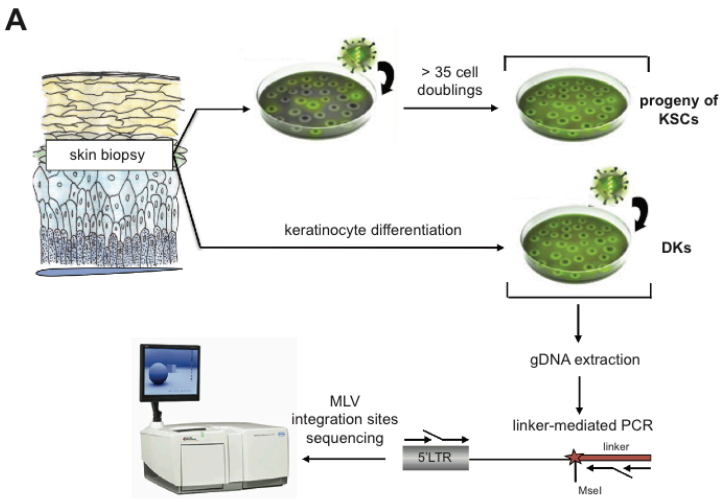
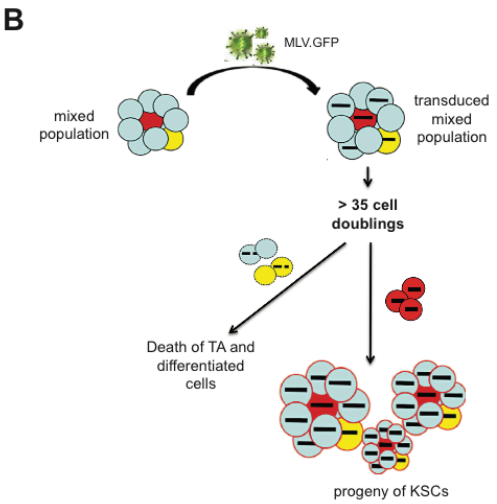
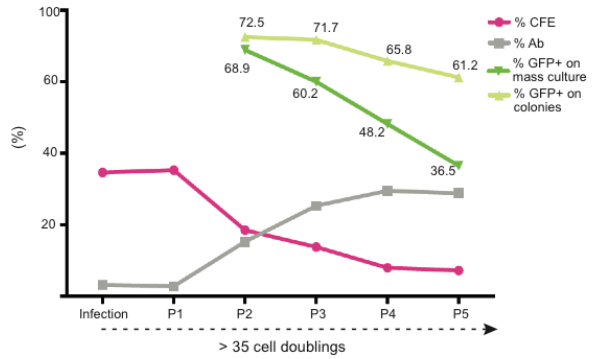
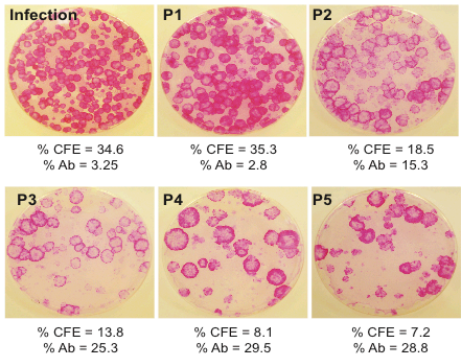

A
p63 BS
Cavazza et al.
p63 BS
Kouwenhoven et al.
-10 kb
+10 kb
-10 kb
+10 kb
p63 BS
Kouwenhoven et al.
p63 BS
Cavazza et al.
B
p63 network in KP unique SEs
p63 network in DK unique SEs
Protein
Binding protein
Receptor ligand
Transcription factor
RNA
Receptor
Metalloprotease
Generic enzyme
Phosphatase
Lipid phosphatase
Generic kinase
C
p63 targets up-regulated in KPs
GO:0007049~cell cycle
GO:0008544~epidermis development
GO:0009611~response to wounding
GO:0031099~regeneration
GO:0008134~transcription factor binding
0
1e-5
1e-10
1e-15
log2 p-value
p63 targets up-regulated in DKs
GO:0030216~keratinocyte differentiation
GO:0006979~response to oxidative stress
GO:0007010~cytoskeleton organization
GO:0030155~regulation of cell adhesion
GO:0030335~regulation of cell migration
0
1e-5
1e-10
1e-15
log2 p-value
D
E
Protein
Binding protein
Receptor ligand
Transcription factor
RNA
Receptor
Metalloprotease
Generic enzyme
Phosphatase
Lipid phosphatase
Generic kinase

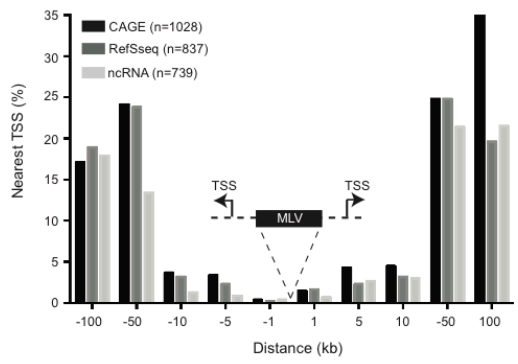
Figure S5. Related to Figure 4. P63-centered transcriptional regulatory circuitries during keratinocyte differentiation. (A) Heatmap of p63 occupancy defined by ChIP-seq datasets obtained either in this study or in Kouwenhoven et al., 2015 at p63 binding sites (genomic regions of a 20-kb window with summits of p63 binding sites retrieved in this study in the middle of each panel). The heatmap and the Venn diagram on the right panel show a complete overlap between p63 binding sites retrieved in this study and in Kouwenhoven et al, 2015. (B) Molecular and transcriptional interaction networks of genes associated to KP- or DK-unique super-enhancers (SEs) and bound by p63. (C) Gene Ontology analysis of p63 target genes in KPs (red bars) and DKs (blue bars). (D) Molecular and transcriptional interaction network of genes associated to KP-unique SEs bound by p63, TCF4 and SMAD3. (E) Molecular and transcriptional interaction network of genes associated to DK-unique SEs bound by p63 and AP1.

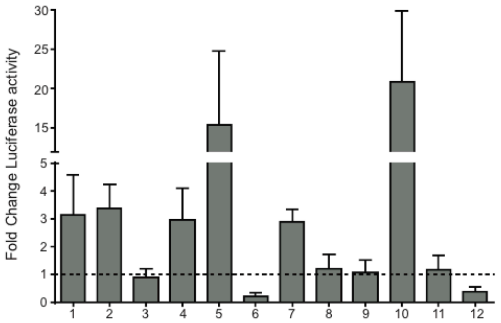




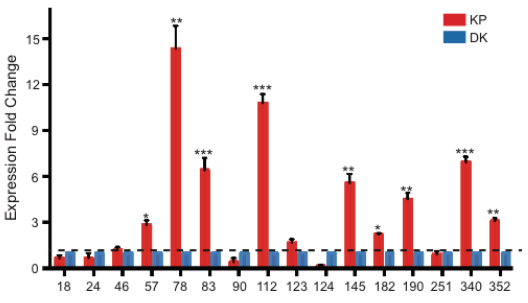








Enhancer #	Cluster Name	Integrations (n)	Size (bp)	Gene	Location
1	DK 1088	6	1372	RNF43	intron 1
2	DK 1252	5	1332	PTPN1	upstream
3	DK 18	4	1163	GRHL3	upstream
4	DK 67	4	1158	NGF	intron 1
5	DK 552	4	595	SUN1	intron 1
6	DK 671	4	887	ncRNA	downstream
7	DK 736	4	461	ST5	intron 3
8	DK 769	4	751	PC	intron 1
9	DK 949	4	641	ITPK1	intron 2
10	DK 1043	4	885	RTN4RL1	intron 1
11	DK 5	3	309	ncRNA	intron 1
12	DK 339	3	493	ncRNA	dowstream



Cluster #	Transcript	Overlapping CAGE
18	ncRNA	no
24	LRIG2	no
46	ncRNA	no
57	ncRNA	yes
78	EST	yes
83	LEKR1	no
90	ncRNA	no
112	NIR_00066175	no
123	ncRNA	no
124	ncRNA	no
145	ncRNA	yes
182	ncRNA	yes
190	SH3PXD2A	no
251	WDFY2	yes
340	CDC25B	no
352	ncRNA	no

Figure S6. Related to Figure 5. Discovering transcriptionally active regulatory elements by retroviral integration site analysis. Overview of the experimental procedure used to identify regulatory elements active in keratinocyte stem cells (KSCs) or differentiated keratinocytes (DKs) by retroviral integration site analysis. To retrospectively map the regulatory regions active in KSCs (panel A and B), early passage foreskin-derived primary keratinocytes were co-cultured on a feeder layer and transduced with a GFP-expressing MLV vector. Cells were sub-cultured for >35 cell doublings (>6 passages) to exhaust the population of transduced transit-amplifying cells (TA, yellow circles) and DKs (blue circles) and enrich for the progeny of culture-maintaining KSCs (red circles). After >6 passages the culture is mainly composed of TA cells and DKs derived by the originally transduced KSCs, stably harboring multiple MLV integration sites at regulatory regions used by KSCs at the time of infection. Cells are collected, sorted for GFP expression. To prospectively define the regulatory regions active in DKs (panel A), early passage foreskin-derived primary keratinocytes were differentiated by contact inhibition and by culturing them in a serum-free medium depleted of several growth factors for 6 days (see Supplemental Experimental Procedures). At day 7, cells were transduced with a GFP-expressing MLV vector and then collected and sorted for GFP expression 72 hours after infection. Genomic DNA is extracted from the two transduced populations and MLV integration sites retrieved by linker-mediated PCR and pyrosequencing. (C and D) Clonal conversion of MLV-transduced KSCs during serial passages as defined by the Colony Forming Unit assay. Early-passage and highly clonogenic keratinocytes were infected (72.5% of GFP-positive cells) and then cultured for >35 cell doublings to eliminate transduced transit amplifying cells and differentiated keratinocytes. The percentage of clonogenic cells (%CFE, in (A) and in (B)) decreased with serial passages and corresponding cell doublings along with an increased fraction of abortive cells ((%Ab, in (A) and in (B)), indicating that the originally transduced KSCs were converting in their differentiated progeny. The percentage of GFP-positive cells (%GFP^+ in (A)) diminished with time, as a result of the decreased number of culture-mantaining cells (KSCs) due to clonal conversion. In line with this evidence, the number of GFP-positive colonies defined by the Colony Forming Unit Assay (%GFP^+colonies in (A)) decreased with serial passages. (E) Distribution of the distance of MLV integration sites from the TSSs of the two closest CAGE promoters (black), RefSeq genes (dark grey) and of annotated noncoding RNAs (light grey) in a 200 kb window. (F) Six out of the twelve tested MLV cluster regions showed an enhancer activity when tested in a Luciferase reporter assay in DKs, while two regions showed a repressor activity. The graph shows relative luciferase reporter activity normalized to reporter construct alone. Information concerning the number of integrations contained in each cluster, the size, the closest or overlapping genes and the location of the cluster with respect to them are reported in the table. Data are reported as average ± SD, with n=3. (G) Confirmation of randomly chosen KSC-specific MLV clusters by assessing the expression of the closest or overlapping gene by RT-PCR in KPs. Fold changes of enrichment of gene expression in KPs were calculated over the expression level detected in DKs. Data are reported as average ± SD, with n=3 (* $p<0.05$; ** $p<0.01$; *** $p<0.001$, t-test). For each MLV-cluster analyzed, we reported information concerning the transcript type or name, and the presence of transcription as detected by CAGE-seq in KP.

TABLE S1. Related to Figure 6. Regulatory regions identified in retrospectively defined KSCs by MLV scanning, (see accompanying Excel file)

SUPPLEMENTAL EXPERIMENTAL PROCEDURES

Cell Culture

Human primary keratinocytes were obtained from three foreskin biopsies of healthy neonatal donors and expanded by cultivation onto lethally irradiated 3T3-J2 cells (a gentle gift from Y. Barrandon's lab) in growth FAD medium, a DMEM and Ham's F12 media mixture (2:1) containing FCS (10%), penicillin-streptomycin (1%), glutamine (2%), insulin (5ug/ml), adenine (24.3ug/ml), hydrocortisone (0.4ug/ml), cholera toxin (50ug/ml), triiodotyronine (2nM). After 3 days FAD medium was replaced and cFAD medium (FAD medium containing 10ng/ml EGF) was added to the culture. Keratinocytes were trypsinized at subconfluence and replated onto a new feeder-layer. Mouse NIH3T3 fibroblast cell line was maintained in Dulbecco's Modified Eagle's medium (Euroclone), supplemented with 10% fetal bovine serum. Differentiation of keratinocytes was induced by cell contact inhibition and by depleting growth factors from standard medium (KGM, Lonza). The differentiation medium consists of KGM, 0.15 mM Ca2+ (Lonza), supplemented with 0.1 mM ethanolamine (Sigma), 0.1 mM phosphoethanolamine (Sigma), 100 U/mL penicillin (Gibco), and 100 µg/mL streptomycin (Gibco). Cells were cultivated at confluence for 6 days. Differentiation medium was changed every second day, and before harvesting of the RNA and chromatin.

Isolation of keratinocyte progenitors from primary keratinocyte cultures

An epidermal progenitor-enriched population from a total culture of keratinocytes was selected by collagen IV adherence assays, as described previously (Jones and Watt, 1993). Briefly, 8 x 10^6 cells were seeded in 5 ml on human placental Collagen IV-coated dishes (Sigma). Rapidly adhering cells, containing mostly progenitor cells (KPs), were isolated by incubating cells for 20 minutes on collage-coated plates. After removal of the supernatant, adherent KPs were washed three times with PBS and detached with trypsin. The supernatants, containing non adhering cells, mostly transient-amplifying cells and terminally differentiated keratinocytes, and isolated KPs were seeded for CFE analysis and collected for RNA extraction.

Real-Time quantitative PCR (RT-qPCR)

Total cellular RNA was extracted from 2-5 x 10^6 KPs, DKs and unsorted keratinocytes cells, using the Rnaeasy Mini Kit (QIAGEN). 500 ng of extracted RNA were loaded on a denaturing 1% agarose gel to check for RNA integrity. The RNA samples were used to set up the retrotranscription reaction (SuperScript III kit, Invitrogen) using Oligo dTs or random examers as primers. Samples with no RT enzyme were processed in parallel with real samples to control for residual DNA contamination. One tenth of the RT reactions were then subjected to PCR, either semi-quantitative or quantitative Real-Time PCR (qRT-PCR). QRT-PCR was performed on an ABI 7900 machine using SYBR green detection chemistry (Applied Biosystem). GAPDH primers were used as an internal control. RT-qPCR data were analysed using the $2^{-\Delta\Delta CT}$ method.

DeepCAGE

DeepCAGE library preparation, sequencing and mapping

RNA from three different KP selection experiments were isolated using RNeasy Plus Mini kit (QIAGEN) and pooled together. DNAFORM Inc. at RIKEN Omics Science Center performed DeepCAGE library preparation, as described previously (Carninci et al., 2006). Briefly, total RNA was concentrated in the presence of trehalose, sorbitol, the template switching (TS) oligonucleotides, and the random (N15) reverse-transcription primers and cDNA synthesis was performed with 5 µg of total RNA and PrimeScript Reverse Transcriptase (TAKARA). RNA/cDNA hybrid was purified with Agencourt RNAClean XP (BECKMAN) and eluted after 10 minute staying at 37°C after the beads were resuspended with 37°C pre-heated H2O. Capped RNA was biotinylated using 15mM Biotin dissolved in H2O. After treated by RNaseOne, cDNA which is hybrid with biotinylated Capped RNA was selected with cap-trapper method. 100 µl of MPG Streptavidin beads (TAKARA) were used after coating by 1.5 µl of tRNA for 30 min. cDNA was released from beads by 60 µl of 50mM NaOH at room temperature for 10 min. On the other hand, quality of cDNA was checked using ATCB as the house_keeping_gene with qRT-PCR and quantification of cDNA was measured. A sample-specific linker, containing a recognition site for the barcode sequence (3 bp) and the type III restriction-modification enzyme EcoP15I was ligated to the single-strand cDNA which was concentrated with speed Vac with T4 DNA ligase (NEB). After ligation, the cDNA was purified with Agencourt AMPure XP (BECKMAN) twice to eliminate the linker dimmers. The second strand synthesis was performed by adding 200 ng of primer, 5'-Bio- CCACCGACAGGTTCAGAGTTC-3', by Hot start (94°C 3min), and the resulting double-stranded cDNA is purified with the Agencourt AMPure XP again to eliminate extra primers. The double-stranded cDNA was cleaved with 1U of EcoP15I (NEB). After heat inactivation, the second linker was ligated to the CAGE tag with T4 DNA ligase (NEB). The CAGE tags were separated from unmodified DNA with MPG Streptavidin beads. The DNA fragments were amplified in a PCR step by using linker-specific primers (1.0 µM final each), with Phusion High-

Fidelity DNA Polymerase (FINNZYMES). After incubation at 98°C for 30 sec, ten cycles of PCR are performed for 10 sec at 98°C, 10 sec at 60°C. The resulting PCR products were pooled and treated with ExonucleaseI (NEB), for 30 min at 37°C. Then PCR solution was purified with MinElute column (QIAGEN). According to the concentration retrieved, the sample was adjusted to 10nM for SOLEXA sequencer.

Samples were sequenced using the Illumina GA II sequencer, with an average read length of 36 bases. Tags were extracted and mapped to human genome version hg19 (NCBI build 37), with a minimum match length of 21 bases and a maximum of one error; tags mapping the human ribosomal DNA sequence were eliminated. The best match for each tag was then calculated as the alignment with the highest score obtained with the alignment algorithm used. For CAGE tags mapping to multiple genome locations, we apply a weighting strategy based on the number of CAGE tags within a 200 bp neighborhood around each candidate mapping location. Equal weights were used if no unique tags are found within the 200 bp region for all candidate mapping locations.

Promoter construction

Level-1 promoters ("transcription start sites") were created by summing the weighted number of CAGE tags at each genome position. We required at least one CAGE tag in at least one experimental condition; other mapping positions were discarded. The level-1 promoters were clustered into Level-2 promoters ("promoters") if they were within 20 bp of each other on the same chromosomal strand. We require that the expression of each level-2 promoter is at least 10 tags per million (tpm) in at least one experimental condition; other promoters are dropped. Level-3 promoters ("promoter regions") were created by joining level-2 promoters if they were within 300 bp of each other on the same chromosomal strand. We calculated the tpm value for each level-1, level-2, and level-3 promoter by dividing the number of CAGE tags of each promoter in each experimental condition by the total number of mapped CAGE tags in that condition, and multiplying by 1,000,000. The names of level-2 and level-3 promoters are based on their most highly expressed genomic position.

Promoter genomic annotation

We annotated level-2 promoters on the base of their vicinity to RefSeq genes, ENSEMBL ncRNA, ncRNA included in publicly available data sets (Cabili et al., 2011; Khalil et al., 2009; Kretz et al.), Vertebrate Genome Annotation (Vega) pseudogenes (Wilming et al., 2008) and Yale Gerstein Group pseudogenes (Zhang et al., 2006). We assigned promoters to repetitive elements defined by RepeatMasker (http://www.repeatmasker.org/). For each dataset, we found the annotation with the smallest distance to the CAGE-defined promoter on the same chromosome strand. We defined the distance as follows: 1. If the 3' end of the promoter is upstream of the 5' end of the annotation, then we considered the distance between the 3' end of the promoter and the 5' end of the annotation. 2. If the 5' end of the promoter is downstream of the 5' end of the annotation, then we considered the distance between the 5' of the annotation and the 5' end of the promoter. 3. Otherwise, the promoter overlaps the 5'end of the annotation and in this case, we defined the distance to be zero. If this distance is less than 400 bp we associated this promoter with the gene or transcript. A total of 16,491 level-2 promoters were retrieved and classified on the basis of their expression level.

CpG island and TATA box prediction

We calculated the normalized CpG content of all CAGE-defined promoters in a region of -1kb/+0.2kb centred around the most expressed tag (TSS) of each level-2 promoter, as described in (Saxonov et al., 2006). Briefly, CpG content for each level-2 promoter was calculated using the program GEECEE in the EMBOSS (Rice et al., 2000) package. The normalized CpG content was calculated by dividing the observed number of CpG dinucleotides by the expected number in a promoter. Normalized CpG contents for promoters followed a bimodal distribution (Figure S3A). We set the cutoff value between high and low CpG to 0.46, where the two peaks in the Gaussian distribution were best separated. Promoters with a normalized CpG content >0.46 were defined as high CpG promoters (HCP), while those with a CpG content <0.46 were classified as low CpG promoters (LCP). Promoters with an observed CpG content equal to the content expected in a promoter(O/E=0) were defined as non-CpG promoters (NCP).

To define TATA box containing promoter, we scanned for the presence of the TBP motif (V$TATA_01) and the TFIID motif (V$TATA_C) in a region of -100/+50bp centred around the most expressed tag (maxTSS) of each level-2 promoter, using the MATRIX-SCAN tool in RSAT (http://rsat.ulb.ac.be/).

Tissue-specificity prediction

Tissue specificity was assessed by calculating a Shannon Entropy score, as described in (Schug et al., 2005). Briefly, given the expression level of a CAGE promoter in *N* tissues, we defined the relative expression of a CAGE promoter *c* in a given cell population *p* as $E_{p/c}=w_{c/p}/\sum_{1\leq p\leq N} w_{c/p}$ where $w_{c/p}$ is the expression level of the promoter in that population. The entropy of a promoter's gene expression distribution is $H_c=\sum_{1\leq p\leq N}-E_{p/c}\log_2(E_{p/c})$, where the values of H_c ranges from 0 to $\log_2(N)$. An entropy score close to zero indicates that the promoter is highly tissue-specific, while an entropy score close to $\log_2(N)$

means that the promoter is expressed more ubiquitously. The catalogue of cell populations used for this analysis included human ES cells, neural stem cells, and $CD34^+$ hematopoietic stem/progenitor cells, for which deepCAGE data were produced by us (results not shown).

Promoter expression analysis

To assess the differential expression of each promoter in KP *versus* DK a pair-wise χ^2 test statistics was applied (Romualdi et al., 2001) and only those promoters significantly different were considered ($p < 0{,}001$). To compare deepCAGE and microarray expression measurement we associated each annotated CAGE promoter with the corresponding microarray probe present for the same gene. We selected all CAGE promoter/Affymetrix probe pairs that were one-to-one associated with each other and calculated the Pearson correlation coefficients of their expression profiles. A background measurement was obtained by calculating the Pearson correlation coefficient of CAGE promoters associated with array probe's expression values of an unrelated cell type.

Gene expression profiling

Total RNA was isolated using Trizol (Invitrogen, Grand Island, NY) and treated sequentially with the RNeasy Mini Kit (Qiagen, Valencia, CA) and Turbo DNA Free (Ambion, Grand Island, NY). RNA-seq libraries were prepared from 300 ng RNA using the TruSeq RNA Sample Preparation kit (Illumina) and 75-bp single-end sequences were obtained on a NextSeq 500 Instrument (Illumina). Sequence tags were mapped to reference genome Hg19 using TopHat v2.0.6 and transcript levels in triplicate samples were calculated as fragments per kb per 10^6 mapped reads (FPKM) using Cufflinks v2.0.2. Differential expression was determined with CuffDiff, using Chi-square tests with 1 degree of freedom and two-tailed *P* values to assess statistical significance (Trapnell et al., 2012).

ChIP-seq

ChIP-assay

We prepared chromatin obtained from three pooled biological replicates of KP and DK after cross-linking for 10 minutes at RT with 1% formaldehyde-containing medium, using *tru*ChIP™ High Cell Chromatin Shearing Kit with SDS Shearing Buffer (Covaris). We sonicated nuclear extracts to obtain DNA fragments averaging 200 bp in length and immunoprecipitated the equivalent of 10^7 cells overnight with 10 µg of rabbit antibodies against H3K4me1 (ab8895, Abcam), H3K4me3 (ab8580, Abcam), and H3K27ac (ab4729, Abcam), as previously described (Cattoglio et al., 2010; Cui et al., 2009).

ChIP-seq library preparation and sequencing

We prepared Illumina libraries, for KP and DK, from 10 ng of immunoprecipitated DNA (IP) and control DNA (INPUT: nuclear extracts sonicated but non-immunoprecipitated) following the Illumina ChIP-seq DNA sample preparation kit. We checked the libraries by capillary electrophoresis by Agilent Bioanalyzer 2100 with the High sensitivity DNA assay and quantified them with Quant-iT™ PicoGreen® dsDNA Kits (Invitrogen) by Nanodrop Fluorometer. We sequenced each library in one lane of a single strand 50 bp Illumina Run.

Bioinformatic ChIP-seq data analysis

We mapped raw reads against the human reference genome (build hg19) using Bowtie(Langmead et al., 2009) allowing up to 2 or 3 mismatches. We then processed each BAM file by using SAMtools (Li et al., 2009), and converted each into a bed file using BEDTools (Quinlan and Hall, 2010). Then we checked the quality of each sequenced sample using cross-correlation analysis implemented in spp R package (Kharchenko et al., 2008). We performed ChIP-seq peak calling using SICER default parameters (Zang et al., 2009) and using each INPUT data to model the background noise. We used NHEK raw H3K4me3, H3K4me1, H3K27ac and H3K27me3 ChIP-seq data from the ENCODE Project (http://genome.ucsc.edu/ENCODE/; GSM733720, GSM733698, GSM733674, GSM733701) and analyzed them as described above for KP.

Identification of cis-regulatory elements

We developed a custom R-workflow to identify promoters and enhancers. The pipeline analyzes the histone modification islands generated by SICER and includes three steps. In the first step, the R script invokes BEDtools (Quinlan and Hall, 2010) to identify regions where H3K4me1 overlap or do not overlap with H3K4me3. $H3K4me3^+$ $H3K4me1^-$ and $H3K4me3^-$ $H3K4me1^+$ regions are classified as putative promoters and enhancers, respectively. In the second step, the R script first normalizes the tag counts of H3K4me3 and H3K4me1 using the sequencing depths of both libraries and then calculates the log-ratios between H3K4me3 and H3K4me1 tag counts for $H3K4me3^+H3K4me1^+$ regions. If the H3K4me3/H3K4me1 ratio

is greater than 0, the region is defined as putative promoter, otherwise as putative enhancer. Finally, we intersected putative promoters and enhancers with H3K27ac^{+} regions to identify active chromatin regions.

Identification of super-enhancers

Enhancers were stitched and super-enhancers were defined using ROSE code (https://bitbucket.org/young_computation/rose), as already described in (Loven et al., 2013; Whyte et al., 2013). Briefly, the algorithm stitches enhancers together if they lie within a certain distance and ranks the enhancers by their input-subtracted signal of H3K27ac. It then separates super-enhancers from typical enhancers by identifying an inflection point of H3K27ac signal versus enhancers rank. ROSE was run with stitching distance of 12,500 bp. In addition, all the enhancers wholly contained in a window ± 2,500 bp around an annotated transcriptional start site (RefSseq, build hg19) were excluded from stitching, allowing for a total 5,000 bp promoter exclusion zone. Super-enhancers were then assigned to the RefSeq gene whose TSS was the nearest to the center of the stitched enhancers.

Retroviral scanning

Retroviral vector construction and production

MLV-derived gamma-retroviral vector containing a green fluorescent protein (GFP) gene, under the control of a wild type MLV LTR (MFG.GFP) was previously described (Cavazza et al., 2013). To generate the MFG.GFPmod construct, the *XhoI* site contained in the LTRs was eliminated by partial digestion and self ligation, in order to obtain an MFG.GFP vector with a single *XhoI* site located in the 3'-LTR. Then, a 70-bp fragment of tha BirA gene was digested with *XhoI* from the pLU-Ub-BirA-hPGK-Cyan vector (a kind gift from D. Trono's lab) and cloned into the *XhoI* sites of MFG.GFP. The 70-bp BirA fragment inserted in the 3'LTR of the MFG.GFP vector allows for the unambiguous identification of MLV integration sites in human primary keratinocytes through LM-PCR, avoiding the contamination of sequences from LTR-containing endogenous retroviruses present in the murine 3T3 cells used as a feeder layer for the culture of keratinocytes.

RV vector supernatants were produced by transient transfection of the amphotropic Phoenix packaging cell line. The supernatant of transfected Phoenix packaging cell line was used to infect the amphotropic murine Am12 cell line, in order to obtain a stable packaging cell line of MGF.GFPmod vector. The vector was produced also as a VSV-G pseudotyped virus, by transient co-transfection of 293T cells; the viral supernatant was then collected and concentrated as described (Di Nunzio et al., 2008), and titrated on 293T cells.

Transduction of human primary keratinocytes

For KSC retroviral scanning library preparation, subconfluent primary skin keratinocytes were trypsinized and 2 x 10^6 cells plated onto a feeder layer of lethally-irradiated 3T3-J2 and Am12 MFG.GFPmod cells in a 1:2 ratio, as previously described (Di Nunzio et al., 2008), with the presence of 8ug/ml polybrene. Transduced keratinocytes were grown for three days, then trypsinized and re-plated onto new feeder-layers. Cells were maintained in culture for >6 passages, replating them at confluence onto feeder layer every 5-6 days. Each passage was monitored by Colony Formation Efficiency (CFE) assay, cytofluorimetric analysis of GFP-positive cells and cell doubling countings.

For the preparation of the DK library, keratinocytes from the same donor were differentiated by contact inhibition and cultured in a serum-free medium depleted of several growth factors for 6 days, until exhaustion of clonogenic ability. 4 x 10^6 keratinocytes were plated at subconfluence and infected by spinoculation (1800 rpm for 35 minutes) with concentrated MFG.GFPmod viral supernatant supplemented with 8ug/ml polybrene. After spinoculation, the supernatant was replaced with fresh medium. Cells were then collected after 72 hours and FACS-sorted.

Sequencing, mapping and annotation of retroviral integration sites

Retroviral integration sites were cloned by linker-mediated PCR (LM-PCR) adapted to the GS-FLX Genome Sequencer (Roche/454 Life Sciences, Branford, CT) pyrosequencing platform, as already described (Cattoglio et al., 2007). Briefly, genomic DNA was extracted from 0.5-5 x 10^6 infected cells and digested with *Mse*I and a second enzyme *EcoRV* to prevent amplification of internal 3' LTR fragments. An *Mse*I double-stranded linker was then ligated and LM-PCR performed with the following nested primers specific for the linker and the 5' LTR, containing a bead-capture tag and a sequencing tag.

Primer sequences:
LINKER NESTED PRIMER:
5'-GCCTTGCCAGCCCGCTCAGAGGGCTCCGCTTAAGGGAC-3'
5'LTR MLV NESTED PRIMER:
5'-GCCTCCCTCGCGCCATCAGTAGCATTGCCCTGTTAGCGAACGGTG-3'

For each transduction, we performed 2 restriction digestions, 6 linker ligations and 18 nested PCRs. Pooled LM-PCR

amplicons were quantified (NanoDrop Technologies, Wilmington, DE), checked by an Agilent Bioanalyzer (Agilent Technologies, Palo Alto, CA), size-fractionated by SPRI beads (Agencourt Bioscience Corporation, Beverly, MA), and sequenced according to the GS-FLX manufacturer's instructions. We performed 7 different sequencing runs for each library. Crude sequence reads were processed and mapped onto the human genome by an automated bioinformatics pipeline. A valid integration contained the MLV nested primer, the entire MLV genome up to a CA dinucleotide and the linker nested primer. Sequences between the 5' LTR and the linker primers were mapped onto the human genome (UCSC Human Genome Project Working Draft, hg19) using Blat sequence alignment tool, requiring a 95% identity over the entire sequence length and selecting the best hit. The absolute genomic coordinates of the integration sites where defined as a result of the combination of genomic alignment and vector relative orientation data. Random genomic sequences were mapped by the same criteria, and used as experimental controls (Cattoglio et al., 2010)

Genomic annotation of MLV integration sites

Insertion sites and experimental control sequences were annotated according to these criteria: sequences were classified as intergenic when occurring at an arbitrarily chosen distance of > 30 kb from any Known Gene (UCSC definition), TSS-proximal when 2,5 kb upstream or downstream of a transcriptional start site (TSS), and intragenic when within the transcribed portion of at least one known gene. In case of multiple transcript variants, we arbitrarily chose the isoform with the nearest TSS to an integration or random site.

For each site, we annotated the genomic features (CpG islands, conserved non-coding sequences, open chromatin) whose hg19 coordinates overlapped for at least 1 nucleotide with the ±50 kb interval around the insertion site. We used UCSC tracks (http://genome.ucsc.edu) for CpG islands (27,639 items). Genomic coordinates of 82,335 mammalian CNCs were described(Kim and Pritchard, 2007). Association with open chromatin sites were generated using publicly available data for DNAse HS and FAIRE-seq on NHEK (GSM1002658). Each integration site was associated with a p63 binding site when overlapping with one or more p63 binding peak. P63 ChIP-seq data were publicly available (GSE59827) or produced in this study. Each integration site was associated to a corresponding epigenetically defined regulatory region (strong and weak promoters and enhancers) when overlapping.

Validation of putative regulatory elements

We amplified by PCR twelve DK MLV-targeted putative enhancers and cloned them upstream of a minimal promoter-Firefly reporter cassette in the pGL4.23 vector (Promega). We used the empty pGL4.23 vector as negative control. pGL4.13 vector (Promega), containing a SV40 promoter-Firefly luciferase reporter cassette, served as positive control. We nucleofected 10^6 DK with 2 µg of test plasmid and 10 ng of pGL4.73 vector, a reporter plasmid expressing Renilla luciferase driven by SV40 promoter (Promega), using the FuGENE HD Transfection Reagent (Promega). 48 hours after transfection, we analyzed cell extracts using the Dual Luciferase assay kit according to manufacturer's instructions (Promega). We normalized Firefly luciferase activity to Renilla luciferase signal. Then, we calculated the fold changes between normalized Firefly luciferase activities of tested and control plasmids.

Other sequencing data analyses

Functional clustering analysis

Functional cluster analysis of differentially expressed genes identified by CAGE-seq and microarrays, and of differentially used promoters was performed using the DAVID 2.1 Functional Annotation Tool (http://david.abcc.ncifcrf.gov). In the DAVID annotation system, a Fisher exact test corrected for multiple comparisons (DAVID's EASE score) is adopted to measure the level of gene-enrichment in Gene Ontology (GO) annotation terms with respect to a background population, and GO categories considered over-represented when yielding an EASE score < 0.05. Genes were also analysed by the Ingenuity Pathways Analysis tool (Ingenuity® Systems, www.ingenuity.com), to search for the most relevant molecular interactions, functions and pathways linking them.

Functional annotation of enhancers

Functional annotation of enhancers defined by the epigenetics marks and by MLV integration sites, and of super-enhancers was obtained with GREAT (http://bejerano.stanford.edu/great/public/html/) (McLean et al., 2010), using the Two Nearest Genes within 100 kb association rules and the whole human genome as a background.

ChIP-seq signal profiles

Average ChIP-seq signal profiles around CAGE promoters and MLV integration sites were generated with the Heatmapper tool (http://deeptools.ie-freiburg.mpg.de/) using bigWig-converted Wig files generated by SICER.

Differential ChIP-seq signals analysis
The differential intensity of H3K4me3, H3K4me1 and H3K27ac ChIP-seq signals in promoter and enhancer regions of KP and DK was calculated using the diffReps package (Shen et al., 2013). Read counts plotted in the box-plots in Figure S3 were calculated using the BamLiquidator package (https://github.com/BradnerLab/pipeline).

Motif analysis
Motif analyses were performed using the MEME-ChIP Suite tool (http://meme.nbcr.net/meme/) and SeqPos motif tool (http://cistrome.dfci.harvard.edu/ap/root). To analyse the enrichment of transcription factor motifs in epigenetically defined active enhancers we scanned a region of 500 bp around the H3K27ac peak, while for super-enhancers we scanned the entire H3K27-defined regions. Motif analysis on MLV integration sites was obtained with MEME and DREME tools by scanning a 500 bp region around each integration site. Molecular and functional interaction networks were calculated using the Metacore software (Thomson Reuters).

Statistical analyses
All statistical analyses were performed using the Rweb1.03 statistical analysis package (www.math.montana.edu/Rweb/).

SUPPLEMENTAL REFERENCES

Cabili, M.N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., and Rinn, J.L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes & development *25*, 1915-1927.

Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., Semple, C.A., Taylor, M.S., Engstrom, P.G., Frith, M.C.*, et al.* (2006). Genome-wide analysis of mammalian promoter architecture and evolution. Nat Genet *38*, 626-635.

Cattoglio, C., Facchini, G., Sartori, D., Antonelli, A., Miccio, A., Cassani, B., Schmidt, M., von Kalle, C., Howe, S., Thrasher, A.J.*, et al.* (2007). Hot spots of retroviral integration in human CD34+ hematopoietic cells. Blood *110*, 1770-1778.

Cattoglio, C., Pellin, D., Rizzi, E., Maruggi, G., Corti, G., Miselli, F., Sartori, D., Guffanti, A., Di Serio, C., Ambrosi, A.*, et al.* (2010). High-definition mapping of retroviral integration sites identifies active regulatory elements in human multipotent hematopoietic progenitors. Blood *116*, 5507-5517.

Cavazza, A., Cocchiarella, F., Bartholomae, C., Schmidt, M., Pincelli, C., Larcher, F., and Mavilio, F. (2013). Self-inactivating MLV vectors have a reduced genotoxic profile in human epidermal keratinocytes. Gene Ther *20*, 949-957.

Cui, K., Zang, C., Roh, T.Y., Schones, D.E., Childs, R.W., Peng, W., and Zhao, K. (2009). Chromatin signatures in multipotent human hematopoietic stem cells indicate the fate of bivalent genes during differentiation. Cell stem cell *4*, 80-93.

Di Nunzio, F., Maruggi, G., Ferrari, S., Di Iorio, E., Poletti, V., Garcia, M., Del Rio, M., De Luca, M., Larcher, F., Pellegrini, G.*, et al.* (2008). Correction of Laminin-5 Deficiency in Human Epidermal Stem Cells by Transcriptionally Targeted Lentiviral Vectors. Mol Ther.

Jones, P.H., and Watt, F.M. (1993). Separation of human epidermal stem cells from transit amplifying cells on the basis of differences in integrin function and expression. Cell *73*, 713-724.

Khalil, A.M., Guttman, M., Huarte, M., Garber, M., Raj, A., Rivea Morales, D., Thomas, K., Presser, A., Bernstein, B.E., van Oudenaarden, A.*, et al.* (2009). Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. Proceedings of the National Academy of Sciences of the United States of America *106*, 11667-11672.

Kharchenko, P.V., Tolstorukov, M.Y., and Park, P.J. (2008). Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nat Biotechnol *26*, 1351-1359.

Kim, S.Y., and Pritchard, J.K. (2007). Adaptive evolution of conserved noncoding elements in mammals. PLoS Genet *3*, 1572-1586.

Kouwenhoven, E.N., Oti, M., Niehues, H., van Heeringen, S.J., Schalkwijk, J., Stunnenberg, H.G., van Bokhoven, H., and Zhou, H. (2015). Transcription factor p63 bookmarks and regulates dynamic enhancers during epidermal differentiation. EMBO reports *16*, 863-878.

Kretz, M., Webster, D.E., Flockhart, R.J., Lee, C.S., Zehnder, A., Lopez-Pajares, V., Qu, K., Zheng, G.X., Chow, J., Kim, G.E.*, et al.* Suppression of progenitor differentiation requires the long noncoding RNA ANCR. Genes Dev *26*, 338-343.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol *10*, R25.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078-2079.

Loven, J., Hoke, H.A., Lin, C.Y., Lau, A., Orlando, D.A., Vakoc, C.R., Bradner, J.E., Lee, T.I., and Young, R.A. (2013). Selective inhibition of tumor oncogenes by disruption of super-enhancers. Cell *153*, 320-334.

McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C.B., Wenger, A.M., and Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol *28*, 495-501.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841-842.

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

Romualdi, C., Bortoluzzi, S., and Danieli, G.A. (2001). Detecting differentially expressed genes in multiple tag sampling experiments: comparative evaluation of statistical tests. Hum Mol Genet *10*, 2133-2141.

Saxonov, S., Berg, P., and Brutlag, D.L. (2006). A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proceedings of the National Academy of Sciences of the United States of America *103*, 1412-1417.

Schug, J., Schuller, W.P., Kappen, C., Salbaum, J.M., Bucan, M., and Stoeckert, C.J., Jr. (2005). Promoter features related to tissue specificity as measured by Shannon entropy. Genome Biol *6*, R33.

Shen, L., Shao, N.Y., Liu, X., Maze, I., Feng, J., and Nestler, E.J. (2013). diffReps: detecting differential chromatin modification sites from ChIP-seq data with biological replicates. PloS one *8*, e65598.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature protocols *7*, 562-578.

Whyte, W.A., Orlando, D.A., Hnisz, D., Abraham, B.J., Lin, C.Y., Kagey, M.H., Rahl, P.B., Lee, T.I., and Young, R.A. (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell *153*, 307-319.

Wilming, L.G., Gilbert, J.G., Howe, K., Trevanion, S., Hubbard, T., and Harrow, J.L. (2008). The vertebrate genome annotation (Vega) database. Nucleic Acids Res *36*, D753-760.

Zang, C., Schones, D.E., Zeng, C., Cui, K., Zhao, K., and Peng, W. (2009). A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. Bioinformatics *25*, 1952-1958.

Zhang, Z., Carriero, N., Zheng, D., Karro, J., Harrison, P.M., and Gerstein, M. (2006). PseudoPipe: an automated pseudogene identification pipeline. Bioinformatics *22*, 1437-1439.