SUPPLEMENTAL INFORMATION

Feedback Regulation between Initiation and Maturation Networks Orchestrates the Chromatin Dynamics of Epidermal Lineage Commitment

Lingjie Li, Yong Wang, Jessica L. Torkelson, Gautam Shankar, Jillian M. Pattison, Hanson H. Zhen, Fengqin Fang, Zhana Duren, Jingxue Xin, Sadhana Gaddam, Sandra P. Melo, Sam N. Piekos, Jiang Li, Eric J. Liaw, Lang Chen, Rui Li, Marius Wernig, Wing H. Wong, Howard Y. Chang, Anthony E. Oro

Supplemental Information includes seven **Supplemental Figures**, three **Supplemental Tables** and **Supplemental Experimental Procedure**.

SUPPLEMENTAL FIGURE AND TABLE LEGENDS

Figure S1. Epigenomic Profiling During Epidermal Lineage Commitment, Related to Figure 1.

(A) Immunostaining of K18, p63, and K14 at selected time points during epidermal differentiation. Scale bar: 75μm.

(B) Experimental design of epidermal differentiation and sample collection. H9 hESCs were treated with differentiation medium 1 (E6 medium with RA/BMP4) for 7 days to induce into surface ectoderm progenitor cells; and then were switched into differentiation medium 2 (DKSFM with EGF/FGF) to induce into mature keratinocytes (60 days). Cells at each indicated time points were collected for ATAC-seq and RNA-seq analysis. Note: D43 is for ATAC-seq; and D45 is for RNA-seq (marked by "*").

(C) Distribution of genomic features in all and differential ATAC-seq peaks from epidermal differentiation.

(D) Representative gene ontology terms identified from the three clusters of differential accessible regions in Figure 1D.

(E) The coordinated gene expression and chromatin accessibility changes in the three defined clusters from epidermal differentiation. The RNA-seq trend plot is for the genes containing differential ATAC-seq signals at their promoters (Figure 1F), and ATAC-seq trend plot is same to Figure 1E.

(F-G) Stepminer analysis reveals the sequential chromatin accessibility changes during epidermal lineage commitment. Diagrams on the left illustrate the process of losing hESC-specific accessible sites (F); and of gaining keratinocyte-specific sites (G) during epidermal differentiation. Middle panel: numbers of hESC- (F) or keratinocyte- (G) specific ATAC-seq peaks retained at each stage during differentiation. Open and closed sites were annotated in yellow and gray color respectively. Right panel: associated gene expression at each stage during differentiation. Here, genes containing hESC- (F) or

keratinocyte- (G) specific peaks at their promoters were selected and $\log_2$(FPKM+1) values in the RNA-seq were plotted.

(H) Genome-wide comparison confirms a higher similarity between hESC-derived keratinocyte (H9KC) and normal human keratinocyte (NHK). The scatter plots of gene expression (RNA-seq, top), chromatin accessibility (ATAC-seq, middle) and transcription factor motif enrichment (defined by -$\log_{10}$ p-value, bottom) from the two cell types are shown.

Figure S2. Genome-wide Profiling and Functional Validation in Other Pluripotent Stem Cells, Related to Figure 1, Figure 2 and Figure 3.

(A) Schematic overview of repeating epidermal differentiation assay in other two pluripotent stem cell lines, i.e. H6 hESCs (same protocol in H9 differentiation) and iPSC (different protocol). Here the three major cell types at representative stages are labeled and used for further analysis.

(B-C) Light image and immunostaining of keratinocytes derived from H6 hESCs (H6KC) and iPSCs (iPS-KC) differentiation. Scale bar: 100μm.

(D) Consistent gene expression patterns during epidermal differentiation in different pluripotent cell lines. The heat maps of gene expression changes at representative stages in H6 and iPSC differentiation are shown. The color bar shows the relative expression value (Z-score of normalized value) from the qRT-PCR analysis. Each group of genes are color labeled, corresponding to the three major clusters from ATAC-seq and RNA-seq analysis in H9 cells.

(E) Consistent open chromatin profiles from epidermal differentiation in different pluripotent cell lines. Hierarchical clustering of cells at representative stages from iPSC and H9 differentiation is shown using their chromatin accessibility similarities from ATAC-seq analysis. The three major clusters are highlighted in different colors.

(F) Genome browser tracks of normalized ATAC-seq signal in early progenitor cells (iPS-D7) vs. mature keratinocytes (iPS-KC) from iPSC differentiation. In early stage, the signal is higher at *K8-K18* locus, but lower at *K5* locus; while in late stage, the signal is increased at *K5* locus, but reduced at *K8-K18* locus.

(G) Scatter plot of differential accessibility between "iPS-D7" and "iPS-KC" cells. y-axis: $\log_2$ fold change in reads per accessible region (iPS-D7 vs. iPS-KC-); x-axis: average $\log_{10}$ count per million reads (CPM) per accessible region (i.e. peak) in ATAC-seq. The values of $\log_2$ fold change >1, or < -1 are labeled as "iPS-D7 high" or "iPS-KC high" with respective colors.

(H) TF motif enrichment in early progenitor cells (iPS-D7, top panel) vs. mature keratinocytes (iPS-KC, bottom panel) from iPSC differentiation. The differential accessible regions are defined from (G). The y-axis is $-\log_{10}$ p-value of a motif enrichment score, which is sorted from largest to smallest. The x-axis is the ranking number of sorted motif. Motifs belong to one TF family are labeled with same color and are indicated in the top corner of each panel.

(I) Chromatin accessibility changes at TFAP2C and p63 binding sites in progenitor vs. mature keratinocyte from iPSC differentiation. Note a reciprocal pattern of TFAP2C and p63 has been shown in H9 differentiation analysis.

(J-K) TFAP2C initiates surface ectoderm commitment in other pluripotent cell lines. Top panel, schematic representation of the TetO-TFAP2C inducible expression system in H6 (J) and iPSC (K); bottom panel, gene expression changes in H6:TetO-TFAP2C (J) and iPSC:TetO-TFAP2C (K) cells. The cells were cultured in E6 basal medium with or without Dox for 7 days for comparison (Dox+ vs. Dox-). qRT-PCR values were normalized to the values of internal control GAPDH. Mean ± S.D. is shown (n=3).

Figure S3. Construction of TF-Chromatin Transcriptional Regulatory Network During Epidermal Lineage Commitment, Related to Figure 2.

(A) Integrated pipeline of TF-accessible regulatory element-TG based network reconstruction.

(B) Description of investigated features, data, and variables for TF, accessible regulatory element, and TG; and public available data sources related to chromatin connectivity.

(C) A full list of transcription factors presented in the TF-chromatin transcriptional regulatory network in Figure 2C. TFs associated with gaining or losing accessibility in the Initiation and Maturation process are listed respectively. Hub TFs in the inferred network (high rank TFs for network connectivity, i.e., the number of target genes) are highlighted with bold font.

(D) Validation of TF-chromatin transcriptional network by investigation of chromatin accessibility of TF binding sites, TF and TG expression of individual factors (OCT4, left panel; TFAP2C, middle panel; p63, right panel). In each panel, top left is the heat map showing the ATAC-seq signal changes within -1/+1kb region from TF binding sites [identified from ChIP-seq summits. OCT4, ENCODE, GEO: GSM803438; p63 (Zarnegar et al., 2012), GEO: GSE33571; TFAP2C, this research]; top right is the heatmap showing the relative expression level of TF in RNA-seq; and bottom trend plot showing the normalized TG expression changes in RNA-seq.

Figure S4. TFAP2C Induces Surface Ectoderm Differentiation, Related to Figure 3.

(A) Light microscopy images showing TFAP2C induces surface ectoderm differentiation upon Dox addition. TFAP2C over-expressed cells (Dox+) gained epithelial like morphology after Dox treatment for 1 day, while the control cells (Dox-) still maintained as tight colonies. Scale bar: 200μm.

(B-C) Loss of TFAP2C reduces the survival and growth of hESCs. (B) Light microscopy images of hESCs transfected with control or TFAP2C-specific siRNA oligos. The total number and size of ES colonies were decreased upon knocking down the expression of TFAP2C. (C) The knock-down efficiency was measured by qRT-PCR with normalization to the control group. Mean ± S.D. is shown (n=3). Scale bar:200 μm.

(D-E) TFAP2C overexpression induces hESC differentiation into surface ectoderm lineage. (D) Gene set enrichment analysis (GSEA) of RNA-seq shows surface ectoderm-specific gene set (Qu et al., 2016) is significantly enriched in TFAP2C-over expressed

cells comparing to control cells (TetO-TFAP2C Dox+ vs. TetO-TFAP2C Dox-). (E) Metagene analysis shows increased chromatin accessibility around the promoter region of surface ectoderm genes upon TFAP2C overexpression. The average ATAC-seq read counts within -2/+2kb region from the transcription start site (TSS) are shown.

(F) Gene Ontology (GO) analysis of the peaks with gained (red) or lost (blue) accessibility. The enriched terms are ranked by $-\log_{10}$ p-value.

(G) Venn diagram showing the overlap of top 20 TF motifs from "TetO-TFAP2C-D7, Dox+" cells with the leading TF motifs from the Initiation network (Figure S3C, top ranked).

(H) Comparison of TF motif enrichment in early differentiated cells by TFAP2C activation (TetO-TFAP2C Dox+, y axis), versus by RA/BMP4 induction (x, axis) at day 7. The Scatter plot of TF motif enrichment level (defined by $-\log_{10}$ p-value) of each cell types with correlation (R) value displayed ($R=0.9395$, $p<0.0001$).

(I) PIQ footprinting analysis indicates an increased likelihood of the occupancy of Initiation network TFs upon TFAP2C overexpression at D7. Significant difference with **** p-value<0.0001 relative to control (DOX-) determined by t-test.

(J) AP2 motifs (AP-2gamma and AP-2alpha) were identified as the top enriched motifs from TFAP2C ChIP-seq analysis.

(K) Distribution of genomic features in TFAP2C-bound regions from ChIP-seq analysis.

(L) Histogram showing the distance to TSS of all TFAP2C-bound regions from ChIP-seq analysis.

(M-N) Functional validation confirms the consistent effect of other Initiation network TFs on surface ectoderm differentiation. The gene expression changes in H9 hESCs containing an inducible over-expression system for GATA3 (i.e. TetO-GATA3, M) and GRHL2 (i.e. TetO-GRHL2, N) are shown. The cells were cultured in E6 basal medium with or without Dox for 7 days for comparison (Dox+ vs. Dox-). qRT-PCR values were normalized to the values of internal control GAPDH. Mean ± S.D. is shown (n=3).

Figure S5. TFAP2C-Induced Surface Ectodermal Cells Can Further Differentiate into Mature Keratinocyte, Related to Figure 4.

(A) FACS analysis comparing the protein expression levels of K18 and K14 in TetO-TFAP2C-D7, TetO-TFAP2C-KC and normal keratinocyte (NHK).

(B-C) TFAP2C-induced basal keratinocytes differentiate into stratified epithelial cells upon calcium treatment. Morphology (B) and gene expression (C) were compared in the cells at D0, D3, and D6 after $CaCL_2$ treatment. Scale bar: 100um. qRT-PCR values were normalized to the value in D0. Mean ± S.D. is shown (n=3).

(D) Venn diagram showing the overlap of top 20 TF motifs from "TetO-TFAP2C-KC" cells with the leading TF motifs from the Maturation network (Figure S3C, top ranked).

(E) Comparison of TF motif enrichment in TFAP2C-induced keratinocytes (TetO-TFAP2C-KC, y axis) versus keratinocytes from normal differentiation (H9KC, x axis). The Scatter plot of TF motif enrichment level (defined by $-\log_{10}$ p-value) of each cell types with correlation (R) value displayed ($R=0.9948$, $p<0.0001$).

(F) PIQ footprinting analysis indicates an increased likelihood of the occupancy of most Maturation network TFs in the keratinocyte from TFAP2C induction (TetO-TFAP2C-KC). Significant difference with **** p-value<0.0001 between TetO-TFAP2C-KC vs. TetO-TFAP2C-D7 determined by t-test.

Figure S6. Functional Study of KLF4 Confirms the Consistent Effect with p63 on Keratinocyte Maturation, Related to Figure 5 and Figure 6.

(A) Loss of p63 resulted in higher level of chromatin accessibility at Initiation-specific peaks, and lower level of that at Maturation-specific peaks. Comparison of the overlap between differential ATAC-seq peaks at D21 (TetO-TFAP2C+p63KO vs. TetO-TFAP2C, identified in Figure 6) with Initiation- (defined from TetO-TFAP2C-D7) and Maturation-(defined from TetO-TFAP2C-KC) specific peaks.

(B-E) Overexpression of KLF4 increases the maturation-associated maker gene expression. (B-C) Constitutively over-expression of KLF4 by lentiviral transfection in H9

hESCs. The cells were differentiated for 14 days. KLF4 expression level was measured by qRT-PCR with normalization to the control group (B). IF staining shows increased level of K14+ cells, but less change of K18+ cells (C). (D-E) Inducible overexpression of KLF4 in H9 hESCs (TetO-KLF4). KLF4 level was elevated by adding doxycycline after 7 days early commitment, and marker gene expression changes were measured during the following differentiation time. (D) qRT-CPR shows increased level of KLF4 upon Dox induction (measured at D14), but its over-expression has less effect on p63. qRT-PCR values were normalized to the values in control group (Dox+ vs. Dox-). (E) Induced-over expression of KLF4 strongly increases the expression level of maturation markers (K5, K14), but has less effect on surface ectoderm markers (K8, K18). The expression level was measure by qRT-PCR and the values were normalized to that in control group (Dox-) at D14. Mean ± S.D. is shown (n=3).

(F) Knock-down of KLF4 expression results in a decreased level of maturation markers, but less effect on surface ectoderm genes or p63 during epidermal differentiation. TetO-TFAP2C H9 hESCs were transfected with lentivirus containing control or KLF4-specific shRNAs after 7 days differentiation and the gene expression level was measure at D21 by qRT-PCR. The expression values were normalize to that in control group. Mean ± S.D. is shown (n=3).

(G) PIQ footprinting analysis indicates a lower likelihood of TF occupancy of KLF4 upon p63 loss of function (TetO-TFAP2C+p63KO). Significant difference with **** $p<0.0001$ relative to control (TetO-TFAP2C) determined by t-test.

Figure S7. Genome-Wide Analysis of TFAP2C and p63 Binding Activities during Epidermal Differentiation, Related to Figure 6 and Discussion.

(A) Genome browser tracks comparing ATAC-seq signal between TFAP2C-induced surface ectoderm progenitor cells (TetO-TFAP2C-D7) and mature keratinocyte (TetO-TFAP2C-KC), relative to p63 ChIP-seq (Zarnegar et al., 2012) at the *TFAP2C* locus. Note: there is a modest decrease of ATAC-seq signal in the promoter region in TetO-TFAP2C-

KC comparing to TetO-TFAP2C-D7; and two p63 binding sites found in the distal region of TFAP2C locus (3' end).

(B) TFAP2C binding sites become less accessible in mature keratinocyte. Left panel, the histogram showing the distribution of ATAC-seq read counts within TFAP2C ChIP-seq peak regions in TFAP2C-induced mature keratinocytes (TetO-TFAP2C-KC) and surface ectoderm progenitor cells (TetO-TFAP2C-D7). Right panel, average enrichment of ATAC-seq chromatin accessibility signal within -1/+1kb region from TFAP2C ChIP-seq peak summits in TFAP2C-induced mature keratinocytes and surface ectoderm progenitor cells.

(C) Distance of p63 binding sites to associated genes in the three clusters identified from Figure 6H (group a, b and c) is shown in top panel. The genome browser tracks of p63 ChIP-seq, and of ATAC-seq for TetO-TFAP2C-D7(Dox-), TetO-TFAP2C-D7(Dox+), and TetO-TFAP2C-KC in the loci of representative genes from the three clusters are shown in bottom panel. The overlapping region between the p63 peak and ATAC-seq peak was highlighted.

(D) Distance of TFAP2C binding sites to associated genes in the two clusters identified in Figure 6L (group a and b) is shown in top panel. The genome browser tracks of TFAP2C ChIP-seq, and of ATAC-seq for TetO-TFAP2C-D7(Dox-), TetO-TFAP2C-D7(Dox+), and TetO-TFAP2C-KC in the loci of representative genes from the two clusters are shown in bottom panel. The overlapping region between the TFAP2C peak and ATAC-seq peak was highlighted.

Table S1. The full list of identified transcription factors during differentiation, Related to Figure 2.

Table S2. The full list of association of union peaks and target genes by HiC and DNase-seq accessibility correlation across tissues, Related to Figure 2.

Table S3. TF-RE-TG list in the networks, Related to Figure 2.

SUPPLEMENTAL EXPERIMENTAL PROCEDURE

Cell culture and differentiation

Human ESC culture

Human ESCs and modified cell lines were seeded in a feeder-free system using Matrigel hESC-Qualified Matrix (BD Corning) and were maintained in Essential 8 media (Life Technologies).

Doxycycline inducible TFAP2C (TetO-TFAP2C) human ESCs

To generate Doxycycline inducible expression cell line, we use PiggyBac Doxycycline inducible plasmid (PB/TW/CRB, courtesy of Yamanaka lab), which contains a CAG promoter driven rtTA-IRES-BSD cassette and pTRE (tetracycline response element) driven transgene expression cassette. Human TFAP2C cDNA (Sino Biological Inc. Cat. No. HG13115-G) was amplified and sub-cloned into the Piggybac plasmid (named as "PB-CRB-TFAP2C") using In-Fusion HD Cloning Kit (Clontech). PB-CRB-TFAP2C expression plasmid and Transposase expression plasmid (System Biosciences) were mixed and co-transfected into H9 hESCs with Lipofectamine™ LTX Reagent with PLUS™ Reagent (Thermo Fisher Scientific,15338100) by following standard protocol. Forty-eight hours after transfection, the cells were selected by blasticidin (4μg/mL) and resistant ES clone were picked and expanded. TFAP2C induction was confirmed by IF and qRT-PCR after Dox treatment over 2 days.

CRISPR/Cas9-mediated targeted gene deletion/replacement via homologous recombination in hESCs

We used CRISPR/Cas9-mediated genome editing to generate our gene knockout hESC lines. We designed two gRNAs to target independent loci surrounding a 5 kb genomic region that we wished to delete. Donor sequences with 700 bp arms flanking left and right of the region to be deleted were also used to promote homologous recombination at the locus. Knockouts were designed to mimic the null mouse models, whereby the conserved DNA binding domains were targeted for deletion: for TFAP2C – exons 4-7; for p63 – exons 6-8 (more detail shown in Melo, et al., Nature Genetics, in press). Both gRNAs and donor sequences were synthesized as 5'-phosphorylated gene blocks (IDT), and the

gRNAs were incorporated into a DNA fragment with all components necessary for gRNA expression. Constructs were transfected into the hESC cells alongside an hCas9 expression plasmid, using Lipofectamine™ LTX Reagent with PLUS™ Reagent (Thermo Fisher Scientific,15338100) and the recommended manufacturer's protocol. After 48 hours, we added puromycin to the cells to select out positive cells for gRNA expression. hESC colonies were picked and expanded individually and knockouts were verified by genomic PCR, Western blotting, and immunofluorescence. The sequences of gRNA are as follows:

TFAP2C-KO gRNA1: ATGCTTAAATGCCTCGTTAC

TFAP2C-KO gRNA2: ACAGAACCTCCACGGGGACT

p63-KO gRNA1: TAGTCATTTGATTCGAGTAG

p63-KO gRNA2: GTAAAGCTGTAGTACATGCC

In vitro differentiation of hESC into keratinocyte

H9 hESCs were passed on to Matrigel coated plate and maintained in Essential 8 media. To induce differentiation into keratinocyte, the cells were first fed with "differentiation medium 1" containing 5ng/ml BMP4 and 1 μM RA in Essential 6 media (Life Technologies) for seven days; thus forming surface ectodermal progenitor cells. After seven days, the medium was changed to "differentiation medium 2", i.e. Defined Keratinocyte Serum Free Medium (DKSFM) with growth supplements containing EGF and FGF (Life Technologies), and the cells underwent selection and expansion for 2 months to get maturation into keratinocyte. The resulting keratinocyte colonies were passed onto Corning PureCoat™ ECM Mimetic 6-well Collagen I Peptide Plate (Corning) and expanded in DKSFM medium. The study of keratinocyte differentiation has been reproduced in other two pluripotent cell lines: H6 hESCs and iPSCs. For H6 hESCs: the cells were maintained and differentiated into keratinocytes using the same protocol for H9 hESC differentiaiton. The iPSC line was derived by sendai virus vector-mediated reprogramming in human fibroblast from Invitrogen (Macarthur et al., 2012). The cells were differentiated into keratinocyte using a different protocol revised from Sebastiano et al., 2014.

In vitro differentiation of TetO-TFAP2C H9 hESCs

To study the function of TFAP2C in surface ectoderm differentiation, TetO-TFAP2C H9 hESCs were induced with 2 μg/ml Doxycycline(Dox) in E6 medium without RA/BMP4 for 7 days. The resulting cells were imaged and collected for subsequent analysis.

To study long term epidermal differentiation ability, TetO-TFAP2C H9 hESCs were first induced into surface ectodermal cells by activation of TFAP2C expression with Dox in E6 medium for 7 days; and then were transitioned to DKSFM with Dox for 14 days and then maintained in DKSFM only until keratinocyte colony was formed. The differentiated keratinocytes were passed and maintained onto Corning PureCoat™ ECM Mimetic 6-well Collagen I Peptide Plate (Corning) in DKSFM, and collected for additional analysis.

Small interfering RNA (siRNA)-mediated gene knockdown in hESCs

To knockdown the expression level of TFAP2C, H9 hESCs were plated onto 6 well plate and were transfected with 30pmol of siRNAs (Sigma-Aldrich) using Lipofectamine RNAiMAX Reagent (Invitrogen) with standard protocol. The siRNA oligonucleotides were designed, synthesized and fluorescent labeled by Sigma-Aldrich. In this study, we use one control siRNA and two independent siRNA for TFAP2C. The sequences of siRNA are as follows:

TFAP2C siRNA #1: sense (5'-3'): GUAAACCAGUGGCAGAAUA[dT][dT];
antisense (5'-3'): UAUUCUGCCACUGGUUUAC[dT][dT][Cyanine5];
TFAP2C siRNA #2: CACAGAACUUCUCAGCCAA[dT][dT];
antisense (5'-3'): UUGGCUGAGAAGUUCUGUG[dT][dT][Cyanine5]
Control siRNA:
siRNA Fluorescent Universal Negative Control #1, Cyanine 5 (cat. No. SIC005-10NMOL, Sigma-Aldrich)

Short hairpin RNA (shRNA)-mediated gene knockdown in cells from hESC differentiation

To knockdown the expression level of KLF4 during epidermal differentiation, the TetO-TFAP2C H9 hESCs were induced into surface ectodermal cells by 7-days Doxycycline treatment ,and then were infected with lentivirus from a pGFP-C-shLenti vector (Origene, cat. No.TL316853) containing two cassettes for shRNA of KLF4 and GFP expression

under different promoters . At differentiation day 21, the infection positive cells were sorted by GFP intensity and were subjected to expression analysis. Lentiviral particles were produced and the cells were infected according to the manual from Origene (HUSH shRNA plasmids and lenti-particles application guide). The sequences of shRNA are as follows:

KLF4 shRNA#1 (5'-3'): TGAGGCAGCCACCTGGCGAGTCTGACATG

KLF4 shRNA#2 (5'-3'): TCAGATGAACTGACCAGGCACTACCGTAA

Control shRNA: Scrambled negative control non-effective shRNA cassette in pGFP-C-shLenti plasmid (Cat. No. TR30021, Origene).

RNA expression analysis

Total RNA from cells was extracted with the RNeasy mini kit (Qiagen). Gene expression was measured using the One-Step RT-PCR SYBR green kit (Stratagene) according to manufactory instructions and normalized to the internal control gene GAPDH.

Immunofluorescence staining

Cells were fixed with 4% paraformaldehyde, permeabilized with 0.2% Triton X-100 in PBS for 30 min, and blocked with 10% horse serum (Vector Laboratories) in 0.2% Triton X-100 in PBS for 30 min at room temperature. Cells were incubated overnight at 4 °C with primary antibodies. Primary antibodies are: K18 (1:800, R&D AF7619), K14 (1:800, BioLegend SIG-3476-100); AP-2γ (1:100, Cell Signaling 2320), p63 (1:100 Gene Tex GTX102425), ITGA6(1:200, Millipore, MAB1378). Then the cells were incubated with Alexa 488, 555 and 647-conjugated secondary antibodies diluted with 0.2% Triton X-100 in PBS (1:500, Life Technologies) for 1 hour at room temperature. Following three washes with PBS, slides were mounted with the Prolong Gold mounting medium (Life Technologies). Prior to mounting, slides were incubated with 1:10000 Hoechst for 10 min. The fluorescence images were taken using the TCS SP2 confocal laser scanning microscope (Leica)

Calcium induced differentiation

To mimic the stratification process in 2-D culture system, keratinocytes derived from TetO-TFAP2C H9 hESCs were induced by the addition of 1.2mM CaCl2 to the medium for 3 or 6 days at full confluence. The cells at each time points were imaged by light microscope and lysed for qRT-PCR analysis on skin stratification associated genes.

In vitro skin reconstitution assay

Generation of organotypic epidermis was performed by following the protocol described previously (Sebastiano et al., 2014) with minor modification. Basically, $1X10^6$ keratinocytes derived from normal differentiation with H9 hESC or from TFAP2C-induced differentiation were collected and seeded on top of devitalized dermis and cultured submerged in DKSF medium for 5 days. The medium was then gradually changed to Keratinocyte Growth Medium (KGM) for 7 days, after which stratification was induced by raising the dermis to air-liquid interface. After 2 weeks, the reconstituted epidermis was collected, fixed in 4% paraformaldehyde, and embedded in optimum cutting temperature compound (OCT) and paraffin for IF staining. The primary antibodies are: K14(1:2000, COVANCE, PRB-155P), K10 (1:500, COVANCE, PRB-159P), Loricrin (1:500, Covance, PRB-145P), Collagen VII (1:250, Millipore MAB1345).

Intracellular staining and FACS analysis

Cells were dissociated and washed twice using FACS buffer (2% BSA /PBS), then stained with Fc blocker (anti-mouse CD16/CD32) and live/dead Aqua 30min on ice. After washing, cells were then fixed and permeabilized with BD reagent (BD cytofix/cytoperm kit 554714). Subsequently, fluorophore conjugated antibodies: anti-K14 (PerCP, CBL197F, Millipore), anti-K18 (FITC, NB120-7797, Novus) antibodies or isotype control were incubated with cells at 4°C for 30min (1:100 dilution). After washing, the samples were applied for FACS analysis.

List of qRT-PCR primers

GAPDH F:5-CTGAGAACGGGAAGCTTGT-3 R:5-GGGTGCTAAGCAGTTGGT-3
TFAP2C F:5-AGATTGGGTTGAATCTTCCG-3 R: 5-GGCTTCACAGACATAGGCAA-3
TFAP2A F: 5-CAGATATGCAAAGAGTTCACCGAC-3 R: 5-TCAAGCAGCTCTGGATGCC-3
GRHL2 F: 5-TCAATACCCGAAGAGCCTACA-3 R: 5-CTTGGCTGTCACTTGCTTTGC-3

GATA3 F: 5-GCGGGCTCTATCACAAAATGA-3 R: 5-GCCTTCGCTTGGGCTTAAT-3
p63 F: 5-TTTCCCACCCCGAGATGA -3 R: 5-TGCGGCGAGCATCCAT-3
KLF4 F: 5-CCTTACCACTGTGACTGGGA-3 R: 5-CCCGTGTGTTTACGGTAGTG-3
K8 F: 5-GATCGCCACCTACAGGAAGCT-3 R: 5-ACTCATGTTCTGCATCCCAGACT-3
K18 F: 5-CCGTCTTGCTGCTGATGACT-3 R: 5-GGCCTTTTACTTCCTCTTCGTG-3
K14 F: 5-GACCATTGAGGACCTGAGGA-3 R: 5-ATTGATGTCGGCTTCCACAC-3
K5 F: 5-ATCTCTGAGATGAACCGGATGATC-3 R:5-CAGATTGGCGCACTGTTTCTT-3
LAMB3 F: 5-GACAGGACTGGAGAAGCGTGTG-3 R:5-CCATTGGCTCAGGCTCAGCT-3
ITGA6 F: 5-GCTGGTTATAATCCTTCAATATCAATTGT-3 R:5-TTGGGCTCAGAACCTTGGTTT-3
ITGB4 F: 5-CTGTACCCGTATTGCGACT-3 R:5-AGGCCATAGCAGACCTCGTA-3
COL7A1 F: 5-GATGACCCACGGACAGAGTT-3 R: 5-ACTTCCCGTCTGTGATCAGG-3
K1 F: 5-TACCTCCACTAGAACCCAT-3 R: 5-GCTGCAAGTTGTCAAGTT-3
K10 F: 5-CGCCTGGCTTCCTACTTGG -3 R: 5-CTGGCGCAGAGCTACCTCA-3
IVL F: 5-AAAGCACCTAGAGCACCC-3 R: 5-GGTTGAATGTCTTGGACCT-3
FLG F: 5-GACATGGCAGCTATGGTA -3 R: 5-AATCCCAGTTGTTTCGATA-3

RNA-seq library preparation and data processing

RNA extraction was performed using RNeasy Plus (Qiagen) from the samples reported in this research. Ribosomal RNA was removed from each RNA extraction using Ribo-Zero Gold rRNA Removal kit (Illumina). The RNA-seq libraries were constructed by TruSeq Stranded mRNA Library Prep kit (Illumina) and sequenced on Illumina Hiseq2000 or NextSeq sequencers. Sequencing reads were mapped to hg19 using TopHat. FPKM values were called using cufflinks. For the time series dataset, the genes having a FPKM value of 0 across all time points were filtered out. Differential gene expression was performed by looking at $\log_2$ fold change > 1. For the TFAP2C samples, differential expression was performed using cuffdiff. Differentially expressed genes had a p-Value < 0.01 as well as a log2 fold change > 1.

ChIP-seq library preparation and data processing

ChIP assay was performed following previously described method (Calo et al., 2015) with minor modification. Briefly, TetO-TFAP2C H9 hESCs were induced with Dox for 7 days, after which cells were crosslinked in 1% formaldehyde for 10 min, followed by quenching with 0.125M glycine for 5min at room temperature (RT). Chromatin was sheared to an average 100-300bp in the Covaris sonicator (Covaris). Sonicated chromatin solution was

aliquoted and incubated overnight at 4°C with the specific antibodies. The antibody used are as follows: AP-2γ (H-77) (sc-8977X, Santa Cruz); H3K4m3 (39159, Active Motif); H3K4m1 (ab8895, Abcam); H3K27ac (39133, Active Motif); H3K27m3 (39155, Active Motif); H3K9m3(ab8898, Abcam). For TFAP2C ChIP-seq, we used chromatin extract from 30 million cells for each experiment. For histone ChIP-seq, we used 10 million cells for each experiment. Immunocomplexes were captured by protein-G Dynal magnetic beads (Life Technologies) followed by stringent washes and elution. The eluted samples were reverse cross-linked at 65°C overnight, and were digested by RNase A at 37°C for 2 hours and proteinase K at 55°C for 30 min to deplete the RNA and proteins. The DNA was purified by QIAquick PCR purification Kit (Qiagen). ChIP-seq libraries were prepared following the NEBNext protocol and sequenced on Illumina NextSeq sequencers. Sequencing reads were mapped to hg19 using Bowtie1.1.2 (Langmead et al., 2009) with parameters –best, --strata and –m 1 to allow for only one alignment. Duplicates are then removed using Samtools rmdup. Peaks were identified using MACS2 (Zhang et al., 2008) with a FDR of 0.01.

ATAC-Seq Library preparation, sequencing, and data preprocessing

The regular ATAC-seq was performed as described (Buenrostro et al., 2013). For iPSC-differentiated cells, we did Omni-ATAC, an improved ATAC-seq protocol described in (Corces et al., 2017). ATAC-seq libraries were sequenced on Illumina Hiseq2000 or NextSeq sequencers. ATAC-Seq pair-end reads were trimmed for Illumina adapter sequences and transposase sequences using a customized script and mapped to hg19 using bowtie (Langmead and Salzberg, 2012) with parameters –S –X2000 –ml. Duplicate reads were discarded with Samtools rmdup (Li et al., 2009). Peaks were identified using HOTSPOT with default parameters (http://www.uwencode.org/proj/hotspot/). HOTSPOT analysis generates two types of peaks: narrow peak and hotspot regions (broad peak). In this study, we used the narrow peaks for all the analysis. Overlapping peaks from all samples were merged into a unique peak list, and raw read counts mapped to each peak for each individual samples were quantified.

Irreproducible Discovery Rate (IDR) analysis

IDR analysis was performed on ATAC-seq and ChIP-seq peaks. In both cases, IDR was called on the p-value of the original peaks.

Differential ATAC-seq peak analysis

Differential peak analysis in paired samples:

Differentially accessible peaks from the merged union peak list were selected with edgeR package from Bioconductor using raw counts of each samples in the union peak list and a fold change threshold of 2, and p-value <0.01.

Differential peak analysis in time-series ATAC-seq from normal differentiation:

Peaks with significantly changing signal were selected using the edgeR package from Bioconductor. All dispersions were estimated. An ANOVA-like test was conducted by specifying all pairwise contrasts with an arbitrarily selected time point, relative to a 1.5-fold change, and a significance threshold of FDR 0.05 used. The read counts of the differential peaks in each sample were further normalized by Z-score transformation. Hierarchical clustering was used to cluster the peaks and samples. The result was presented as a heatmap by Java TreeView.

Genomic annotation of peaks from ATAC-seq or ChIP-seq

Peak genomic annotation was performed by a customized script assigning peaks to specific regions using the following criteria: “promoter”: regions from -2kb to 1kb of transcription start sites (TSS); “TSS-proximal enhancer”: regions from -10kb to -2kb of TSS; “gene tail”: regions from -2kb to transcription end sites; “exon” and “intron” regions in the gene body and defined in Refseq annotation; “intergenic”: any other genomic regions.

Gene ontology analysis

Gene ontology enrichment analysis were performed using GREAT (McLean et al., 2010) to calculate statistics by associating genomic regions with nearby genes (here we chose the parameter: “two nearest genes within 100kb”) and applying the gene annotations to the regions.

Motif analysis of peaks from ATAC-seq and ChIP-seq

Motif analysis on peak regions was performed using HOMER function (http://homer.salk.edu/homer/motif/) "findMotifsGenome.pl" with default parameters to calculate the occurrence of a TF motif in peak regions compared to that in background regions. We used "-$\log_{10}$ (p-value)" to rank the enrichment level of TF motifs.

PIQ footprinting analysis

We used PIQ, a machine-learning algorithm (Sherwood et al., 2014), to identify the TF binding footprints at corresponding motifs from ATAC-seq data. PIQ outputs the purity scores to evaluate the probability of true TF binding. In this study, we investigated the binding activity of a list of specific TFs in several group comparisons. For each TF, the motif position weight matrix (PWM) were from JASPAR database (http://jaspar.genereg.net) and HOMER. To maximize the detection sensitivity, we merged the ATAC-seq reads from replicates. To eliminate the detection bias, we sampled from the higher-depth merged library to achieve equal depth for next-step analysis if the sequencing depth was quite different among the samples from one group of comparison. After running PIQ algorithm, the purity score of one given TF from different samples were compared by paired t-test.

ATAC-seq signal intensity around TSS

A window from -2k upstream to +2k downstream of TSS was taken and divided into eighty 50bp equally sized bins. The number of uniquely mapped and properly paired ATAC-seq tags overlapping each bin was counted by a customized script. The average fragment count in each bin of all the TSS regions was plotted.

ATAC-seq signal intensity around transcription factor bound regions

TF ChIP-seq data were either generated in this study or download from public resource. In each TF analysis, a window from -1k upstream to +1k downstream of ChIP-seq peak summits was taken and divided into forty 50bp equally sized bins. The number of uniquely mapped and properly paired ATAC-seq tags overlapping each bin was counted, normalized by library size and log-transformed using an in-house script. Therefore, we

generated a matrix with each row representing a peak region, and each column containing the normalized tag counts from a 50 bp bin in a consecutive manner within the 2kb window. To visualize the signal intensity across all the TF bound regions, the data matrix was presented as a heatmap by Java TreeView. To get an average intensity, we took the mean of fragment count in each bin of all the binding region and plot the result as a line graph or heatmap. To compare the intensity among different samples, we merged the individual data matrix from each group into one single file, and performed k-means clustering on peak regions and presented the result in a heatmap. Each cluster of peak regions was taken and compared among all the samples. The associated GO terms and distance to TSS were analyzed based on each cluster of peak regions.

Gene set enrichment analysis (GSEA)

GSEA was performed to determine whether a priori defined set of genes shows a statistically significant difference between biological samples (Subramanian et al., 2005). In this study, we generated a gene set of surface ectoderm from previous research (Qu et al., 2016). Gene expression value from the RNA-seq analysis of the samples “TeTO-TFAP2C Dox+” and “TetO-TFAP2C Dox-” were used for comparison. GSEA was performed following the developer’s protocol (http://www.broad.mit.edu/gsea/).

Selecting key TFs by integrating expression and motif enrichment

The following procedure is developed to identify key TFs from the paired time-series ATAC-seq and RNA-seq data in differentiation. We started by curating a TF-motif mapping to associate binding motifs in Homer database with its corresponding TF (Zamanighomi et al., 2017). We created a list of enriched TF binding motifs in the accessible regions from the ATAC-seq data at each time point by HOMER software (Heinz et al., 2010). We then filtered the motif list by checking TFs’ expression level from RNA-seq data at each time point. Those TFs were divided into 5 levels by FPKM value, Level I: <2; Level II: 2-7; Level III: 7-12, Level IV: 12-20; Level V: >20. TFs were divided into 5 levels by enrichment score, Level I: <45; Level II: 45-250; Level III: 250-400; Level IV: 400-900; Level V: >900. TFs’ Motif enrichment was assessed by the enrichment score –log(P) in open regions. Here P is the p-value derived from hypergeometric test. Based

on the above binning data for FPKM and enrichment score, we drew the bubble heatmap for the TFs shown in Figure 2A. The circle point in heatmap denotes the TF. The size of a circle point denotes the level of TF's motif enrichment score. The color of a circle point denotes TF's FPKM value. The x-axis is the time point along differentiation.

We first derived a long list of 39 TFs (Table S1). To extract the most dynamically and highly expressed TFs in the time series, we applied the following filter to obtain a short list. For expression value, we filtered those TFs whose FPKM values were not changed (at the same expression level): MAFK, TCF7L2, ATF7, and ATF4. We also filtered TFs whose level of FPKM values were almost invariant: BACH2, USF2, ATF1, and NF1. Finally, we removed those TFs whose maximal FPKM across time points was less than 14: MAFK, BACH2, FOSL1, TCF7L2, ZBTB33, ATF7, HOXA2, IRF1, IRF2, CEBPB, GABPA, ELK1, ELK4, and NF1. We did similar filter by the motif enrichment score [-log(P)]. We filtered those TFs whose enrichment scores were at the same level: ZBTB33 and PBX1; whose level of enrichment scores were almost invariant: NR4A1, TCF3, TCF7L2, IRF1, ELK1, ELK4, ELF1, and MAFF. As a result, we obtained a short list of 17 TFs. This includes the key TFs highly expressed and dynamic in both expression and chromatin accessible site binding during the differentiation: OCT4 (POU5F1), SOX2, CTCF, GATA3, TFAP2C, GRHL2, p63, and KLF4.

In addition to the visualization by binning data in heatmap, we defined one driver score to quantitatively assess the TF importance by integrating its expression value and motif enrichment score. Suppose TF's expression at time point t is $\mathrm{X}_{TF,t}$, and its motif enrichment score in open regions at time point t is $\mathrm{E}_{TF,t}$. Then we defined the driver score by taking the geometrical mean as follows,

$$\mathrm{D}_{TF,t} = \sqrt{\mathrm{X}_{TF,t} \times \mathrm{E}_{TF,t}}$$

This driver score allows us to illustrate the functional pattern of for several important TFs. Those patterns indicate the driver score can efficiently depict TF's regulatory roles. The expression level, enrichment score, and driver score for 39 TFs are summarized in Table S1.

Predicting target genes for ATAC-seq union peaks

To assign the target genes for the union peaks, we obtained the associations for the union peaks with the promoter of target genes from two sources: physical interaction from HiC experimental data and inferred from cross-tissue accessibility correlation. One was the high resolution HiC data in NHEK cell line (Rao et al., 2014). We detected if the target gene's promoter and the union peak were within the two interacting chromatin regions by HiC. In total, we have 11,083 union peak and target gene interactions (Union peaks 188,850, target genes 2,006, each gene is averagely regulated by 5.52 union peaks). The second source was the co-accessibility relationship from cross-tissue correlation of DNase-seq data in ENCODE (Consortium, 2012; Thurman et al., 2012). We detected if the promoter DHSs and union peak DHSs within ±500 kb correlate with each other at threshold 0.7. We obtained 50,129 union peak and target gene relationships (Union peaks 188,850, target genes 27,110, each gene is averagely regulated by 1.85 union peaks). By pooling these two sources together, we had a draft 56,397 associations to connect union peaks with possible target genes (Union peaks 188,850, target genes 28,553, each gene was averagely regulated by 1.98 union peaks). The number of overlapping associations between HiC and accessibility correlation is 4,815 and the number of overlapping target genes is 563. This will serve as a prior network to further integrate with our paired ATAC-seq and RNA-seq. The associations among union peaks and target genes are summarized in Table S2.

TF-RE-TG Triplet Inference Modeling

We developed a statistical model to integrate ATAC-seq and RNA-seq data and to infer, from the observed expression and accessibility data in time-course cellular context, how each RE interacts with relevant TFs to affect the expression of its TGs. We started with an assembled union peak list called from ATAC-seq across all time samples as our REs in the following analysis. Next, we identified the upstream TFs and downstream genes for a RE, treated each TF-RE-TG triplet as the basic regulatory unit (Figure 2B), ranked them by integrating genomic features, and extracted the significant regulatory relations by maximizing the joint probability $P(TF, RE, TG)=P(TG|RE)P(RE|TF)P(TF)$.

Our model is based on several assumptions. RE openness was defined as the fold enrichment of the read starts in this region versus the read starts in a 1Mb background window. Each TF was described by its motif binding score to the RE and its expression level from the FPKM value of RNA-seq experiments. The regulation of a TF on a RE was quantified as the regression of TF expression to RE openness. Finally, regulation from a RE to a TG correlates the RE openness and TG expression plus some constraints to identify direct regulation. In the time course differentiation dataset, we observed three major stages by the global dynamics of chromatin state. Stage 1 has one sample (D0), Stage 2 has three samples (D7, D14, D21) and Stage 3 has two samples (D43 and H9KC). The expression level for each stage was taken as the maximum when there were multiple samples. We took this global dynamics into account and proposed a model to infer the RE's openness change from Stage 1 to 2 (surface ectoderm initiation) and Stage 2 to 3 (keratinocyte maturation). In this way, we reconstructed the TF-RE-TG network by integrating dynamic change of driver TFs and downstream TGs with REs (union peak elements). Our computation has four major steps (details illustrated in Figure S3):

Step 1: Predicting RE (union peak from ATAC-seq)'s target gene.

We used HiC and cross-tissue accessibility correlation of DNase-seq signal to assign target genes to the union peaks. The details are in the above section.

Step 2: Collecting genomic features from chromatin state and expression level.

The experimental data includes two levels of experiments. One is the chromatin level data (ATAC-seq data). The other is the transcription level data (RNA-seq data). ATAC-seq data will indicate the RE openness, i.e., the chromatin state of cis-regulatory element (union peaks). From the paired data, we collected genomic features includes TF binding derived from motif occurrence, conservation for the REs, openness change of REs, TF expression change, TF's promoter openness change, TG expression change, and promoter openness change.

Our aim is to model how a TF will regulate a TG via REs in multiple conditions with matched ATAC-seq and RNA-seq data. Give one TF-RE-TG triplet, we started by enumerating all the REs $e_0, e_1, e_2, \ldots e_m$ around TG to serve as cis-regulatory elements. Those REs were grouped into two classes by their distance to the TG. Those include the promoter of TG (denoted by e_0) and *m* potential enhancers ($e_1, e_2, \ldots e_m$). We assumed that TF may regulate this TG's expression by physically binding to those cis-regulatory elements. We then used an unsupervised model to rank the TF-RE-TG triplets and infer if TG is likely to be regulated by this TF via REs. Our model provides a resource to extract regulatory relations among REs, TFs, and TGs.

We assumed that target genes regulated by TF via REs tend to differ in multiple ways from target genes that were not regulated by this TF via REs. The evidence we considered are as follows.

1. TF binds to TG's cis-elements. We scanned TG's cis-regulatory for all positions with substantial similarity to TF's sequence motif or position weight matrix (PWM). Considering the proximity of those binding sites (we used the proximity of binding sites to its center for RE), for RE $\mathrm{e_i}$, we derived a binding strength {$\mathrm{B_i}$, i=0,1,…m}, here $\mathrm{B_i}$ denoted the strength that TF regulating TG via RE $\mathrm{e_i}$.

We assumed that the association strength of TF-TG pair *r* via RE e_i is a weighted sum of intensities of all of the motif binding sites:

$$\mathrm{B}_i = \sum_{l=1}^{k_i} \mathrm{PWM\ Score}_{il} e^{-\frac{d_{il}}{d_{i0}}}$$

where $\mathrm{PWM\ Score}_{il}$ is the intensity of the l-th binding site in RE e_i for TF-TG pair r. For enhancer, d_{il} is the distance between the center of RE and the l-th binding site. d_{i0} is a constant. We set $d_{i0} = 500bp\ \mathrm{i} \neq 0$ for union peaks in our implement. This models the effect of a binding intensity decays exponentially when d_{il} increases where the speed depends on d_{il}.

2. Evolutionary sequence conservation of motif binding sites. The true binding sites are more likely to show evolutionary sequence conservation. By utilizing the sequence comparison results with other species, we derived a conservation score $\{C_i, i=0,1,\ldots,m\}$, here C_i denotes the average sequence conservation of the motif binding sites for TF regulating TG in RE e_i.

Conservation information is useful to predict potential binding site in cis-regulatory element. The average PhastCons conservation score is used on the motif matched region across the placental mammals on the 44-way multiple alignment. $\mathrm{Cons.Score}_i$ is then calculated for each TF-TG pair r by averaging all the k_i binding sites in RE e_i.

$C_i = \sum_{l=1}^{k_i} \frac{1}{k_i} \mathrm{Cons.Score}_{il}$

3. REs' openness. The REs are more likely to be in open region by DNA accessibility data if TF utilizes this RE to regulate TG (i.e., RE is active for TF-TG regulation). The openness data $\{O_i, i=0,1,\ldots m\}$, for RE e_i is obtained from ATAC-seq data.

As mentioned before, ATAC-seq will measure the count of reads in a given cis-regulatory element (promoter or enhancer). We can quantify the openness for the RE e_i by a simple fold change score, which computes the enrichment of read counts in e_i by comparing with a large background region. Briefly, let N_i be the number of reads in RE e_i of length L_i and G_i that in the W background window (1Mb in our case) around this RE. The openness of RE e_i can be defined as

$$O_i = \frac{N_i/L_i}{G_i/W}$$

4. The expression of TF and TG. The gene expression levels of TF and TG, X_{TF}, X_{TG}, in this condition or cell type can be measured by RNA-seq.

For the RNA-seq, the expression level is quantified by FKPM value which is the fold change of total number of reads mapped to TG (M_i) with gene length L_{TG} to the totally mapped reads N in the experiment,

$$\mathrm{X}_{TG} = \frac{\mathrm{M}_i/\mathrm{L}_{TG}}{N/10^9}$$

$$\mathrm{X}_{TF} = \frac{\mathrm{M}_i/\mathrm{L}_{TF}}{N/10^9}$$

Step 3: Integrating genomic features and rank TF-RE-TG triplets

For each RE e_i of TF-TG pair *r*, we introduced random variable Z_i (i=1，2，…, m) to denote the regulation of TF on TG via RE e_i,

$$Z_i = \begin{cases} 1 & \text{if TF utilizes RE } e_i \text{ to regulate TG} \\ 0 & \text{Otherwise} \end{cases}$$

We then modeled the probabilities of observing the triplet given motif binding strength B_i, binding sites conservation C_i, the openness of the element O_i, TF expression level X_{TF}, TF motif enrichment E_{TF}, TG expression level X_{TG}, and TG's promoter openness O_{TG}. For each potential TF-TG regulation via RE e_i, we calculated a conditional probability $P(Z_i = 1|B_i, C_i, O_i, \mathrm{X}_{TF}, \mathrm{E}_{TF}, \mathrm{X}_{\mathrm{TG}}, O_{\mathrm{TG}})$. Accurately modelling the conditional probability needs to introduce parameters and the parameter inference requires large amount of data (Duren et al., 2017). Here we fully took into account the three-stage global chromatin dynamics from the short time series data and utilized the following simplified inference procedure.

Our aim is to model how a TF will regulate a TG via REs with multiple conditions measurement in matched ATAC-seq and RNA-seq data. Give one TF-RE-TG triplet, we assumed that TF may regulate this TG's expression by REs. Since we can collect the genomic features for TF-union peak-TG triplet in multiple conditions (time points), this allows us to calculate the fold change score for O_i, $\mathrm{X}_{\mathrm{TG}}, O_{\mathrm{TG}}$ as,

$$F_{O_i} = O_i^{t+1}/O_i^t$$
$$F_{\mathrm{X}_{TG}} = {\mathrm{X}_{TG}}^{t+1}/{\mathrm{X}_{\mathrm{TG}}}^t$$
$$F_{\mathrm{O}_{TG}} = {\mathrm{O}_{TG}}^{t+1}/{\mathrm{O}_{\mathrm{TG}}}^t$$

TF's feature is its driver score defined as

$$F_{\mathrm{X}_{TF}} = \sqrt[2]{{\mathrm{X}_{TF}}^t \times {\mathrm{E}_{TF}}^t}$$

Here $O_i^{t+1}, {\mathrm{X}_{TF}}^{t+1}, {\mathrm{E}_{TF}}^{t+1}, {X_{TG}}^{t+1}, {\mathrm{O}_{TG}}^{t+1}$ are the data observed in time point *t+1* and $O_i^t, {\mathrm{X}_{TF}}^t, {\mathrm{E}_{TF}}^t, {\mathrm{X}_{TG}}^t, {\mathrm{O}_{TG}}^t$ are the data observed in time point *t*.

Then we had seven features $B_i, C_i, F_{O_i}, F_{\mathrm{X}_{TF}}, F_{\mathrm{X}_{\mathrm{TG}}}, F_{\mathrm{O}_{\mathrm{TG}}}$ for each TF-RE-TG triplet. These features are not independent. We used the following feature transformation to model the regulation,

$$B_i F_{O_{TG}}, B_i F_{\mathrm{X}_{TF}}, \mathrm{C}_i F_{\mathrm{X}_{TF}}, F_{\mathrm{X}_{TG}}, F_{\mathrm{O}_{TG}}$$

$B_i F_{O_{TG}}$ indicates direct regulation by motif binding and open RE. $B_i F_{\mathrm{X}_{TF}}$, $\mathrm{C}_i F_{\mathrm{X}_{TF}}$ includes the indirect regulation that TF is highly expressed and may regulate other TFs to binding to RE of TG. Together with $F_{\mathrm{X}_{\mathrm{TG}}}$ and $F_{\mathrm{O}_{TG}}$, the combination of those transformed features indicates the differential regulatory patterns for TF-RE-TG pair and we can predict TF-RE-TG regulations.

With those transformed features, we assumed normal distribution for each feature across all the triplets and those transformed features are independent to each other. Using Fisher's method, we can combine the features into one single score S,

$$S = -2\sum_{i=1}^{K} \ln(P_i)$$

where P_i is the p-value for the i-th hypothesis test to assess the significance level of feature *i*. When the p-values tend to be small, the test statistic *S* will be large, which

suggests that TF-RE-TG regulation by combining information from every test. When all the partial tests are independent with their corresponding test statistics P_i, all these p-values can be combined into a joint test whether there is a global effect. S follows a chi-squared distribution with 2K degrees of freedom, from which a p-value for the global hypothesis can be easily obtained. K is the number of features being combined (K=5 for our transformed features). As a result, all the triplets can be ranked by score *S* and each score associates to a p-value.

Fisher's combination assumes equal weights for all the features, Stouffer's Z-score method allows us to introduce prior to weight the transformed features.

Step 4: Controlling FDR and extracting significant TF-RE-TG triplets into network

We performed the Benjamini–Hochberg procedure to control the false discovery rate (FDR) for multiple test (Benjamini and Hochberg, 1995). By taking a cutoff FDR<0.01, we predicted a set of TF-RE-TG triplet (Table S3). Pooling all the triplets together, we then have a TF-RE-TG network, where TF and TG are nodes and RE is the edge. This network contains a large number of TF-TG relations (i.e., TF and TG connected through a RE) and can be extracted from triplets. The TF-TG networks are visualized by Cytoscape (Shannon et al., 2003). Our network models the matched expression and accessibility dynamics across different stages in differentiation and holds the great promise to recover a significant portion of the information in the missing data on binding locations and chromatin states and to achieve accurate inference of gene regulatory relations (Duren et al., 2017; Wang et al., 2016).

REFERENCE

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. J Roy Stat Soc B Met *57*, 289-300.

Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., and Greenleaf, W.J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods *10*, 1213-1218.

Calo, E., Flynn, R.A., Martin, L., Spitale, R.C., Chang, H.Y., and Wysocka, J. (2015). RNA helicase DDX21 coordinates transcription and ribosomal RNA processing. Nature *518*, 249-253.

Consortium, E.P. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57-74.

Corces, M.R., Trevino, A.E., Hamilton, E.G., Greenside, P.G., Sinnott-Armstrong, N.A., Vesuna, S., Satpathy, A.T., Rubin, A.J., Montine, K.S., Wu, B.*, et al.* (2017). An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. Nat Methods *14*, 959-962.

Duren, Z., Chen, X., Jiang, R., Wang, Y., and Wong, W.H. (2017). Modeling gene regulation from paired expression and chromatin accessibility data. Proc Natl Acad Sci U S A *114*, E4914-E4923.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell *38*, 576-589.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Methods *9*, 357-359.

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.

Macarthur, C.C., Fontes, A., Ravinder, N., Kuninger, D., Kaur, J., Bailey, M., Taliana, A., Vemuri, M.C., and Lieu, P.T. (2012). Generation of human-induced pluripotent stem cells by a nonintegrating RNA Sendai virus vector in feeder-free or xeno-free conditions. Stem Cells Int *2012*, 564612.

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.

Qu, Y., Zhou, B., Yang, W., Han, B., Yu-Rice, Y., Gao, B., Johnson, J., Svendsen, C.N., Freeman, M.R., Giuliano, A.E.*, et al.* (2016). Transcriptome and proteome characterization of surface ectoderm cells differentiated from human iPSCs. Sci Rep *6*, 32007.

Rao, S.S., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S.*, et al.* (2014). A 3D map of the human

genome at kilobase resolution reveals principles of chromatin looping. Cell *159*, 1665-1680.

Sebastiano, V., Zhen, H.H., Haddad, B., Bashkirova, E., Melo, S.P., Wang, P., Leung, T.L., Siprashvili, Z., Tichy, A., Li, J.*, et al.* (2014). Human COL7A1-corrected induced pluripotent stem cells for the treatment of recessive dystrophic epidermolysis bullosa. Science translational medicine *6*, 264ra163.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res *13*, 2498-2504.

Sherwood, R.I., Hashimoto, T., O'Donnell, C.W., Lewis, S., Barkal, A.A., van Hoff, J.P., Karun, V., Jaakkola, T., and Gifford, D.K. (2014). Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. Nat Biotechnol *32*, 171-178.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S.*, et al.* (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A *102*, 15545-15550.

Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B.*, et al.* (2012). The accessible chromatin landscape of the human genome. Nature *489*, 75-82.

Wang, Y., Jiang, R., and Wong, W.H. (2016). Modeling the causal regulatory network by integrating chromatin accessibility and transcriptome data. Natl Sci Rev *3*, 240-251.

Zamanighomi, M., Lin, Z., Wang, Y., Jiang, R., and Wong, W.H. (2017). Predicting transcription factor binding motifs from DNA-binding domains, chromatin accessibility and gene expression data. Nucleic Acids Res *45*, 5666-5677.

Zarnegar, B.J., Webster, D.E., Lopez-Pajares, V., Vander Stoep Hunt, B., Qu, K., Yan, K.J., Berk, D.R., Sen, G.L., and Khavari, P.A. (2012). Genomic profiling of a human organotypic model of AEC syndrome reveals ZNF750 as an essential downstream target of mutant TP63. Am J Hum Genet *91*, 435-443.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W.*, et al.* (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol *9*, R137.