Additional file 3: Supplementary Methods and Supplementary References

for

Single-cell transcriptomics reveals spatial and temporal turnover of keratinocyte differentiation regulators

Alex Finnegan[1,#], Raymond J. Cho[2,#], Alan Luu[1], Paymann Harirchian[3], Jerry Lee[3], Jeffrey B. Cheng[3,*,†], Jun S. Song[1,*,†]

[1]Department of Physics, Carl R. Woese Institute of Genomic Biology, University of Illinois at Urbana-Champaign, Champaign, IL, USA

[2]Department of Dermatology, University of California, San Francisco, San Francisco, CA, USA

[3]Department of Dermatology, University of California, San Francisco and Veterans Affairs Medical Center, San Francisco, CA, USA

[#] These authors contributed equally to this work
[*] These senior authors contributed equally to this work
[†] Correspondence: Jeffrey.Cheng@ucsf.edu, songj@illinois.edu

SUPPLEMENTARY METHODS

Keratinocyte isolation and primary culture

Primary human keratinocytes were isolated from neonatal foreskin surgical tissue discards obtained with written informed consent using protocols approved by the UCSF institutional review board. Following the method of Lowdon *et al.* [1], skin was incubated overnight at 4 °C in 25 U/mL dispase solution (Corning Life Sciences, Corning, NY). Next, epidermis was mechanically separated from the dermis and incubated in 0.05% trypsin for 15 minutes at 37 °C. Dissociated epidermal cells were filtered with a 100 µm nylon cell strainer (Corning Life Sciences) and then cultured in keratinocyte growth media (KGM; Medium 154CF supplemented with 0.07 mM $CaCl_2$ and Human Keratinocyte Growth Supplement; Life Technologies, Waltham, MA).

Expression level of Candidate TFs in cell culture

To assess concordance between Candidate TF's differentiation-promoting scores calculated from epidermal scRNA-seq data (Results: "Knockdown of ETV4 and ZBED2, predicted promoters of the BK state, induces differentiation"; Supplementary Methods: "Prioritization of knockdown targets"; Figure S4) and changes in bulk RNA expression of these TFs during in vitro differentiation, we generated RNA-seq expression for primary cultured human keratinocytes cultured in basal/proliferating (0.07 mM Ca) or high calcium-induced differentiation (1.2 mM Ca) conditions. Negative control siRNA treated keratinocytes were used a proxy for normal cultured keratinocytes. Keratinocytes were initially seeded at a density of 100,000 and 150,000 cells in 12-well plates using KGM with 0.07 mM Calcium. Within 30 minutes of plating, 10 nM of either ON-TARGETplus Non-targeting Control siRNA #1 or 2 mixed with 2.5 uL/well of Hiperfect transfection reagent was added. At ~48 hours post-transfection, subconfluent 100,000 cell wells were harvested using 0.5 mL TRIzol reagent (Life Technologies) for RNA extraction as per manufacturer's protocol. At ~48 hours post-transfection, the 150,000 cell wells had reached confluency and the media was replaced with 1 mL fresh KGM with 1.2 mM Calcium. After 24 hours of exposure to high 1.2 mM calcium, the confluent cells were also harvested using 0.5 mL TRIzol reagent and RNA-seq was performed. RNA-seq library preparation was performed using KAPA Biosystems Stranded RNA-Seq Kits and RiboErase HMR (Roche, Pleasanton, CA) with 300-1000 ng of total RNA. To minimize batch effects, technical duplicate libraries were generated for each sample. Ribosomal RNA (rRNA) was depleted by hybridization of complementary DNA oligonucleotides plus treatment with RNase H and DNase to remove rRNA duplexed to DNA and original DNA oligonucleotides, respectively. RNA fragmentation was conducted using heat and magnesium. Using random primers, first strand complementary DNA (cDNA) synthesis was conducted followed by second strand synthesis and A-tailing was added to the 3' ends using dAMP. Fragments were amplified using appropriate adapter sequences via ligation-mediated PCR. Then the libraries were quantitated with either Quant-iT™ dsDNA or Qubit™ dsDNA HS assay kits (Life Technologies). Quality assessment was performed using the LabChip GX Touch HT microfluidics platform (Perkin Elmer, Waltham, MA). 2 X 150 base pair sequencing on a NovaSeq 6000 instrument was performed on libraries with a PhiX Control v3 (Illumina, San Diego, CA). The RNA-Seq by Expectation Maximization (RSEM) algorithm [2] was used to quantify gene expression in terms of FPKM for technical replicates in both biological conditions. Change in expression between the differentiation-promoting (1.2 mM Ca) and non-differentiation-promoting (0.07 mM Ca) conditions was quantified as the $\log_2$ ratio of gene expression averaged over technical replicates (Figure S4).

Identification of keratinocyte stages

To identify clusters of functionally distinct keratinocytes, we used our previously published imputation, dimensionality reduction and spectral clustering techniques [3]. Imputation mitigates the effect of scRNA-seq dropout by sharing expression information among similar cells. We used the MAGIC imputation algorithm (version 0.0) [4] with cell similarity matrix obtained from ZINB-WaVE dimensionality reduction [5]. In this method, ZINB-WaVE fits a model predicting the mean expression and probability of dropout for each gene in each cell from cell-level covariates (percent mitochondrial UMI, total UMI and batch) and from 20 latent cell-level covariates that are learned from the data, yielding a low dimensional, bias-corrected representation of each cell. The resulting 92889×20 matrix of low dimensional cell representations were used to calculate cell-cell distances needed for constructing the affinity matrix for MAGIC's diffusion-based imputation. Our application of MAGIC used default adaptive distance parameters *ka=10, k=30* and diffusion time *t=10* as chosen previously based on recovery of simulated dropout events [3] (Supplementary Methods: Calculation of gene correlations). Finally, because the imputed expression values output by MAGIC are not normalized to a common cell library size, we renormalized MAGIC output for each cell to units of imputed UMI per 10,000. These imputed and renormalized expression values were used in downstream cell clustering.

Analysis downstream of imputation focused on 22,338 foreskin keratinocytes, identified based on anatomic location of samples and membership in expression-based clusters identified as keratocytes in Cheng *et al.* [3]. To Identify differentiation stages within these cells, we performed principal component analysis (PCA) representing each cell by the $\log_2$-transformed (with pseudocount 1) imputed expression of a set of genes robustly expressed in the full data set (at least 5 UMI in at least 100 of the 92,889 cells passing quality control). The first 20 PCs sufficed to capture nearly all the variation in our imputed data (Figure S2), and we clustered the foreskin keratinocytes in this 20 dimensional space using an adaptive distance implementation [3] of the k-means-based approximate spectral clustering (KASP) algorithm [6]. KASP clustering with adaptive parameters *ka=10, k=30* was used to identify 8 keratinocyte clusters which were then ordered and named stages 1-8 based on mean cluster expression of known marker genes (Figure S3).

Calculation of gene correlations

Our construction of regulatory networks used co-expression, measured by Pearson correlation, as a proxy for gene-TF regulatory relationships. Pearson correlations were calculated between log-transformed imputed counts per million (cpm) using

$$\log \text{imputed cpm} = \log_{10}(100x + 1)$$

Where x denotes the output of our imputation algorithm (units of imputed counts per 10,000). We took two steps to prevent introduction of large-magnitude, spurious correlations that could lead to false positive regulatory relationships. First, we performed stage-wise filtering of cells with outlier expression. Second, we reduced MAGIC's diffusion time parameter *t* to prevent over-smoothing of imputed expression values used to calculate correlations.

Estimates of Pearson correlation are strongly affected by outliers. In our study, these outliers were removed by filtering out cells lowly expressing genes expressed by the bulk of keratinocytes. Specifically, for each foreskin keratinocyte, we calculated the sum of imputed expression across genes expressed ($\geq$ 1 UMI raw data) in at least 1% of all keratinocytes. Stage-wise distributions of these summed expression values identified outlier cells in each stage (Figure S10); by removing cells in the lowest percentiles (see Table S7 for stage-wise thresholds), we mitigated a skew in the distribution of gene correlations (Figure S11, top row).

MAGIC's diffusion-based imputation algorithm mitigates dropout effects by replacing raw expression values with a weighted average of expression values of cells with similar low dimensional representation. The extent of local averaging increases with diffusion time *t*, and large *t* can over-smooth expression values, thereby averaging out true biological variation and thus strengthening spurious correlations. We observed this effect in the broadening of distributions of Pearson correlations calculated using *t=10*, compared to the same distribution calculated using *t=4* (Figure S11, middle row). The diffusion time parameter *t=10* was previously selected for this dataset based on recovery of simulated dropout events [3], a useful metric for assuring that key expression values are not lost at the single cell level. Recognizing that the optimal value of the *t* parameter may depend on the type of downstream analysis and wishing to reduce spurious correlations, we used the expression values obtained from our imputation pipeline with *t=4* and filtered for outliers using the above summed expression criteria (Figure S11, bottom row and Table S7) as imputed expression values in all analyses downstream of keratinocyte stage identification.

Clustering transcription factor expression trajectories and super-enhancer differential motif enrichment.

We performed hierarchical clustering of stage-wise mean expression values to identify dynamic TFs showing similar differentiation trajectories. Keratinocyte TFs (Methods: Identification of keratinocyte-specific genes and transcription factors) were filtered to include only those whose maximum value of mean imputed expression across stages 1-7 was at least 1.75-fold higher than the minimum across the same set; to discard lowly expressed TFs, the minimum was set to 5 counts per million (cpm) when it was less than this threshold. The stage-wise mean expression values of these dynamic TFs were converted to log cpm with pseudocount 1 and then clustered using Pearson correlation distance and average linkage.

To relate regulatory activity measured by TF expression to regulatory activity measured by abundance of functional TF binding sites, we performed differential motif enrichment analysis in super-enhancers (SEs) characterizing BK vs. DK states. We obtained hg19 coordinates of BK and DK SEs from the authors of Klein *et al.* [7] (referred to as NHEK-P SE and NHEK-D SE in that publication) and used Bedtools [8] to define BK-specific SEs not overlapping any DK SEs and DK-specific SEs not overlapping BK SEs. Next, we collected position-specific scoring matrices associated with our Keratinocyte TFs from the JASPAR [9], TRANSFAC [10], and Hocomoco [11] databases, as well as those published in Jolma *et al.* [12]. FIMO (version 5.0.1) [13] was used to scan BK- and DK-specific SEs with each motif using default parameters plus the max-strand option and a 0th order Markov background model given by the background frequencies of single nucleotides in the union of BK- and DK-specific SEs. This produced a table of motif hit counts for each TF motif in each BK- or DK-specific SE. Motifs were tested for differential enrichment of hit counts per unit length between BK-specific and DK-specific SEs using the Mann-Whitney U test followed by Benjamini-Hochberg multiple hypothesis correction. We accepted motifs with adjusted p-values less than 10^{-3} as differentially enriched and used the asymptotic normality of the U statistic under the null hypothesis to measure the magnitude and direction of enrichment as the z-score of the U statistic. When motifs from multiple databases yielded differential enrichment for the same TF or TF dimer, we selected the strongest motif, by calling the length of the shortest candidate motif ℓ and then ranking the motifs by the sum of Kullback-Leibler divergence from the 0th-order background across ℓ most divergent bases. Finally, some TFs, such as JUN, FOS and FOSL1, were associated with several different motifs either as monomers or components of heterodimers; in the case of differential enrichment for these functionally distinct motifs, we assigned to the TF the mean of the U statistic z-scores for these enriched motifs.

Prioritization of knockdown targets

We prioritized Candidate Keratinocyte TFs according to log-fold change, during differentiation, of putative targets selected from the set of Keratinocyte Genes (Methods: Identification of keratinocyte specific genes and transcription factors). To identify regulatory targets, we first partitioned Candidate Keratinocyte TFs, denoted here by the set T, according to their pattern of differential expression between the BK and DK states (Methods: Differential expression). The set $T^{(BK)}$ contained TFs differentially upregulated in the BK state; the set $T^{(DK)}$ contained TFs differentially upregulated in the DK state; and, the set $T^{(nDE)}$ contained TFs not differentially expressed between the two states.

Next, we considered as potential targets the set *G* of Keratinocyte Genes differentially expressed between the BK and DK state. Activating and inhibiting relationships between elements of *T* and *G* were assigned based on the strength of correlation calculated across cells specific to each partition of T. More precisely, for TFs in the partitions $T^{(BK)}, T^{(DK)}$ and $T^{(nDE)}$, we computed correlations across cells in stages 1-4, 4-7 and 1-7, respectively, leading to the Pearson correlation coefficients $r_{i,j}^{(BK)}$, $r_{i,j}^{(DK)}$, $r_{i,j}^{(nDE)}$ between the log-transformed imputed expression of TF i and gene j (Supplementary Methods: Calculation of gene correlations). Next, for each partition $k \in \{BK, DK, nDE\}$, we constructed thresholds, $r_{+}^{(k)}$ and $r_{-}^{(k)}$, on correlation strength using

$$r_{+}^{(k)} = \max(0\,, \text{percentile}(95\,, \{r_{i,j}^{(k)} : i \in T \text{ and } j \in (G \cup T) - \{i\}\}\ )\,)$$

$$r_{-}^{(k)} = \min(0\,, \text{percentile}(5\,, \{r_{i,j}^{(k)} : i \in T \text{ and } j \in (G \cup T) - \{i\}\}\ )\,)$$

where $\text{percentile}(x, A)$ denotes the x^{th} percentile of the set *A*. Finally, each TF i in each partition $T^{(k)}$ was assigned a differentiation-promoting score, $\text{score}(i)$, by summing $\log_2$ expression fold-changes between DK and BK states for all elements of *G* passing the thresholds $r_{+}^{(k)}$ and $r_{-}^{(k)}$:

$$\text{score}(i) = \sum_{j \in G - \{i\}} \text{sign}\left(r_{i,j}^{(k)}\right) L(j) \left(I\left(\, r_{i,j}^{(k)} \geq r_{+}^{(k)}\right) + I\left(\, r_{i,j}^{(k)} \leq r_{-}^{(k)}\right)\right).$$

In this equation, *I* denotes the indicator function, $L(j)$ is the $\log_2$ fold-change of expression for gene j between the DK and BK states (Methods: Differential Expression) and $\text{sign}\left(r_{i,j}^{(k)}\right)$ accounts for the activating or inhibiting effect of TF i on gene j. Figure S4 shows the resulting differentiation-promoting scores along with $\log_2$-fold expression changes between imputed single-cell data averaged over the DK and BK states and between keratinocytes cultured in high (1.2 mM Ca) and low (0.07 mM Ca) calcium conditions. RNAi knockdown experiments tested the basal promoting function of TFs with the six most negative differentiation-promoting scores, after removing HOXA1 which was lowly expressed (less than 5 FPKM) in the keratinocytes cultured in *in-vitro* basal/proliferative conditions.

Regulatory network construction

Regulatory networks were constructed for the BK and DK states as follows. For the BK state, we considered Keratinocyte TFs with motifs enriched in BK SEs compared to DK SEs as putative BK regulators. Similarly, we took Keratinocyte Genes not downregulated in the BK state compared to the DK state as putative BK targets. Signed similarity scores $S_{i,j}$ between genes i and j were calculated using the soft thresholding method of [14]:

$$S_{i,j} \ = \operatorname{sign}\left(r_{i,j}^{(BK)}\right)\left|r_{i,j}^{(BK)}\right|^{\beta}$$

where $r_{i,j}^{(BK)}$ denotes the Pearson correlation of log-transformed imputed expression for genes *i* and *j* across single cells in stages 1-4, and $\beta = 4$. Putative BK regulators were organized by hierarchical agglomerative clustering using the distance

$$d(i,j) = 1 - S_{i,j}$$

and average linkage. TF modules were called using the “inconsistent” criteria in SciPy’s fcluster function with parameters depth=2 and threshold=0.75 [15]. Putative BK target genes were also organized by hierarchical agglomerative clustering. Each target gene was represented by a vector of similarity scores between the gene and all putative BK regulators. These vectors were clustered using Euclidean distance and average linkage. Like TF modules, gene modules were called using the “inconsistent” criteria of the fcluster function with parameters depth=4 and threshold=2.15 (Figure S6A). We identified regulatory relationships between pairs of identified Gene and TF Modules by applying thresholding to the distribution of magnitude of mean similarity scores between all pairs:

$$\left\{ \left| \underset{i \in A\,, j \in B}{\text{mean}} S_{i,j} \right| : \; A \in \text{TF Modules}, \text{B} \in \text{Gene Modules} \right\},$$

(Figure S6B). TF-Gene Module pairs with mean signed similarly score magnitude exceeding the threshold of Figure S6C were identified as having activating or inhibiting regulatory relationships (Figure S5D) and were the focus of further investigation.

The DK state network was constructed in the same manner as the BK state subject to the following changes: putative DK regulators were selected for motif enrichment in DK SEs compared to BK SEs; putative DK targets were Keratinocyte Genes not downregulated in the DK state compared to the BK state; calculation of Pearson correlations used single cells in stages 4-7; identification of TF modules used the fcluster function with parameters depth=2 and threshold=0.75; and identification of target gene modules used the fcluster function with parameters depth=16 and threshold=3.2 (Figure S9A-D).

Antioxidant analysis

Genes annotated for antioxidant function were downloaded from the AmiGO2 database (version 2.5.12) [16] and filtered to include only those genes expressed in more than 1% of all keratinocytes in scRNAseq data (Table S2). Genes with dynamic expression in foreskin keratinocytes ($\log_2$ fold-change between minimum and maximum stage-wise mean expression for stages 1-7 greater than 1, with the minimum set to 5 imputed cpm when it was less than this threshold) were selected for hierarchical agglomerative clustering. We clustered genes represented as vectors of $\log_2$ stage-wise mean imputed cpm with pseudocount 1 using Pearson correlation distance and average linkage.

To test the significance of size enrichment of the cluster showing peak expression in stages 1-3, we generated a null distribution of maximum cluster sizes using a permutation approach. For each of 10,000 iterations, we independently permuted the elements of each $\log_2$ stage-wise mean expression vector and repeated the hierarchical clustering procedure identifying four clusters. The *p*-value was calculated from the percentile of the observed cluster size in the distribution of simulated maximum cluster sizes.

BCC and SCC similarity analysis

To perform stage-wise similarity analysis of BCC and SCC, we first compiled a list of differentially expressed genes for each cancer type. Lists of upregulated and downregulated genes and their corresponding q-values for SCC were obtained from The Cancer Genome Atlas N [17] data file S5.1, which were generated by performing differential expression analysis on 16 samples of SCC tumor samples matched with adjacent normal tissue from five different sites: floor of mouth, larynx, oral cavity, tongue, and tongue/base of tongue. We constructed our own BCC gene list from the expression data obtained from Atwood *et al.* [18] by performing differential expression analysis on 9 Smoothened inhibitor resistant BCC samples vs. 8 normal skin samples using the *t*-test. All samples were unpaired and obtained from unspecified locations. The *t*-test *p*-values were converted into corresponding *q*-values using the R package "qvalue" version 2.2.2 [19]. Next, for each stage, using the results from the stage-specific differential expression analysis (Methods: Differential expression), we ranked the keratinocyte-specific genes in order of increasing logFC of expression, resulting in downregulated genes occupying lower ranks and upregulated genes occupying higher ranks. More formally, given the set of keratinocyte stage specific genes G and the function f such that for any $g \in G$, $f(g)$ is the logFC of expression of g compared to other stages, we ordered the elements of G as $\{g_1, g_2, \ldots, g_N\}$ such that $f(g_i) < f(g_j)$ if $i < j$ (rank i is considered lower than rank j if $i < j$) and N is the number of genes in the set. We then defined an indexing scheme for genes that displayed both stage-specific differential expression and differential regulation in cancer compared to normal tissue. Given set D consisting of genes downregulated in cancer $(q < 0.05)$, set U

composed of genes upregulated in cancer (q < 0.05), and set $H = (U \cup D) \cap G$, we define the function h that maps the elements of H to an index as follows:

$$h(g_i) = \begin{cases} i, & g_i \in D \\ N - i, & g_i \in U \end{cases}.$$

This definition of h ensures that genes with the direction of differential regulation in cancer matching the direction of stage-specific differential expression are assigned low indices. To compare the similarity of a cancer type to a particular differentiation stage, we then devised a statistic for measuring the association between stage-specific logFC of gene expression values and probability of differential expression in cancer. Namely, we used the one-sided Mann-Whitney U statistic testing the null hypothesis that a randomly selected element of $\{h(g_i);\ g_i \in H\}$ is greater than a randomly selected element of $\{1, 2, \ldots, N\}$. To visualize the similarity of keratinocyte stages to a cancer type, we plotted a kernel density estimate of the probability that keratinocyte stage-specific gene $g_i \in G$ at gene rank i according to stage-specific logFC of expression is downregulated in cancer $(g_i \in D)$ or upregulated in cancer $(g_i \in U)$ using a Gaussian kernel with radius 1000 (Figure 5A-B). That is, given the kernel function $K(x) = e^{-\frac{x}{2\sigma^2}}$ with $\sigma = 1000$ and indicator functions for downregulation and upregulation in cancer defined as

$$d(i) = \begin{cases} 1, & g_i \in D \\ 0, & g_i \notin D \end{cases}$$

and

$$u(i) = \begin{cases} 1, & g_i \in U \\ 0, & g_i \notin U \end{cases},$$

respectively, we defined the kernel density estimate of probability of downregulation in cancer at gene rank i as

$$f_d(i) = \frac{\sum_{x=a}^{b} K(i - x) d(i)}{\sum_{x=a}^{b} K(i - x)}$$

with $a = max(1, i - 3\sigma)$ and $b = min(N, i + 3\sigma)$. The kernel density estimate of probability of upregulation in cancer was defined similarly but with the indicator function $u(i)$. The degree of similarity was assessed by the skew of probability of upregulation in cancer towards higher ranks and skew of probability of downregulation in cancer towards lower ranks.

To implicate keratinocyte TFs associated with cancer, we first identified stages 2 and 4 to be candidate stages of cell origin for BCC and SCC. For each pairing between candidate stage and cancer, we identified the set of genes that are upregulated in both the cancer ($q < 0.1$) and stage (top 10% of genes with respect to logFC of expression) to be stage-specific cancer-associated genes. Keratinocyte TFs were then ranked by their correlation to the stage-specific cancer-associated genes, using correlation across single cells in the corresponding stage. The top 10 positively correlated TFs for each stage-cancer pair are displayed in Table S6.

SUPPLEMENTARY REFERENCES

1. Lowdon RF, Zhang B, Bilenky M, Mauro T, Li D, Gascard P, Sigaroudinia M, Farnham PJ, Bastian BC, Tlsty TD, et al: **Regulatory network decoded from epigenomes of surface ectoderm-derived cell types.** *Nat Commun* 2014, **5:**5442.
2. Li B, Dewey CN: **RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome.** *BMC Bioinformatics* 2011, **12:**323.
3. Cheng JB, Sedgewick AJ, Finnegan AI, Harirchian P, Lee J, Kwon S, Fassett MS, Golovato J, Gray M, Ghadially R, et al: **Transcriptional Programming of Normal and Inflamed Human Epidermis at Single-Cell Resolution.** *Cell Rep* 2018, **25:**871-883.
4. van Dijk D, Sharma R, Nainys J, Yim K, Kathail P, Carr AJ, Burdziak C, Moon KR, Chaffer CL, Pattabiraman D, et al: **Recovering Gene Interactions from Single-Cell Data Using Data Diffusion.** *Cell* 2018, **174:**716-729 e727.
5. Risso D, Perraudeau F, Gribkova S, Dudoit S, Vert JP: **A general and flexible method for signal extraction from single-cell RNA-seq data.** *Nat Commun* 2018, **9:**284.
6. Yan DH, Huang L, Jordan MI: **Fast Approximate Spectral Clustering.** *Kdd-09: 15th Acm Sigkdd Conference on Knowledge Discovery and Data Mining* 2009**:**907-915.
7. Klein RH, Lin Z, Hopkin AS, Gordon W, Tsoi LC, Liang Y, Gudjonsson JE, Andersen B: **GRHL3 binding and enhancers rearrange as epidermal keratinocytes transition between functional states.** *PLoS Genet* 2017, **13:**e1006745.
8. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features.** *Bioinformatics* 2010, **26:**841-842.
9. Mathelier A, Fornes O, Arenillas DJ, Chen CY, Denay G, Lee J, Shi W, Shyr C, Tan G, Worsley-Hunt R, et al: **JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles.** *Nucleic Acids Res* 2016, **44:**D110-115.

10. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, et al: **TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes.** *Nucleic Acids Res* 2006, **34:**D108-110.

11. Kulakovskiy IV, Vorontsov IE, Yevshin IS, Sharipov RN, Fedorova AD, Rumynskiy EI, Medvedeva YA, Magana-Mora A, Bajic VB, Papatsenko DA, et al: **HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis.** *Nucleic Acids Res* 2018, **46:**D252-D259.

12. Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, Morgunova E, Enge M, Taipale M, Wei G, et al: **DNA-binding specificities of human transcription factors.** *Cell* 2013, **152:**327-339.

13. Grant CE, Bailey TL, Noble WS: **FIMO: scanning for occurrences of a given motif.** *Bioinformatics* 2011, **27:**1017-1018.

14. Zhang B, Horvath S: **A general framework for weighted gene co-expression network analysis.** *Stat Appl Genet Mol Biol* 2005, **4:**Article17.

15. Jones E, Oliphant T, Peterson P, others: **SciPy: Open source scientific tools for Python.** 1.0.0 edition; 2001-.

16. Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S, Ami GOH, Web Presence Working G: **AmiGO: online access to ontology and annotation data.** *Bioinformatics* 2009, **25:**288-289.

17. Cancer Genome Atlas N: **Comprehensive genomic characterization of head and neck squamous cell carcinomas.** *Nature* 2015, **517:**576-582.

18. Atwood SX, Sarin KY, Whitson RJ, Li JR, Kim G, Rezaee M, Ally MS, Kim J, Yao C, Chang AL, et al: **Smoothened variants explain the majority of drug resistance in basal cell carcinoma.** *Cancer Cell* 2015, **27:**342-353.

19. Storey JD: **A direct approach to false discovery rates.** *Journal of the Royal Statistical Society Series B-Statistical Methodology* 2002, **64:**479-498.