ANANSE: An enhancer network-based computational approach for predicting key transcription factors in cell fate determination

Quan Xu[1], Georgios Georgiou[1,2], Gert Jan C. Veenstra[1], Huiqing Zhou[1,3,*] and Simon J. van Heeringen[1,*]

[1] Department of Molecular Developmental Biology, Faculty of Science, Radboud Institute for Molecular Life Sciences, Radboud University, 6525GA Nijmegen, The Netherlands.

[2] Current address: Viroclinics Biosciences BV, Rotterdam 3029 AK, The Netherlands.

[3] Department of Human Genetics, Radboud Institute for Molecular Life Sciences, Radboud University Medical Center, 6525GA Nijmegen, The Netherlands.

[*] Correspondence: s.vanheeringen@science.ru.nl; j.zhou@science.ru.nl; Tel +31-24-3616850

Abstract

Proper cell fate determination is largely orchestrated by complex gene regulatory networks centered around transcription factors. However, experimental elucidation of key transcription factors that drive cellular identity is often intractable. Here, we present ANANSE (ANalysis Algorithm for Networks Specified by Enhancers), a network-based method that exploits enhancer-encoded regulatory information to identify the key transcription factors in cell fate determination. As cell type-specific transcription factors predominantly bind to enhancers, we use regulatory networks based on enhancer properties to prioritize transcription factors. First, we predict genome-wide binding profiles of transcription factors in various cell types using enhancer activity and transcription factor binding motifs. Subsequently, applying these inferred binding profiles, we construct cell type-specific gene regulatory networks, and then predict key transcription factors controlling cell fate conversions using differential gene networks between cell types. Compared to other existing methods, ANANSE correctly predicts the highest number of transcription factors experimentally demonstrated to be sufficient for trans-differentiation. Finally, we apply ANANSE to define an atlas of key transcription factors in 18 normal human tissues. In conclusion, we present a ready-to-implement computational tool for efficient prediction of transcription factors in cell fate determination and to study transcription factor-mediated regulatory mechanisms. ANANSE is freely available at https://github.com/vanheeringen-lab/ANANSE.

Keywords: Enhancer, Transcription Factor, Gene Regulatory Network, Cell Fate Determination

Introduction

Every multicellular organism develops from a single cell. During this process, cells undergo division and differentiation, eventually forming a diversity of cell types that are organized into organs and tissues. How one cell develops into different cell types, a process known as cell fate determination, is critical during development. It has been shown that transcription factors (TFs) play key roles in cell fate determination (Davis et al., 1987; Jopling et al., 2011; Pang et al., 2011; Stadhouders et al., 2019; Takahashi et al., 2007; Vierbuchen et al., 2010). TFs bind to specific cis-regulatory sequences in the genome, including enhancers and promoters, and regulate expression of their target genes (Lambert et al., 2018; Vaquerizas et al., 2009). The interactions between TFs and their downstream target genes form gene regulatory networks (GRNs), controlling a dynamic cascade of cellular information processing (Davidson, 2010; Tegner and Bjorkegren, 2007). Cell fate determination is orchestrated by a series of TF regulatory events, largely by complex GRNs (Wilkinson et al., 2017). The key role of TFs and GRNs in cell fate determination is further corroborated by examples of cell fate conversions, often referred as cellular reprogramming (Iwafuchi-Doi and Zaret, 2016; Peñalosa-Ruiz et al., 2019). Cellular reprogramming includes generating induced pluripotent stem cells (iPSCs) from somatic cells, and trans-differentiation that converts one mature somatic cell type to another without undergoing an intermediate pluripotent state (Davis et al., 1987; Jopling et al., 2011; Pang et al., 2011; Stadhouders et al., 2019; Takahashi et al., 2007; Vierbuchen et al., 2010). These reprogramming processes are initiated by enforced expression of combinations of different key TFs, which is believed to alter the output of GRNs in the cell, namely gene expression and the epigenetic landscape (Buschbeck and Hake, 2017; Qu et al., 2018; Reik et al., 2001).

In the past, cellular reprogramming was often studied with experimental approaches by screening or testing individual genes or TFs based on the known function of these genes, which is labor-intensive and inefficient. Therefore, there is a need for better predictions of key TFs in cell fate determination that can instruct experimental cellular reprogramming approaches. Several computational methods for predicting key TFs in cell fate conversions have been reported. Many methods only consider GRNs based on gene expression levels and differences in gene expression between cell types, and identify candidate key TFs with high expression in the target cell types (Cahan et

al., 2014; D'Alessio et al., 2015; Hartmann et al., 2018; Heinaniemi et al., 2013; Lang et al., 2014; Morris et al., 2014; Roost et al., 2015). Mogrify uses not only gene expression but also GRNs constructed based on TF binding motifs in promoters to predict TFs that are capable of inducing conversions between cell types (Rackham et al., 2016). Although these methods can perform reasonably well and prioritize a short list of predicted key TFs for some cell fate conversions, e.g. SOX2 during fibroblasts to iPSCs, the current computational tools do not take enhancers into consideration. It has been well established that TFs that control tissue- and cell type-specific gene expression in cell fate determination and development often bind to enhancers (Andersson et al., 2014; Bulger and Groudine, 2011; Qu et al., 2018; Spitz and Furlong, 2012). Binding of tissue- and cell type-specific TFs largely to enhancers is also confirmed by a large number of genome-wide chromatin immunoprecipitation followed by sequencing analyses (ChIP-seq) (Davis et al., 2018; Valouev et al., 2008), e.g. TP63 in keratinocytes and ZIC2 in embryonic stem cells (Luo et al., 2015; Qu et al., 2018). Therefore, a computational method that uses enhancer properties and enhancer-based GRNs may improve the prediction of key TFs for cell fate determination. Furthermore, most current computational tools require comprehensive training or background data, such as cell/tissue expression data or pre-constructed networks. These datasets and the computational algorithms are not always publicly accessible, which prevents the general usage of these methods in studying transcriptional regulation or designing new trans-differentiation strategies.

In this study, we established an enhancer GRN-based method, ANalysis Algorithm for Networks Specified by Enhancers (ANANSE), that infers genome-wide regulatory programs and identifies key TFs for cell fate determination. First, we systematically analyzed the peaks of all available human TF ChIP-seq experiments (Chèneby et al., 2017), confirming that most cell type- and tissue-specific TFs predominantly bind to enhancers. We then predicted cell type-specific TF binding profiles with a model based on activities and sequence features of enhancers. Second, combining TF binding profiles and gene expression data, we built cell type-specific enhancer GRNs in each cell type or tissue. We used reference GRNs based on gene expression, Gene Ontology terms and TF-target gene interactions to evaluate the quality of our inferred networks. Third, we predicted the key TFs underlying cell fate conversions based on a differential network analysis. Compared with other reported prediction algorithms, ANANSE recovers the largest fraction of TFs that were validated by experimental trans-differentiation approaches. The results demonstrate that ANANSE accurately prioritizes TFs that drive cellular fate changes. Finally, to demonstrate the wide utility of ANANSE, we applied it to 18 human tissues and generated an atlas of key TFs underlying human tissue identity.

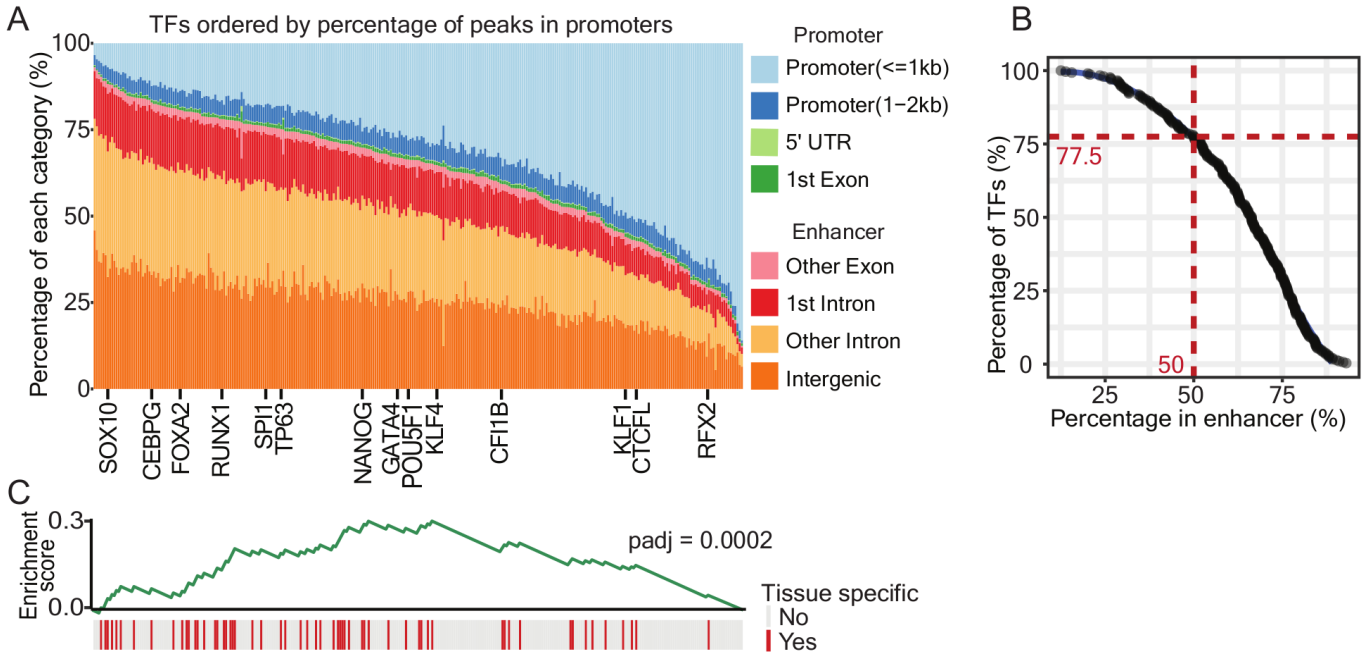


Figure 1. Tissue-specific TFs predominantly bind to enhancers. (A) Genomic location analysis of binding sites of 296 human TFs. The human genome was split into several categories: Promoter (<=1kb), Promoter (1-2kb), 5' UTR, and 1st Exon, Other Exons, 1st Intron, Other Introns, and Intergenic; these categories were further grouped into a promoter-proximal class (Promoter (<=1kb), Promoter (1-2kb), 5' UTR, and 1st Exon) and an enhancer class (Other Exons, 1st Intron, Other Introns, and Intergenic). The percentage of binding sites of each TF in different categories was calculated, and indicated with different colors. TFs were ordered by the percentage of binding sites within the promoter-proximal class. Several example TFs are marked at the bottom of the figure. **(B)** The percentage of TF binding sites in the enhancer class. 77.5% of TFs have 50% of its binding sites in the enhancer class of the genome. **(C)** Gene Set Enrichment Analysis (GSEA) on tissue-specific TFs and their enhancer binding. The red bars mark the tissue-specific TFs. The order of TFs is consistent with (A). Grey bars represent TFs that do not show tissue-specific gene expression. The GSEA enrichment score is represented by the green line (padj: 0.0002).

Results

Cell type-specific transcription factors predominantly bind to enhancers

To systematically examine TF binding patterns in the genome in relation to cell type specificity, we analyzed all available human TF binding sites in various cell types and tissues. We downloaded the binding sites of 296 TFs from the ReMap project that re-analyzed publicly available ChIP-seq data (Chèneby et al., 2017). To determine the genomic distribution of these binding sites, we divided the genome into different genomic categories according to human UCSC known gene annotation (Hsu et al., 2006), and assigned binding sites to these categories based on the locations of the binding sites (Figure 1A). We grouped these categories into two main classes: 1) a promoter-proximal class, containing promoter (<=2kb), 5' UTR and 1st exon peaks, and 2) an enhancer class, containing all exons except the first, the 1st intron, other introns and intergenic categories. The percentage of TF binding sites in each genomic category was calculated, and TFs were ordered according to their percentages in the promoter-proximal class (Figure 1A) (Supplementary Table S4). We found that different TFs have a binding preference in either the promoter range or in the enhancer range (Figure 1A). However, the majority of TFs (77.5%) mainly bind in cis-regulatory regions that are distal from the promoter (Figure 1B). While these binding sites are not necessarily all functional, we will refer to them as enhancers.

To further dissect the binding pattern of cell type- or tissue-specific TFs, we performed Gene Set Enrichment Analysis (GSEA) (Sergushichev, 2016) on TF expression in different tissues. We defined tissue-specific TFs based upon their expression levels in human tissues from the Human Protein Atlas (Uhlén et al., 2015) (Figure 1C). GSEA showed that TFs mostly binding to enhancers are indeed enriched for tissue-specific expression (adjusted p value = 2.0e-4) (Figure 1C) (Supplementary Table S4). For example, SOX10 is a critical TF during neural crest and peripheral nervous system development (Kim et al., 2014), while TP63 is a master regulator in epithelial development (Soares et al., 2019). Both of these tissue-specific TFs showed a very high percentage of enhancer-binding, 93% for SOX10 and 82% for TP63 (Figure 1A).

Taken together, our analysis of transcription binding sites revealed that most TFs, and specifically tissue-specific TFs, predominantly bind to enhancers that are not proximal to gene promoters. This demonstrates that it is essential to include the enhancer information in computational methods for predicting key TFs in cell fate determination.

ANANSE: an enhancer network-based method to identify transcription factors in cell fate changes

Starting from the premise that the majority of TFs predominantly bind to enhancer regions, we developed ANANSE, a network-based method that uses properties of enhancers and their GRNs to predict key TFs in cell fate determination. As trans-differentiation is an ideal model for studying cell fate conversions controlled by key TFs, we set out to use this model to validate our computational approach. In the following paragraphs a conceptual overview of ANANSE is provided. Subsequently we will validate each of the steps involved.

First, we inferred cell type-specific TF binding profiles for each cell type. The input data of ANANSE consists of genome-wide measurements of enhancer activity (defined below), transcription factor motifs and expression data. We inferred the TF binding probability based on a supervised model that integrates the enhancer activity combined with TF motif scores.

Second, we constructed cell type-specific GRNs based on the inferred TF binding probability and expression levels of predicted target genes (Figure 2B, 2C). The nodes in the network represent the TF or gene. The TF-gene interaction scores, represented by edges of the network, are calculated based on the predicted TF binding probability, the distance between the enhancer and the target gene, and expression of both TF and the target gene. By integrating these data, ANANSE determines the interaction score of each TF-gene pair.

Third, we used the GRN that is differential between two cell types to calculate the TF 'influence' score (Cahan et al., 2014; Rackham et al., 2016) (Figure 2D, 2E). This score represents a measure of importance of a TF in explaining transcriptional differences between the two cell types (Figure 2D, 2E). In this step, the difference in gene regulatory interactions represented by TF-gene interaction scores between the source and the target cell types is calculated.

The details of the algorithms are described in the following sections.

Transcription factor binding can be predicted by the motif score in combination with the enhancer activity

Sequence-specific TFs bind to their cognate DNA motifs in the genome and activate or repress their target genes. To infer the target genes of a TF, the genomic binding sites of this TF are informative. ChIP-seq has been broadly used to identify TF binding sites at genome-wide scale. However, it is unfeasible to perform ChIP-seq for every TF in all cell types, e.g. due to the availability and quality of the TF antibodies. Therefore, it would be highly beneficial to be able to predict binding sites of individual TFs in a given cell type.

Here, we used a conceptually simple general logistic regression model to predict TF binding sites based on the TF motif z-score and the enhancer activity. In this model, we used EP300 ChIP-seq peak summits as the enhancer location. The enhancer activity is based on the number of reads in regions centered at the enhancer summit. For each enhancer, we scanned for motifs in a 200bp region centered at the peak summit using GimmeMotifs (Bruse and Heeringen, 2018; van Heeringen and Veenstra, 2010) with a non-redundant database of 1,796 motifs (see Methods). The motif z-score was calculated by GimmeMotifs with the GC%-normalization option. The log-odds score based on the positional frequency matrix is normalized by using the mean and standard deviation of scores of random genomic regions. These random regions are selected to have a similar GC% as the input sequence.

To train and evaluate our model, we used data from the ENCODE-DREAM in vivo Transcription Factor Binding Site Prediction Challenge (https://www.synapse.org/ENCODE). The ChIP-seq data from this challenge has been consistently processed using the ENCODE pipelines (Davis et al., 2018).

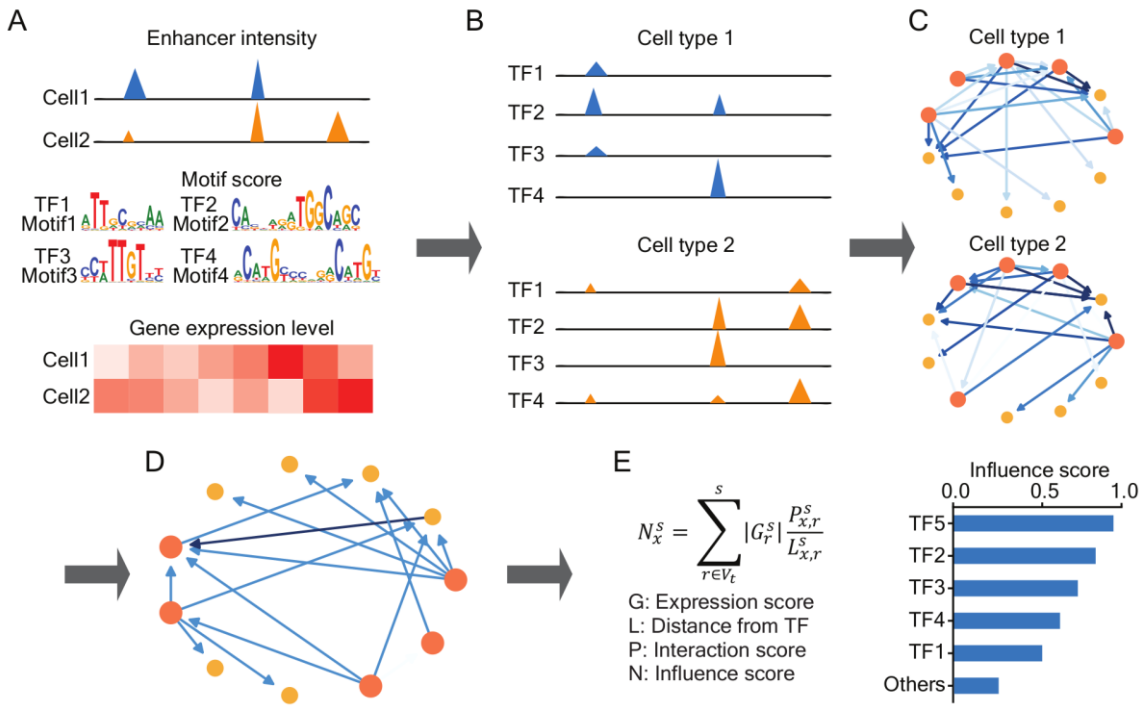


Figure 2. An overview of the workflow of the ANANSE method. (A) Data types used by ANANSE. These data include motif score of all TFs, gene expression data (e.g. RNA-seq) and enhancer data that can be obtained by ATAC-seq, EP300 ChIP-seq or H3K27ac ChIP-seq from each cell type. The blue and orange peaks represent enhancers in two cell types. The four sequence logos represent the motifs of four TFs. The heatmap represents gene expression level in two different cell types. **(B)**The TF binding profiles predicted from the enhancer data and TF motif scores in each cell type. The two GRNs below show cell type-specific TF binding profiles in two cell types (source and target cell types). **(C)** The cell type-specific GRN predicted based on TF-Gene binding and TF/Gene expression. The two networks show cell type-specific GRNs in two cell types. The orange circle represents a TF or a gene, and the size of the circle indicates the target gene number of the corresponding TF. The blue arrow indicates regulation between two TFs, and the color intensity represents the relative strength of the predicted regulatory interaction. **(D)** The differential GRN between the two cell types. In this step, the interaction specific for the target cell type is kept constant, and if the interaction score of the target cell type is higher than that of the source cell type, the interaction score is further used. **(E)** The barplot shows the ranked influence score of all TFs calculated from the differential GRN. The influence score is calculated based on the gene expression score, the distance from the TF to the gene in the predicted network, and the interaction score between TF and gene.

We chose EP300 ChIP-seq data from five cell types (GM12878, hESC, HeLa-S3, HepG2 and K562) to represent enhancer activities in our method. ChIP-seq peaks of 30 TFs in these five cell types were used as the training data (Supplementary Table S5). We created a reference of putative enhancers by combining the peaks from all five ENCODE cell types. We examined the locations of the TF peaks by overlapping with the union of the EP300-bound enhancers of the five cell types, and only the subset of peaks that overlapped with these enhancer regions was kept for further analysis (see Methods for details).

To test the prediction performance of our model, we established a leave-one-out cross-validation procedure. We excluded each TF in turn for testing, and trained the model on the remaining 29 TFs. Using the trained model, we evaluated whether binding of the test TF was correctly predicted in the five cell types. We evaluated the performance of the model using the AUC (Area Under Curve) of the ROC (Receiver Operating Characteristic) (Supplementary Table S5). The median ROC AUC of all 30 TFs was 0.85, which is significantly higher than expected by random chance (AUC = 0.5) (Figure 3A, Supplementary Figure S1). In addition, the regression model outperforms the baseline of motif z-score (Figure 3B, Supplementary Figure S1). Generally, ROC curves are appropriate when the observations are balanced between groups, however, they can be misleading in the case of unbalanced data (Saito and Rehmsmeier, 2015). Therefore, we calculated the PR AUC score for the binding prediction of all 30 TFs (Figure 3C) (Supplementary Table S5), which shows a similar pattern as the ROC AUC. The median PR AUC of the model (0.81) is significantly better than both the random situation (median 0.52) and the motif z-score alone (median 0.69).

An example of binding site predictions of several TFs at the HNF4A locus in HepG2 cell line is shown in Figure 3D, which demonstrates that our method is able to predict most TF binding sites where strong enhancer signals are present. These analyses illustrated that we established a precise TF binding site prediction method.

ANANSE predicts cell type-specific gene regulatory networks

Using the inferred cell type-specific binding profiles, we sought to determine the interactions of TFs and their target gene (TF-gene) to establish cell type-specific GRNs. To calculate the TF-gene interaction score we first identified all enhancers for each gene. In our TF binding prediction model, we used EP300 ChIP-seq as training data. Since EP300 ChIP-seq is not available for a wide variety of cell lines and tissues, we expanded the input data of our method. In this method, we used either EP300 ChIP-seq or ATAC-seq peak summits as the enhancer location. The enhancer activity is based on the number of reads in regions centered at the enhancer summit. For this, we used either 200 bp for narrow signals from ChIP-seq of

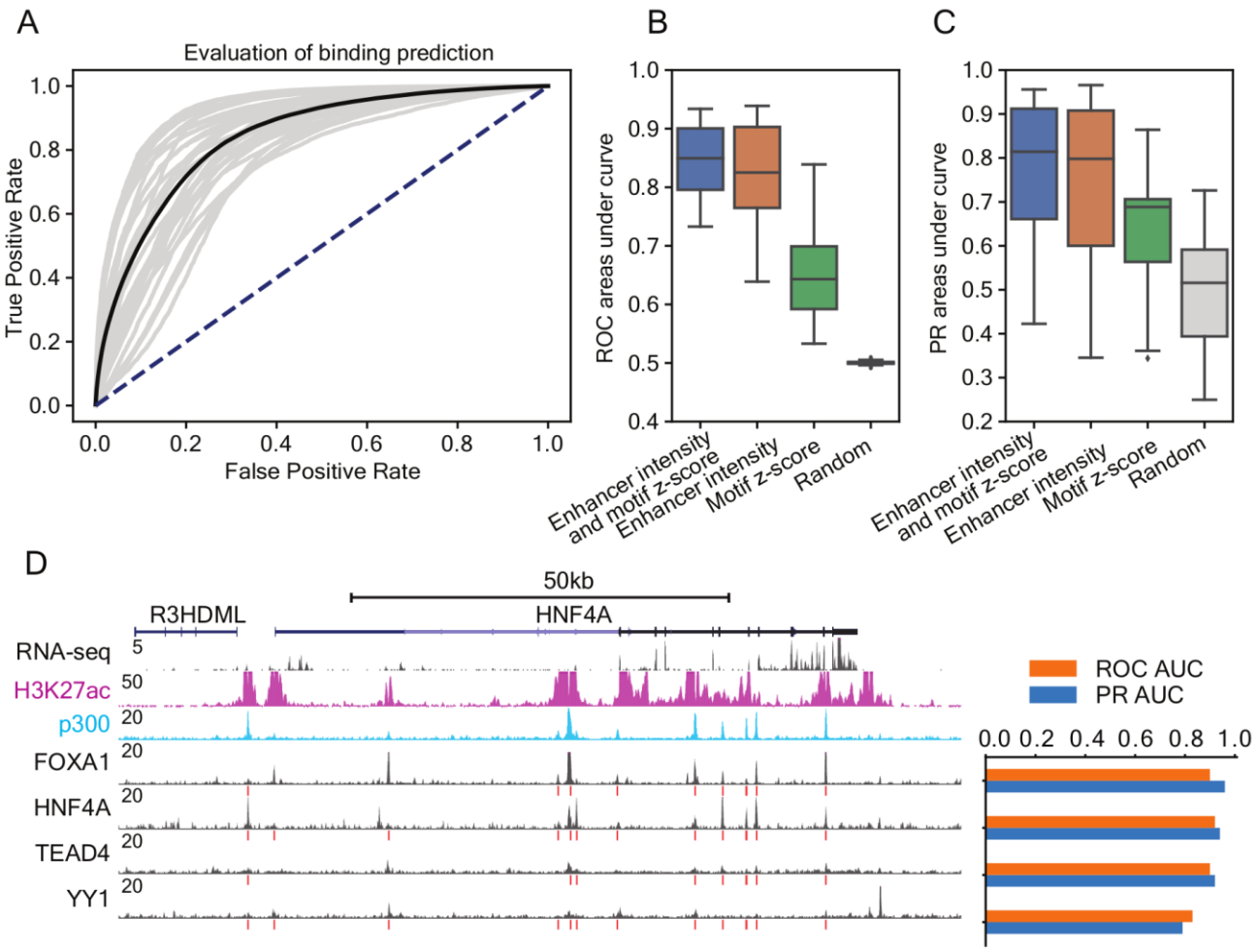


Figure 3. The performance of predicting TF binding sites using TF motif scores and enhancer activities. (A) Receiver-operator characteristic (ROC) curves representing the performance evaluation of the regression model in predicting genome-wide binding for 29 TFs. The ROC AUC score was calculated using leave-one-out cross validation. Each gray line represents the prediction of one TF, based on the model trained on all other 28 TFs. The black line indicates the mean prediction performance of all 29 TFs. **(B)** The prediction performance of the regression model (blue) compared to the baseline of enhancer signal (orange), motif z-score (red), and random data (green). The boxplot shows the ROC AUC of 29 TFS. **(C)** The same data as in B), quantified using the PR AUC. **(D)**. A UCSC Genome Browser screenshot of predicted TF binding sites (red bars) and ChIP-seq signals (black peaks) of 4 TFs (FOXA1, HNF4A, TEAD4, and YY1) in the HepG2 cell line at the HNF4A gene locus. The RNA-seq signals (black), ChIP-seq signals of EP300 (light blue) and H3K27ac (magenta) from ENCODE in the HepG2 cell line are shown for comparison. The barplot on the right shows the ROC (orange) and PR (blue) AUC of four TFs, with the same order of TFs shown in ChIP-seq signals on the left.

EP300, or 2 kb for broader signals from ChIP-seq of H3K27ac. In the case of ATAC-seq data, the enhancer intensity is based on H3K27ac ChIP-seq, as we expect that H3K27ac signal better correlates with enhancer activity. However, in principle the method would also work with ATAC-seq signal.

For each gene, we took all enhancers that are located at least 1kb from the gene TSS and within a maximum distance of 100 kb. Subsequently, the strength of a TF-gene interaction in the network was defined by the sum of the predicted TF binding strength in all identified enhancers of the target gene, weighted by the distance (Figure 4A). The distance weight was calculated from the linear genomic distance between the enhancer and the TSS of a gene, such that distal enhancers receive a low weight and nearby enhancers have a high weight (Wang et al., 2016) (Figure 4A). This model resulted in a TF-gene binding score, indicating the TF-target gene binding intensity for all combinations of TFs and target genes pairs.

Based on the assumption that the interaction of every TF-gene pair in a specific cell type is proportional to their relative expression, we included the expression level of the TF and the target gene. We scaled the expression level of the TF and the target gene, initially expressed as transcripts per million (TPM), to a normalized expression of 0 to 1, with the lowest expression as 0 and highest as 1. Combining the TF-gene binding score and TF and target expression scores by taking the mean, we obtained a TF-gene interaction score that represents the predicted strength of the TF-gene regulation (Figure 4B). Together, all TF-gene interaction scores represent a cell type-specific GRN.

To evaluate the quality of the inferred GRNs, we used three different types of reference datasets: gene co-expression, Gene Ontology (GO) annotation (The Gene Ontology, 2019) and regulatory interaction databases containing known TF-target gene interactions.

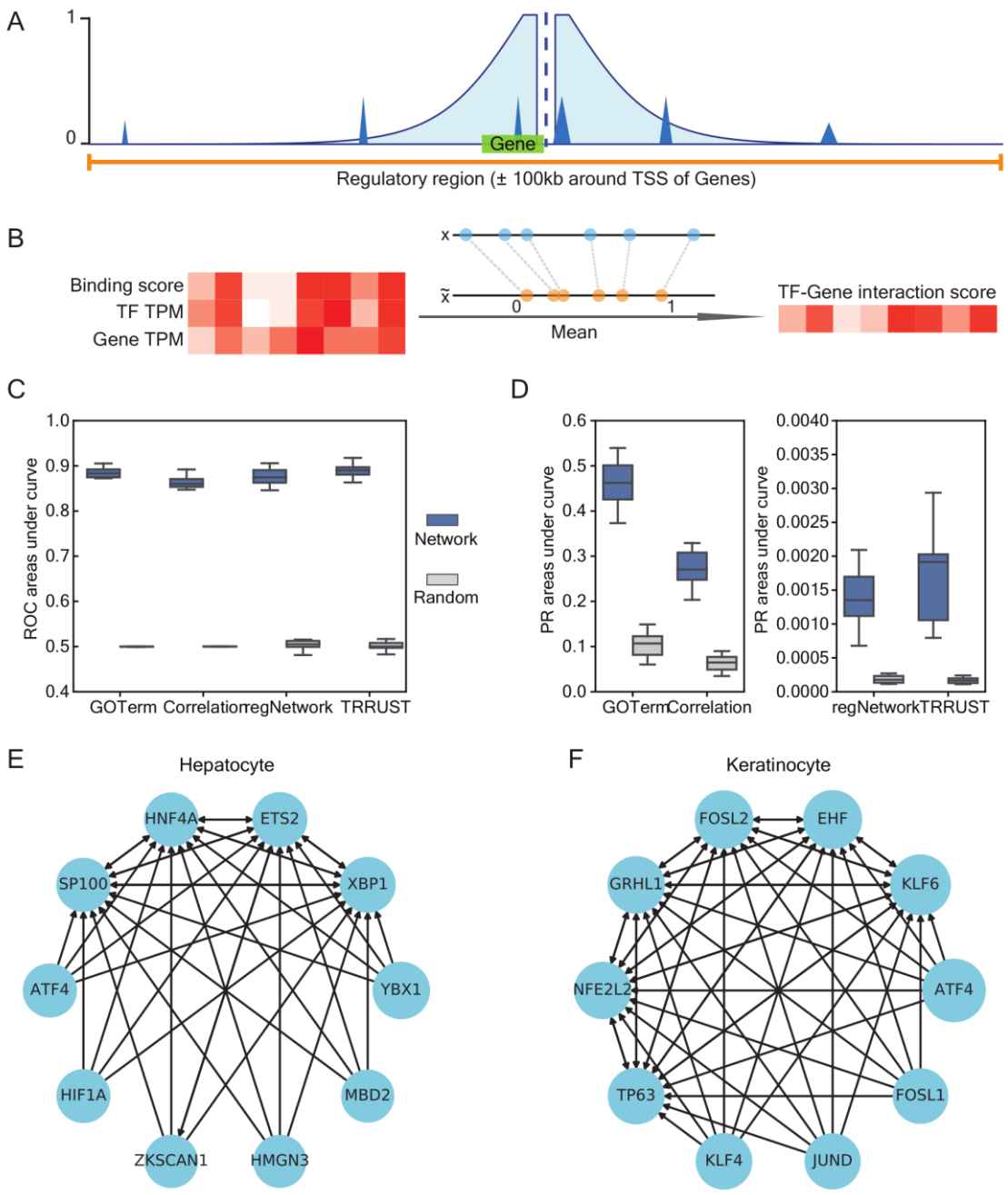


Figure 4. Establishment of cell type-specific enhancer gene regulatory networks. (A) A schematic overview of the calculation of TF-Gene binding score from the predicted binding intensity of TFs. The green rectangle represents one target gene and the orange line shows 100kb up- and downstream of the TSS of the corresponding gene, the range that is used to include enhancers for calculation. The dark blue triangles represent all predicted TF binding peaks within the 100kb range around the gene. The height of the shaded light blue area represents the weight calculated based on the linear genomic distance from TSS of the target gene to the enhancers. For example, the distance weight for the distance of 1kb from the TSS is 0, and for the distance of 100kb from the TSS is 5 (Wang et al., 2016). **(B)** A schematic overview of the calculation of the TF-Gene interaction score. The heatmaps on the left represent the TF-Gene binding score, TF expression level (TPM) and target gene expression level (TPM) of each TF-Gene regulatory pair. All three scores are scaled from 0 to 1, and the mean of three scores of each TF-Gene pair is defined as the interaction score (right heatmap) of the corresponding TF-Gene pair. **(C)** Evaluation of the predicted networks using different types of data: a gene ontology (GO) term co-regulatory network, a cell type-specific gene correlation network, and two TF-Gene regulatory networks based on interaction databases (regNetwork and TRRUST). The boxplots show the AUC of ROC for 8 different cell type. ROC AUC of the predicted networks is shown in blue; the random networks are indicated in green. **(D)** The same evaluation as in C), with the PR AUC shown as a boxplot. **(E)** Example network predicted for hepatocytes. The blue circles show the top 10 TFs in this cell type. The size of the circle indicates the target gene number of the corresponding TF. The black arrows indicate the interaction score between the two TFs. **(F)** Example network predicted for keratinocytes, visualized as in E).

We downloaded co-expression data for human genes from the COXPRESdb database (Obayashi et al., 2019). All TF-gene pairs with a correlation >= 0.6 were used as true positives. For the GO validation set, we used TF-gene pairs that were annotated with at least one common GO term as true positives. Finally, we used the TF-gene interactions from two databases of regulatory interactions, RegNetwork (Liu et al., 2015) and TRRUST (Han et al., 2017). RegNetwork is an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse, and TRRUST is an expanded reference database of human and mouse transcriptional regulatory interactions.

We evaluated cell type-specific GRNs predicted by ANANSE in eight cell type: astrocyte, cardiomyocyte, fibroblast, hepatocyte, iPSC, keratinocyte, macrophage and osteocyte. As the previously described reference datasets contain regulatory interactions for all genes, regardless of cell type, we created high confidence cell type-specific reference data by filtering for TFs and genes that are expressed, using a stringent cutoff of TPM > 10. We first evaluated the model by calculating the area under the curve of the Receiver Operator using the different reference sets for the nine cell types (Figure 4C). The mean AUC ranges from 0.86 using the co-expression reference to 0.89 using the TRRUST database while the median AUC of randomized networks is close to 0.5 (Figure 4C). Some of the reference databases contain very few interactions (the positives in this evaluation) as compared to all possible interactions (which determine the negatives). For instance, the fraction of positive interactions is 0.03% in TRRUST and RegNetwork. Therefore, we also evaluated the predicted networks using the Precision-Recall area under the curve (PR AUC) (Figure 4D). In absolute terms, the PR AUC is considerably lower than the ROC AUC, especially for the TRRUST and RegNetwork reference sets (median PR AUC of 0.0017 and 0.0014, respectively), but for all cell types there is a relatively large and statistically significant difference between the predicted GRN and the random network (TRRUST p-value = 0.006 and RegNetwork p-value = 0.006). We also repeated the evaluation with a lower cutoff TPM >1 for both TFs and target genes, which does not change the interpretation of the results (Supplemental Figure S2).

To qualitatively assess the cell type-specific GRNs predicted by ANANSE, we chose two well-studied cell types, hepatocytes and epidermal keratinocytes, and constructed their GRNs using the top ten predicted TFs of each cell type, as ranked by outdegree. The GRN of hepatocyte contains classical hepatocytes marker genes and reprogramming factors, such as HNF4A, HIF1A and ATF4 (Figure 4E) (Simeonov and Uppal, 2014).. The GRN of keratinocytes includes TFs that are known to regulate keratinocyte proliferation and epidermal differentiation, such as KLF5, KLF6 and MYC (Figure 4F) (Nair et al., 2006; Oberbeck et al., 2019; Richardson et al., 2006; Sen et al., 2012; Tsuji et al., 2018).

Taken together, our benchmarks and examples demonstrate that GRNs generated by ANANSE allow for meaningful cell type-specific prioritization of TFs.

ANANSE accurately predicts transcription factors for trans-differentiation

Having established that ANANSE-inferred GRNs can enrich for biologically relevant regulatory interactions, we aimed to use these GRNs to identify key TFs that regulate cell fate determination. To this end, trans-differentiation is a good model for this purpose, as experimentally validated TFs have been determined for various trans-differentiation strategies. Here, we first inferred the GRNs for all cell types using our ANANSE approach. The ANANSE-inferred GRN differences between two cell states, a *source* cell type and a *target* cell type, was calculated to represent the differential GRN between two cell types, which contains the GRN interactions that are specific for or higher in the target cell type. Subsequently, using an approach inspired by Mogrify (Rackham et al., 2016), we calculated the influence score of TFs for these trans-differentiations by determining the differential expression score of its targets weighted by the regulatory distance (see Methods for details).

To evaluate the prediction by ANANSE, we used experimentally validated TFs for several trans-differentiation strategies. For this, we collected TFs for eight trans-differentiation strategies with fibroblasts as the source cell type. The target cell types include astrocytes (Caiazzo et al., 2015), cardiomyocytes (Fu et al., 2013), hepatocytes (Simeonov and Uppal, 2014), iPSCs (Takahashi et al., 2007), keratinocytes (Kurita et al., 2018), macrophages (Xie et al., 2004), neural crest cells (Kim et al., 2014), and osteocytes (Li et al., 2017) (Table 1). The complete list of the relevant TFs with literature evidence is shown in Supplementary Table S6. We used EP300 ChIP-seq data or the combination of ATAC-seq and H3K27ac ChIP-seq data of these cell types to create cell type-specific GRNs (Table 2), and then calculated TF influence scores and ranked the TFs in each cell type.

Target cell type	Experimentally validated TFs	TFs predicted by ANANSE	Reference
Astrocyte	NFIA, NFIB, SOX9	SOX9	(Caiazzo et al., 2015)
Cardiomyocyte	GATA4, MEF2C, TBX5, ESRRG, MESP1	GATA4	(Fu et al., 2013)
Hepatocyte	FOXA1, FOXA3, HNF4A	HNF4A	(Simeonov and Uppal, 2014)
iPSC	SOX2, OCT4, KLF4, MYC	SOX2, OCT4	(Takahashi et al., 2007)
Keratinocyte	TP63, GRHL2, TFAP2A, MYC	TP63	(Kurita et al., 2018)
Macrophage	CEBPA, SPI1	CEBPA, SPI1	(Xie et al., 2004)
Osteocyte	RUNX2	RUNX2	(Li et al., 2017)

Table 1: The summary of eight experimentally validated trans-differentiations from fibroblast to target cell types.

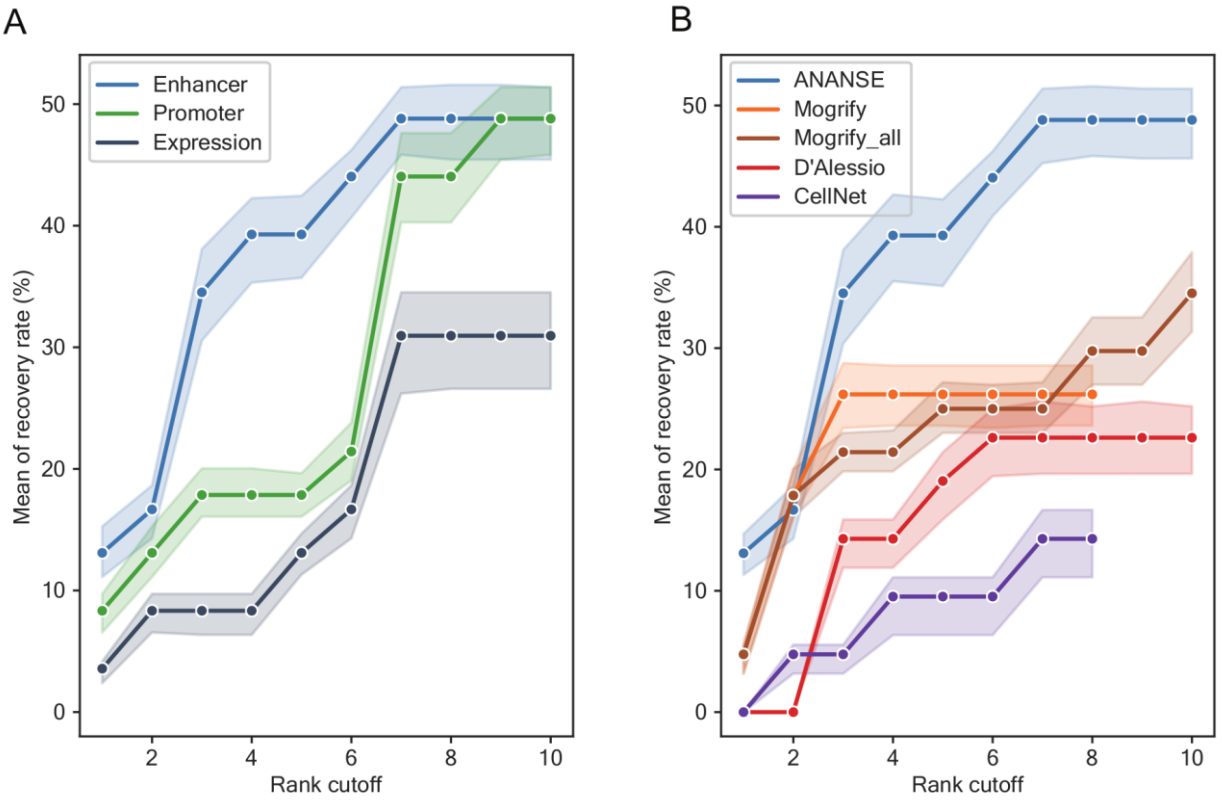


Figure 5. Evaluation of the performance of ANANSE using experimentally validated trans-differentiation strategies. (A) The line plots show the comparison of the predicted top TFs for trans-differentiation from cell type-specific networks. Based on the difference between two networks, TFs were prioritized using the influence score calculation implemented in ANANSE. Shown is the fraction of predicted TFs compared to all known TFs based on trans-differentiation protocols described in the literature (y-axis) as a function of the top number of TFs selected (x-axis). The shaded area represents the minimum and maximum percentage of corresponding recovered TFs when using seven out of eight trans-differentiations. Three different types of networks were used: gene expression (deep blue), promoter-based TF binding in combination with expression (green), and enhancer-based TF binding in combination with expression (blue). **(B)** The line plots show the comparison of the predicted top TFs for trans-differentiation based on different computational methods. The y-axis indicates the percentage of experimentally validated cell TFs that are recovered as a function of the number of top predictions, similar as in A). Five different methods are shown: ANANSE (blue), Mogrify (orange), Mogrify without using their selection algorithm, the method of d'Alessio et al. (red) and CellNet (purple). The shaded area represents the minimum and maximum percentage of corresponding recovered TFs when using seven out of eight trans-differentiations. CellNet only contains data from fibroblast to ESC, Hepatocyte, and Macrophage; and Mogrify and CellNet only contain the top 8 predicted factors.

When we calculate TF influence scores from cell type-specific GRNs, it is important to decide what size of GRN should be chosen in terms of the top number of edges. We inferred the key TFs for the eight trans-differentiations using six different sizes of GRNs (10K, 50K, 100K, 200K, 500K, and 1M edges) (Supplementary Table S7). Supplementary Figure S3 and S4 show the percentage of known (experimentally validated) cell-fate determinants that are recovered as a function of the number of top predictions that are included. These results show that using a GRN size of 100K or 200K interactions results in the best recovery ability. Therefore, we chose a GRN size of 100K interactions for all following analyses. In four out of the eight cases, ANANSE includes the complete set of experimentally defined TFs in the top 10 predicted factors (Table 1). For example, ANANSE predicts CEBPA and SPI1 for reprograming fibroblasts to macrophages (Xie et al., 2004) and FOXA1, FOXA3 and HNF4A for reprogramming to hepatocytes, which are consistent with the experimental trans-differentiation strategies (Simeonov and Uppal, 2014). For the other examples, ANANSE prioritizes up to half of the experimentally defined TFs (Table 1).

As cell type-specific TFs predominantly bind to enhancers (Figure 1), and the inclusion of enhancer information is one of the unique features of ANANSE, we anticipated that ANANSE would have more accurate predictions of TFs for trans-differentiation. To investigate this, we compared ANANSE with previously reported expression- and promoter- based GRNs (Cahan et al., 2014; D'Alessio et al., 2015; Rackham et al., 2016) (Figure 5A). We created both expression and promoter based GRNs of the eight source and target cell type combinations. For expression-based GRNs, we used only the mean of the scaled TPM of TFs and genes together as the interaction score of TFs and genes. For the promoter-based GRNs, we selected the highest binding score of TFs within 2kb of the TSS of the corresponding gene as the binding score of the TF-gene pair. Subsequently, the mean of the scaled TPM of the TF and the gene together with the binding score determines the interaction score of the TF and gene (Figure 4B). We then

inferred the key TFs for the eight trans-differentiations using ANANSE and these two types of GRNs. The ANANSE influence score based on the enhancer GRNs includes 40% of the known TFs in the top 4 predictions (Supplementary Figure S5 and Supplementary Table S6 and S8). In contrast, using the influence score based on the promoter GRN or the expression GRN, we could recover only 18% and 9% of the known TFs in the top 4 predictions (Figure 5A and Supplementary Figure S5). These results demonstrate that using enhancers in the construction of GRNs significantly improves the prediction of relevant TFs in cell fate determination.

Next, we further quantified the performance difference between ANANSE and previously reported methods, namely, Mogrify, CellNet and the method of D'Alessio et al (Cahan et al., 2014; D'Alessio et al., 2015; Rackham et al., 2016) (Figure 5B and Supplementary Figure S6). For Mogrify, we downloaded both the prioritized list of TFs based on TF expression in source cell types and GRN overlap, as well as the full unfiltered list of TFs. For these comparisons, we aimed to include all eight trans-differentiation strategies. In some cases, as data for the exact cell type is unavailable, similar cell or tissue types were used as surrogates. For example, the osteoblast-Sciencell was used to substitute for osteoblast. For CellNet, we used the previously described results of three cell types: hepatocytes, iPSCs and macrophages (Rackham et al., 2016). Using the eight cell type conversions as a reference, ANANSE has the highest recovery when more than two TFs are used for trans-differentiation (Figure 5B and Supplementary Figure S6). ANANSE predicts a mean of 40% TFs using the top four TFs ranked by influence score, while other methods predict a maximum of 27% TFs with this rank cutoff (Figure 5B and Supplementary Figure S6). When the number of predicted TFs was increased to ten, ANANSE could increase its recovery rate to 50%, while the maximum mean recovery of other methods is 35% (Figure 5B and Supplementary Figure S6).

In summary, these analyses show that including enhancers in the GRN construction significantly improves the prediction TFs in cell fate conversion and that ANANSE outperforms other established methods based on experimentally validated trans-differentiation TFs. Our results demonstrate that ANANSE can accurately prioritize TFs in cell fate determination.

ANANSE identified an atlas of key transcription factors in normal human tissues

The gene expression programs that drive the cellular differentiation programs of different tissues are largely controlled by TFs. To find out which key TFs drive cell fate determination in different tissues, we applied ANANSE to human tissue data. We downloaded enhancer activity data of 18 human tissues from the dbCoRC database (Huang et al., 2017) and the RNA-seq data of corresponding tissues from the Human Protein Atlas project (Uhlén et al., 2015). Using these enhancer and gene expression data, we constructed tissue-specific GRNs using ANANSE, and then calculated the TF influence scores for each of the tissues when taking the combination of all other tissues as the source tissue (Supplementary table S9). We clustered the 18 tissues based on the correlation between TF influence scores using hierarchical clustering, showing that the influence score captures regulatory similarities and differences between tissues (Figure 6A and Supplementary Figure S4). For example, the esophagus and the skin cluster together, as these tissues are composed mostly of stratified squamous epithelial cells, and skeletal muscle and heart tissue are clustered together as both tissues contain striated muscle tissues.

For all studied tissues, we have provided a rich resource of key TFs of each tissue, with a list of top ten key TFs (Figure 6B). Many TFs in this list are known to play important functions for specific tissues, e.g. ELF3 and KLF5 for stomach, colon, and small intestine (Jedlicka et al., 2008; Katz et al., 2002); TFAP2A, TFAP2C, TP63, and GRHL2 for the skin and esophagus (Dollé, 2009; Qu et al., 2018; Wilanowski et al., 2008); SOX2, SOX8 and OLIG1/2 for brain (Bani-Yaghoub et al., 2006; Meijer et al., 2012; Muto et al., 2009); and SPI1 and IRF1 for lung, spleen and bone marrow (Ohteki et al., 2001) (Figure 6A).

The gastrointestinal tract tissues share a number of highly influence score TFs such as ELF3, KLF4, and HNF4A, which play roles in stomach, colon, and small intestine development, and are consistent with the current research on gastrointestinal tract tissues (Figure 6A) (Jedlicka et al., 2008; Katz et al., 2002; Thompson et al., 2018). ELF3 is important in intestinal morphogenesis, homeostasis, and disease (Jedlicka et al., 2008). KLF4 is required for terminal differentiation of goblet cells in the colon (Katz et al., 2002).

Our analysis showed that TP63, TFAP2A, TFAP2C, RARG, and GRHL1 are common important TFs in the skin and esophagus (Figure 6B). The function of these TFs has been well studied in the skin. TP63 is one of the TFs that is important in both skin and esophagus development (Daniely et al., 2004; Kurita et al., 2018; Qu et al., 2018). TP63 and TFAP2A have been used in *in vivo* reprogramming of wound-resident cells to generate skin epithelial tissue (Kurita et al., 2018). Both TFAP2A and TFAP2C are required for proper early morphogenesis and development as well as terminal differentiation of the skin epidermis (Budirahardja et al., 2016; Kousa et al., 2018; Wang et al., 2008). GRHL1 is important for the functioning of the epidermis. Grhl1 knockout mice exhibit palmoplantar keratoderma, impaired hair anchoring, and desmosomal abnormalities (Wilanowski et al., 2008). It would be interesting to investigate how they play roles in esophagus. RARG is expressed in the developing skin epithelium, as well as in all prospective squamous keratinizing epithelia, including the esophagus and left wall of the stomach (Dollé, 2009; Ruberte et al., 1990). PAX9 regulates squamous cell differentiation and carcinogenesis in the oro-oesophageal epithelium (Xiong et al., 2018).

In summary, using ANANSE, we predicted key TFs for 18 human normal tissues. Many of these predicted TFs correlate well with the known literature of these tissues. In addition, the predicted key TFs in each tissue also provide us a rich resource to unveil novel TFs in specific tissues.

Discussion

Lineage specification and cell fate determination are critical processes during development. They are necessary to form the diversity of cell types that are organized into organs and tissues. TFs form a central component in the regulatory networks that control lineage choice and differentiation. Indeed, cell fate can be switched *in vitro* through manipulation of TF expression (Caiazzo et al., 2015; Fu et al., 2013; Kurita et al., 2018; Li et

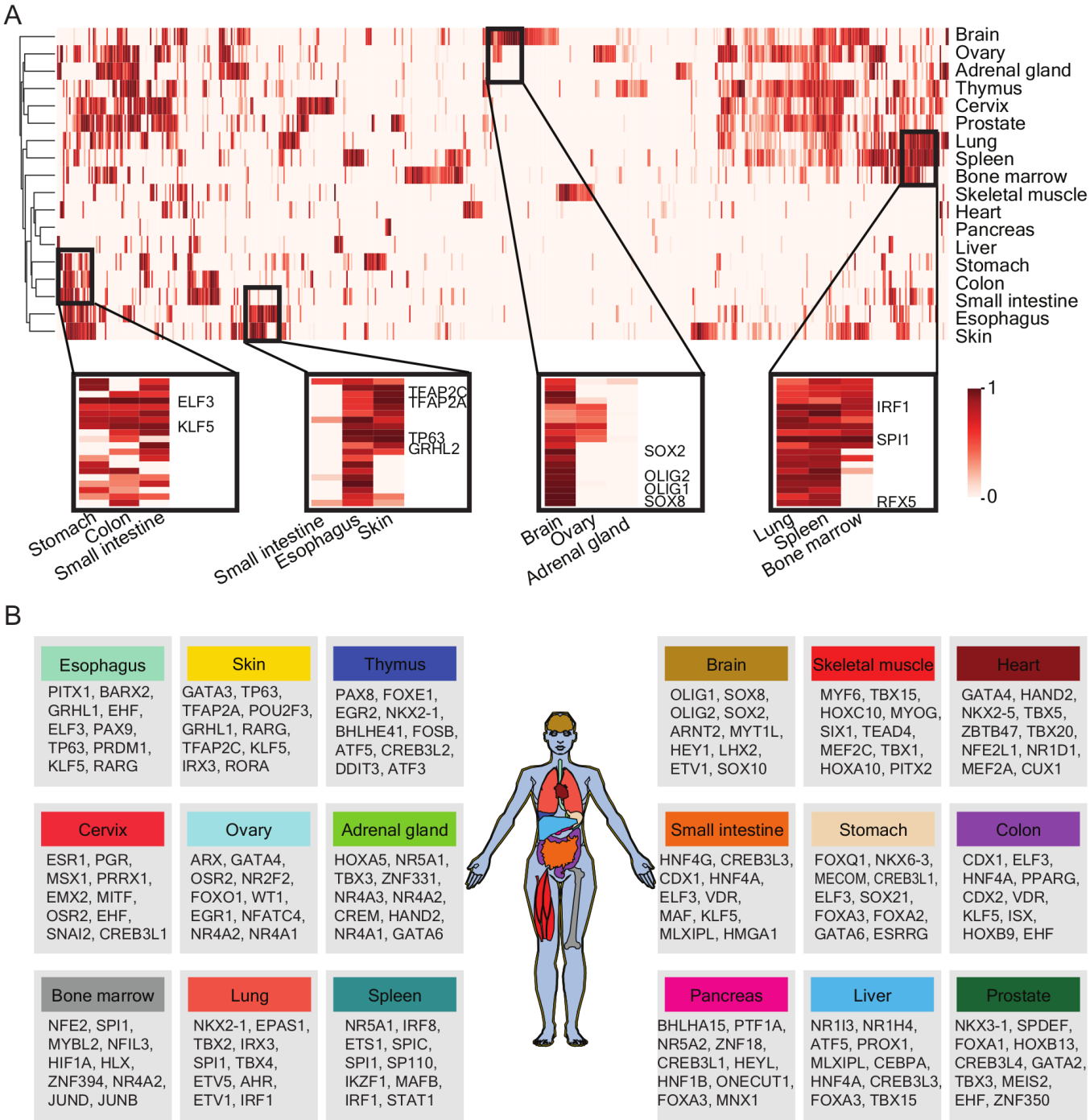


Figure 6. Applying ANANSE to expression data of human tissues to identify key transcription factors. (A) Heatmap of the predicted influence scores of all TFs using ANANASE on data from 18 human tissues. The color in the heatmap indicates the relative influence score, from low to high. The four small heatmaps highlighted below show important TFs in related tissues. **(B)** The top 10 key TFs of 18 tissues inferred by ANANSE. The color of the tissue is consistent with the tissue name in the box. The order of TF of each tissue is based on the influence score of the TF ranked from high to low.

al., 2017; Simeonov and Uppal, 2014; Takahashi et al., 2007; Xie et al., 2004). However, the regulatory factors that determine cell identity remain unknown for many cell types. To address this issue, we developed ANANSE, a new computational method to predict the key TFs that regulate cellular fate changes.

Using an ensemble approach, we established a general model that leverages genome-wide, cell type-specific enhancer signals from ATAC-seq or Ep300 or H3K27ac ChIP-seq data, and TF motif data to reconstruct TF binding networks for each cell type. ANANSE takes a two-step approach. First, TF binding is imputed for all enhancers using a TF-agnostic model. In contrast to existing methods that aim to predict binding by training TF-specific models (Batsis et al., 2019; Keilwagen et al., 2019; Li et al., 2019; Quang and Xie, 2019), we used a more general model. Our model will be less accurate in predicting TF

binding compared to these models specifically trained for individual TFs. However, the advantage is that our model can predict binding for every TF as long as it is known to which motif the TF can bind. In addition, it can be used for factors for which there is no training data available, and for instance, it can also be applied to non-model organisms that lack comprehensive ChIP-seq assays. We found that the enhancer signal intensity is the most important predictive feature for binding (Figure 3A-C and Supplementary Figure S1). The TF-specific motif z-score modestly increases the performance.

Second, we summarized the imputed TF signals per gene, using a distance-weighted decay function (Wang et al., 2016), and combine this measure with TF and target gene expression to infer cell type-specific GRNs. There is a lack of gold standards, especially to evaluate cell type-specific regulatory networks. To evaluate the GRNs, we used two orthogonal types of benchmarks: a database of known, experimentally identified TF-gene interactions and functional enrichment using Gene Ontology annotation. The databases with known interactions that we used (TRRUST (Han et al., 2017) and regNetwork (Liu et al., 2015)) contain only a fraction of true regulatory interactions, and therefore this benchmark is affected by a large fraction of false negatives. All our benchmark evaluations demonstrate that ANANSE significantly enriches for true regulatory interactions. However, it also highlights that GRN inference is far from a solved problem. The PR-AUC values are low, as is generally the case in eukaryotic GRN inference (Chen and Mar, 2018).

In contrast to previous approaches, our method takes advantage of TF binding in enhancers, instead of only gene expression differences or TF binding to proximal promoters. This resulted in significantly improved performance, as benchmarked on experimentally validated trans-differentiation protocols. It has been previously shown that cell type-specific regulation is much better captured by enhancers as compared to promoter-proximal regulatory elements. For instance, TF binding and chromatin accessibility in distal elements better reflect the cell type identity of hematopoietic lineages than in promoters (Corces et al., 2016; Heinz et al., 2010). Many important transcriptional regulators mainly bind at regulatory regions that are not proximal to the promoter. Indeed, our analysis of the genomic binding distribution of ~300 human TFs showed that the large majority of cell type-specific TFs mainly binds in enhancer regions (Figure 1C). Therefore, we reasoned that TF binding at enhancers would be essential to model cell fate and lineage decisions. We tested the application of the networks inferred by ANANSE to human *in vitro* trans-differentiation approaches. Seminal work showed that computational algorithms can help to characterize cellular fate transitions, and to provide rational prioritization of TF candidates for trans-differentiation (Cahan et al., 2014; Morris et al., 2014; Rackham et al., 2016). We implemented a network-based approach to prioritize TFs that determine cell fate changes. Using a collection of known, experimentally validated trans-differentiation protocols, we demonstrated that ANANSE consistently outperforms other published approaches. This means that cellular trajectories can be characterized using ANANSE to identify the TFs that are involved in cell fate changes. In comparison with a promoter-based approach, we show that using enhancer-based regulatory information contributes significantly to this increased performance (Figure 5). One noticeable example is the trans-differentiation from fibroblasts or mesenchymal cells to keratinocytes. In current experimentally validated trans-differentiation methods, the epithelial master regulator TP63 is essential for establishing the keratinocyte cell fate (Chen et al., 2014; Kurita et al., 2018). However, TP63 was not predicted in the previously published computational methods (Cahan et al., 2014; Morris et al., 2014; Rackham et al., 2016). One plausible explanation is that TP63 is a TF for specific epithelial cells and tissues and it binds predominantly (87%) to enhancers (Andersson et al., 2014; Bulger and Groudine, 2011; Qu et al., 2018; Spitz and Furlong, 2012), whereas previous computational tools do not take enhancer properties into consideration.

We used ANANSE to identify tissue-specific TFs for different human tissues. We predicted the top 10 key TFs for all studied tissues. Many TFs in this list are known for important functions in these specific tissues. For example, some NK homeodomain, GATA, and T-box TFs are found in normal cardiac development, which have important functions during heart specification, patterning, and differentiation (Bruneau, 2013; Kathiriya et al., 2015; Stefanovic and Christoffels, 2015). Many SOX family of TFs are known critical for neural system development in brain tissue (Bani-Yaghoub et al., 2006; Muto et al., 2009). Although not all predicted TFs are known to have important role in specific tissues, further research is warranted. The TFs in the TF atlas predicted by ANANSE are may also be good candidates for studying tissue development and engineering in regenerative medicine

Another large benefit of the model that we implemented in ANANSE is the wide applicability. The source code of ANANSE is publicly available under a liberal license. It is straightforward to run ANANSE on new data, such as different cell types or even species. Types of data required for this analysis are the following: gene expression data (RNA-seq) and genome-wide measure of enhancer activity. Enhancer data can be either EP300 or H3K27ac ChIP-seq. Both types of data can be relatively easily obtained, not only in human cell types or in common model species, but also often in non-model species (Villar et al., 2015). This means that this method can be easily applied in a wide variety of biological studies in gene regulation during development and cellular fate changes.

We also acknowledge limitations in our approach. In ANANSE, we link enhancer regions to genes on basis of distance. For each TF and gene interaction pair, ANANSE only considers TF binding information located at most 100kb up and downstream of the corresponding gene. Although data from a recent CRISPR enhancer interference screen showed that genomic distance is largely informative in predicting enhancer-target interactions (Fulco et al., 2019), this approach may be limited when applying to genes regulated through long-range regulation, especially those via inter-chromosomal regulation (Olivares-Chauvet et al., 2016). This limitation of our method can potentially be addressed using chromosome conformation capture techniques (3C) or other adaptations as circular 3C (4C), chromosome conformation capture carbon copy (5C), chromatin immunoprecipitation using PET (ChIA-PET) and Hi-C (Kempfer and Pombo, 2019). However, these types of data are currently only available for a limited number of cell

types, therefore incorporation of topology data would limit the broad utility and application of our approach.

Conclusion

Here we presented ANANSE, a computational tool for efficient predictions of TFs in cell fate determination. It outperforms other published methods in predicting TFs that can induce trans-differentiation. In addition, it is open source, freely available and can be easily used on custom data. In summary, ANANSE exploits the powerful impact enhancers have on gene regulatory networks, and it provides insights into TF mediated regulatory mechanisms underlying cell fate determination and development.

Methods

Datasets

The 296 ChIP-seq datasets for the peak location analysis were download from the ReMap project (Chèneby et al., 2017). For training and evaluation of the regression model to predict TF binding EP300 ChIP-seq data and TF ChIP-seq data from 5 different cell lines (GM12878, hESC, HeLa-S3, HepG2 and K562) downloaded from the ENCODE-DREAM project (https://www.synapse.org/ENCODE) (ENCODE-DREAM, 2017) were used. The non-redundant database of 1,690 motifs for 1,164 TFs was created by clustering all vertebrate motifs from the CIS-BP database using GimmeMotifs (van Heeringen and Veenstra, 2010; Weirauch et al., 2014). The data used to predict key TFs for trans-differentiation is shown in Table 2, including RNA-seq, EP300 and H3K27ac ChIP-seq, and assay for transposase-accessible chromatin sequencing (ATAC-seq) (Buenrostro et al., 2013) data in 8 cell types.

ChIP-seq, ATAC-seq and RNA-seq analyses

The reads of the ChIP-seq and ATAC-seq experiments were mapped to the human genome (hg38) using STAR (version 2.5.3a) with default settings (Dobin et al., 2013). Duplicate reads were marked and removed using Picard. Peaks were called on the ChIP-seq and ATAC-seq data with only the uniquely mapped reads using MACS2 (version 2.7) relative to the Input track using the standard settings and a q-value of 0.01 (Zhang et al., 2008). The measurement of consistent peaks between replicates was identified by IDR (version 2.0.3) (Li et al., 2011). Quantification of expression levels was performed on RNA-seq data, using salmon (version 0.43.0) (Patro et al., 2017) with default settings and Ensembl transcript sequences (version GRCh37) (Cunningham et al., 2018). Salmon's transcript-level quantifications results were imported and aggregated to gene level counts by the tximport R package (version 1.12.3) (Soneson et al., 2015). The expression level (transcript-per-million, TPM) of each cell type and the differential expression fold change between two cell types were calculated using the DESeq2 R package (version 1.24.0) (Love et al., 2014). The expression TPM data used to predict key TFs for trans-differentiation is shown in Supplementary Table S1, deferential expression genes data is shown in Supplementary Table S2, and the enhancer data is shown in Supplementary Table S3.

Analysis of the genomic distribution of TF binding sites

For every TF, we combined all the peaks in the ReMap database (Chèneby et al., 2017) by taking of the peaks in all cell types and tissues for this specific TF. TFs that had less than 600 peaks were removed. The percentage of peaks in each genomic location was calculated using the ChIPseeker R package (version 1.20.0) (Yu et al., 2015). The fgsea R package (version 1.10.1) was used to do the gene set enrichment analysis (GSEA) (Sergushichev, 2016).

Defining putative enhancer regions

The EP300 ChIP-seq or ATAC-seq peaks were used to define putative enhancer regions for specific cell types. The summits of the MACS2 peaks were chosen first, then extended (+/- 100bp) to a total size of 200bp. The EP300 or H3K27ac ChIP-seq peak intensity generated by MACS2 (bedGraph file) was used to represent enhancer intensity. The bedGraphToBigWig tool (Kent, 2014) was used to convert bedGraph files to bigWig files. The intensity of enhancer peak was calculated using thebigWigSummary tool (Kent, 2014), which selects the highest signal of EP300 (200 bp around the peak summit) or H3K27ac (2,000 bp around the peak summit) ChIP-seq.

Binding network inference

The GimmeMotifs package (version 0.13.1+216.g7f81a8a) (Bruse and Heeringen, 2018; van Heeringen and Veenstra, 2010) was used to scan for motifs in enhancer regions. The GC normalization setting in GimmeMotifs package was used to normalize the GC% bias in different

Cell type	RNA-seq	EP300	ATAC-seq	H3K27ac
Astrocyte	GSE104232 (Tchieu et al., 2019)	/	GSE104232 (Tchieu et al., 2019)	ENCODE (ENCODE Project Consortium, 2012)
Cardiomyocyte	GSE116862 (Zhang et al., 2019)	/	GSE85330 (Liu et al., 2017)	GSE116862 (Zhang et al., 2019)
Fibroblast	GSE120081 (van der Raadt et al., 2019)	GSE97033 (Zhao et al., 2017)	/	/
Hepatocyte	ENCODE (ENCODE Project Consortium, 2012)	PRJNA239635	/	/
iPSC	GSE120107 (Soares et al., 2019)	ENCODE (ENCODE Project Consortium, 2012)	/	/
Keratinocyte	GSE97033 (Zhao et al., 2017)	GSE67382 (Bao et al., 2015)	/	/
Macrophage	GSE85243 (Novakovic et al., 2016)	/	GSE85243 (Novakovic et al., 2016)	GSE85243 (Novakovic et al., 2016)
Osteocyte	GSE29611 (ENCODE Project Consortium, 2012)	GSE29611 (ENCODE Project Consortium, 2012)	/	/

Table 2: The data used to predict key TFs for trans-differentiation

enhancers. To correct for the bias of motif length, z-score normalization was performed on the motif scores. Normalization was done per motif, based on motif matches in random genomic regions using the same motif scan settings. The highest z-score was chosen if a TF had more than one motif. The enhancer intensity levels were scaled from 0 to 1, with 1 being the highest and 0 the lowest. The binding intensity of each TF in each enhancer was predicted by logistic regression using motif z-score and scaled enhancer intensity as input.

Gene regulatory network inference

The weighted sum of TF predicted enhancer intensity within 100kb around TSS is defined as the TF-gene binding score (Eq. 1). The distance weight is based on a linear genomic distance between the enhancer and the TSS of a gene according to equation 2.

$$B_{x,r} = \sum_{k} w_k s_k \qquad (1)$$

where $B_{x,r}$ is the binding score between TF x and target gene r, w_k is the weighted distance between an enhancer and the target gene and where s_k is predicted binding intensity at genomic position k of TF x. The distance weight calculation was similar to the method previously described in (Wang et al., 2016), except that only signal in enhancers is used, enhancers within 2kb around TSS are removed and the weight of enhancers within 2kb to 5kb is set to 1.

$$w_k = \begin{cases} 0, & k \in (0kb, 2kb] \\ 1, & k \in (2kb, 5kb] \\ \dfrac{2e^{-\mu|k-t_r|}}{1+e^{-\mu|k-t_r|}}, & k \in (5kb, 100kb] \end{cases} \qquad (2)$$

where t_r is the genomic position of the TSS of gene r and the parameter μ, which determines the decay rate as a function of distance from the TSS, is set such that an enhancer 10 kb from the TSS contributes one-half of that at the TSS.

We scaled the expression level of the TF and the target gene, expressed as transcripts per million (TPM), and the TF-gene binding score $B_{x,r}$ we calculated in the first step from 0 to 1, with 1 being the highest and 0 the lowest. Combining the TF-gene binding score and TF and target expression scores by taking the mean, we obtained a TF-gene interaction score.

Gene regulatory network evaluation

To evaluate the quality of the predicted GRNs, three different types of reference datasets were used: gene co-expression, Gene Ontology (GO) annotation (The Gene Ontology, 2019) and two regulatory interaction databases (RegNetwork(Liu et al., 2015) and TRRUST (Han et al., 2017)). The expression correlation database was downloaded from COXPRESdb (Obayashi et al., 2019), and the original mutual rank correlation score was scaled to 0 to 1 for each TF, with 1 being the highest and 0 the lowest, and all scaled correlation score higher than 0.6 were considered as true interaction pairs. The human GO validation Gene Association File (GAF) (version 2.1) was download from http://geneontology.org. We used all TF-gene pairs that were annotated with at least one common GO term as true positives. In each cell type, four types of cell type-specific reference datasets were selected from previously described reference datasets by only selecting only interaction for which both the TF and its target gene are expressed (TPM > 10 or TPM > 1) in the corresponding cell type. For the random network we used the same network interaction structure, but has randomized the interaction score (the edge weight). The AUC of ROC and PR for each cell type GRN and corresponding random GRN were calculated.

Influence score inference

To calculate the influence score for the transition from a source cell type to a target cell type, we used the GRNs for both cell types. In each network, we selected the top 100k interactions based on the rank of its interaction score. We obtained a differential GRN by taking the interactions only located in the target cell type and those with an interaction score that was at least 0.3 higher in the target cell type than in the source cell type. The difference of the interaction score was used as the edge weight for the differential GRN.

Based upon the differential GRN a local network was built for each TF, up to a maximal number of three edges. Using equation (3), a target score was calculated for each node in the network, based on 1) its edge distance from the TF of interest, 2) the interaction score and 3) the change in expression between the source cell type and the target cell type.

$$N_x^s = \sum_{r \in V_t}^{s} |G_r^s| \frac{P_{x,r}^s}{L_{x,r}^s} \qquad (3)$$

Where $r \in V_t$ is each gene (r) in the set of nodes (V_t) that make up the local sub-network of TF x and $L_{x,r}^s$ is the level (or the number of steps) that gene r is away from TF x in the network s. Nodes located further from the TF have less effect on the target score. $P_{x,r}^s$ is the interaction score between TF x and target gene r and G_r^s, the expression score, is the log-transformed fold change of the expression of gene r.

The target score (N_x^s) for each TF is the sum of the scores from all the nodes in its local network. Nodes present in multiple edges are calculated only for the edge closet to the TF of interest. Self-regulating nodes are not considered. The target score and the G_r^s of each TF are scaled to 0 to 1, and the mean of them was defined as the influence score of this TF. Subsequently, all TFs are ranked by their influence score.

Regulatory profile analysis of human tissues

The RNA-seq data of 18 human tissues were downloaded from Human Protein Atlas project (Uhlen et al., 2010). Enhancer intensity (H3K27ac signal) was downloaded from dbCoRC database, and enhancer peaks were from ReMap project (Chèneby et al., 2017; Huang et al., 2017). The gene expression score of each tissue was calculated by log2 TPM fold change between a tissue and the average of all other tissues. The GRN of each tissue was inferred using ANANSE. For prediction for TFs of one tissue, GRN interaction scores of all other tissues were averaged as the source GRN. All correlation analyses were clustered by hierarchical clustering method. The modular visualization of anatograms and tissues was done using the gganatogram package (version 1.1.1) (Maag, 2018).

Acknowledgements

We would like to thank the ENCODE DREAM challenge organizers for making consistently processed data available, and Jos Smits and Siebren Frölich for testing the ANANSE software package. We used the style from the Finkelstein lab at https://github.com/finkelsteinlab/BioRxiv-Template.

Notes

This article contains supporting information online.

Funding

QX was funded by the Chinese Scholarship Council (grant 201606230213). SJvH was supported by the Netherlands Organization for Scientific research (NWO grant 016.Vidi.189.081). Early work on this project was supported by a US National Institutes of Health grant (NICHD, R01HD069344) to GJCV.

Conflict of interest statement. None declared.

References

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.

Bani-Yaghoub, M., Tremblay, R.G., Lei, J.X., Zhang, D., Zurakowski, B., Sandhu, J.K., Smith, B., Ribecco-Lutkiewicz, M., Kennedy, J.,

Walker, P.R., *et al.* (2006). Role of Sox2 in the development of the mouse neocortex. Dev Biol *295*, 52-66.

Bao, X., Rubin, A.J., Qu, K., Zhang, J., Giresi, P.G., Chang, H.Y., and Khavari, P.A. (2015). A novel ATAC-seq approach reveals lineage-specific reinforcement of the open chromatin landscape via cooperation between BAF and p63. Genome biology *16*, 284.

Batsis, J.A., Boateng, G.G., Seo, L.M., Petersen, C.L., Fortuna, K.L., Wechsler, E.V., Peterson, R.J., Cook, S.B., Pidgeon, D., Dokko, R.S., *et al.* (2019). Development and Usability Assessment of a Connected Resistance Exercise Band Application for Strength-Monitoring. World Acad Sci Eng Technol *13*, 340-348.

Bruneau, B.G. (2013). Signaling and transcriptional networks in heart development and regeneration. Cold Spring Harb Perspect Biol *5*, a008292.

Bruse, N., and Heeringen, S.J.v. (2018). GimmeMotifs: an analysis framework for transcription factor motif analysis. BioRxiv.

Budirahardja, Y., Tan, P.Y., Doan, T., Weisdepp, P., and Zaidel-Bar, R. (2016). The AP-2 Transcription Factor APTF-2 Is Required for Neuroblast and Epidermal Morphogenesis in Caenorhabditis elegans Embryogenesis. PLoS genetics *12*, e1006048.

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.

Bulger, M., and Groudine, M. (2011). Functional and mechanistic diversity of distal transcription enhancers. Cell *144*, 327-339.

Buschbeck, M., and Hake, S.B. (2017). Variants of core histones and their roles in cell fate decisions, development and cancer. Nat Rev Mol Cell Biol *18*, 299-314.

Cahan, P., Li, H., Morris, S.A., Da Rocha, E.L., Daley, G.Q., and Collins, J.J. (2014). CellNet: network biology applied to stem cell engineering. Cell *158*, 903-915.

Caiazzo, M., Giannelli, S., Valente, P., Lignani, G., Carissimo, A., Sessa, A., Colasante, G., Bartolomeo, R., Massimino, L., and Ferroni, S. (2015). Direct conversion of fibroblasts into functional astrocytes by defined transcription factors. Stem cell reports *4*, 25-36.

Chen, S., and Mar, J.C. (2018). Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. BMC Bioinformatics *19*, 232-232.

Chen, Y., Mistry, D.S., and Sen, G.L. (2014). Highly rapid and efficient conversion of human fibroblasts to keratinocyte-like cells. J Invest Dermatol *134*, 335-344.

Chèneby, J., Gheorghe, M., Artufel, M., Mathelier, A., and Ballester, B. (2017). ReMap 2018: an updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments. Nucleic acids research *46*, D267-D275.

Corces, M.R., Buenrostro, J.D., Wu, B., Greenside, P.G., Chan, S.M., Koenig, J.L., Snyder, M.P., Pritchard, J.K., Kundaje, A., Greenleaf, W.J., *et al.* (2016). Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. Nature genetics *48*, 1193-1203.

Cunningham, F., Achuthan, P., Akanni, W., Allen, J., Amode, M R., Armean, I.M., Bennett, R., Bhai, J., Billis, K., Boddu, S., *et al.* (2018). Ensembl 2019. Nucleic Acids Research *47*, D745-D751.

D'Alessio, A.C., Fan, Z.P., Wert, K.J., Baranov, P., Cohen, M.A., Saini, J.S., Cohick, E., Charniga, C., Dadon, D., and Hannett, N.M. (2015). A systematic approach to identify candidate transcription factors that control cell identity. Stem cell reports *5*, 763-775.

Daniely, Y., Liao, G., Dixon, D., Linnoila, R.I., Lori, A., Randell, S.H., Oren, M., and Jetten, A.M. (2004). Critical role of p63 in the development of a normal esophageal and tracheobronchial epithelium. Am J Physiol, Cell Physiol *287*, C171-C181.

Davidson, E.H. (2010). Emerging properties of animal gene regulatory networks. Nature *468*, 911-920.

Davis, C.A., Hitz, B.C., Sloan, C.A., Chan, E.T., Davidson, J.M., Gabdank, I., Hilton, J.A., Jain, K., Baymuradov, U.K., Narayanan, A.K., *et al.* (2018). The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res *46*, D794-D801.

Davis, R.L., Weintraub, H., and Lassar, A.B. (1987). Expression of a single transfected cDNA converts fibroblasts to myoblasts. Cell *51*.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15-21.

Dollé, P.J.N.r.s. (2009). Developmental expression of retinoic acid receptors (RARs). Nuclear receptor signaling *7*, nrs. 07006.

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

ENCODE-DREAM (2017). ENCODE-DREAM in vivo Transcription Factor Binding Site Prediction Challenge. Available from: https://www.synapse.org/ENCODE.

Fu, J.-D., Stone, N.R., Liu, L., Spencer, C.I., Qian, L., Hayashi, Y., Delgado-Olguin, P., Ding, S., Bruneau, B.G., and Srivastava, D. (2013). Direct reprogramming of human fibroblasts toward a cardiomyocyte-like state. Stem cell reports *1*, 235-247.

Fulco, C.P., Nasser, J., Jones, T.R., Munson, G., Bergman, D.T., Subramanian, V., Grossman, S.R., Anyoha, R., Patwardhan, T.A., Nguyen, T.H., *et al.* (2019). Activity-by-Contact model of enhancer specificity from thousands of CRISPR perturbations. 529990.

Han, H., Cho, J.-W., Lee, S., Yun, A., Kim, H., Bae, D., Yang, S., Kim, C.Y., Lee, M., and Kim, E. (2017). TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. Nucleic acids research *46*, D380-D386.

Hartmann, A., Okawa, S., Zaffaroni, G., and Del Sol, A. (2018). SeesawPred: A Web Application for Predicting Cell-fate Determinants in Cell Differentiation. Sci Rep *8*, 13355.

Heinaniemi, M., Nykter, M., Kramer, R., Wienecke-Baldacchino, A., Sinkkonen, L., Zhou, J.X., Kreisberg, R., Kauffman, S.A., Huang, S., and Shmulevich, I. (2013). Gene-pair expression signatures reveal lineage control. Nat Methods *10*, 577-583.

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. Molecular cell *38*, 576-589.

Hsu, F., Kent, W.J., Clawson, H., Kuhn, R.M., Diekhans, M., and Haussler, D. (2006). The UCSC known genes. Bioinformatics *22*, 1036-1046.

Huang, M., Chen, Y., Yang, M., Guo, A., Xu, Y., Xu, L., and Koeffler, H.P. (2017). dbCoRC: a database of core transcriptional regulatory circuitries modeled by H3K27ac ChIP-seq signals. Nucleic acids research *46*, D71-D77.

Iwafuchi-Doi, M., and Zaret, K.S. (2016). Cell fate control by pioneer transcription factors. Development *143*, 1833-1837.

Jedlicka, P., Gutierrez-Hartmann, A.J.H., and histopathology (2008). Ets transcription factors in intestinal morphogenesis, homeostasis and disease. Histology and histopathology *23*, 1417.

Jopling, C., Boue, S., and Izpisua Belmonte, J.C. (2011). Dedifferentiation, transdifferentiation and reprogramming: three routes to regeneration. Nat Rev Mol Cell Biol *12*, 79-89.

Kathiriya, I.S., Nora, E.P., and Bruneau, B.G. (2015). Investigating the transcriptional control of cardiovascular development. Circulation research *116*, 700-714.

Katz, J.P., Perreault, N., Goldstein, B.G., Lee, C.S., Labosky, P.A., Yang, V.W., and Kaestner, K.H. (2002). The zinc-finger transcription factor Klf4 is required for terminal differentiation of goblet cells in the colon. Development *129*, 2619-2628.

Keilwagen, J., Posch, S., and Grau, J. (2019). Accurate prediction of cell type-specific transcription factor binding. Genome Biol *20*, 9.

Kempfer, R., and Pombo, A. (2019). Methods for mapping 3D chromosome architecture. Nat Rev Genet.

Kent, J., ENCODE DCC. (2014). kentUtils: Jim Kent command line bioinformatic utilities. Available from: https://github.com/ENCODE-DCC/kentUtils.

Kim, Y.J., Lim, H., Li, Z., Oh, Y., Kovlyagina, I., Choi, I.Y., Dong, X., and Lee, G. (2014). Generation of multipotent induced neural crest by direct reprogramming of human postnatal fibroblasts with a single transcription factor. Cell stem cell *15*, 497-506.

Kousa, Y.A., Fuller, E., and Schutte, B.C. (2018). IRF6 and AP2A Interaction Regulates Epidermal Development. The Journal of investigative dermatology *138*, 2578-2588.

Kurita, M., Araoka, T., Hishida, T., O'Keefe, D.D., Takahashi, Y., Sakamoto, A., Sakurai, M., Suzuki, K., Wu, J., and Yamamoto, M. (2018). In vivo reprogramming of wound-resident cells generates skin epithelial tissue. Nature *561*, 243.

Lambert, S.A., Jolma, A., Campitelli, L.F., Das, P.K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T.R., and Weirauch, M.T. (2018). The Human Transcription Factors. Cell *172*, 650-665.

Lang, A.H., Li, H., Collins, J.J., and Mehta, P. (2014). Epigenetic landscapes explain partially reprogrammed cells and identify key reprogramming genes. PLoS Comput Biol *10*, e1003734.

Li, H., Quang, D., and Guan, Y. (2019). Anchor: trans-cell type prediction of transcription factor binding sites. Genome Res *29*, 281-292.

Li, Q., Brown, J.B., Huang, H., and Bickel, P.J. (2011). Measuring reproducibility of high-throughput experiments. The annals of applied statistics *5*, 1752-1779.

Li, Y., Wang, Y., Yu, J., Ma, Z., Bai, Q., Wu, X., Bao, P., Li, L., Ma, D., and Liu, J. (2017). Direct conversion of human fibroblasts into osteoblasts and osteocytes with small molecules and a single factor, Runx2. bioRxiv, 127480.

Liu, Q., Jiang, C., Xu, J., Zhao, M.-T., Van Bortle, K., Cheng, X., Wang, G., Chang, H.Y., Wu, J.C., and Snyder, M.P. (2017). Genome-wide temporal profiling of transcriptome and open chromatin of early cardiomyocyte differentiation derived from hiPSCs and hESCs. Circulation research *121*, 376-391.

Liu, Z.-P., Wu, C., Miao, H., and Wu, H. (2015). RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. Database *2015*.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology *15*, 550.

Luo, Z., Gao, X., Lin, C., Smith, E.R., Marshall, S.A., Swanson, S.K., Florens, L., Washburn, M.P., and Shilatifard, A. (2015). Zic2 is an enhancer-binding factor required for embryonic stem cell specification. Molecular cell *57*, 685-694.

Maag, J.L.V. (2018). gganatogram: An R package for modular visualisation of anatograms and tissues based on ggplot2. F1000Res *7*, 1576-1576.

Meijer, D.H., Kane, M.F., Mehta, S., Liu, H., Harrington, E., Taylor, C.M., Stiles, C.D., and Rowitch, D.H. (2012). Separated at birth? The functional and molecular divergence of OLIG1 and OLIG2. Nat Rev Neurosci *13*, 819-831.

Morris, S.A., Cahan, P., Li, H., Zhao, A.M., San Roman, A.K., Shivdasani, R.A., Collins, J.J., and Daley, G.Q. (2014). Dissecting engineered cell types and enhancing cell fate conversion via CellNet. Cell *158*, 889-902.

Muto, A., Iida, A., Satoh, S., and Watanabe, S. (2009). The group E Sox genes Sox8 and Sox9 are regulated by Notch signaling and are required for Muller glial cell development in mouse retina. Exp Eye Res *89*, 549-558.

Nair, M., Teng, A., Bilanchone, V., Agrawal, A., Li, B., and Dai, X. (2006). Ovol1 regulates the growth arrest of embryonic epidermal progenitor cells and represses c-myc transcription. J Cell Biol *173*, 253-264.

Novakovic, B., Habibi, E., Wang, S.-Y., Arts, R.J., Davar, R., Megchelenbrink, W., Kim, B., Kuznetsova, T., Kox, M., and Zwaag, J. (2016). β-Glucan reverses the epigenetic state of LPS-induced immunological tolerance. Cell *167*, 1354-1368. e1314.

Obayashi, T., Kagaya, Y., Aoki, Y., Tadaka, S., and Kinoshita, K. (2019). COXPRESdb v7: a gene coexpression database for 11 animal species supported by 23 coexpression platforms for technical evaluation and evolutionary inference. Nucleic Acids Res *47*, D55-D62.

Oberbeck, N., Pham, V.C., Webster, J.D., Reja, R., Huang, C.S., Zhang, Y., Roose-Girma, M., Warming, S., Li, Q., Birnberg, A*., et al.* (2019). The RIPK4-IRF6 signalling axis safeguards epidermal differentiation and barrier function. Nature *574*, 249-253.

Ohteki, T., Maki, C., and Koyasu, S. (2001). Overexpression of Bcl-2 Differentially Restores Development of Thymus-Derived CD4[−]8⁺ T Cells and Intestinal Intraepithelial T Cells in IFN-Regulatory Factor-1-Deficient Mice. *166*, 6509-6513.

Olivares-Chauvet, P., Mukamel, Z., Lifshitz, A., Schwartzman, O., Elkayam, N.O., Lubling, Y., Deikus, G., Sebra, R.P., and Tanay, A. (2016). Capturing pairwise and multi-way chromosomal conformations using chromosomal walks. Nature *540*, 296-300.

Pang, Z.P., Yang, N., Vierbuchen, T., Ostermeier, A., Fuentes, D.R., Yang, T.Q., Citri, A., Sebastiano, V., Marro, S., Südhof, T.C*., et al.* (2011). Induction of human neuronal cells by defined transcription factors. Nature *476*, 220-223.

Patro, R., Duggal, G., Love, M.I., Irizarry, R.A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. Nature methods *14*, 417-419.

Peñalosa-Ruiz, G., Bright, A.R., Mulder, K.W., and Veenstra, G.J.C. (2019). The interplay of chromatin and transcription factors during cell fate transitions in development and reprogramming. Biochim Biophys Acta Gene Regul Mech *1862*, 194407.

Qu, J., Tanis, S.E., Smits, J.P., Kouwenhoven, E.N., Oti, M., van den Bogaard, E.H., Logie, C., Stunnenberg, H.G., van Bokhoven, H., and Mulder, K.W. (2018). Mutant p63 affects epidermal cell identity through rewiring the enhancer landscape. Cell reports *25*, 3490-3503. e3494.

Quang, D., and Xie, X. (2019). FactorNet: A deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data. Methods *166*, 40-47.

Rackham, O.J., Firas, J., Fang, H., Oates, M.E., Holmes, M.L., Knaupp, A.S., Suzuki, H., Nefzger, C.M., Daub, C.O., and Shin, J.W. (2016). A predictive computational framework for direct reprogramming between human cell types. Nature genetics *48*, 331.

Reik, W., Dean, W., and Walter, J. (2001). Epigenetic reprogramming in mammalian development. Science (New York, NY) *293*, 1089-1093.

Richardson, R.J., Dixon, J., Malhotra, S., Hardman, M.J., Knowles, L., Boot-Handford, R.P., Shore, P., Whitmarsh, A., and Dixon, M.J. (2006). Irf6 is a key determinant of the keratinocyte proliferation-differentiation switch. Nature genetics *38*, 1329-1334.

Roost, M.S., van Iperen, L., Ariyurek, Y., Buermans, H.P., Arindrarto, W., Devalla, H.D., Passier, R., Mummery, C.L., Carlotti, F., de Koning, E.J*., et al.* (2015). KeyGenes, a Tool to Probe Tissue Differentiation Using a Human Fetal Transcriptional Atlas. Stem Cell Reports *4*, 1112-1124.

Ruberte, E., Dolle, P., Krust, A., Zelent, A., Morriss-Kay, G., and Chambon, P. (1990). Specific spatial and temporal distribution of retinoic acid receptor gamma transcripts during mouse embryogenesis. *108*, 213-222.

Saito, T., and Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One *10*, e0118432.

Sen, G.L., Boxer, L.D., Webster, D.E., Bussat, R.T., Qu, K., Zarnegar, B.J., Johnston, D., Siprashvili, Z., and Khavari, P.A. (2012). ZNF750 is a p63 target gene that induces KLF4 to drive terminal epidermal differentiation. Dev Cell *22*, 669-677.

Sergushichev, A. (2016). An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. BioRxiv, 060012.

Simeonov, K.P., and Uppal, H. (2014). Direct reprogramming of human fibroblasts to hepatocyte-like cells by synthetic modified mRNAs. PloS one *9*, e100134.

Soares, E., Xu, Q., Li, Q., Qu, J., Zheng, Y., Raeven, H.H., Brandao, K.O., Petit, I., van den Akker, W.M., and van Heeringen, S.J. (2019). Single-cell RNA-seq identifies a reversible mesodermal activation in abnormally specified epithelia of p63 EEC syndrome. Proceedings of the National Academy of Sciences *116*, 17361-17370.

Soneson, C., Love, M.I., and Robinson, M.D. (2015). Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Res *4*, 1521.

Spitz, F., and Furlong, E.E.M. (2012). Transcription factors: from enhancer binding to developmental control. Nat Rev Genet *13*, 613-626.

Stadhouders, R., Filion, G.J., and Graf, T. (2019). Transcription factors and 3D genome conformation in cell-fate decisions. Nature *569*, 345-354.

Stefanovic, S., and Christoffels, V.M. (2015). GATA-dependent transcriptional and epigenetic control of cardiac lineage specification and differentiation. Cell Mol Life Sci *72*, 3871-3881.

Takahashi, K., Tanabe, K., Ohnuki, M., Narita, M., Ichisaka, T., Tomoda, K., and Yamanaka, S. (2007). Induction of pluripotent stem cells from adult human fibroblasts by defined factors. Cell *131*, 861-872.

Tchieu, J., Calder, E.L., Guttikonda, S.R., Gutzwiller, E.M., Aromolaran, K.A., Steinbeck, J.A., Goldstein, P.A., and Studer, L. (2019). NFIA is a gliogenic switch enabling rapid derivation of functional human astrocytes from pluripotent stem cells. Nature biotechnology *37*, 267.

Tegner, J., and Bjorkegren, J. (2007). Perturbations to uncover gene networks. Trends Genet *23*, 34-41.

The Gene Ontology, C. (2019). The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Res *47*, D330-D338.

Thompson, C.A., DeLaForest, A., and Battle, M.A. (2018). Patterning the gastrointestinal epithelium to confer regional-specific functions. Dev Biol *435*, 97-108.

Tsuji, G., Ito, T., Chiba, T., Mitoma, C., Nakahara, T., Uchi, H., and Furue, M. (2018). The role of the OVOL1-OVOL2 axis in normal and diseased human skin. J Dermatol Sci *90*, 227-231.

Uhlén, M., Fagerberg, L., Hallström, B.M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., and Asplund, A. (2015). Tissue-based map of the human proteome. Science *347*, 1260419.

Uhlen, M., Oksvold, P., Fagerberg, L., Lundberg, E., Jonasson, K., Forsberg, M., Zwahlen, M., Kampf, C., Wester, K., and Hober, S. (2010). Towards a knowledge-based human protein atlas. Nature biotechnology *28*, 1248.

Valouev, A., Johnson, D.S., Sundquist, A., Medina, C., Anton, E., Batzoglou, S., Myers, R.M., and Sidow, A. (2008). Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. Nature methods *5*, 829.

van der Raadt, J., van Gestel, S.H., Nadif Kasri, N., and Albers, C.A. (2019). ONECUT transcription factors induce neuronal characteristics and remodel chromatin accessibility. Nucleic acids research *47*, 5587-5602.

van Heeringen, S.J., and Veenstra, G.J.C. (2010). GimmeMotifs: a de novo motif prediction pipeline for ChIP-sequencing experiments. Bioinformatics *27*, 270-271.

Vaquerizas, J.M., Kummerfeld, S.K., Teichmann, S.A., and Luscombe, N.M. (2009). A census of human transcription factors: function, expression and evolution. Nat Rev Genet *10*, 252-263.

Vierbuchen, T., Ostermeier, A., Pang, Z.P., Kokubu, Y., Südhof, T.C., and Wernig, M. (2010). Direct conversion of fibroblasts to functional neurons by defined factors. Nature *463*, 1035-1041.

Villar, D., Berthelot, C., Aldridge, S., Rayner, T.F., Lukk, M., Pignatelli, M., Park, T.J., Deaville, R., Erichsen, J.T., Jasinska, A.J.*, et al.* (2015). Enhancer evolution across 20 mammalian species. Cell *160*, 554-566.

Wang, S., Zang, C., Xiao, T., Fan, J., Mei, S., Qin, Q., Wu, Q., Li, X., Xu, K., and He, H.H. (2016). Modeling cis-regulation with a compendium of genome-wide histone H3K27ac profiles. Genome research *26*, 1417-1429.

Wang, X., Pasolli, H.A., Williams, T., and Fuchs, E. (2008). AP-2 factors act in concert with Notch to orchestrate terminal differentiation in skin epidermis. J Cell Biol *183*, 37-48.

Weirauch, M.T., Yang, A., Albu, M., Cote, A.G., Montenegro-Montero, A., Drewe, P., Najafabadi, H.S., Lambert, S.A., Mann, I., and Cook, K. (2014). Determination and inference of eukaryotic transcription factor sequence specificity. Cell *158*, 1431-1443.

Wilanowski, T., Caddy, J., Ting, S.B., Hislop, N.R., Cerruti, L., Auden, A., Zhao, L.L., Asquith, S., Ellis, S., Sinclair, R.*, et al.* (2008). Perturbed desmosomal cadherin expression in grainy head-like 1-null mice. EMBO J *27*, 886-897.

Wilkinson, A.C., Nakauchi, H., and Göttgens, B. (2017). Mammalian Transcription Factor Networks: Recent Advances in Interrogating Biological Complexity. Cell Syst *5*, 319-331.

Xie, H., Ye, M., Feng, R., and Graf, T. (2004). Stepwise reprogramming of B cells into macrophages. Cell *117*, 663-676.

Xiong, Z., Ren, S., Chen, H., Liu, Y., Huang, C., Zhang, Y.L., Odera, J.O., Chen, T., Kist, R., Peters, H.*, et al.* (2018). PAX9 regulates squamous cell differentiation and carcinogenesis in the oro-oesophageal epithelium. J Pathol *244*, 164-175.

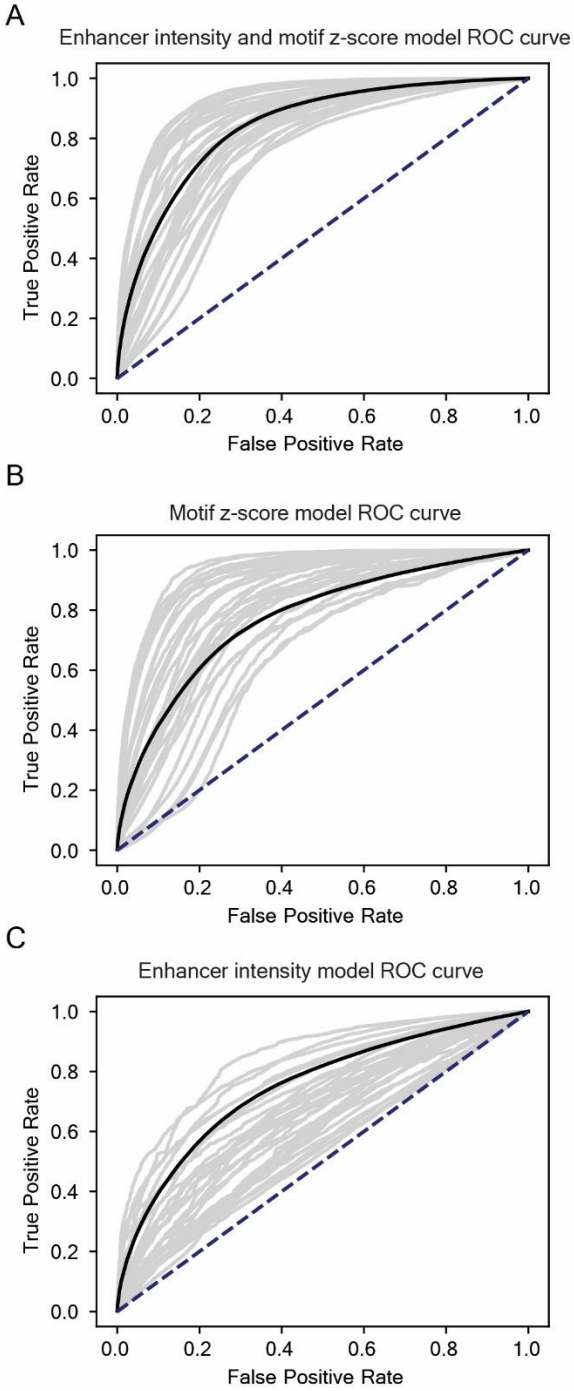
Yu, G., Wang, L.-G., and He, Q.-Y. (2015). ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. Bioinformatics (Oxford, England) *31*, 2382-2383.

Zhang, Y., Li, T., Preissl, S., Grinstein, J., Farah, E., Destici, E., Lee, A.Y., Chee, S., Qiu, Y., and Ma, K. (2019). 3D Chromatin Architecture Remodeling during Human Cardiomyocyte Differentiation Reveals A Role Of HERV-H In Demarcating Chromatin Domains. bioRxiv, 485961.

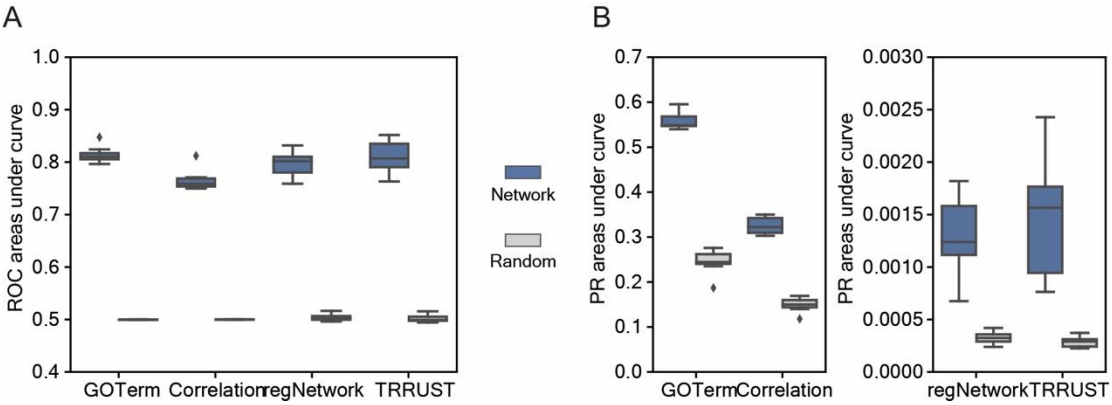
Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., and Li, W. (2008). Model-based analysis of ChIP-Seq (MACS). Genome biology *9*, R137.

Zhao, M.-T., Shao, N.-Y., Hu, S., Ma, N., Srinivasan, R., Jahanbani, F., Lee, J., Zhang, S.L., Snyder, M.P., and Wu, J.C. (2017). Cell type-specific chromatin signatures underline regulatory DNA elements in human induced pluripotent stem cells and somatic cells. Circulation research *121*, 1237-1250.

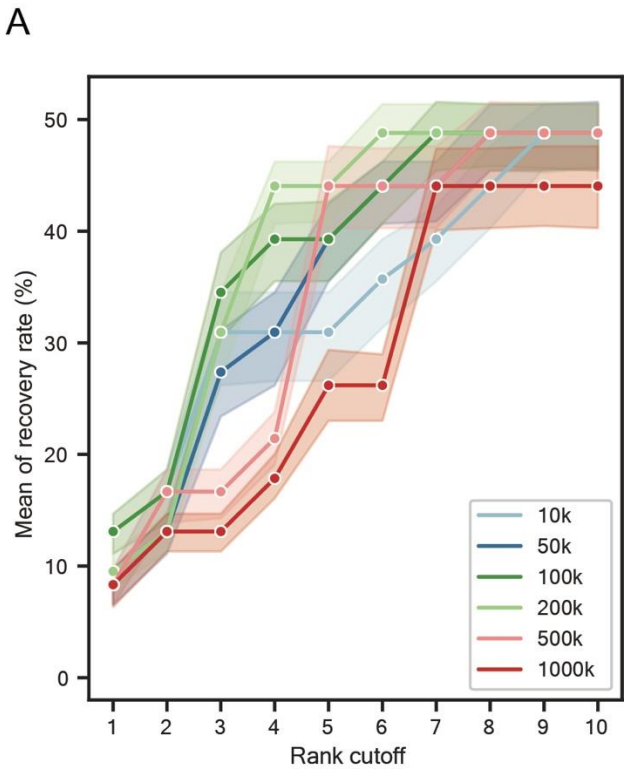
Supplementary figures



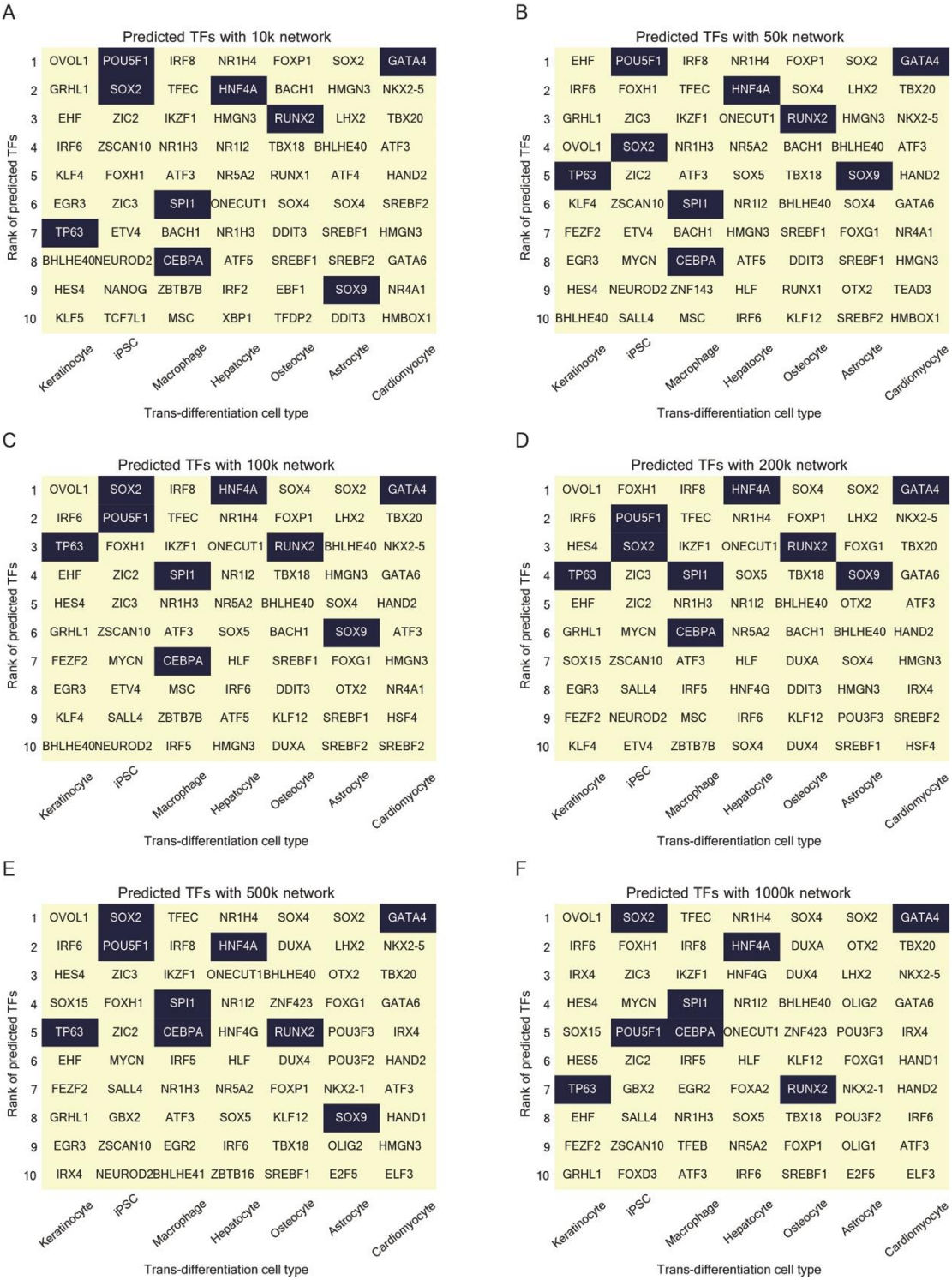
Supplementary Figure S1. ROC curve analysis of TF binding prediction performance.
(A) ROC curve of TF binding prediction based on enhancer activities. Each line represents one TF prediction result based on the model trained on data from all the other 28 TFs. **(B)** ROC curve of TF binding prediction based on motif z-scores. **(C)** ROC curve of TF binding prediction based on enhancer intensities together with motif z-scores.



Supplementary Figure S2. Evaluation of cell type-specific enhancer gene regulatory networks.
(A) Evaluation of the predicted networks using different types of data: a gene ontology (GO) term co-regulatory network, a cell type-specific gene correlation network, and two TF-Gene regulatory networks based on interaction databases (regNetwork and TRRUST). This is the same evaluation as presented in Figure 4C-D, but using a TPM cutoff of 1 instead of 10. The boxplots show the AUC of ROC for 8 different cell type. ROC AUC of the predicted networks is shown in blue; the random networks are indicated in green. **(B)** The same evaluation as in A), with the PR AUC shown as a boxplot.

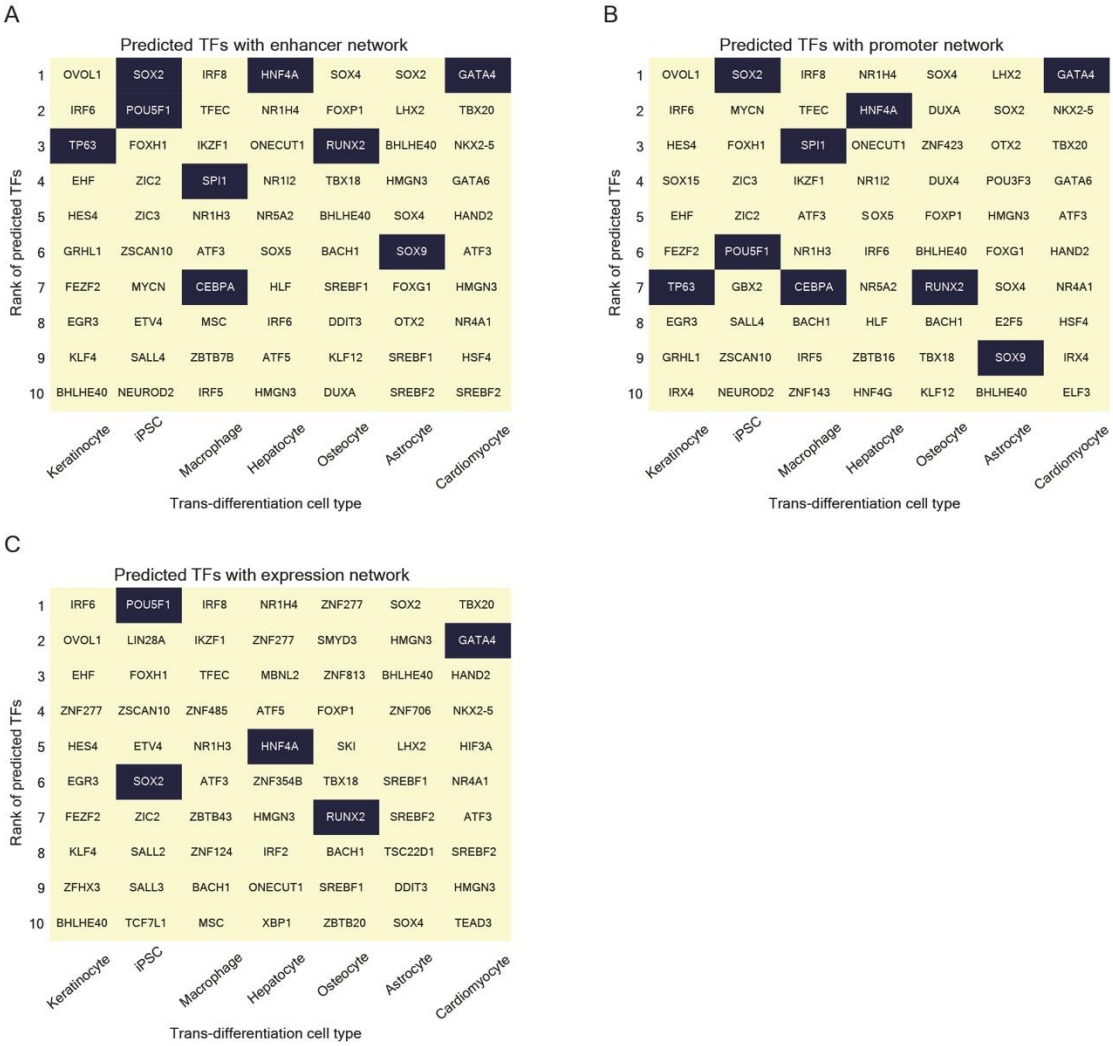


Supplementary Figure S3. Comparison of different GRN sizes used in the ANANSE prediction in seven experimentally validated trans-differentiation strategies.
(A) The line plots show the comparison of the predicted key TFs for six different sizes of GRNs Shown is the fraction of predicted TFs compared to all known TFs based on trans-differentiation protocols described in the literature (y-axis) as a function of the top number of TFs selected (x-axis). The shaded area represents the minimum and maximum percentage of corresponding recovered TFs when using seven out of eight trans-differentiations.

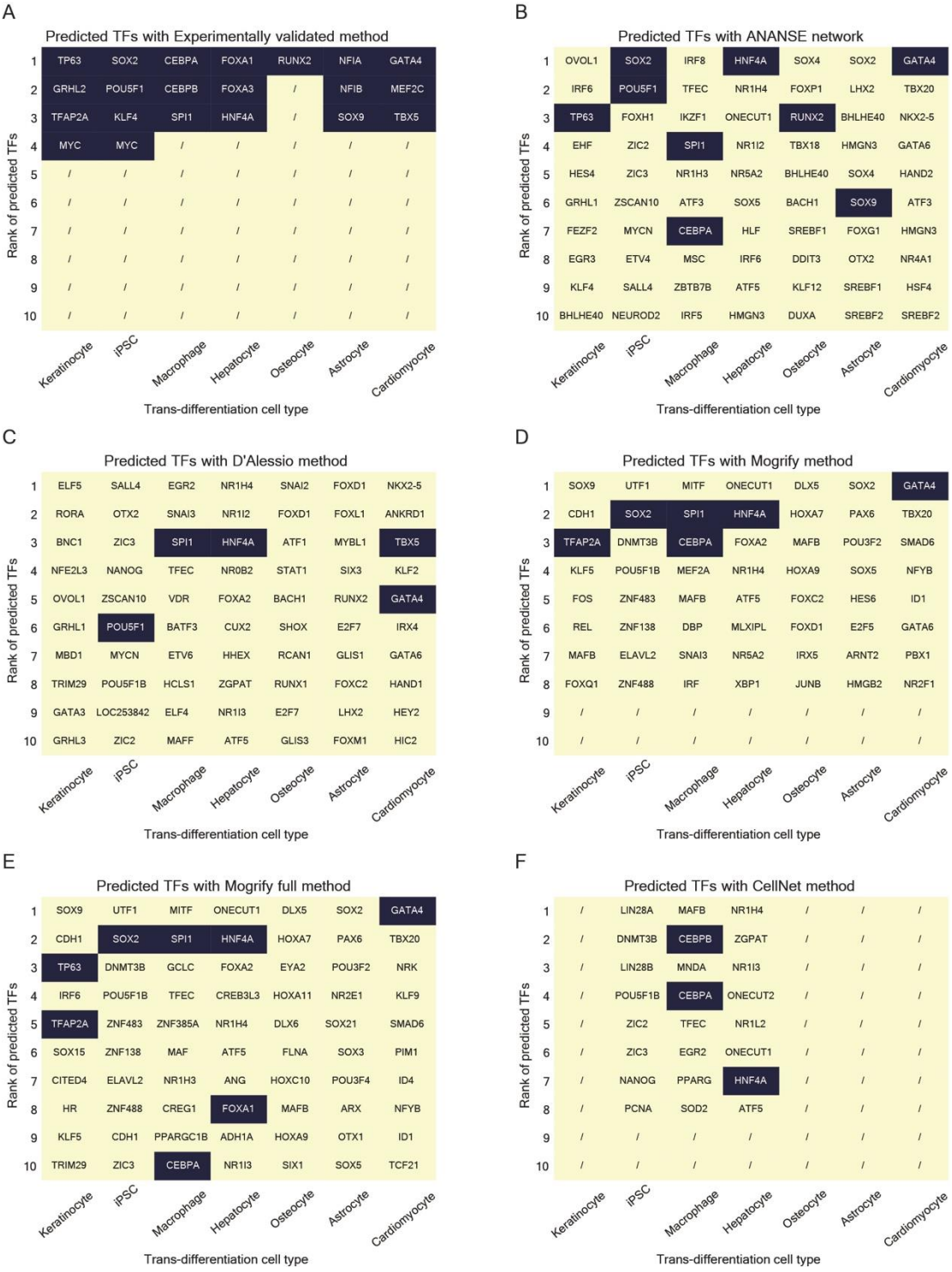


Supplementary Figure S4. Comparison of top 10 key TFs predicted by different GRN sizes in seven experimentally validated trans-differentiation strategies.

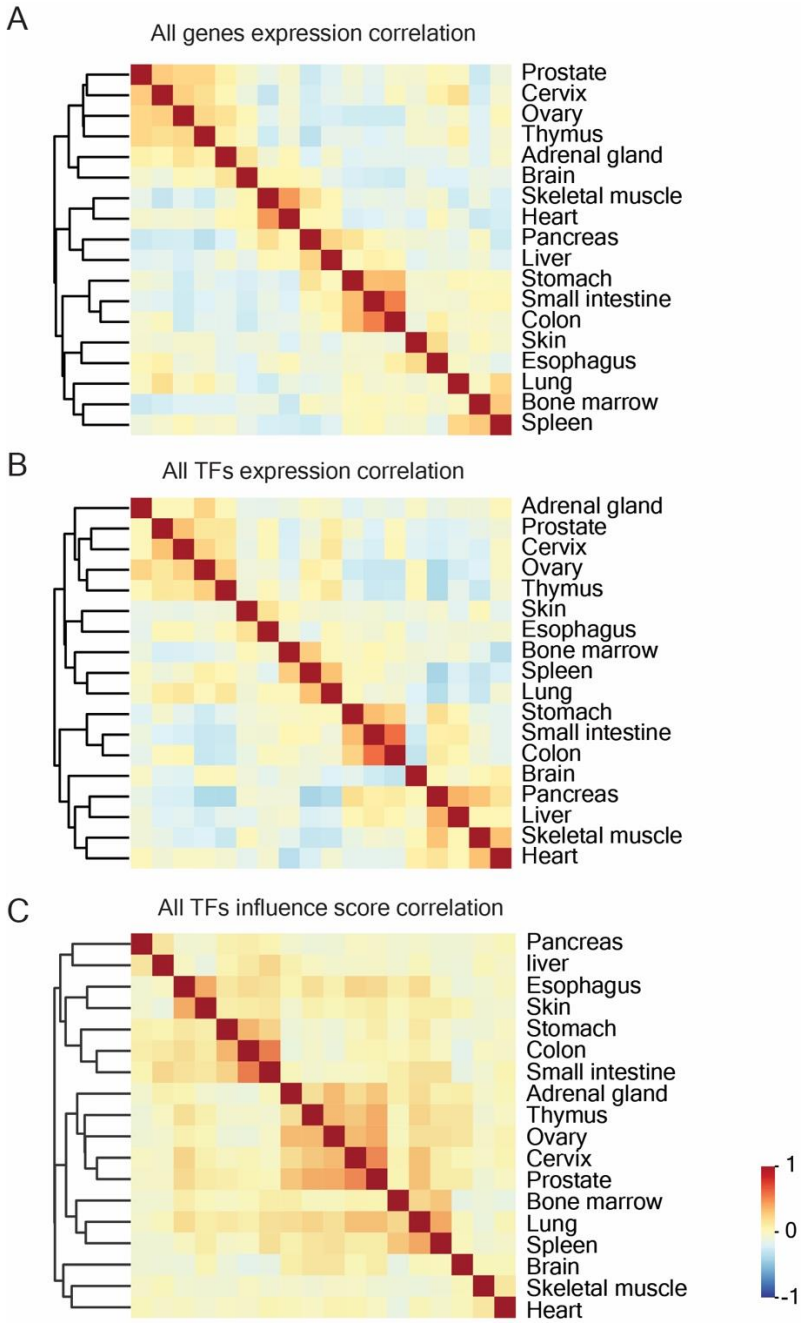
The x-axis shows seven experimentally validated trans-differentiations, and the y-axis shows the top 10 predicted key TFs ranked by their influence score. Black boxes highlight the TFs that were used in trans-differentiation experiments. **(A)** 10k network. **(B)** 50k network. **(C)** 100k network. **(D)** 200k network. **(E)** 500k network. **(F)** 1000k network.



Supplementary Figure S5. Comparison of top 10 key TFs predicted by different GRNs (enhancer, promoter and expression) in seven experimentally validated trans-differentiation strategies.
The x-axis shows seven experimentally validated trans-differentiations, and the y-axis shows the top 10 predicted key TFs ranked by their influence score. Black boxes highlight the TFs that were used in trans-differentiation experiments. **(A)** The results for ANANSE, based on a GRN that was inferred using peaks, regardless of their genomic location (includes both promoters and enhancers). **(B)** The results for ANANSE, based on a GRN that was inferred using the highest peak in the gene promoter, defined as < 2kb from the gene transcription start site. **(C)** The results for ANANSE, based on a GRN that was inferred using only the expression levels of the TFs and target genes.



Supplementary Figure S6. Comparison of top 10 key TFs predicted by different methods in seven experimentally validated trans-differentiation strategies.
The x-axis shows seven experimentally validated trans-differentiations, and the y-axis shows the top 10 predicted key TFs ranked by their influence score. Black boxes highlight the TFs that were used in trans-differentiation experiments. **(A)** All experimentally validated TFs. **(B)** ANANSE. **(C)** D'Alessio. **(D)** Mogrify. **(E)** Mogrify full list. **(F)** CellNet.



Supplementary Figure S7. Classification of the human gene expression and TF influence score.
(A) Heatmap showing the pairwise correlation between all 18 tissues based on gene expression. The colors in the heatmap indicate high (red) or low (blue) correlation across the tissue set. **(B)** Heatmap showing the pairwise correlation between all 18 tissues based on the expression of all TFs. **(C)** Heatmap showing the pairwise correlation between all 18 tissues based on TF influence scores.