



# Total Functional Score of Enhancer Elements Identifies Lineage-Specific Enhancers That Drive Differentiation of Pancreatic Cells

Bioinformatics and Biology Insights  
Volume 14: 1–12  
© The Author(s) 2020  
Article reuse guidelines:  
sagepub.com/journals-permissions  
DOI: 10.1177/1177932220938063



Venkat S. Malladi<sup>1,2</sup> , Anusha Nagari<sup>1</sup>, Hector L Franco<sup>1,3</sup>   
and W Lee Kraus<sup>1</sup> 

<sup>1</sup>Laboratory of Signaling and Gene Regulation, Cecil H. and Ida Green Center for Reproductive Biology Sciences, The University of Texas Southwestern Medical Center, Dallas, TX, USA.

<sup>2</sup>Department of Bioinformatics, The University of Texas Southwestern Medical Center, Dallas, TX, USA. <sup>3</sup>Department of Genetics and Lineberger Comprehensive Cancer Center, The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA.

**ABSTRACT:** The differentiation of embryonic stem cells into various lineages is highly dependent on the chromatin state of the genome and patterns of gene expression. To identify lineage-specific enhancers driving the differentiation of progenitors into pancreatic cells, we used a previously described computational framework called Total Functional Score of Enhancer Elements (TFSEE), which integrates multiple genomic assays that probe both transcriptional and epigenomic states. First, we evaluated and compared TFSEE as an enhancer-calling algorithm with enhancers called using GRO-seq-defined enhancer transcripts (method 1) versus enhancers called using histone modification ChIP-seq data (method 2). Second, we used TFSEE to define the enhancer landscape and identify transcription factors (TFs) that maintain the multipotency of a subpopulation of endodermal stem cells during differentiation into pancreatic lineages. Collectively, our results demonstrate that TFSEE is a robust enhancer-calling algorithm that can be used to perform multilayer genomic data integration to uncover cell type-specific TFs that control lineage-specific enhancers.

**KEYWORDS:** Enhancer, epigenome, gene regulation, pancreas, tissue-specific transcription, transcription factor

**RECEIVED:** May 25, 2020. **ACCEPTED:** June 2, 2020.

**TYPE:** Methods and Protocols

**FUNDING:** The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the following: (1) a grant from the Cancer Prevention and Research Institute of Texas (CPRIT; RP100417) to the LONESTAR Oncology Network for Epigenetics Therapy and Research, (2) a grant from CPRIT (RP160319) to W.L.K., (3) a grant from the NIH/NIDDK (R01 DK058110) to W.L.K., and (4) funds from the Cecil H. and Ida Green Center for Reproductive Biology Sciences Endowment to W.L.K.

**DECLARATION OF CONFLICTING INTERESTS:** The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**CORRESPONDING AUTHOR:** W Lee Kraus, Laboratory of Signaling and Gene Regulation, Cecil H. and Ida Green Center for Reproductive Biology Sciences, The University of Texas Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, TX 75390-8511, USA. Email: LEE.KRAUS@utsouthwestern.edu

## Introduction

Embryonic development is a complex process during which pluripotent cells differentiate into specialized cells of various lineages. The mechanisms underlying developmental competence are highly complex and are dependent on the tissue type and precise timing of signaling cues.<sup>1</sup> Lineage specification is dependent on the interactions of transcription factors (TFs) and chromatin states at enhancers. Enhancers and the TFs regulating their formation have been shown to play an important role in cell type-specific activation of gene expression.<sup>2,3</sup> Although thousands of potential enhancers have been identified in cell types derived from various lineages and tissues, identification of the enhancers that are active (versus inactive or poised) remains a major challenge.<sup>4</sup> In addition, the ability to identify the TFs acting at numerous enhancers in each cell type is challenging.<sup>5,6</sup>

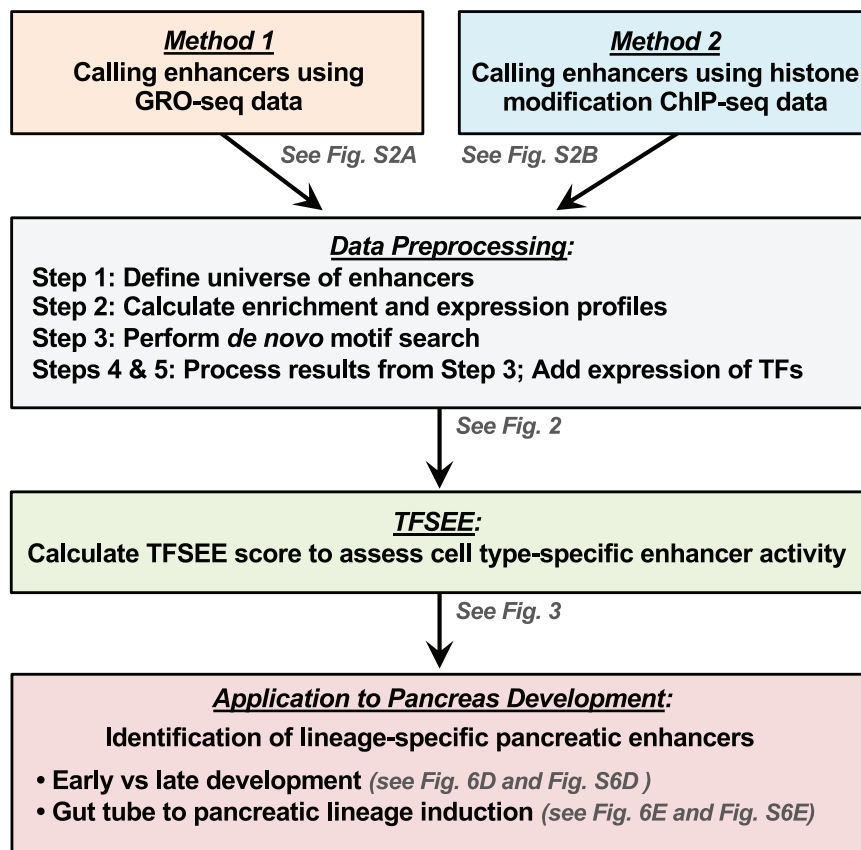
Enhancers have been shown to share several common features, such as increased chromatin accessibility (as measured by DNase-seq or ATAC-seq)<sup>7-9</sup> and enrichment of posttranslational modification of the amino-terminal tails of core histone proteins (as assessed by ChIP-seq), including histone H3 lysine 4 monomethyl (H3K4me1) and histone H3 lysine 27 acetyl (H3K27ac).<sup>10-12</sup> While these epigenomic features can reveal

the location of many enhancers across the genome, they cannot readily differentiate between active and inactive enhancers.<sup>13,14</sup> Recent genomic assays have shown that active enhancers are bound by RNA polymerase II (Pol II) and are transcribed, producing noncoding RNAs known as enhancer RNAs (eRNAs).<sup>14-16</sup> While the full breadth of functions of eRNAs are unknown, we and others have shown that enhancer transcription (as measured by total RNA-seq, GRO-seq, or PRO-seq) can be used in the absence of any other genomic information to predict enhancer activity.<sup>3,15-23</sup>

In recent years, advances in technology have facilitated the large-scale functional characterization of enhancer activity<sup>22,24-26</sup> and the annotation of TF-binding sites (TFBSs) genome-wide in various cell types and tissues.<sup>5,27</sup> However, due to the large number of cell types, TFs, and experimental conditions,<sup>28</sup> integration of these independent data sets to achieve a comprehensive analysis of gene expression and actionable predictions of TFs driving cell type-specific gene expression is challenging. Analyses that predict TFBSs, which are usually 4 to 12 nucleotides in length,<sup>29</sup> using TF-binding profile databases<sup>29-31</sup> fail to consider that such sequences occur frequently by chance throughout the genome and that TF binding is cell type specific.<sup>32</sup> To overcome these limitations, we previously



Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (<https://creativecommons.org/licenses/by-nc/4.0/>) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (<https://us.sagepub.com/en-us/nam/open-access-at-sage>).



**Figure 1.** Flowchart of data analysis and inputs into TFSEE. Flowchart describing the preprocessing steps, methods, and subsequent analysis described in this article. TFSEE accepts enhancer calls from different inputs: Method 1, enhancers called using enhancer transcription based on GRO-seq data. Method 2, enhancers called using enrichment of histone modifications. TF indicates transcription factors; TFSEE, Total Functional Score of Enhancer Elements.

established a computational pipeline and tool called Total Functional Score of Enhancer Elements (TFSEE), which can be used to identify location and activity of enhancers in any cell or tissue type together with their cognate TFs.<sup>33</sup>

In this study, we aimed to (1) evaluate TFSEE as an enhancer-calling algorithm and (2) understand the TF-driven transcriptional programs differentiating human embryonic stem cells (hESCs) into pancreatic cells.<sup>1,34</sup> This developmental model allowed us to explore spatiotemporal gene regulation during development by enhancers and TFs. In the studies presented herein, we provide a detailed characterization of TFSEE and demonstrate the broader use of TFSEE to identify enhancers and TFs during the differentiation of embryonic stem cells into pancreatic progenitor cells to uncover cell type-specific TFs that control lineage-specific enhancers (Figure 1).

## Materials and Methods

### Genomic data curation

We used previously published GRO-seq, ChIP-seq, and RNA-seq data from<sup>1,34</sup> time course differentiation of hESCs to pancreatic endoderm (PE). All data sets are available from NCBI's Gene Expression Omnibus (<https://www.ncbi.nlm.nih.gov/geo/>) or EMBL-EBI's ArrayExpress (<http://www.ebi.ac.uk/>

[arrayexpress/](http://arrayexpress/)) repositories using the accession numbers listed in Table S1.

### Analysis of ChIP-seq data

The raw reads were aligned to the human reference genome (GRCh37/hg19) using default parameters in Bowtie version 1.0.0.<sup>35</sup> The aligned reads were subsequently filtered for quality and uniquely mappable reads were retained for further analysis using Samtools version 0.1.19<sup>36</sup> and Picard version 1.127 (<http://broadinstitute.github.io/picard/>). Library complexity was measured using BEDTools version 2.17.0<sup>37</sup> and meets ENCODE data quality standards.<sup>38</sup> Relaxed peaks were called using MACS version 2.1.0<sup>39</sup> with a  $P$  value of  $1 \times 10^{-2}$  for each replicate, pooled replicates' reads, and pseudoreplicates. Peak calls from the pooled replicates that were observed in either both replicates or in both pseudoreplicates were used for subsequent analysis.

### Analysis of RNA-seq data

The raw reads were aligned to the human reference genome (GRCh37/hg19) using default parameters in STAR version 2.4.2a.<sup>40</sup> Quantification of genes against Gencode version 19<sup>41</sup>

annotations was done using default parameters in RSEM version 1.2.31.<sup>42</sup>

### *Analysis of GRO-seq data*

The GRO-seq reads were trimmed to the first 36 bases to trim adapter and low-quality sequence, using default parameters of `fastx_trimmer` in `fastx-toolkit` version 0.0.13.2 ([http://hannon-lab.cshl.edu/fastx\\_toolkit/](http://hannon-lab.cshl.edu/fastx_toolkit/)). The trimmed reads were aligned to the human reference genome (GRCh37/hg19) using default parameters in `BWA` version 0.7.12.<sup>36</sup>

### *Kernel density*

Kernel density plot representations were used to express the univariate distribution of ChIP-seq reads under peaks, RNA-seq reads for protein-coding genes, and GRO-seq reads for short paired and short unpaired eRNAs. The kernel density plots were calculated in Python (ver. 2.7.11) using the `kdeplot` function from `seaborn` version 0.7.1 (<http://seaborn.pydata.org/>) with default parameters.

### *Defining transcription start sites and promoters*

We made distinct transcription start sites (TSSs) for protein-coding genes from Gencode version 19<sup>41</sup> annotations using `MakeGencodeTSS` (<https://github.com/sdjebali/MakeGencodeTSS>). We identified active promoters using enrichment of H3K4me3.<sup>43</sup> An RPKM cutoff of  $\geq 1$  for H3K4me3 in at least one cell line was used to identify a peak as an active enhancer (Figure S1A).

### *Enhancer calling by GRO-seq*

*Calling a universe of transcripts from GRO-seq data.* Transcript calling was performed using a 2-state hidden Markov model using the `groHMM` data analysis package version 3.4<sup>16,21</sup> (<https://bioconductor.org/packages/release/bioc/html/groHMM.html>) on each individual cell line. The negative log transition probability of the switch between transcribed state to nontranscribed state and the variance in read counts in the nontranscribed state that are used to predict the transcription units for the cell lines in this study are listed in Table S2. We then built a universe of transcripts by merging the `groHMM`-called transcripts from individual cell lines and stratifying the boundaries to remove overlaps/redundancies occurring from the union of all transcripts.

*Calling active enhancers using GRO-seq-defined enhancer transcripts.* We filtered and collected a subset of short intergenic transcripts  $< 9$  kb in length and  $> 3$  kb away from known TSSs of protein-coding genes from Gencode version 19 annotations<sup>41</sup> and H3K4me3 peaks. These were further classified into (1) short paired eRNAs and (2) short unpaired eRNAs as described previously.<sup>19</sup> For the short paired eRNAs, the sum of

the GRO-seq RPKM values for both strands of DNA was used to determine whether an enhancer transcript pair is expressed using a cutoff of  $\text{RPKM} \geq 0.5$  (Figure S1B). An RPKM cutoff of  $\geq 1$  was used to determine the universe of expressed short unpaired eRNAs (Figure S1C). The comprehensive universe of expressed eRNAs (short paired and short unpaired) assembled using the cutoffs noted above for each cell line was used for further analyses.

*Motif analyses for GRO-seq-defined enhancers.* De novo motif analyses was performed on a 1 kb region ( $\pm 500$  bp [base pairs]) surrounding the overlap center or the TSS for short paired and short unpaired eRNAs, respectively, using the command-line version of MEME from MEME Suite version 4.11.1.<sup>44</sup> The following parameters were used for motif prediction: (1) zero or one occurrence per sequence (`-mod zoops`); (2) number of motifs (`-nmotifs 15`); (3) minimum, maximum width of the motif (`-minw 8, -maxw 15`); and (4) search for motif in given strand and reverse complement strand (`-revcomp`). The predicted motifs from MEME were matched to known motifs in the JASPAR database (JASPAR\_CORE\_2016 Vertebrates. `meme`)<sup>30</sup> using TOMTOM.<sup>31</sup>

### *Enhancer calling by ChIP-seq*

*Calling active enhancers using histone modification ChIP-seq data.* We built a universe of peak calls by merging the peaks from individual cell lines for histone modifications (H3K4me1 and H3K27ac) and stratifying the boundaries to remove overlaps/redundancies occurring from the union of all peaks. Potential enhancers were defined as peaks that were  $> 3$  kb from known TSSs, protein-coding genes from Gencode version 19 annotations,<sup>41</sup> and H3K4me3 peaks. An RPKM cutoff of  $\geq 1$  for H3K4me1 and H3K27ac (Figure S1D and E) in at least one cell line was used to identify a peak as an active enhancer. The universe of active enhancers was assembled using the cutoffs noted above for each cell line and was used for further analyses.

*Motif analyses for ChIP-seq-defined enhancers.* De novo motif analyses were performed on a 1 kb region ( $\pm 500$  bp) surrounding the peak summit for the top 10,000 enhancers, using the command-line version of MEME-ChIP from MEME Suite version 4.11.1.<sup>44,45</sup> The following parameters were used for motif prediction: (1) zero or one occurrence per sequence (`-mod zoops`); (2) number of motifs (`-nmotifs 15`); (3) minimum, maximum width of the motif (`-minw 8, -maxw 15`). All the other parameters were set at the default. The predicted motifs from MEME were matched to known motifs in the JASPAR database (JASPAR\_CORE\_2016 Vertebrates. `meme`)<sup>30</sup> using TOMTOM.<sup>31</sup>

### *Generating heatmaps and clusters*

For each cell line, the functional scores were  $Z$ -score normalized. To identify cognate TFs by cell type, we performed

hierarchical clustering by calculating the Euclidean distance using clustermap from seaborn version 0.7.1 (<https://seaborn.pydata.org/>). For visualization of the multidimensional TFSEE scores, we performed *t*-distributed stochastic neighbor embedding analysis (t-SNE)<sup>46</sup> using the TSNE function and labeled the clusters by calculating K-means clustering using the KMeans function with the expectation-maximization algorithm in scikit-learn version 0.17.1 (<http://scikit-learn.org/>).

#### Nearest neighboring gene analyses and box plots

The universe of expressed genes in each cell line was determined from the RNA-seq data using a FPKM cutoff of >0.4 (Figure S1F). The set of nearest neighboring expressed genes for each enhancer defined by an expressed eRNA or the enrichment of active histone marks was determined for each cell line. Box plot representations were used to express the levels of transcription or enrichment for each called enhancer and transcription of their nearest neighboring expressed genes. The read distribution (RPKM) for each enhancer or (FPKM) gene was calculated and plotted using the boxplot function from matplotlib version 2.0.2 (<https://matplotlib.org/>). Wilcoxon rank sum tests were performed to determine the statistical significance of all comparisons.

#### Overlapping enhancer analysis

Quantification of overlapping enhancers was assessed using a universe of enhancers for each identification method and counting the overlapping enhancers with other methods using BEDTools version 2.17.0.<sup>37</sup> The percentage of overlapping enhancers was calculated in Python (ver. 2.7.11) and plotted using the barplot function from matplotlib version 2.0.2 (<https://matplotlib.org/>). To visualize the intersection of enhancer marks, we used UpsetR version 1.4.<sup>47</sup>

## Results

### The TFSEE model

The TFSEE model integrates data from multiple genomic assays, such as GRO-seq, RNA-seq, and ChIP-seq, with TF expression and motif information to predict (1) TFs driving the formation of active enhancers in a particular cell type and (2) the locations of their cognate enhancers. Enhancer identification methods using enrichment of enhancer histone modifications (eg, H3K4me1 and H3K27ac)<sup>10-12</sup> or enhancer transcription<sup>16</sup> are quite well established. To explore the utility of TFSEE, we have analyzed using both methods of enhancer identification as inputs for TFSEE (Figure 1). In step 1 (Figure 2), a universe of active enhancers across the different constituent cell types was identified based on enhancer transcription as assessed by GRO-seq or total RNA-seq (method 1) (Figure S2A). One could also substitute this with the enhancers identified using the enrichment of epigenomic marks that are known

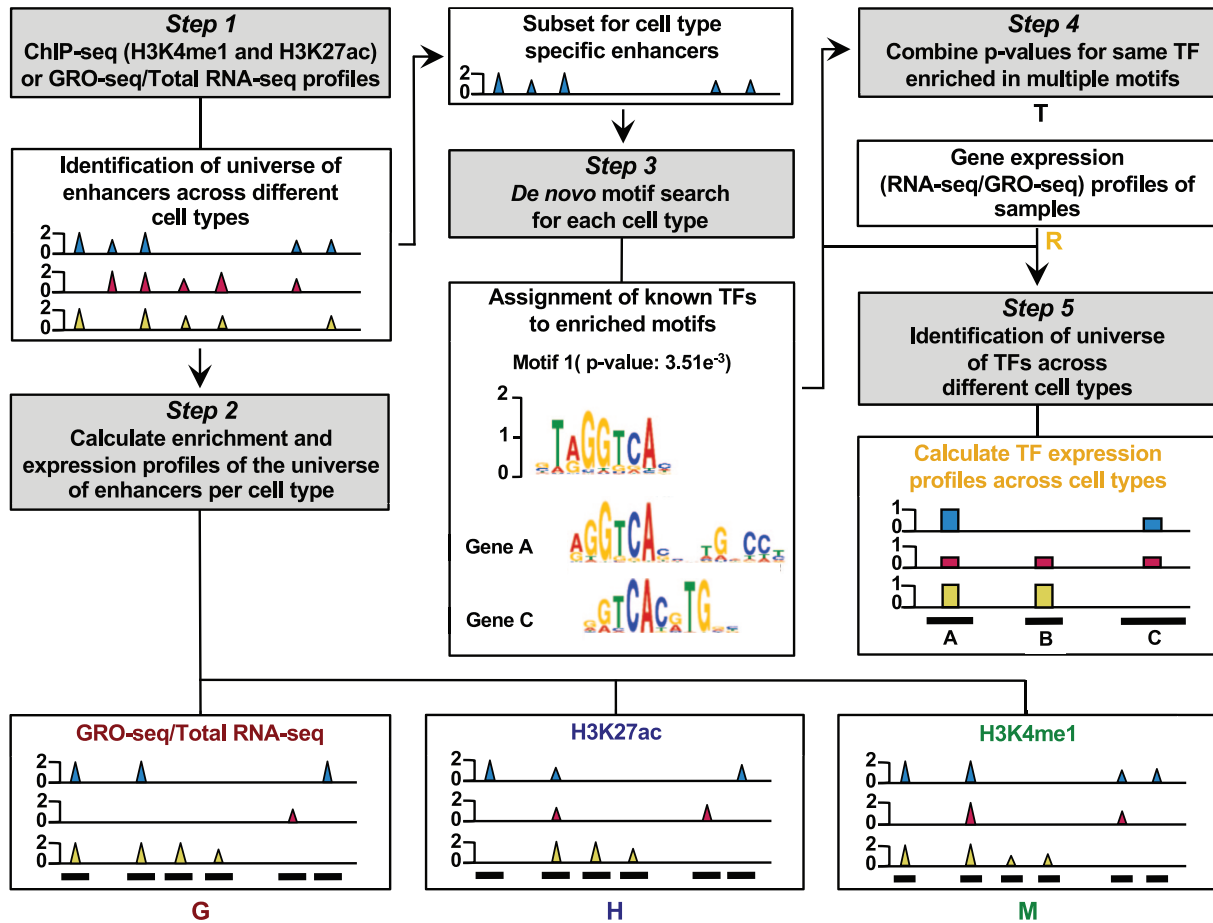
to be enriched at enhancers, such as H3K4me1 and H3K27ac (method 2) (Figure S2B). After the enhancer calling step (by method 1 or method 2), the TFSEE model includes 5 key data processing steps (Figure 2), followed by data integration to calculate the enrichment and activity profiles, that is, the TFSEE score (Figure 3).

*Step 1—Method 1: enhancer calling based on enhancer transcripts defined by GRO-seq.* In this approach, the active enhancers were identified based on enhancer transcripts (ie, eRNAs) called using GRO-seq data. The GRO-seq data were analyzed using groHMM version 3.4<sup>16,21</sup> and the transcript calling was performed as described in the methods section. The transcript calls were further filtered to identify a universe of short intergenic transcripts <9 kb in length and >3 kb away from the known TSSs of protein-coding genes and H3K4me3 peaks, which mark promoters. A final universe of expressed short paired eRNAs and short unpaired eRNAs for each cell type was identified with the cutoffs as mentioned in the methods section. Next, the expression profiles (using GRO-seq data) and histone mark enrichment profiles (using ChIP-seq data) were calculated at these active enhancers for each cell type to calculate the enhancer activity (*A*) (Figure 3).

*Step 1—Method 2: enhancer calling based on histone modification defined by ChIP-seq.* In this approach, the enhancers were identified based on histone modifications (H3K4me1 and H3K27ac) using ChIP-seq data. The enhancers called based on histone modification peak calls for each cell type were merged and the redundancies were removed as described in the methods section. Next, the potential intergenic enhancers were defined as the merged peaks that are >3 kb from known TSSs, protein-coding gene bodies, and H3K4me3 peaks. An RPKM cutoff of  $\geq 1$  for H3K4me1 and H3K27ac (Figure S1D and E) in at least one cell type was used to call a peak as an active enhancer. The universe of active enhancers was then assembled for each cell type.

*Step 2: Calculating enrichment and activity profiles.* For the universe of enhancers for each cell type (identified in step 1), the enhancer activity levels were assessed genome-wide by calculating the enrichment of histone modifications (ie, H3K4me1 and H3K27ac) and enhancer transcription (GRO-seq or total RNA-seq) (Figure 2). Next, we generated an enhancer activity matrix  $A_{C \times E}$  for all cell types, *C*, for the universe of active enhancers, *E*. For this analysis, we assumed that the enhancer activity of each cell type is linearly correlated to the amount of enhancer transcription (GRO-seq or total RNA-seq, *G*) and to the epigenomic marks (H3K4me1, *M*, and H3K27ac, *H*). To reduce bias, the enrichment for each individual enhancer was scaled between 0 and 1. Enhancer activity, *A*, was calculated using the following formula:

$$A = G + M + H$$



**Figure 2.** Data processing for Total Functional Score of Enhancer Elements (TFSEE) method. The TFSEE method has 5 data processing steps that are used to identify enhancer location and activity and their cognate transcription factors (TFs). In step 1, epigenomic (ChIP-seq) or the transcriptional (GRO-seq or total RNA-seq) profiles are used to generate a universe of active enhancers across the different constituent lineages. In step 2, TFSEE calculates the enrichment (H3K4me1 and H3K27ac) and enhancer transcription (GRO-seq and total RNA-seq) profiles under all identified active enhancers per lineage. Lineage-specific enhancers are used as input for step 3, where a de novo motif search is performed to identify potential TFs at each enhancer. If a motif is represented multiple times for a given enhancer location, TFSEE combines the probability of that motif into a single  $P$  value in step 4. Step 5 integrates the amount of enhancer transcription (GRO-seq or total RNA-seq) and the expression of the TFs whose motifs were predicted in step 3 and 4 for all cell types, to provide an output of TF expression profiles across every cell type.

When using only histone modifications (step 1, method 2), enhancer activity can be calculated using the following formula:

$$A = M + H$$

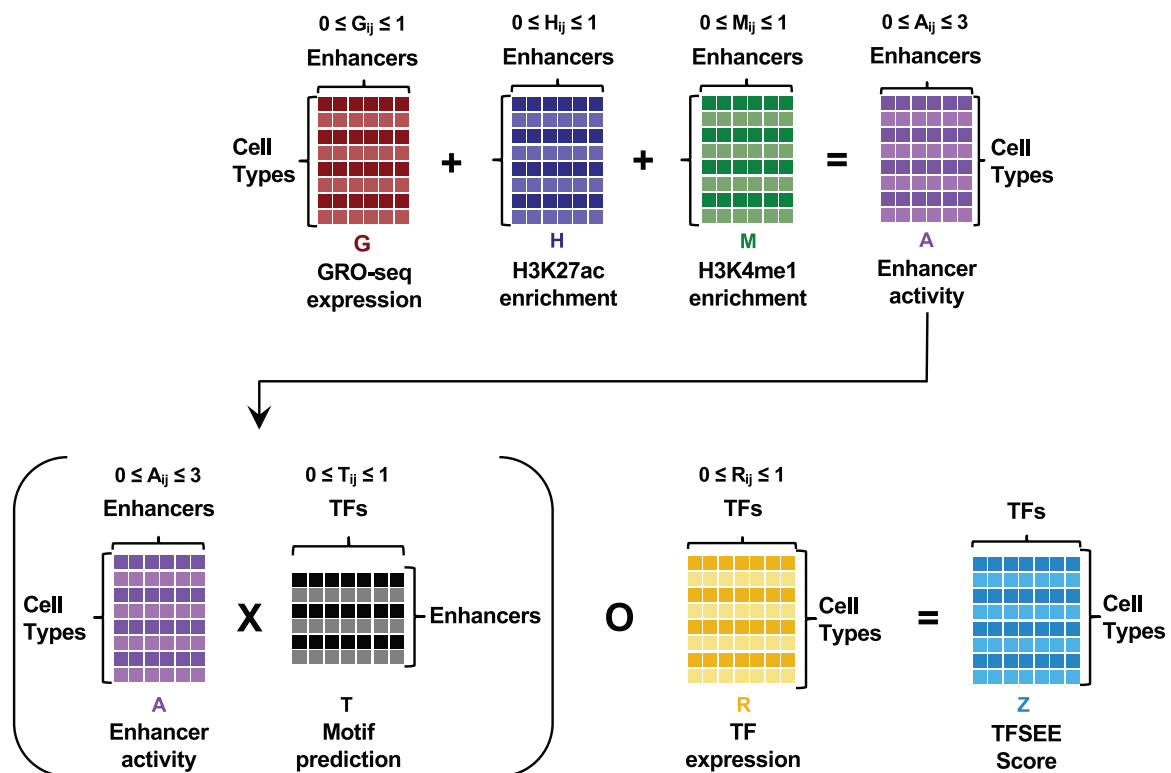
*Steps 3 to 5: De novo motif searching and TF expression.* The TFSEE was designed primarily to detect enhancer activity changes and TF-enhancer relationships for each cell type. The steps in this section are common for both the methods of enhancer calling, which includes de novo motif searching and postprocessing of the motif search results. In steps 3 to 4, the TF-enhancer relationships were determined using a de novo motif search, and a matrix of probabilities of the TFs was created by annotating every enhancer to TF relationships for each cell type (Figures 2 and 3). If a motif is represented multiple times for a given enhancer location, TFSEE combines the probability of that motif into a single  $P$  value using the Stouffer method.<sup>48</sup> In step 5, the expression profile of all the TFs from step 4 (Figures 2 and 3) was calculated from GRO-seq or RNA-seq data across all the cell types.

*Calculating the TFSEE score by data integration.* The final stage integrates all of the data compiled in steps 1 to 5 (Figure 3) to determine the TFSEE score matrix and generate a heatmap of TFSEE scores. First, the enhancer activity matrix,  $A_{C \times E}$ , was combined with the motif prediction matrix,  $T_{E \times P}$  to generate a scaled motif prediction  $P$  value,  $T$ , for each enhancer,  $E$ , to form an intermediate matrix product. This matrix product is combined entrywise with the TF expression matrix,  $R$ , from step 5, and the expression of each TF,  $F$ , for each cell type,  $C$ , into a resulting matrix,  $Z$ , composed of  $C$  cell types and  $F$  TFs. The TFSEE scores can be expressed as the following formula:

$$Z = (A \times T) \cdot R$$

*Using TFSEE for the unbiased identification of enhancers during pancreatic differentiation*

To demonstrate the utility of TFSEE, we used it to define the enhancer landscape and identify TFs that maintain the multipotency of a subpopulation of endodermal stem cells during



**Figure 3.** Overview of Total Functional Score of Enhancer Elements (TFSEE) method. TFSEE combines diverse data sets to identify enhancer location and activity and their cognate transcription factors (TFs). An illustration of TFSEE data integration stage, taking the outputs generated at each step to identify the location, activity level, and predicted TFs at each enhancer across all cell types. (Top) All matrices represent scaled enhancer activity for each cell type in each enhancer prediction method ( $G$ ,  $H$ , and  $M$ ). All matrices are linearly combined into a resulting matrix  $A$ , to provide a total enhancer activity score. (Bottom) Enhancer activity matrix  $A$ , combined with motif prediction matrix  $T$ , represents scaled motif prediction  $P$  values for each enhancer, to form an intermediate matrix product. This matrix product is entrywise combined with TF expression matrix  $R$  (scaled TF expression for each cell type), into a resulting matrix  $Z$ , on which TFSEE clustering is performed.

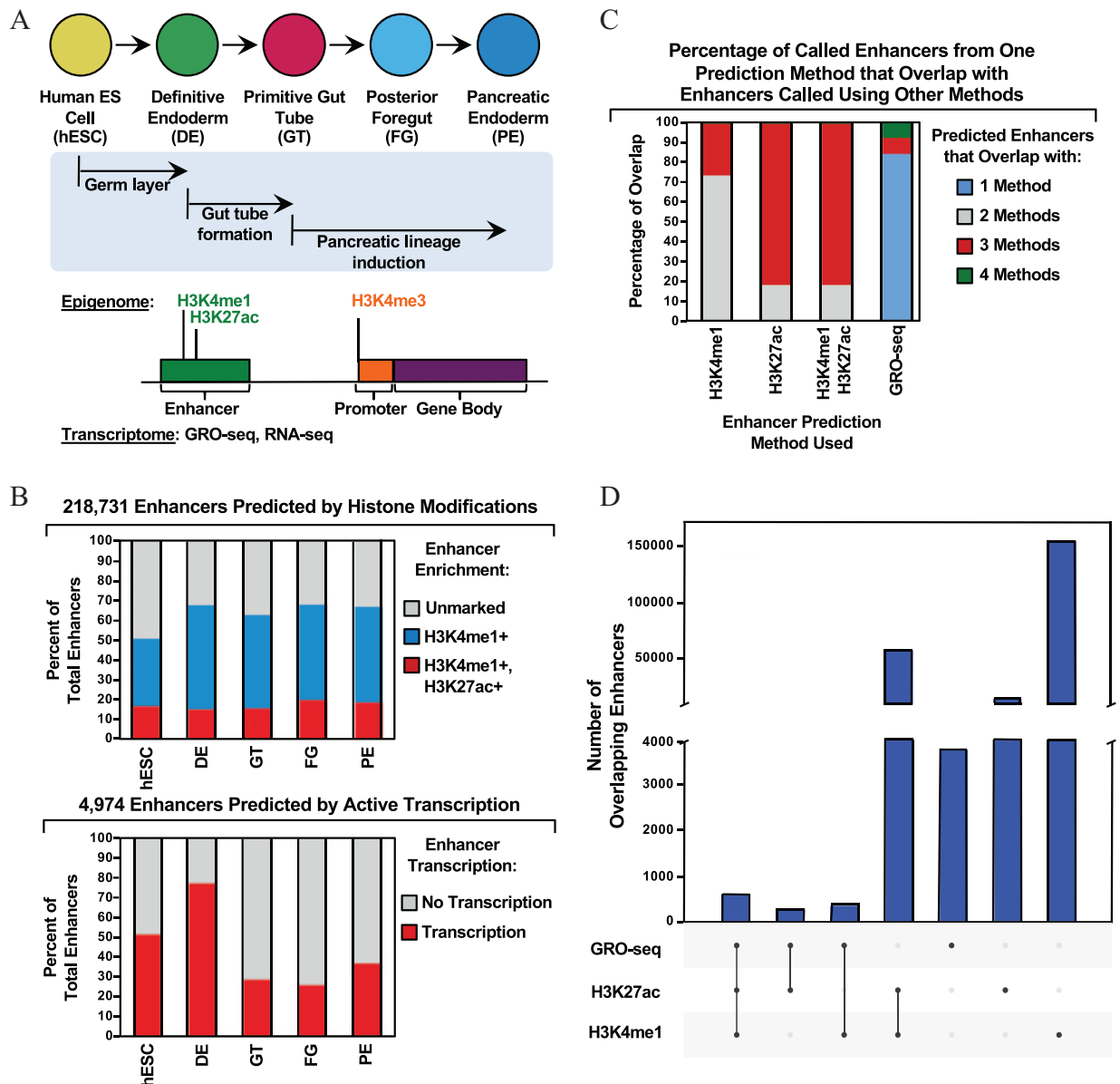
differentiation into pancreatic lineages. For these analyses, we mined previously published ChIP-seq data sets for 3 different histone modifications (ie, H3K4me1, H3K4me3, and H3K27ac), in addition to GRO-seq and RNA-seq data sets, at 5 defined stages of endoderm lineage differentiation: hESCs, definitive endoderm (DE), primitive gut tube (GT), posterior foregut (FG), and PE (Figure 4A, Table S1).

**Calling enhancers by independent methods.** Using differentiation of pancreatic stem cells as a biological model (Figure 4A), we identified the enhancer universe for the cell types by 2 methods: (1) enhancer transcription signatures from GRO-seq data (Figure S2A) and (2) enrichment of epigenomic marks (ie, H3K4me1 or H3K27ac) (Figure S2B). To avoid complications associated with overlaps between enhancer transcription and promoter transcription, we only considered candidate enhancers  $>3$  kb away from the annotated TSSs of active protein-coding genes, as identified by the enrichment of H3K4me3<sup>43</sup> (using GENCODE version 19 annotations<sup>41</sup>) (Figure S2A and B).

We then predicted candidate enhancers using methods 1 and 2. We identified a set of 4974 candidate enhancers (Figure 4B) by method 1 using GRO-seq data, as described previously,<sup>19</sup> with RPKM cutoffs of  $\geq 0.5$  or  $\geq 1$  (Figure S1B and C) in at least one cell lineage. We also identified a set of enhancers by method 2 using histone modifications, by filtering the

enhancer universe based on the enrichment of H3K4me1 and H3K27ac (RPKM cutoff of  $\geq 1$  [Figure S1D and E] for both marks in at least one cell line), and identified a set of 218731 candidate enhancers across all stages of pancreatic differentiation (Figure 4B). This stringent filter is necessary to reduce the false-positive enhancers from method 2 that could easily be annotated as alternative chromatin states using ChromHMM.<sup>49</sup> In addition, the majority of histone called enhancers were marked by only H3K4me1 (Figure 4B). These results confirm the enhancer landscape across pancreatic differentiation reported by Wang et al.<sup>1</sup>

**Comparison of enhancer calls by methods 1 and 2.** Next, we compared the enhancer universes called by enhancer transcription (method 1) and histone modifications (method 2). We found that 12% of enhancers called based on enhancer transcription using GRO-seq data were also identified based on enrichment of H3K4me1, H3K27ac, or both marks combined (Figure 4C, Figure S3A). Interestingly, greater than 84% of the enhancers identified based solely on enhancer transcription were not called based on enrichment of H3K27ac or H3K4me1 (Figure 4C and D, Figure S3B). This is likely due to the fact that these may not be the primary or only chromatin marks denoting active enhancers in these systems, and other marks (ie, H4K8ac or H3K9ac) or combinations of marks might serve as better



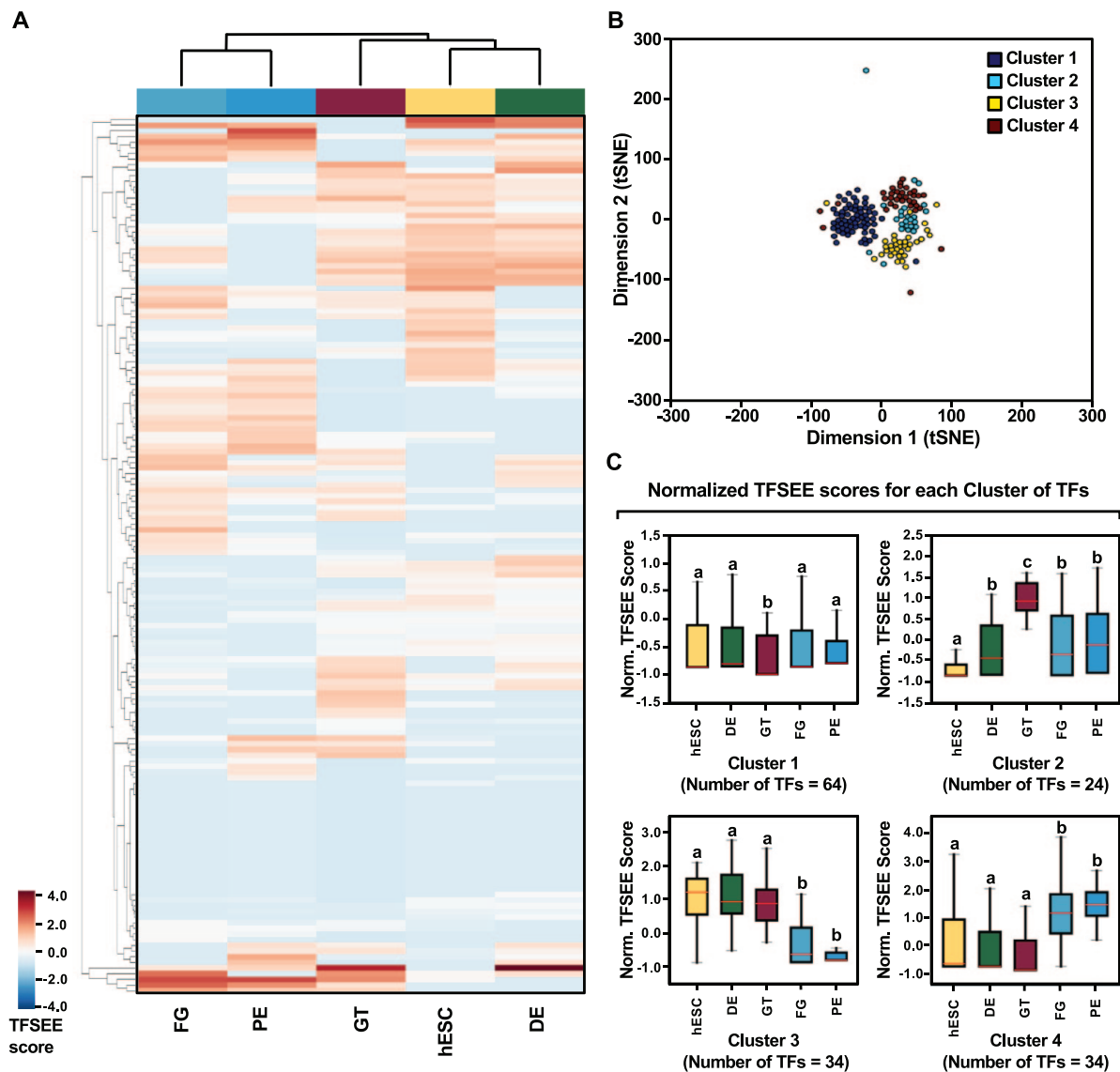
**Figure 4.** Comparison of approaches for genome-wide prediction of enhancers during pancreatic differentiation. (A) (Top) Schematic diagram of pancreatic differentiation starting from human embryonic stem cells (hESCs) to pancreatic endoderm (PE). (Bottom) Depiction of epigenomic (ChIP-seq) and transcriptional (GRO-seq and RNA-seq) profiles for each cell line used for analysis. (B) Stacked bar chart comparing the predicted activity of candidate enhancers categorized by (Top) H3K4me1 and H3K27ac enrichment or (bottom) enhancer transcription (GRO-seq). (C) Stacked bar chart comparing enhancer prediction methods in pancreatic differentiation. Enhancers were called using enhancer transcription (GRO-seq) or using H3K4me1 enrichment, H3K27ac enrichment, or a combination of both histone marks. The percentage of called enhancers from one prediction method that overlap with enhancers called using other methods is shown. (D) UpSet plot showing the set intersection of enhancer identification methods shown in panel (C). DE indicates definitive endoderm; FG, posterior foregut; GT, primitive gut tube; hESCs, human embryonic stem cells; PE, pancreatic endoderm; TF, transcription factors; TFSEE, Total Functional Score of Enhancer Elements.

identifiers.<sup>50</sup> In contrast, less than 1% of enhancers called based on the enrichment of H3K4me1 and H3K27ac overlapped with the enhancers identified based on enhancer transcription or both H3K4me1 and H3K27ac combined (Figure 4C). This may be due, in part, to the fact that enhancer calling based on H3K4me1 or H3K27ac enrichment yields much larger numbers of putative enhancers (Figure 4D), many of which may be false positives or inactive as true regulatory elements (Figure S3C and D). Based on these findings, we decided to focus on the enhancers identified based on enhancer transcription using GRO-seq data (method 1), which had the highest percentage

of enhancers that were called by all 3 methods, as an input to TFSEE for the subsequent analysis.

#### *TFSEE identifies lineage-specific enhancers and their cognate TFs during pancreatic differentiation*

*TFSEE scores determined by using inputs from method 1.* After determining the TFSEE scores globally across the lineage of pancreatic differentiation, we performed unsupervised hierarchical clustering on the enhancers predicted by method 1 based on enhancer transcription, which grouped the



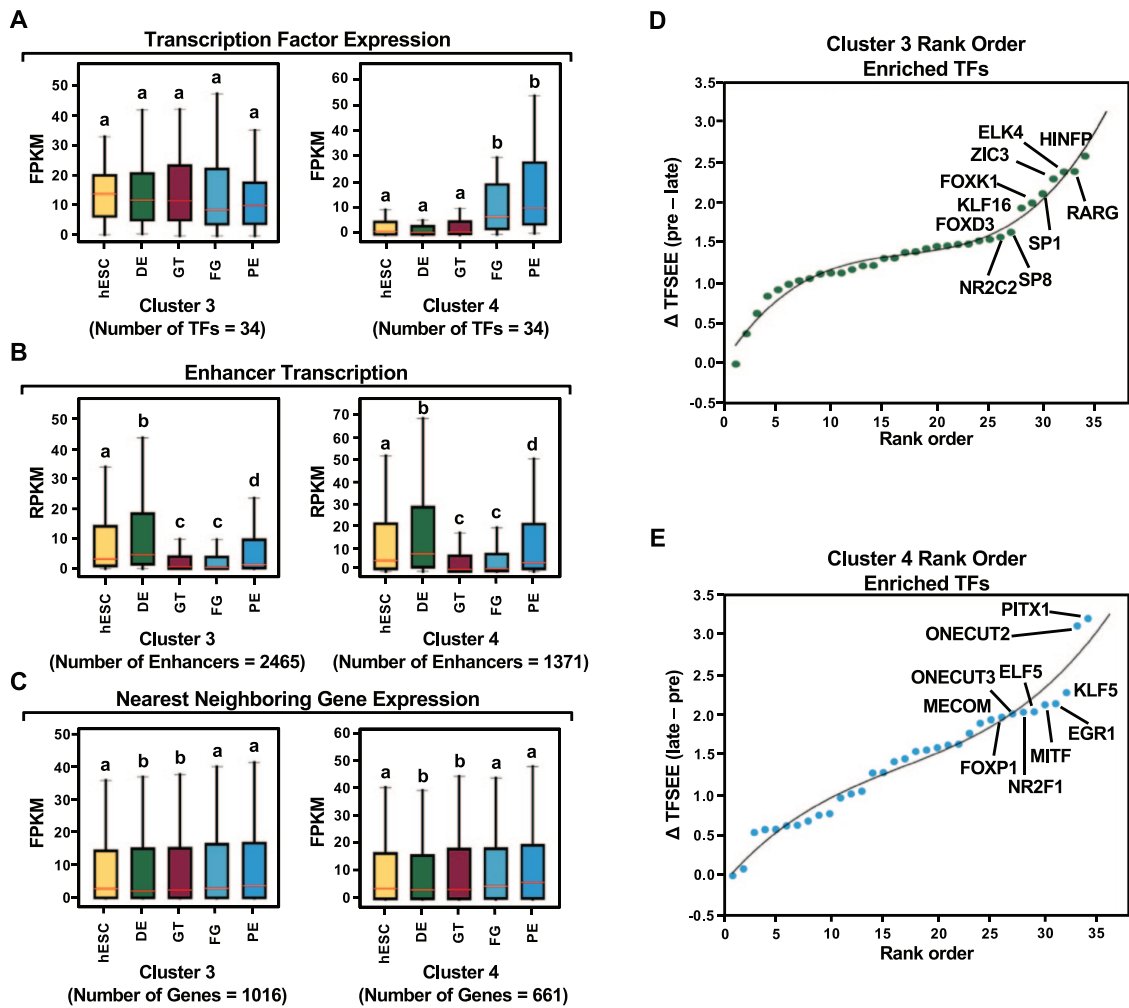
**Figure 5.** TFSEE identifies cell type-specific enhancers and their cognate TFs that drive gene expression during pancreatic differentiation. (A) Unsupervised hierarchical clustering of cell type-normalized TFSEE scores shown in a heatmap representation. hESC (human embryonic stem cell); DE (definitive endoderm); GT (primitive gut tube); FG (posterior foregut); PE (pancreatic endoderm). (B) Biaxial t-SNE clustering plot of cell type-normalized TFSEE scores showing evidence of 4 distinct clusters, each point represents an individual TF. (C) Box plots of normalized TFSEE score for clusters identified in pancreatic differentiation (panel B), number of TFs are indicated in each cluster. Bars marked with different letters are significantly different (Wilcoxon rank sum test,  $P < 1 \times 10^{-4}$ ). Cluster 1, TFs associated with early (hESC, DE) and late pancreatic differentiation (FG and PE). Cluster 2, TFs associated with GT pluripotency. Cluster 3, TFs associated with pre-pancreatic lineage induction (hESC, DE, and GT). Cluster 4, TFs associated with late pancreatic differentiation (FG and PE). TF, transcription factors; TFSEE, Total Functional Score of Enhancer Elements.

lineage-specific cell types into 2 major clades: (1) FG and PE and (2) hESC, DE, and GT (Figure 5A). To better understand the TF-enhancer dynamics across the 4 represented cell types in pancreatic differentiation, we clustered the TFSEE scores across all the differentiation stages, revealing 4 major clusters (Figure 5B). We then examined the enrichment of putative enhancers and their associated TFs across stages by quantifying their normalized TFSEE scores. This analysis revealed 4 major enhancer clusters: (1) those driving early (hESC, DE) and late pancreatic differentiation (FG and PE), (2) those enriched in GT, (3) those driving early pancreatic endodermal lineage formation

(hESC, DE and GT), and (4) those driving late pancreatic differentiation (FG and PE) (Figure 5C).

To investigate the distinct roles of lineage-specific enhancers and their cognate TFs, focusing on those that provide a clear demarcation of enrichment between early and late pancreatic differentiation, we first examined the expression levels of the messenger RNAs encoding the predicted TFs for each cluster in each of the stages. Our analysis revealed that TFs identified in early pancreatic differentiation show similar expression across the later stages of development, whereas TFs identified in late pancreatic differentiation are expressed most predominantly in the FG and PE stages (Figure 6A) coinciding with pancreatic





**Figure 6.** TFSEE-predicted TFs are enriched in pre- and late pancreatic differentiation. (A to C) Box plots of normalized TF expression (panel A), enhancer transcription (panel B), and gene expression for the nearest neighboring genes to active enhancers (panel C) in pre- (cluster 3) and late pancreatic (cluster 4) differentiation across the different cell types. Bars marked with different letters are significantly different from each other (Wilcoxon rank sum test). hESC (human embryonic stem cell); DE (definitive endoderm); GT (primitive gut tube); FG (posterior foregut); PE (pancreatic endoderm). (A) TFs identified in cluster 3 by TFSEE show equal expression across differentiation, whereas cluster 4 highlights TFs highly expressed in FG and PE. TF expression as measured by RNA-seq. The number of TFs in each cluster are in parenthesis ( $P < 1 \times 10^{-4}$ ). (B) Enhancer transcription as measured by GRO-seq. The number of enhancers in each cluster are in parenthesis ( $P < 1 \times 10^{-4}$ ). (C) Gene expression as measured by RNA-seq. The number of genes in each cluster are in parenthesis ( $P < .05$ ). (D and E) Rank order of TFs enriched in cluster 3 and cluster 4 identified using TFSEE. The top 10 TFs in each cluster are noted. TF, transcription factors; TFSEE, Total Functional Score of Enhancer Elements.

induction at the FG stage (Figure 4A). In addition, we observed an enrichment of TFs in a stage-specific manner for TFs enriched early (hESC, DE) and late (FG and PE), but not for those maintaining GT pluripotency (Figure S4A).

Next, we determined whether enhancer transcription corresponding to the enriched TFs in each cluster, using the TFSEE scores, correlates with the regulation of nearby genes. To do so, we identified the enhancers corresponding to the predicted TFs using motif enrichment and binding site prediction. We then determined the level of transcription for each enhancer using GRO-seq (Figure 6B, Figure S4B) and the level of expression for the nearest neighboring gene (upstream or downstream) using RNA-seq (Figure 6C, Figure S4C). Interestingly, transcribed enhancers exhibited stage-specific enrichment, which did not correspond to the patterns observed

based on TFSEE scores (Figure 6B, Figure S4B). This result likely reflects the fact that 1364 of the enhancers are shared between clusters (55%,  $n = 2465$ , for cluster 3; 99%,  $n = 1371$ , for cluster 4), and the variation between clusters is due to differences in TF expression and their affinity for motifs. Likewise, the expression of the nearest neighboring gene for each transcribed enhancer did not exhibit stage-specific enrichment (Figure 6C, Figure S4C) due to the vast abundance of enhancers with neighboring genes shared between the clusters. However, without further high-throughput data to study enhancer-promoter interactions (as measured by 4C, ChIA-PET, or Hi-C),<sup>51-53</sup> it is difficult to discern the logic of the stage-specific regulatory network.

To further understand the potential regulators of each cluster, we determined a rank order frequency distribution for all

TFs within each cluster (Figure 6D and E, Figure S4D and E). This analysis revealed enrichment of HINFP, RARG, ZIC3, and SP1-like family TFs (SP1 and SP8), which are important regulators of embryonic development<sup>54–57</sup> (Figure 6D). In addition, the Onecut family (ONECUT2 and ONECUT3), EGR1, MITF, and FOXP1 TFs, which were enriched in cluster 4, have been shown to function in pancreatic and islet cell development<sup>58–61</sup> (Figure 6E).

*TFSEE scores determined using inputs from method 2.* To determine the robustness of the method, we calculated TFSEE scores using enhancers predicted based on histone enrichment alone. We performed unsupervised hierarchical clustering and retrieved only 3 clusters, in contrast to the 4 clusters from enhancers predicted based on enhancer transcription (Figure S5A and B). These results highlight the potential role of TF-enhancer interactions driving the formation of the early pancreatic endodermal lineage (hESC, DE, and GT), as well as late pancreatic differentiation (FG and PE), but do not reveal any other stage-specific drivers (Figure S5C).

For comparable clusters from early and late pancreatic differentiation, we determined whether enhancer “activity” based on H3K27ac enrichment correlates with the regulation of nearby genes. We monitored TF expression by RNA-seq (Figure S6A) and then calculated the level of activity for each enhancer using H3K27ac ChIP-seq data (Figure S6B) and the expression of the nearest neighboring gene (upstream or downstream) using RNA-seq (Figure S6C). We found that H3K27ac-enriched enhancers exhibit stage-specific enrichment, as observed for method 1, which does not correspond to the patterns found from TFSEE enrichment (Figure S6B). This result reflects the fact that 1017 of the enhancers are shared between clusters (74%,  $n = 1375$  for cluster 2; 62%,  $n = 1640$  for cluster 3), and the variation between clusters is due to differences in TF expression and their affinity for the motifs. Furthermore, similar to method 1, the nearest neighboring gene for each transcribed enhancer does not exhibit stage-specific enrichment (Figures S6C). The potential regulators of clusters 2 and 3 were determined using a rank order frequency distribution for all TFs within each cluster (Figure S6D and E).

*Comparison of TF identification using inputs from method 1 or method 2.* To determine the robustness of the approaches, we compared TFs identified by TFSEE using inputs from method 1 or method 2 for pre- and late pancreatic differentiation. We found 9 and 12 TFs (out of the total of 44 and 46 unique TFs identified) enriched in common for early and late pancreatic differentiation, respectively (Figure S6F). These differences in the enriched TFs may be, in part, due to the much larger numbers of putative enhancers called using H3K4me1 and H3K27ac enrichment, many of which may be false positives or inactive as regulatory elements, producing a greater assortment of enriched TF motifs. Taken together, our results show that

TFSEE can be used to identify cell type-specific TFs that control lineage-specific enhancers.

## Discussion

The carefully choreographed mechanisms involved in driving the lineage-specific transcriptional responses during development remain poorly understood. In this study, we integrated a variety of publicly available high-throughput sequencing data from pancreatic lineage development to identify the potential regulators driving the early or late pancreatic lineage development. Our analysis revealed the enrichment of the Onecut family (ONECUT2 and ONECUT3), EGR1, MITF, and FOXP1 TFs in the late pancreatic differentiation phase (FG, PE), which have been shown to function in pancreatic and islet cell development.<sup>58–61</sup> All but the Onecut family were identified as top-ranked TFs by both methods. Our analyses provide a detailed operational description of TFSEE, a previously published computational method<sup>33</sup> for identification of active enhancers and associated cognate TFs, during differentiation of hESCs toward pancreatic cell type. TFSEE employs a multiview clustering of multiple genomic assays that directly models changes in the transcriptional and epigenetic states across cell types. This approach allowed us to directly integrate disparate data while encoding assumptions and dependencies between data types in an interpretable and extendable model.

## Evaluation and use of TFSEE

The TFSEE model gains power by both explicitly modeling the enhancer landscape for each cell type and detecting the enhancer activity changes and TF-enhancer relationships across all cell types. To date, we have applied TFSEE to transcriptional and epigenomic data from hESC differentiation time course experiments<sup>1,34</sup> (analyses described herein) and a variety of breast cancer cell lines.<sup>33,62</sup> Our results show that this method can identify cell type-specific TFs and their cognate enhancers that are biologically relevant and are good candidates for further biological validation. In particular, this method identifies TFs bound at active enhancers, which regulate gene expression, supporting the biological relevance of TFSEE predictions. In this study, we identified enrichment of HINFP, RARG, ZIC3, and SP1-like family TFs (SP1 and SP8), which are known important regulators of embryonic development,<sup>54–57</sup> 3 of which (ie, ZIC3, SP1, and SP8) were identified as the top TFs by both methods.

In addition, TFSEE enables analysis of driver TFs using a limited amount of data. The model was able to identify lineage-specific TFs with as little as 5 cell types and with only 2 data types, RNA-seq and ChIP-seq (for H3K4me3, H3K4me1, and H3K27ac) (Figure S5A and B). A limitation of the TFSEE method is that while the model can be used with a reduced number of data types for enhancer identification, it fails to identify additional subtype- or stage-specific drivers with

reduced data input (Figure S5C). In this case, the overlapping clusters identified only a subset of TFs that are jointly enriched (Figure S6F).

### Integrating additional genomic data into TFSEE

TFSEE could potentially be applied to any cell type with limited data, either GRO-seq or total RNA-seq (for enhancer calling by method 1), or histone modifications (for enhancer calling using H3K27ac and H3K4me1 by method 2). The integration of additional data into TFSEE could allow extension of the model, providing greater granularity about subtype-specific TFs and a better understanding of gene regulatory networks. Genomic data indicating open regions of chromatin (eg, ATAC-seq,<sup>63</sup> DNase-seq,<sup>7</sup> or MNase-seq<sup>64</sup>) could extend the dynamic range of “enhancer activity” (Figure 3) and help eliminate false-positive enhancers in each cell type. Likewise, adding enrichment of transcriptional co-regulator p300<sup>65</sup> could serve the same role. Furthermore, integrating additional histone modifications, which could be used to annotate alternate chromatin states by ChromHMM,<sup>49</sup> may provide a finer filter for enhancer identification by method 2 than achieved here when limiting the analysis to the histone modifications used herein. Finally, to better understand the cluster-specific regulatory networks, the addition of chromatin looping data for enhancer-promoter interactions (as measured by 4C, ChIA-PET, or Hi-C)<sup>51–53</sup> may provide some advantages. The looping data would provide a better understanding of how enhancers shared between clusters determine cell type-specific expression profiles. We believe that including any or all of data described above into TFSEE would improve the model and help to discern cell type-specific regulatory networks and can easily be added due to the flexibility of the model.

### Conclusions

The increasing availability of different types of genomic data sets provides an opportunity to perform data integration to uncover cell type-specific TF drivers. To facilitate identification of these drivers, we developed and further evaluated TFSEE, which systemically identifies active enhancers and their cognate TFs. We showed that TFSEE can identify stage-specific TFs during differentiation of hESCs into pancreatic lineages. Collectively, our results show how TFSEE can be used to predict molecular drivers maintaining cell type-specific function and biology.

### Acknowledgements

The authors thank members of the Kraus Laboratory for providing critical feedback on this work and helpful comments on the manuscript.

### Author Contributions

HLF, AN, VSM, and WLK designed the project. WLK conceived of the conceptual framework for TFSEE and VM, AN,

and HLF developed it further. VM processed and analyzed the GRO-seq, RNA-seq, and ChIP-seq data and performed the integrative computational analyses of the genomic data. VM and AN prepared an initial draft of the figures and text, which were edited and finalized by WLK and AN, with input from HLF. WLK secured funding for the work and provided overall project direction.

### ORCID iDs

Venkat S. Malladi  <https://orcid.org/0000-0002-0144-0564>

Hector L Franco  <https://orcid.org/0000-0001-9354-0679>

W Lee Kraus  <https://orcid.org/0000-0002-8786-2986>

### Data Availability and Implementation Statement

The data sets analyzed in this study were obtained from the NCBI's Gene Expression Omnibus (<https://www.ncbi.nlm.nih.gov/geo/>) or EMBL-EBI's ArrayExpress (<http://www.ebi.ac.uk/arrayexpress/>) repositories using the accession numbers listed in Table S1. Implementation of this method is available at: <https://git.biohpc.swmed.edu/gcrb/tfsee>.

### Supplemental Material

Supplemental material for this article is available online.

### REFERENCES

1. Wang A, Yue F, Li Y, et al. Epigenetic priming of enhancers predicts developmental competence of hESC-derived endodermal lineage intermediates. *Cell Stem Cell*. 2015;16:386–399.
2. Shlyueva D, Stampfel G, Stark A. Transcriptional enhancers: from properties to genome-wide predictions. *Nat Rev Genet*. 2014;15:272–286.
3. Heinz S, Romanoski CE, Benner C, Glass CK. The selection and function of cell type-specific enhancers. *Nat Rev Mol Cell Biol*. 2015;16:144–154.
4. Kleftogiannis D, Kalnis P, Bajic VB. Progress and challenges in bioinformatics approaches for enhancer identification. *Brief Bioinform*. 2016;17:967–979.
5. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012;489:57–74.
6. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, et al. Integrative analysis of 111 reference human epigenomes. *Nature*. 2015;518:317–330.
7. Crawford GE, Holt IE, Whittle J, et al. Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Res*. 2006;16:123–131.
8. Sheffield NC, Thurman RE, Song L, et al. Patterns of regulatory activity across diverse human cell types predict tissue identity, transcription factor binding, and long-range interactions. *Genome Res*. 2013;23:777–788.
9. Davie K, Jacobs J, Atkins M, et al. Discovery of transcription factors and regulatory regions driving in vivo tumor development by ATAC-seq and FAIRE-seq open chromatin profiling. *PLoS Genet*. 2015;11:e1004994.
10. Heintzman ND, Stuart RK, Hon G, et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet*. 2007;39:311–318.
11. Heintzman ND, Hon GC, Hawkins RD, et al. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*. 2009;459:108–112.
12. Rajagopal N, Ernst J, Ray P, et al. Distinct and predictive histone lysine acetylation patterns at promoters, enhancers, and gene bodies. *G3 (Bethesda)*. 2014;4:2051–2063.
13. Tak YG, Hung Y, Yao L, et al. Effects on the transcriptome upon deletion of a distal element cannot be predicted by the size of the H3K27Ac peak in human cells. *Nucleic Acids Res*. 2016;44:4123–4133.
14. Dogan N, Wu W, Morrissey CS, et al. Occupancy by key transcription factors is a more accurate predictor of enhancer activity than histone modifications or chromatin accessibility. *Epigenetics Chromatin*. 2015;8:16.
15. Kim TK, Shiekhhattar R. Architectural and functional commonalities between enhancers and promoters. *Cell*. 2015;162:948–959.

16. Hah N, Danko CG, Core L, et al. A rapid, extensive, and transient transcriptional response to estrogen signaling in breast cancer cells. *Cell*. 2011;145:622-634.
17. Wu H, Nord AS, Akiyama JA, et al. Tissue-specific RNA expression marks distant-acting developmental enhancers. *PLoS Genet*. 2014;10:e1004610.
18. Wang D, Garcia-Bassets I, Benner C, et al. Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. *Nature*. 2011;474:390-394.
19. Hah N, Murakami S, Nagari A, Danko CG, Kraus WL. Enhancer transcripts mark active estrogen receptor binding sites. *Genome Res*. 2013;23:1210-1223.
20. Core LJ, Martins AL, Danko CG, Waters CT, Siepel A, Lis JT. Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat Genet*. 2014;46:1311-1320.
21. Chae M, Danko CG, Kraus WL. groHMM: a computational tool for identifying unannotated and cell type-specific transcription units from global run-on sequencing data. *BMC Bioinformatics*. 2015;16:222.
22. Kheradpour P, Ernst J, Melnikov A, et al. Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Res*. 2013;23:800-811.
23. Franco HL, Nagari A, Kraus WL. TNF $\alpha$  signaling exposes latent estrogen receptor binding sites to alter the breast cancer cell transcriptome. *Mol Cell*. 2015;58:21-34.
24. Visel A, Minovitsky S, Dubchak I, Pennacchio LA. VISTA enhancer browser—a database of tissue-specific human enhancers. *Nucleic Acids Res*. 2007;35:D88-D92.
25. Arnold CD, Gerlach D, Stelzer C, Boryn LM, Rath M, Stark A. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science*. 2013;339:1074-1077.
26. Kwasniewski JC, Fiore C, Chaudhari HG, Cohen BA. High-throughput functional testing of ENCODE segmentation predictions. *Genome Res*. 2014;24:1595-1602.
27. Gerstein MB, Kundaje A, Hariharan M, et al. Architecture of the human regulatory network derived from ENCODE data. *Nature*. 2012;489:91-100.
28. Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM. A census of human transcription factors: function, expression and evolution. *Nat Rev Genet*. 2009;10:252-263.
29. Jolma A, Yin Y, Nitta KR, et al. DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature*. 2015;527:384-388.
30. Mathelier A, Fornes O, Arenillas DJ, et al. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res*. 2016;44:D110-D115.
31. Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS. Quantifying similarity between motifs. *Genome Biol*. 2007;8:R24.
32. Whitfield TW, Wang J, Collins PJ, et al. Functional analysis of transcription factor binding sites in human promoters. *Genome Biol*. 2012;13:R50.
33. Franco HL, Nagari A, Malladi VS, et al. Enhancer transcription reveals subtype-specific gene expression programs controlling breast cancer pathogenesis. *Genome Res*. 2018;28:159-170.
34. Xie R, Everett LJ, Lim HW, et al. Dynamic chromatin remodeling mediated by polycomb proteins orchestrates pancreatic differentiation of human embryonic stem cells. *Cell Stem Cell*. 2013;12:224-237.
35. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*. 2009;10:R25.
36. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25:1754-1760.
37. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26:841-842.
38. Landt SG, Marinov GK, Kundaje A, et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res*. 2012;22:1813-1831.
39. Feng J, Liu T, Qin B, Zhang Y, Liu XS. Identifying ChIP-seq enrichment using MACS. *Nat Protoc*. 2012;7:1728-1740.
40. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29:15-21.
41. Harrow J, Frankish A, Gonzalez JM, et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res*. 2012;22:1760-1774.
42. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*. 2011;12:323.
43. Guenther MG, Levine SS, Boyer LA, Jaenisch R, Young RA. A chromatin landmark and transcription initiation at most promoters in human cells. *Cell*. 2007;130:77-88.
44. Bailey TL, Boden M, Buske FA, et al. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res*. 2009;37:W202-W208.
45. Machanick P, Bailey TL. MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics*. 2011;27:1696-1697.
46. van der Maaten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res*. 2008;9:2579-2605.
47. Conway JR, Lex A, Gehlenborg N. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics*. 2017;33:2938-2940.
48. Stouffer SA, Suchman EA, DeVinney LC, Star SA, Williams RM. *The American Soldier: Adjustment During Army Life*. Princeton, NJ: Princeton University Press; 1949.
49. Ernst J, Kellis M. Chromatin-state discovery and genome annotation with ChromHMM. *Nat Protoc*. 2017;12:2478-2492.
50. Zhu Y, Sun L, Chen Z, Whitaker JW, Wang T, Wang W. Predicting enhancer transcription and activity from chromatin modifications. *Nucleic Acids Res*. 2013;41:10032-10043.
51. Simonis M, Klous P, Splinter E, et al. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat Genet*. 2006;38:1348-1354.
52. Zhang J, Poh HM, Peh SQ, et al. ChIA-PET analysis of transcriptional chromatin interactions. *Methods*. 2012;58:289-299.
53. Mifsud B, Tavares-Cadete F, Young AN, et al. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat Genet*. 2015;47:598-606.
54. Xie R, Medina R, Zhang Y, et al. The histone gene activator HINFP is a nonredundant cyclin E/CDK2 effector during early embryonic cell cycles. *Proc Natl Acad Sci USA*. 2009;106:12359-12364.
55. Jeong Y, Mangelsdorf DJ. Nuclear receptor regulation of stemness and stem cell differentiation. *Exp Mol Med*. 2009;41:525-537.
56. Lim LS, Loh YH, Zhang W, et al. Zic3 is required for maintenance of pluripotency in embryonic stem cells. *Mol Biol Cell*. 2007;18:1348-1358.
57. Zhao C, Meng A. Sp1-like transcription factors are regulators of embryonic development in vertebrates. *Dev Growth Differ*. 2005;47:201-211.
58. Vanhorenbeeck V, Jenny M, Cornut JF, et al. Role of the Onecut transcription factors in pancreas morphogenesis and in pancreatic and enteric endocrine differentiation. *Dev Biol*. 2007;305:685-694.
59. Muller I, Rossler OG, Wittig C, Menger MD, Thiel G. Critical role of Egr transcription factors in regulating insulin biosynthesis, blood glucose homeostasis, and islet size. *Endocrinology*. 2012;153:3040-3053.
60. Mazur MA, Winkler M, Ganic E, et al. Microphthalmia transcription factor regulates pancreatic beta-cell function. *Diabetes*. 2013;62:2834-2842.
61. Spaeth JM, Hunter CS, Bonatakis L, et al. The FOXP1, FOXP2 and FOXP4 transcription factors are required for islet alpha cell proliferation and function in mice. *Diabetologia*. 2015;58:1836-1844.
62. Xi Y, Shi J, Li W, et al. Histone modification profiling in breast cancer cell lines highlights commonalities and differences among subtypes. *BMC Genomics*. 2018;19:150.
63. Lee K, Cho H, Rickert RW, et al. FOXA2 is required for enhancer priming during pancreatic differentiation. *Cell Rep*. 2019;28:382-393.e387.
64. Schones DE, Cui K, Cuddapah S, et al. Dynamic regulation of nucleosome positioning in the human genome. *Cell*. 2008;132:887-898.
65. Visel A, Blow MJ, Li Z, et al. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*. 2009;457:854-858.