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

*This manuscript (permalink) was automatically generated from vsmalladi/tfsee-manuscript@0c8b456 on April 18, 2018.*

## Authors


- **Venkat S. Malladi**
  0000-0002-0144-0564 · vsmalladi · katatonikkat
  The Laboratory of Signaling and Gene Regulation, Cecil H. and Ida Green Center for Reproductive Biology Sciences and Division of Basic Reproductive Biology Research, Department of Obstetrics and Gynecology, University of Texas Southwestern Medical Center; Department of Bioinformatics, University of Texas Southwestern Medical Center"

- **Anusha Nagari**

  The Laboratory of Signaling and Gene Regulation, Cecil H. and Ida Green Center for Reproductive Biology Sciences and Division of Basic Reproductive Biology Research, Department of Obstetrics and Gynecology, University of Texas Southwestern Medical Center

- **Hector L. Franco**

  The Laboratory of Signaling and Gene Regulation, Cecil H. and Ida Green Center for Reproductive Biology Sciences and Division of Basic Reproductive Biology Research, Department of Obstetrics and Gynecology, University of Texas Southwestern Medical Center; Department of Genetics and Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill"

- **W. Lee Kraus**

  The Laboratory of Signaling and Gene Regulation, Cecil H. and Ida Green Center for Reproductive Biology Sciences and Division of Basic Reproductive Biology Research, Department of Obstetrics and Gynecology, University of Texas Southwestern Medical Center

# Abstract

The ability to integrate different genome data sets to systematically identify active enhancers together with their cognate transcription factors (TF) remains a difficult and somewhat arbitratry process. We have developed a computational framework that systematically identifies active enhancers in any cell or tissue type together with the TFs bound at the enhancers by integrating multiple genomic assays that probe the transcriptional (GRO-seq and RNA-seq) and epigenetic (ChIP-seq) states of the cells. Our method, called Total Functional Score of Enhancer Elements (TFSEE), integrates the magnitude of enhancer transcription as a measure of enhancer activitity, enrichment of histone modifications typically associated with enhancers (H3K4me1 and H3K27ac), TF expression levels, and TF motif p-values to compute a probability score of TF binding events at active enhancers across the genome. This method has allowed us to define the enhancer landscape during differentiation of embryonic stem cells into pancreatic lineages and in breast cancer cells to define the regulatory pathways of the distinct molecular subtypes of breast cancer. Using TFSEE, we have identified key breast cancer subtype-specific transcription factors that are bound at active enhancers and dictate gene expression patterns determining growth outcomes. To demonstrate the broader utility of our approach, we have used this algorithm to identify transcription factors during the differentiation of embryonic stem cells into pancreatic cells. The analysis has revealed transcription factors maintaining the multipotency of endoderm stem cells and promoting differentiation into pancreatic progenitor cells. Taken together our results show that TFSEE can be used to perform multilayer genomic data integration to uncover novel cell type-specific transcription factors that control lineage-specific enhancers.

# Background

Enhancers and the transcription factors (TFs) regulating their formation have been shown to play an important role in cell type-specific activation of gene expression [1,2]. Although thousands of potential enhancers have been identified in cell lines and tissues, identification of the enhancers that are active versus not active or poised remains remains a major challenge [3]. In addition, the ability to identify the TFs acting at the numerous enhancers in each cell type is technically challenging [4,5].

Active enhancers have been shown to share several common features; such as increased chromatin accessibility (as measured by DNase-seq or ATAC-seq) [6,7,8] and enrichment of post-translational modification of histone tails (as assessed by ChIP-seq), including H3K4me1 and H3K27ac [10,11,9]. While these epigenetic features can successfully identify the location of many enhancers across the genome, they cannot readily differentiate between active and non-active enhancers [12,13]. However, recent genomic assays have shown that enhancers tend to be bound by RNA polymerase II (Pol II) and transcribed, producing non-coding RNAs known as enhancer RNAs ('eRNAs') [14,15,16]. While the functions of the enhancer RNA transcripts 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 as a predictor of enhancer activity [15,16,17,18,19,2,20,21,22,23].

In recent years, advances in technology have facilitated the large scale functional characterization of enhancer activity [24,25,26,27] and annotation of genome-wide binding sites of TFs in various cell types and tissues [28,4]. However, due to countless cell types, experimental conditions and the large number of TFs [29], 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 can be very challenging. Furthermore, analyses that predict TF binding sites (TFBSs), which are usually 4-12 nucleotides in length [30], using TF binding profile databases [31,32,33], fail to consider that such sequences occur frequently by chance throughout the genome and that TF binding is cell type specific [34]. To overcome these limitations, we established a novel method 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.

In TFSEE, we integrate enhancer location and activity, TF motif prediction for each enhancer and the level of TF expression (Figure 2). We have previously demonstrated TFSEE in the identification of key breast cancer subtype-specific transcription factors determining growth outcomes [35]. In the studies presented herein, we demonstrate the broader use of TFSEE to identify enhancers and TFs during the differentiation of embryonic stem cells into pancreatic progenitor cells. Taken together our results show that TFSEE can be used to perform multilayer genomic data integration to uncover novel cell type-specific transcription factors that control lineage-specific enhancers (Figure 3A).

# Results

## The TFSEE model

The TFSEE model integrates multiple genomics assays, GRO-seq, RNA-seq, and ChIP-seq, data with TF motif information to predict TFs driving the formation of active enhancers and the locations of their cognate enhancers. The TFSEE model consists of five key data processing steps (Figure 1) followed by a data integration stage (Figure 2). In step 1 of TFSEE, a universe of active enhancers across the different constituent cell types are identified. The enhancers can be identified either by enhancer transcription (GRO-seq or total RNA-seq) (Figure S1A) or enrichment of epigenomic marks (H3K4me1 and H3K27ac) (Figure S1B). In step 2 of TFSEE, genome-wide enhancer activity levels are assessed by calculating the enrichment (H3K4me1 and H3K27ac) and eRNA transcription (GRO-seq or total RNA-seq) profiles under the universe of enhancers per cell type. TFSEE was designed to detect enhancer activity changes and TF:enhancer links for each cell type. All TF to enhancer links are determined by a de novo motif search and summarizing the probability of that TF using the tools in steps 3-4 of TFSEE, which creates a table annotating enhancer to TF for each cell type. For all TFs identified TFSEE calculates the expression profile using (GRO-seq or RNA-seq) across every cell type in step 5.

The final stage integrates all the data in steps 1-5 (Figure 2) to determine TFSEE score matrix and heatmap. First, we generate an enhancer activity matrix $A_{CxE}$ for all cell types C for the universe active of E enhancers, as determined from step 2. We assume that the enhancer activity of each cell type is linearly correlated to the amount enhancer transcription (GRO-seq or total RNA-seq, G), and to the epigenomic marks (H3K4me1, M and H3K27ac, H). To reduce bias each individual enhancer enrichment is scaled between 0 and 1. Enhancer activity can be expressed as the following formula:

$$A = G + M + H$$

Next, the enhancer activity matrix $A_{CxE}$, is combined with motif prediction matrix $T_{ExF}$, represent scaled motif prediction p-values, T, for each enhancer E, to form an intermediate matrix product. This matrix product is entrywise combined with TF expression matrix R, from step 5, the expression of each TF F for each cell type C, into a resulting matrix Z composed of C cell types and F TFs. TFSEE can be expressed as the following formula:

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

## Choice of biological model system and data

To better understand the TF-driven transcriptional programs using TFSEE, we used previously published transcriptional and epigenomic data from time course differentiation of human embryonic

stem cells (hESC) to pancreatic endoderm (PE) [36,37] (Figure 3A). For these analyses, we used GRO-seq and RNA-seq, as well as ChIP-seq for 3 different histone modifications (Figure 3A).

### Predicting Enhancers by Active Transcription and Histone Modification

## Discussion

## Acknowledgments

# Material and Methods

## Genomic Data Curation

We used previously published GRO-seq, ChIP-seq and RNA-seq data from [36,37] time course differentiation of human embryonic stem cells (hESC) to pancreatic endoderm (PE). All data sets are available from NCBI's Gene Expression Omnibus [38] or EMBL-EBI's ArrayExpress [39] repositories using the accession numbers listed in Table S1.

Table S1: **Description and accession numbers of GRO-seq, ChIP-seq and RNA-seq datasets.**

| Assay | Accessions |
|---|---|
| GRO-seq | GSM1316306, GSM1316313, GSM1316320, GSM1316327, GSM1316334 |
| H3K4me3 ChIP-seq | ERR208008, ERR208014, ERR207998, ERR20798, ERR207999 |
| H3K4me1 ChIP-seq | GSM1316302, GSM1316303, GSM1316309, GSM1316316, GSM1316317, GSM1316310, GSM1316323, GSM1316324, GSM1316330, GSM1316331 |
| H3K27ac ChIP-seq | GSM1316300, GSM1316301, GSM1316307, GSM1316308, GSM1316314, GSM1316315, GSM1316321, GSM1316322, GSM1316328, GSM1316329 |
| Input ChIP-seq | ERR208001, ERR208012, ERR207984, ERR208011, ERR207986, GSM1316304, GSM1316305, GSM1316311, GSM1316312, GSM1316318, GSM1316319, GSM1316325, GSM1316326, GSM1316332, GSM1316333 |
| RNA-seq | ERR266333, ERR266335, ERR266337, ERR266338, ERR266341, ERR266342, ERR266344, ERR266346, ERR266349, ERR266351 |

## Analysis of ChIP-seq Data Sets

The raw reads were aligned to the human reference genome (GRCh37/hg19) using default parameters in Bowtie version 1.0.0 [40]. The aligned reads were subsequently filtered for quality and uniquely mappable reads were retained for further analysis using Samtools version 0.1.19 [41] and Picard version 1.127 [42]. Library complexity was measured using BEDTools version 2.17.0 [43] and meets ENCODE data quality standards [44]. Relaxed peaks were called using MACS

version 2.1.0 [45] with a p-value of $1 \times 10^{-2}$ for each replicate, pooled replicates' reads and pseudoreplicates. Peak calls from the pooled replicates that are either observed in both replicates, or in both pseudoreplicates were used for subsequent analysis.

## Analysis of RNA-seq Data Sets

The raw reads were aligned to the human reference genome (GRCh37/hg19) using default parameters in STAR version 2.4.2a [46]. Quantification of genes against Gencode version 19 [47] annotations was done using default parameters in RSEM version 1.2.31 [48].

## 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_trimer in fastx-toolkit version 0.0.13.2 [49]. The trimmed reads were aligned to the human reference genome (GRCh37/hg19) using default parameters in BWA version 0.7.12 [50].

## 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 [51] with default parameters.

## Defining Transcription Start Sites

We made distinct transcription start sites (TSS) for protein-coding genes from Gencode version 19 [47] annotations using MakeGencodeTSS [52].

## Enhancer calling by GRO-seq

### *Transcript calling*.

Transcript calling was performed using a two-state hidden Markov model using the groHMM data analysis package version 3.4 [16,21,53] on each individual cell lines. The negative log transition probability of the switch between transcribed state to non-transcribed state and the variance in read counts in the non-transcribed state that are used to predict the transcription units for the cell lines in this study are listed Table S2.

Table S2: **groHMM tunning parameters.**

| Cell Line | -Log Transition Probability | Variance in read counts |
|---|---|---|
| hES | 50 | 45 |
| DE | 50 | 35 |
| GT | 50 | 50 |
| FG | 50 | 35 |
| PE | 50 | 35 |

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 Enhancer Transcripts*.

We filtered and collected a subset of short intergenic transcripts $<$ 9 kb in length and $>$ 3 kb away from known transcription start sites (TSSs) of protein-coding genes from Gencode version 19 annotations [47], and H3K4me3 peaks. These were further classified into (1) short paired eRNAs and (2) short unpaired eRNAs as described previously [19]. For the short paired eRNAs, the sum of the GRO-seq RPKM values for both strands of DNA was used to determine if an enhancer transcript pair is expressed using a cutoff of RPKM $\geq$ 0.5. An RPKM cutoff of $\geq$ 1 was used to determine the universe expressed short unpaired eRNAs. The comprehensive of expressed eRNAs (short paired and short unpaired) was assembled using the cutoffs noted above for each cell line was used for further analyses.

### *Motif Analyses*.

De novo motif analyses was performed on a 1 kb region ($\pm$ 500 bp) surrounding the overlap center or the transcription start site for short paired and short unpaired eRNAs, respectively, using the command-line version of MEME from MEME Suite version 4.11.1 [54]. 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) [32] using TOMTOM [33].

## Enhancer calling by ChIP-seq

### *Calling Active Enhancers*.

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 $>$ 3kb from known TSS, protein coding genes from Gencode version 19 annotations [47], and H3K4me3 peaks. A RPKM cutoff of $\geq$ 1 for H3K4me1 and H3K27ac 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*.

De novo motif analyses were performed on a 1 kb region ($\pm$ 500 bp) surrounding the peak summit for the top 10000 enhancers, using the command-line version of MEME-ChIP from MEME Suite version 4.11.1 [54,55]. 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) [32] using TOMTOM [33].

## Generating Heatmaps and Clusters

For each cell line, the functional scores were Z-score normalized. To identify cognate transcription factors by cell type, we performed hierarchical clustering by calculating the Euclidean distance using clustermap from seaborn version 0.7.1 [51]. For visualization of the multidimensional TFSEE scores, we performed t-distributed stochastic neighbor embedding analysis (t-SNE) [56] 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 [57,58,59].

## 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. 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 [60,61]. Wilcoxon rank sum tests were performed to determine the statistical significance of all comparisons.

# Figures and Figure Legends

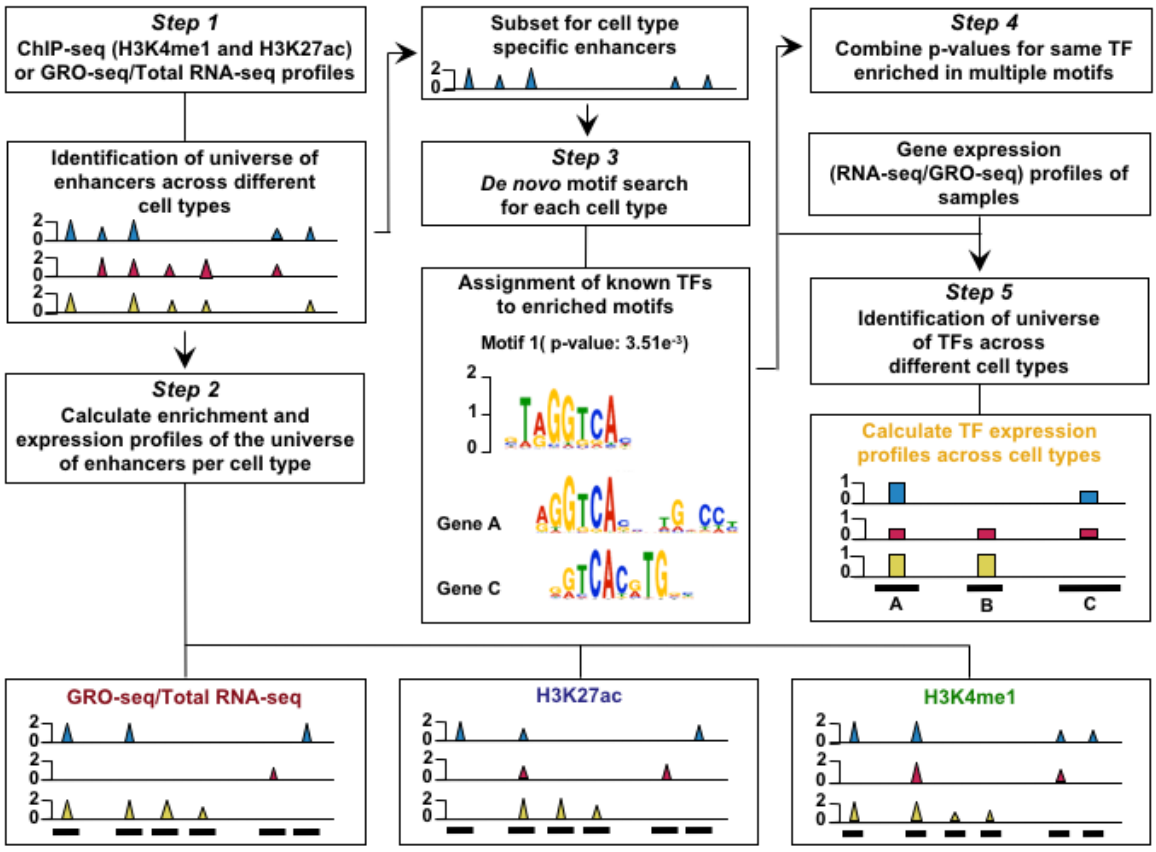



**Figure 1 – Malladi et al. (2018)**

Figure 1: **Data Processing for Total Functional Score of Enhancer Elements (TFSEE) Method.** The TFSEE method has five 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 cell types. In step 2, TFSEE calculates the enrichment (H3K4me1 and H3K27ac) and eRNA transcription (GRO-seq and total RNA-seq) profiles under all identified active enhancers per cell type. Cell type-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 eRNA 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.

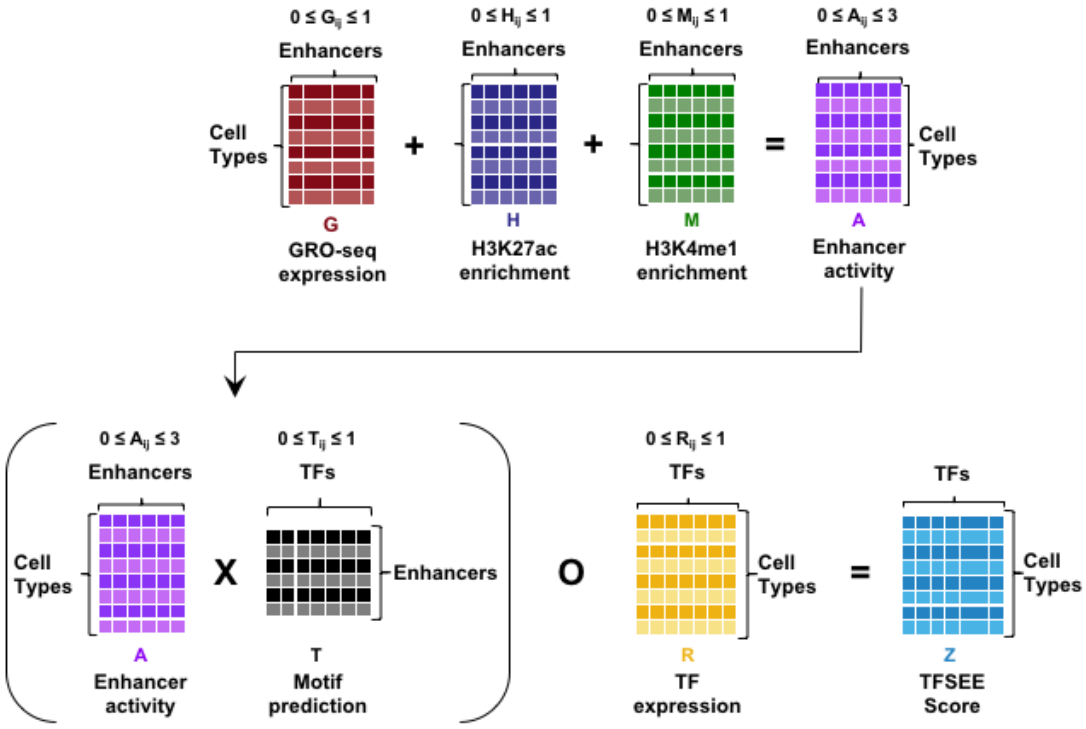



**Figure 2 – Malladi et al. (2018)**

Figure 2: **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 in panel A, 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, is combined with motif prediction matrix T, represent 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.

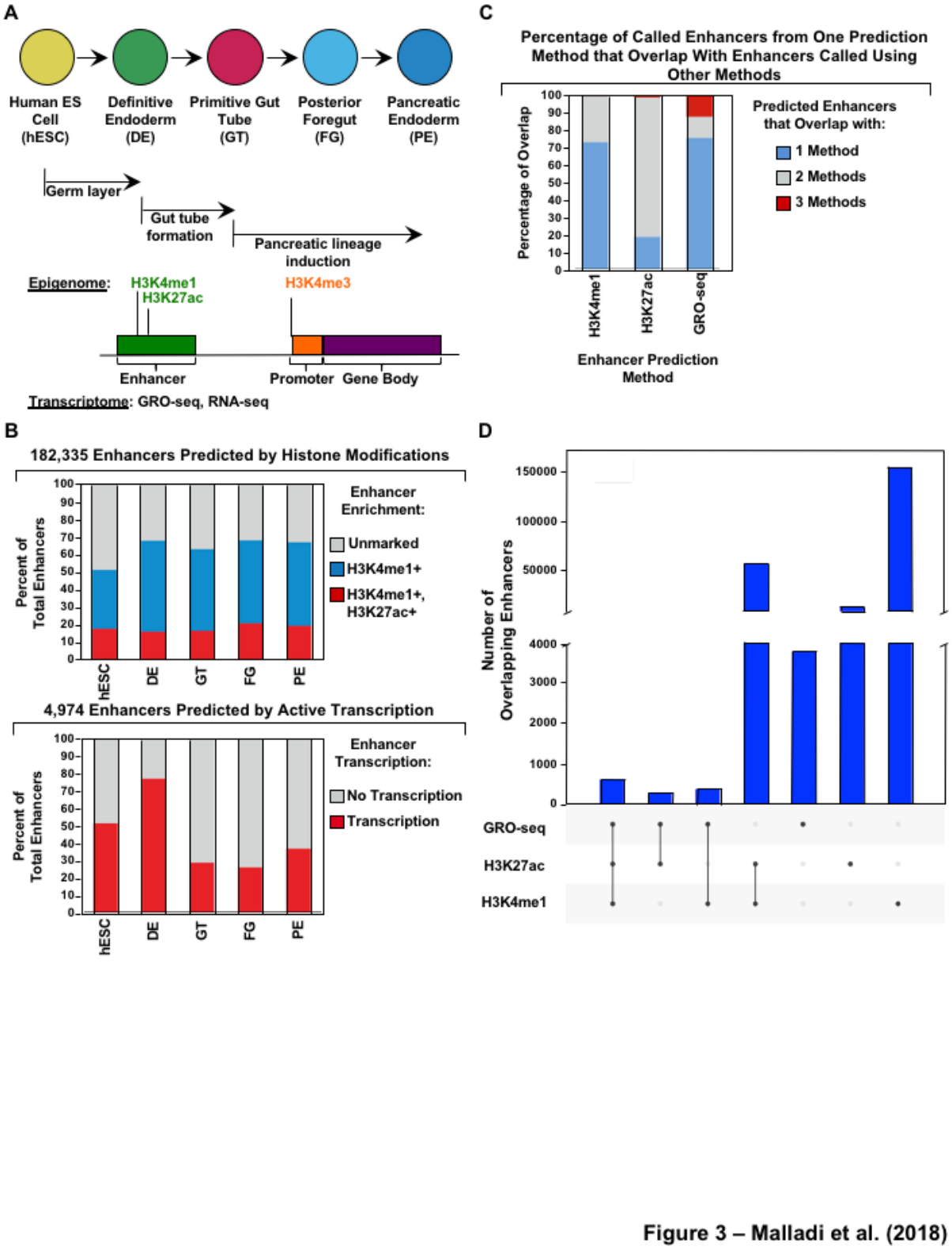



**Figure 3 – Malladi et al. (2018)**

Figure 3: **Comparison of genome-wide prediction of enhancers during pancreatic differentiation. (A)** *(Top)* Schematic 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 expression 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 by using H3K4me1 enrichment, or H3K27ac enrichment. 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.

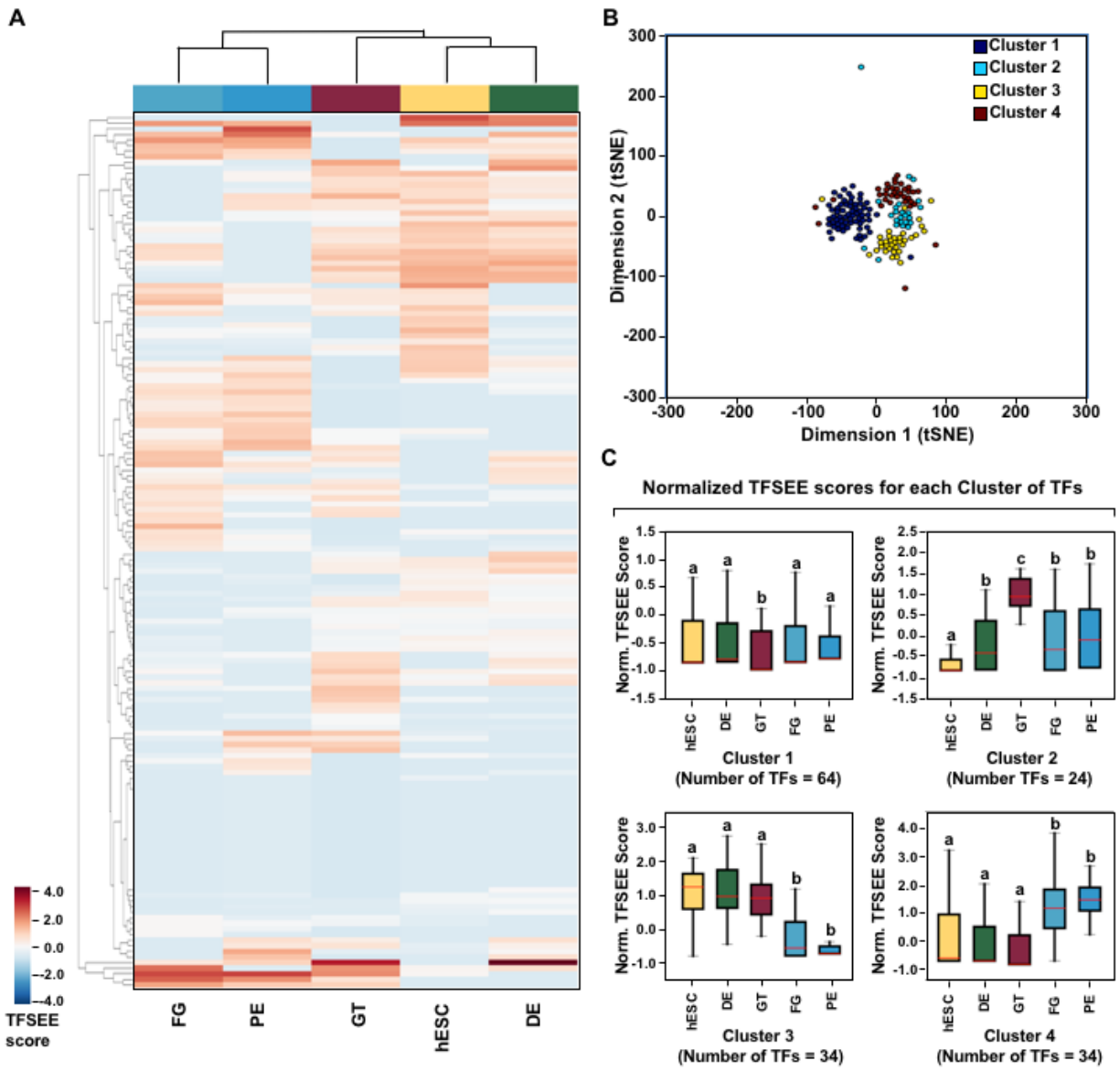



**Figure 4 – Malladi et al. (2018)**

Figure 4: **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 four 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, $p1 \times 10^{-4}$). 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).

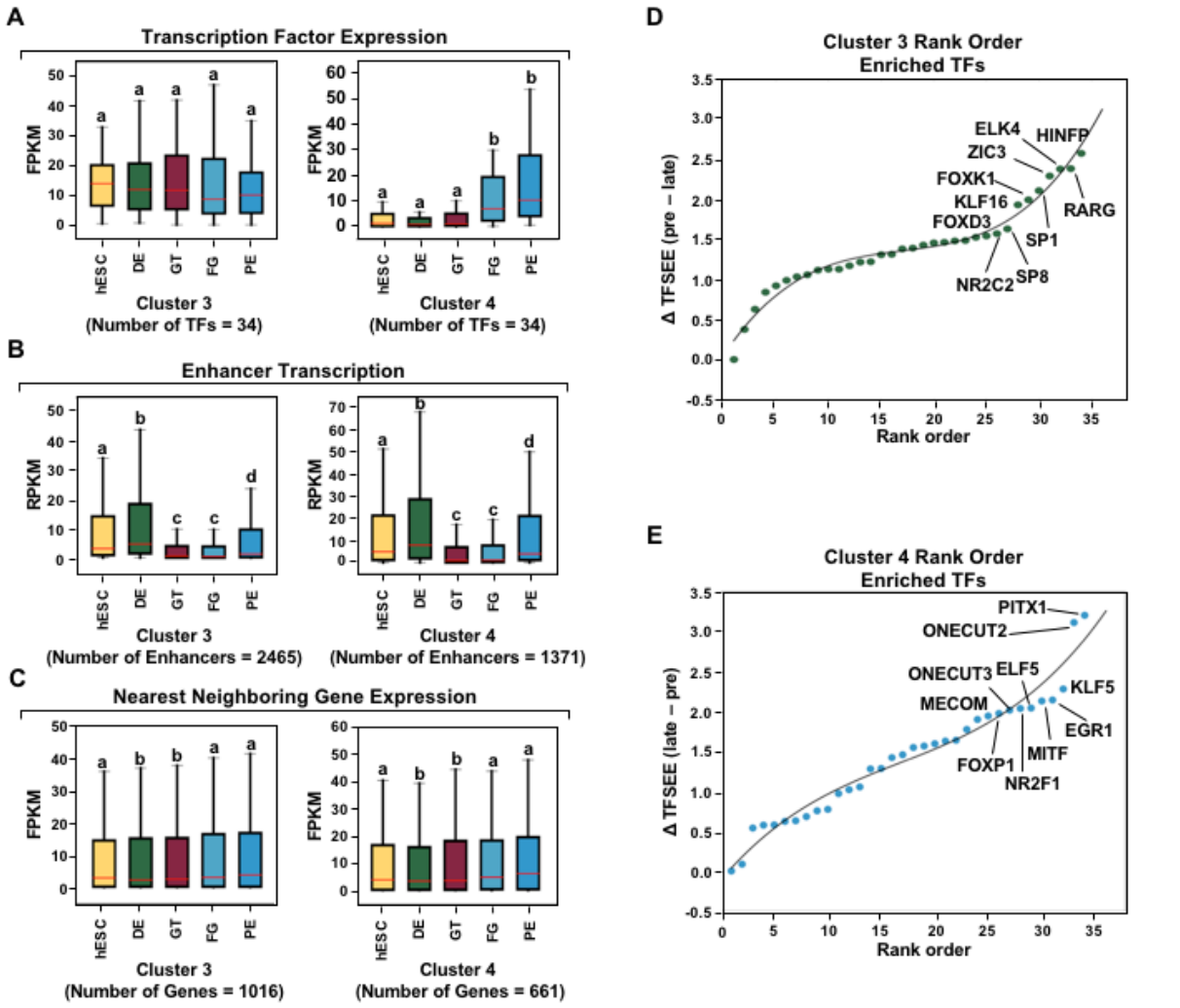
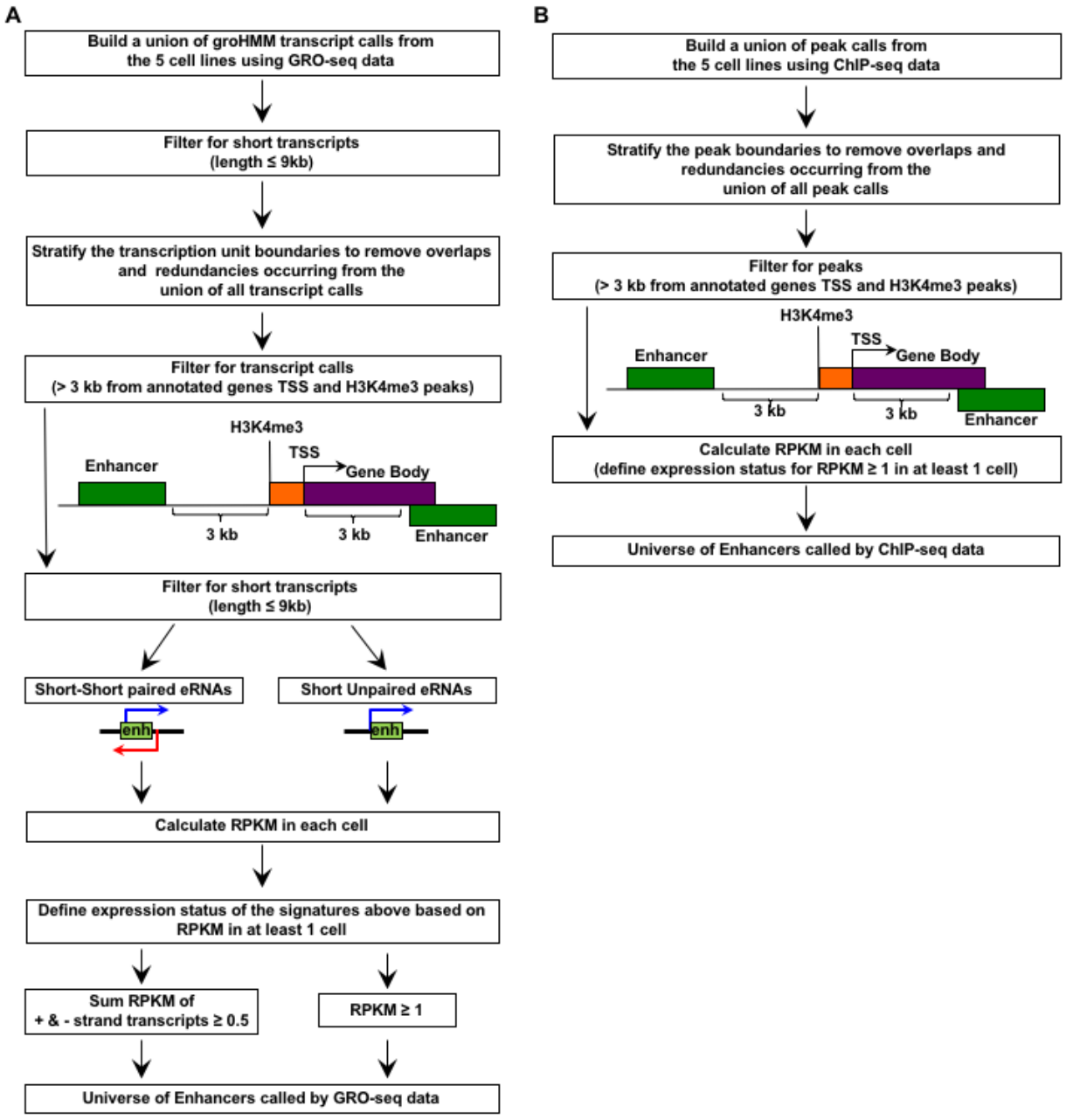



**Figure 5 – Malladi et al. (2018)**

Figure 5: **TFSEE-Predicted TFs are enriched in pre- and late- pancreatic differentiation. (A-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

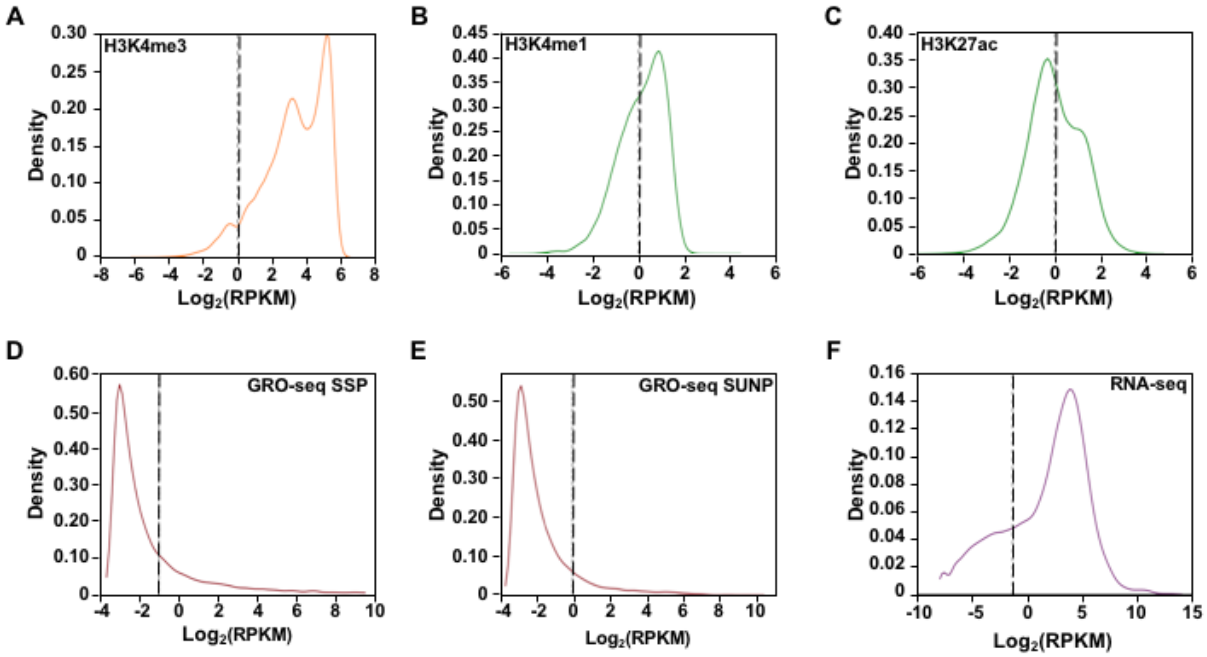
letters are significantly different from each other (Wilcoxon rank sum test). hESC (human embryonic stem cell); DE (definitive endoderm); GT (primitive gut tube); FG (posterior foregut); PE (pancreatic endoderm). **(A)** TFs identified in cluster 3 by TFSEE show equal expression across differentiation. While, cluster 4 highlights TFs highly expressed in FG and PE. TF expression as measured by RNA-seq. Number of TFs in each cluster are in parenthesis. ($p1 \times 10^{-4}$) **(B)** Enhancer transcription as measured by GRO-seq. Number of enhancers in each cluster are in parenthesis. $p1 \times 10^{-4}$). **(C)** Gene expression as measured by RNA-seq. Number of genes in each cluster are in parenthesis. ($p0.05$) **(D and E)** Rank order of TFs enriched in the Cluster 3 and the Cluster 4 identified using TFSEE. The top ten TFs in each Cluster are noted.




**Supp. Figure 1 – Malladi et al. (2018)**

Figure S1: **Unbiased, genome-wide prediction of active enhancers. (A)** Overview of the computational pipeline used for the genome-wide annotation of enhancer transcripts (eRNAs) and prediction of active enhancers using GRO-seq data. **(B)** Overview of the computational pipeline

used for the genome-wide annotation of and prediction of active enhancers using ChIP-seq (H3K4me1 and H3K27ac) data.




**Supp. Figure 2 – Malladi et al. (2018)**

Figure S2: **Density plots of enhancer and gene expression levels across all cell types.** Kernel density plots of log-transformed RPKM and FPKM values for determining active enhancers and genes. The dashed grey line represents the minimum expression cutoff. **(A)** Density plot of H3K4me3 (promoter mark) cutoff RPKM $\geq 1$. **(B)** Density plot of H3K4me1 (enhancer mark) cutoff

RPKM $\geq 1$. **(C)** Density plot of H3K27ac (enhancer mark) cutoff RPKM $\geq 1$. **(D)** Density plot of short-short paired GRO-seq transcription (SSP) (enhancer mark) cutoff RPKM $\geq 1$. **(E)** Density plot of short-unpaired GRO-seq transcription (SUNP) ( enhancer mark) cutoff RPKM $\geq 0.5$. **(F)** Density plot of RNA-seq (gene expression) cutoff FPKM $\geq 0.4$

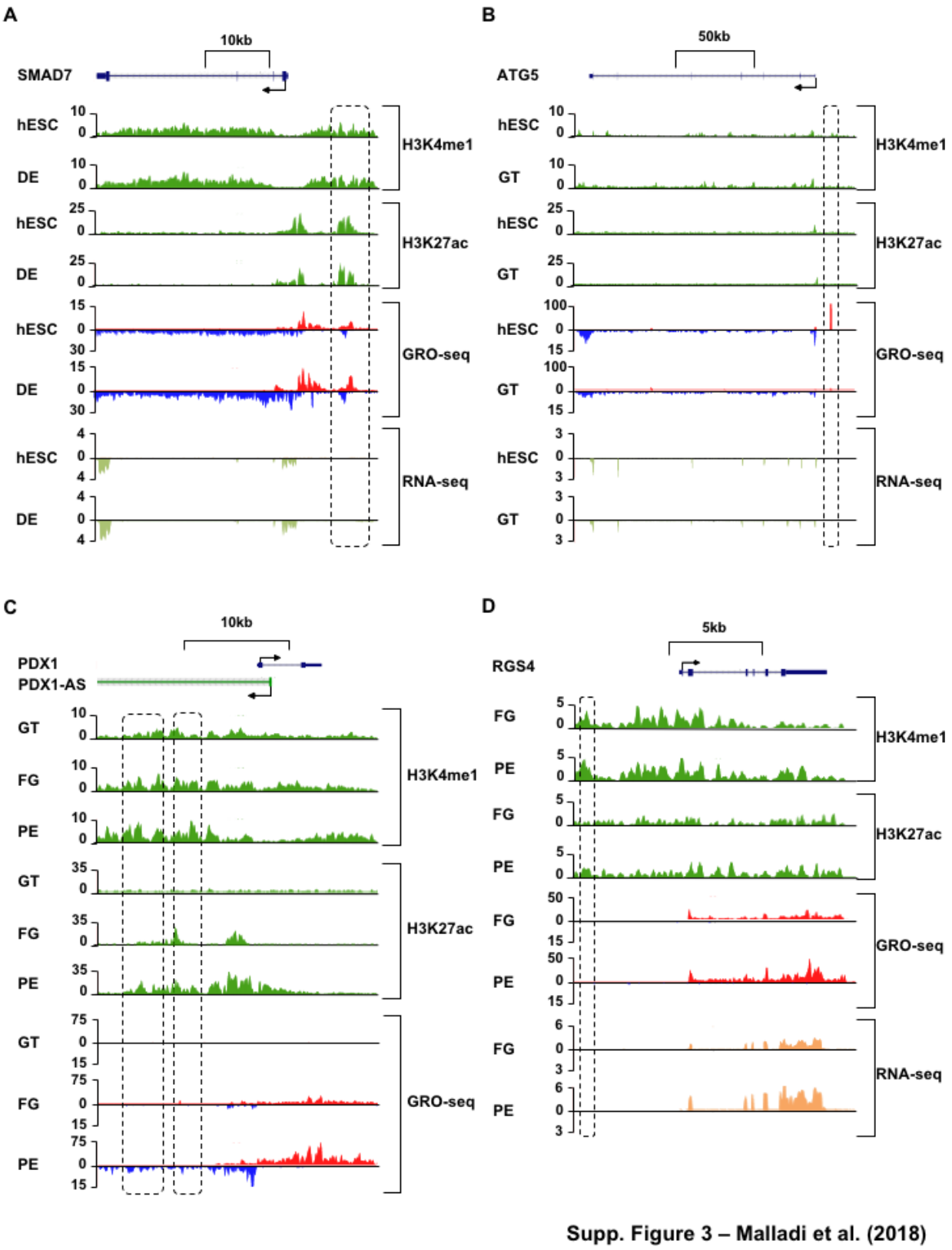
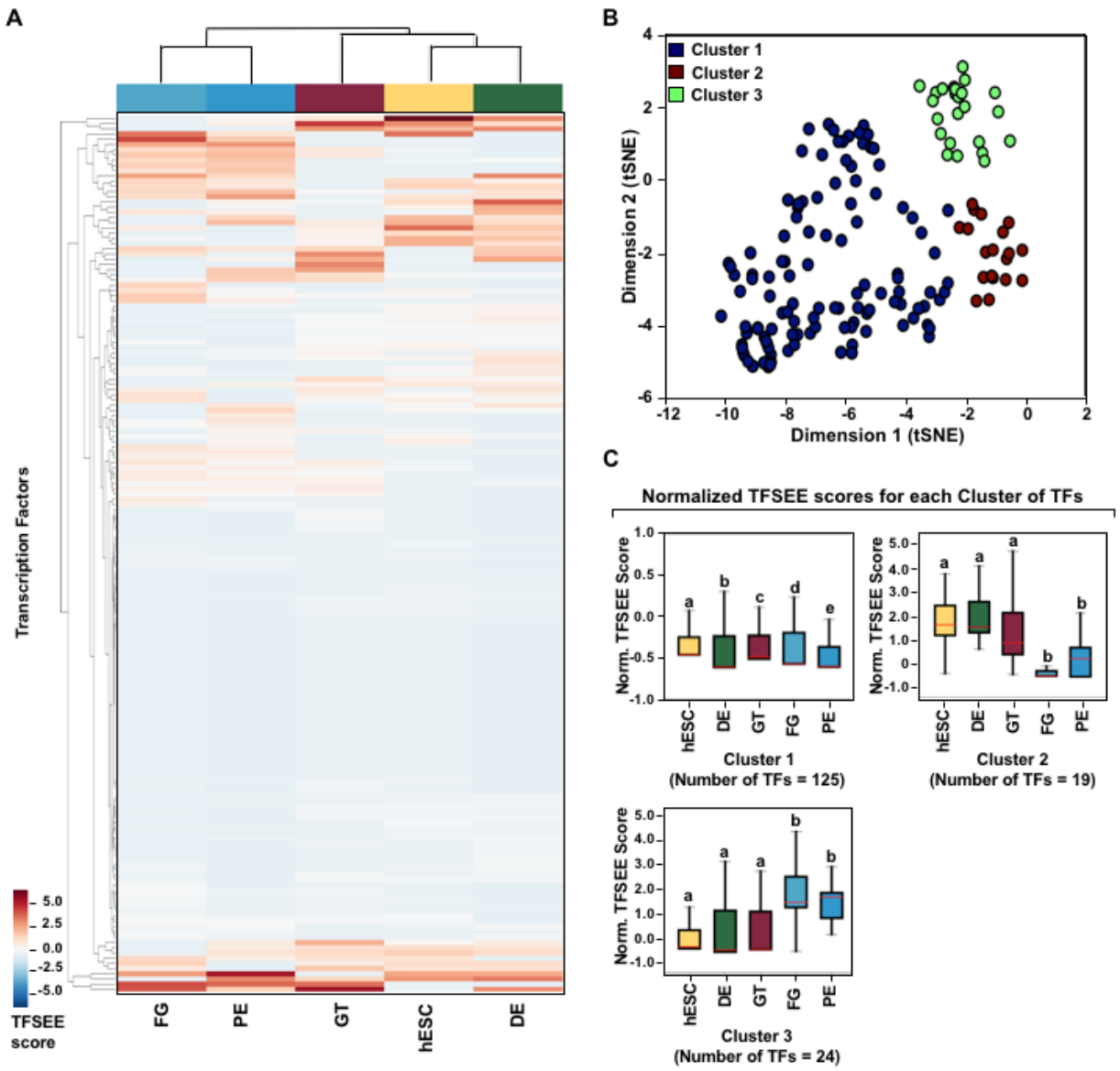



Figure S3: **Enhancer transcription is a better predictor of enhancer activity and target gene expression than other features of active chromatin. (A-D)** UCSC Genome browser views of GRO-seq, histone modification ChIP-seq and RNA-seq data showing a transcribed enhancer *(black box with dashed line)* and its nearest neighboring gene. hESC (human embryonic stem cell);

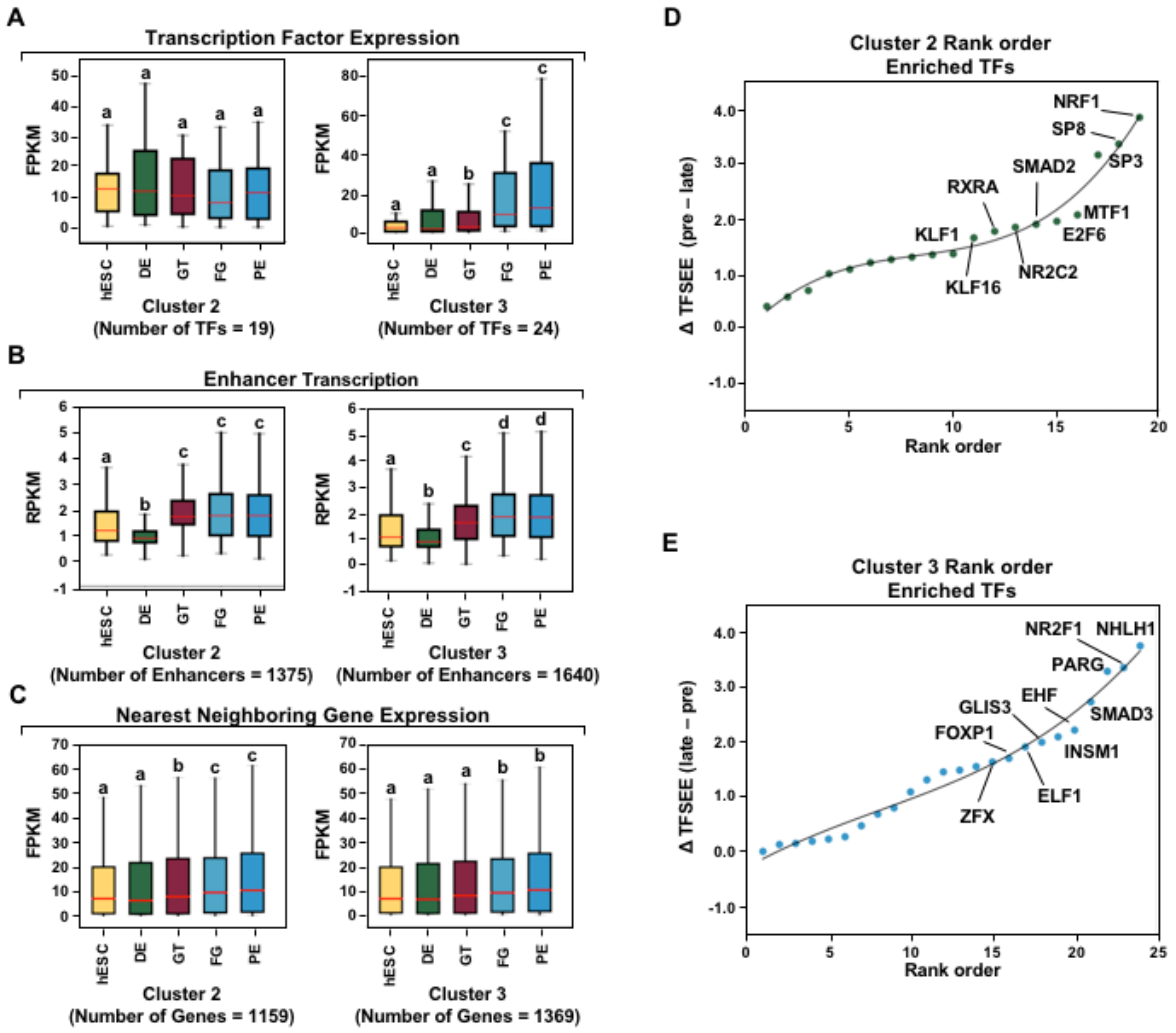
DE (definitive endoderm); GT (primitive gut tube); FG (posterior foregut); PE (pancreatic endoderm). **(A)** Browser view showing a transcribed enhancer and its nearest neighboring gene (SMAD7). The data highlights histone modifications typically enriched at enhancers *(green)*, however the increased transcription determined by GRO-seq *(red/blue)* for DE correlates to expression of nearest genes determined by RNA-seq *(orange/light green)*. **(B)** Browser view showing a transcribed enhancer and its nearest neighboring gene (ATG5. The data highlights an enhancer identified by GRO-seq *(red/blue)*, however lacks typical histone modifications enriched at enhancers *(green*). The increased transcription determined by GRO-seq for hESC correlates to expression of nearest genes determined by RNA-seq *(orange/light green)*. **(C)** Browser view showing a transcribed enhancer and its nearest neighboring gene (PDX1). The data highlights an enhancer identified by histone modifications enriched at enhancers *(green)*, however increased transcription determined by GRO-seq *(red/blue)* correlates with antisense gene (AS-PDX1). **(D)** Browser view showing a transcribed enhancer and its nearest neighboring gene (RGS4). The data highlights an enhancer identified by histone modifications enriched at enhancers *(green)*, however lacks enhancer transcription identified by GRO-seq *(red/blue)*. The increased enhancer signal determined by histone modifications for PE shows correlates to expression of nearest genes determined by RNA-seq *(orange/light green)* and GRO-seq *(red/blue)*.




**Supp. Figure 4 – Malladi et al. (2018)**

Figure S4: **TFSEE defined by histone modifications identifies cell type-specific enhancers and their cognate TFs that drive gene expression in pancreatic differentiation. (A)** Unsupervised hierarchical clustering of cell line normalized TFSEE scores shown in a heatmap representation.

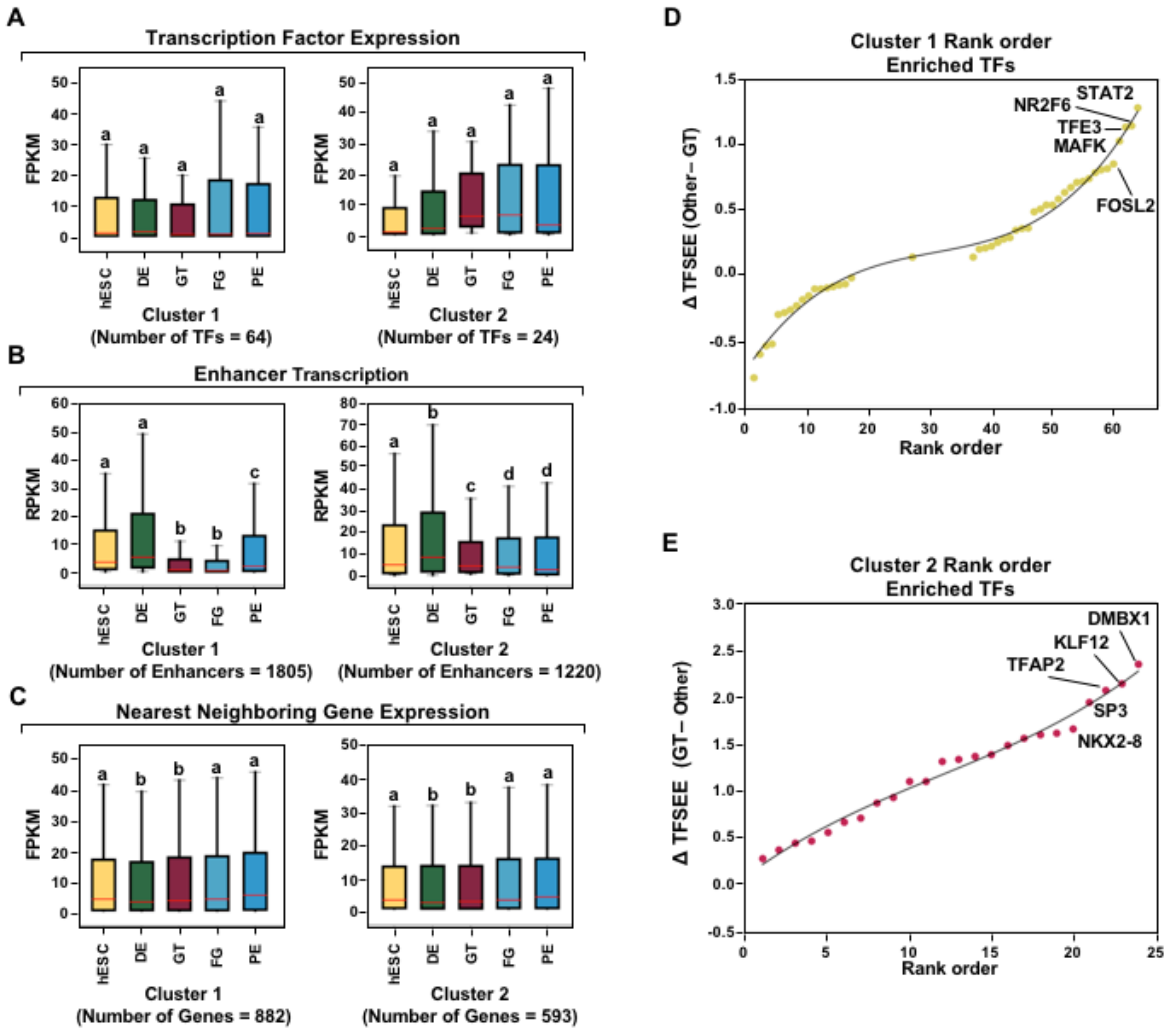
**(B)** Biaxial t-SNE clustering plot of cell type-normalized TFSEE scores showing evidence of three distinct clusters, each point represents an individual TF. **(C)** Boxplots of normalized TFSEE score for clusters identified in pancreatic differentiation. Bars marked with different letters are significantly different from each other (Wilcoxon rank sum test, $p1 \times 10^{-2}$). Number of TFs in each cluster are in parenthesis. Cluster 1, TFs associated across pancreatic lineage Cluster 2, TFs associated with pre-pancreatic lineage induction (hESC, DE and GT). Cluster 3, TFs associated with late-pancreatic differentiation (FG and PE).




**Supp. Figure 5 – Malladi et al. (2018)**

Figure S5: **TFSEE-Predicted TFs, by histone modifications, are enriched in pre- and late-pancreatic differentiation. (A-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 2) and late-pancreatic (cluster 3) 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 2 by TFSEE show equal expression across differentiation. While, cluster 3 highlights TFs highly expressed in FG and PE. TF expression as measured by RNA-seq. Number of TFs in each cluster are in parenthesis. ( $p1 \times 10^{-4}$) **(B)** Enhancer transcriptions as measured by GRO-seq. Number of enhancers in each cluster are in parenthesis. ($p1 \times 10^{-4}$). **(C)** Gene expression as measured by RNA-seq. Number of genes in each cluster are in parenthesis. ($p0.05$). **(D and E)** Rank order of TFs enriched in the Cluster 2 and the Cluster 3 identified using TFSEE. The top ten TFs in each Cluster are noted.




**Supp. Figure 6 – Malladi et al. (2018)**

Figure S6: **TFSEE-Predicted TFs are enriched and depleted in Primitive Gut Tube during pancreatic differentiation. (A-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 depleted (cluster 1) and enriched (cluster 2) in primitive gut tube during pancreatic

differentiation across 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)** TF expression as measured by RNA-seq. Number of TFs in each cluster are in parenthesis. ( $p1 \times 10^{-2}$) **(B)** Enhancer transcriptions as measured by GRO-seq. Number of enhancers in each cluster are in parenthesis. ($p1 \times 10^{-4}$). **(C)** Gene expression as measured by RNA-seq. Number of genes in each cluster are in parenthesis. ($p0.05$). **(D and E)** Rank order of TFs enriched in the Cluster 1 and the Cluster 2 identified using TFSEE. The top five TFs in each Cluster are noted.

# References


1. **Transcriptional enhancers: from properties to genome-wide predictions**
Daria Shlyueva, Gerald Stampfel, Alexander Stark
*Nature Reviews Genetics* (2014-03-11) https://doi.org/10.1038/nrg3682

2. **The selection and function of cell type-specific enhancers**
Sven Heinz, Casey E. Romanoski, Christopher Benner, Christopher K. Glass
*Nature Reviews Molecular Cell Biology* (2015-02-04) https://doi.org/10.1038/nrm3949

3. **Progress and challenges in bioinformatics approaches for enhancer identification**
Dimitrios Kleftogiannis, Panos Kalnis, Vladimir B. Bajic
*Briefings in Bioinformatics* (2015-12-03) https://doi.org/10.1093/bib/bbv101

4. **An integrated encyclopedia of DNA elements in the human genome***Nature* (2012-09)
https://doi.org/10.1038/nature11247

5. **Integrative analysis of 111 reference human epigenomes**
Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, Pouya Kheradpour, Zhizhuo Zhang, Jianrong Wang, Michael J. Ziller, … Manolis Kellis
*Nature* (2015-02-18) https://doi.org/10.1038/nature14248

6. **Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS)**
G. E. Crawford
*Genome Research* (2005-12-12) https://doi.org/10.1101/gr.4074106

7. **Patterns of regulatory activity across diverse human cell types predict tissue identity, transcription factor binding, and long-range interactions**
N. C. Sheffield, R. E. Thurman, L. Song, A. Safi, J. A. Stamatoyannopoulos, B. Lenhard, G. E. Crawford, T. S. Furey
*Genome Research* (2013-03-12) https://doi.org/10.1101/gr.152140.112

8. **Discovery of Transcription Factors and Regulatory Regions Driving In Vivo Tumor Development by ATAC-seq and FAIRE-seq Open Chromatin Profiling**
Kristofer Davie, Jelle Jacobs, Mardelle Atkins, Delphine Potier, Valerie Christiaens, Georg Halder, Stein Aerts
*PLOS Genetics* (2015-02-13) https://doi.org/10.1371/journal.pgen.1004994

9. **Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome**
Nathaniel D Heintzman, Rhona K Stuart, Gary Hon, Yutao Fu, Christina W Ching, R David

Hawkins, Leah O Barrera, Sara Van Calcar, Chunxu Qu, Keith A Ching, … Bing Ren
*Nature Genetics* (2007-02-04) https://doi.org/10.1038/ng1966

10. **Histone modifications at human enhancers reflect global cell-type-specific gene expression**
Nathaniel D. Heintzman, Gary C. Hon, R. David Hawkins, Pouya Kheradpour, Alexander Stark, Lindsey F. Harp, Zhen Ye, Leonard K. Lee, Rhona K. Stuart, Christina W. Ching, … Bing Ren
*Nature* (2009-03-18) https://doi.org/10.1038/nature07829

11. **Distinct and Predictive Histone Lysine Acetylation Patterns at Promoters, Enhancers, and Gene Bodies**
Nisha Rajagopal, Jason Ernst, Pradipta Ray, Jie Wu, Michael Zhang, Manolis Kellis, Bing Ren
*G3: Genes|Genomes|Genetics* (2014-08-12) https://doi.org/10.1534/g3.114.013565

12. **Effects on the transcriptome upon deletion of a distal element cannot be predicted by the size of the H3K27Ac peak in human cells**
Yu Gyoung Tak, Yuli Hung, Lijing Yao, Matthew R. Grimmer, Albert Do, Mital S. Bhakta, Henriette O'Geen, David J. Segal, Peggy J. Farnham
*Nucleic Acids Research* (2016-01-06) https://doi.org/10.1093/nar/gkv1530

13. **Occupancy by key transcription factors is a more accurate predictor of enhancer activity than histone modifications or chromatin accessibility**
Nergiz Dogan, Weisheng Wu, Christapher S Morrissey, Kuan-Bei Chen, Aaron Stonestrom, Maria Long, Cheryl A Keller, Yong Cheng, Deepti Jain, Axel Visel, … Ross C Hardison
*Epigenetics & Chromatin* (2015-04-23) https://doi.org/10.1186/s13072-015-0009-5

14. **A Large Fraction of Extragenic RNA Pol II Transcription Sites Overlap Enhancers**
Francesca De Santa, Iros Barozzi, Flore Mietton, Serena Ghisletti, Sara Polletti, Betsabeh Khoramian Tusi, Heiko Muller, Jiannis Ragoussis, Chia-Lin Wei, Gioacchino Natoli
*PLoS Biology* (2010-05-11) https://doi.org/10.1371/journal.pbio.1000384

15. **Architectural and Functional Commonalities between Enhancers and Promoters**
Tae-Kyung Kim, Ramin Shiekhattar
*Cell* (2015-08) https://doi.org/10.1016/j.cell.2015.08.008

16. **A Rapid, Extensive, and Transient Transcriptional Response to Estrogen Signaling in Breast Cancer Cells**
Nasun Hah, Charles G. Danko, Leighton Core, Joshua J. Waterfall, Adam Siepel, John T. Lis, W. Lee Kraus
*Cell* (2011-05) https://doi.org/10.1016/j.cell.2011.03.042

17. **Tissue-Specific RNA Expression Marks Distant-Acting Developmental Enhancers**
Han Wu, Alex S. Nord, Jennifer A. Akiyama, Malak Shoukry, Veena Afzal, Edward M. Rubin, Len A.

Pennacchio, Axel Visel
*PLoS Genetics* (2014-09-04) https://doi.org/10.1371/journal.pgen.1004610

18. **Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA**
Dong Wang, Ivan Garcia-Bassets, Chris Benner, Wenbo Li, Xue Su, Yiming Zhou, Jinsong Qiu, Wen Liu, Minna U. Kaikkonen, Kenneth A. Ohgi, … Xiang-Dong Fu
*Nature* (2011-05-15) https://doi.org/10.1038/nature10006

19. **Enhancer transcripts mark active estrogen receptor binding sites**
N. Hah, S. Murakami, A. Nagari, C. G. Danko, W. L. Kraus
*Genome Research* (2013-05-01) https://doi.org/10.1101/gr.152306.112

20. **Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers**
Leighton J Core, André L Martins, Charles G Danko, Colin T Waters, Adam Siepel, John T Lis
*Nature Genetics* (2014-11-10) https://doi.org/10.1038/ng.3142

21. **groHMM: a computational tool for identifying unannotated and cell type-specific transcription units from global run-on sequencing data**
Minho Chae, Charles G. Danko, W. Lee Kraus
*BMC Bioinformatics* (2015-07-16) https://doi.org/10.1186/s12859-015-0656-3

22. **Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation**
Wenbo Li, Dimple Notani, Qi Ma, Bogdan Tanasa, Esperanza Nunez, Aaron Yun Chen, Daria Merkurjev, Jie Zhang, Kenneth Ohgi, Xiaoyuan Song, … Michael G. Rosenfeld
*Nature* (2013-06) https://doi.org/10.1038/nature12210

23. **TNFα Signaling Exposes Latent Estrogen Receptor Binding Sites to Alter the Breast Cancer Cell Transcriptome**
Hector L. Franco, Anusha Nagari, W. Lee Kraus
*Molecular Cell* (2015-04) https://doi.org/10.1016/j.molcel.2015.02.001

24. **VISTA Enhancer Browser–a database of tissue-specific human enhancers**
A. Visel, S. Minovitsky, I. Dubchak, L. A. Pennacchio
*Nucleic Acids Research* (2007-01-03) https://doi.org/10.1093/nar/gkl822

25. **Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay**
P. Kheradpour, J. Ernst, A. Melnikov, P. Rogov, L. Wang, X. Zhang, J. Alston, T. S. Mikkelsen, M. Kellis
*Genome Research* (2013-03-19) https://doi.org/10.1101/gr.144899.112

26. **Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-seq**
C. D. Arnold, D. Gerlach, C. Stelzer, L. M. Boryn, M. Rath, A. Stark
*Science* (2013-01-17) https://doi.org/10.1126/science.1232542

27. **High-throughput functional testing of ENCODE segmentation predictions**
Jamie C. Kwasnieski, Christopher Fiore, Hemangi G. Chaudhari, Barak A. Cohen
*Genome Research* (2014-07-17) https://doi.org/10.1101/gr.173518.114

28. **Architecture of the human regulatory network derived from ENCODE data**
Mark B. Gerstein, Anshul Kundaje, Manoj Hariharan, Stephen G. Landt, Koon-Kiu Yan, Chao Cheng, Xinmeng Jasmine Mu, Ekta Khurana, Joel Rozowsky, Roger Alexander, … Michael Snyder
*Nature* (2012-09) https://doi.org/10.1038/nature11245

29. **A census of human transcription factors: function, expression and evolution**
Juan M. Vaquerizas, Sarah K. Kummerfeld, Sarah A. Teichmann, Nicholas M. Luscombe
*Nature Reviews Genetics* (2009-04) https://doi.org/10.1038/nrg2538

30. **DNA-dependent formation of transcription factor pairs alters their binding specificity**
Arttu Jolma, Yimeng Yin, Kazuhiro R. Nitta, Kashyap Dave, Alexander Popov, Minna Taipale, Martin Enge, Teemu Kivioja, Ekaterina Morgunova, Jussi Taipale
*Nature* (2015-11-09) https://doi.org/10.1038/nature15518

31. **DNA-Binding Specificities of Human Transcription Factors**
Arttu Jolma, Jian Yan, Thomas Whitington, Jarkko Toivonen, Kazuhiro R. Nitta, Pasi Rastas, Ekaterina Morgunova, Martin Enge, Mikko Taipale, Gonghong Wei, … Jussi Taipale
*Cell* (2013-01) https://doi.org/10.1016/j.cell.2012.12.009

32. **JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles**
Anthony Mathelier, Oriol Fornes, David J. Arenillas, Chih-yu Chen, Grégoire Denay, Jessica Lee, Wenqiang Shi, Casper Shyr, Ge Tan, Rebecca Worsley-Hunt, … Wyeth W. Wasserman
*Nucleic Acids Research* (2015-11-03) https://doi.org/10.1093/nar/gkv1176

33. **Quantifying similarity between motifs**
Shobhit Gupta, John A Stamatoyannopoulos, Timothy L Bailey, William Noble
*Genome Biology* (2007) https://doi.org/10.1186/gb-2007-8-2-r24

34. **Functional analysis of transcription factor binding sites in human promoters**
Troy W Whitfield, Jie Wang, Patrick J Collins, E Christopher Partridge, Shelley Aldred, Nathan D Trinklein, Richard M Myers, Zhiping Weng
*Genome Biology* (2012) https://doi.org/10.1186/gb-2012-13-9-r50

35. **Enhancer transcription reveals subtype-specific gene expression programs controlling breast cancer pathogenesis**
Hector L. Franco, Anusha Nagari, Venkat S. Malladi, Wenqian Li, Yuanxin Xi, Dana Richardson, Kendra L. Allton, Kaori Tanaka, Jing Li, Shino Murakami, … W. Lee Kraus
*Genome Research* (2017-12-22) https://doi.org/10.1101/gr.226019.117

36. **Dynamic Chromatin Remodeling Mediated by Polycomb Proteins Orchestrates Pancreatic Differentiation of Human Embryonic Stem Cells**
Ruiyu Xie, Logan J. Everett, Hee-Woong Lim, Nisha A. Patel, Jonathan Schug, Evert Kroon, Olivia G. Kelly, Allen Wang, Kevin A. D'Amour, Allan J. Robins, … Maike Sander
*Cell Stem Cell* (2013-02) https://doi.org/10.1016/j.stem.2012.11.023

37. **Epigenetic Priming of Enhancers Predicts Developmental Competence of hESC-Derived Endodermal Lineage Intermediates**
Allen Wang, Feng Yue, Yan Li, Ruiyu Xie, Thomas Harper, Nisha A. Patel, Kayla Muth, Jeffrey Palmer, Yunjiang Qiu, Jinzhao Wang, … Maike Sander
*Cell Stem Cell* (2015-04) https://doi.org/10.1016/j.stem.2015.02.013

38. **GEO**
Gene Expression Omnibus
https://www.ncbi.nlm.nih.gov/geo/

39. **ArrayExpress**
ArrayExpress – functional genomics data
http://www.ebi.ac.uk/arrayexpress/

40. **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome**
Ben Langmead, Cole Trapnell, Mihai Pop, Steven L Salzberg
*Genome Biology* (2009) https://doi.org/10.1186/gb-2009-10-3-r25

41. **The Sequence Alignment/Map format and SAMtools**
H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin,
*Bioinformatics* (2009-06-08) https://doi.org/10.1093/bioinformatics/btp352

42. **Picard**
Broad Institute
*GitHub* http://broadinstitute.github.io/picard/

43. **BEDTools: a flexible suite of utilities for comparing genomic features**
Aaron R. Quinlan, Ira M. Hall
*Bioinformatics* (2010-01-28) https://doi.org/10.1093/bioinformatics/btq033

44. **ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia**
S. G. Landt, G. K. Marinov, A. Kundaje, P. Kheradpour, F. Pauli, S. Batzoglou, B. E. Bernstein, P. Bickel, J. B. Brown, P. Cayting, … M. Snyder
*Genome Research* (2012-09-01) https://doi.org/10.1101/gr.136184.111

45. **Identifying ChIP-seq enrichment using MACS**
Jianxing Feng, Tao Liu, Bo Qin, Yong Zhang, Xiaole Shirley Liu
*Nature Protocols* (2012-08-30) https://doi.org/10.1038/nprot.2012.101

46. **STAR: ultrafast universal RNA-seq aligner**
Alexander Dobin, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, Thomas R. Gingeras
*Bioinformatics* (2012-10-25) https://doi.org/10.1093/bioinformatics/bts635

47. **GENCODE: The reference human genome annotation for The ENCODE Project**
J. Harrow, A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, B. L. Aken, D. Barrell, A. Zadissa, S. Searle, … T. J. Hubbard
*Genome Research* (2012-09-01) https://doi.org/10.1101/gr.135350.111

48. **RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome**
Bo Li, Colin N Dewey
*BMC Bioinformatics* (2011) https://doi.org/10.1186/1471-2105-12-323

49. **FASTX-Toolkit**
Hannon Lab
http://hannonlab.cshl.edu/fastx_toolkit/

50. **Fast and accurate short read alignment with Burrows-Wheeler transform**
H. Li, R. Durbin
*Bioinformatics* (2009-05-18) https://doi.org/10.1093/bioinformatics/btp324

51. **Seaborn: V0.7.1 (June 2016)**
Michael Waskom, Olga Botvinnik, Drewokane, Paul Hobson, David, Yaroslav Halchenko, Saulius Lukauskas, John B. Cole, Jordi Warmenhoven, Julian De Ruiter, … Antony Lee
*Zenodo* (2016-06-05) https://doi.org/10.5281/zenodo.54844

52. **MakeGenecodeTSS**
Sarah Djebali
*GitHub* https://github.com/sdjebali/MakeGencodeTSS

53. **groHMM**
Minho Chae Charles G. Danko
*Bioconductor* (2017) https://doi.org/10.18129/b9.bioc.grohmm

54. **MEME SUITE: tools for motif discovery and searching**
T. L. Bailey, M. Boden, F. A. Buske, M. Frith, C. E. Grant, L. Clementi, J. Ren, W. W. Li, W. S. Noble
*Nucleic Acids Research* (2009-05-20) https://doi.org/10.1093/nar/gkp335

55. **MEME-ChIP: motif analysis of large DNA datasets**
Philip Machanick, Timothy L. Bailey
*Bioinformatics* (2011-04-12) https://doi.org/10.1093/bioinformatics/btr189

56. **Visualizing data using t-SNE**
Laurens van der Maaten, Geoffrey Hinton
*Journal of Machine Learning Research 9* (2008-11)

57. **Scikit-learn: Machine Learning in Python**
Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Gilles Louppe, Peter Prettenhofer, Ron Weiss, … Édouard Duchesnay
*arXiv* (2012-01-02) https://arxiv.org/abs/1201.0490v3

58. **Visualizing Large-scale and High-dimensional Data**
Jian Tang, Jingzhou Liu, Ming Zhang, Qiaozhu Mei
*Proceedings of the 25th International Conference on World Wide Web - WWW '16* (2016) https://doi.org/10.1145/2872427.2883041

59. **Scikit-Learn: 0.17.1 Release Tag For Doi**
Olivier Grisel, Andreas Mueller, Fabian Pedregosa, Lars, Alexandre Gramfort, Gilles Louppe, Peter Prettenhofer, Mathieu Blondel, Vlad Niculae, Arnaud Joly, … Maheshakya Wijewardena
*Zenodo* (2016-04-17) https://doi.org/10.5281/zenodo.49911

60. **Matplotlib: A 2D Graphics Environment**
John D. Hunter
*Computing in Science & Engineering* (2007) https://doi.org/10.1109/mcse.2007.55

61. **Matplotlib/Matplotlib V2.0.2**
Michael Droettboom, Thomas A Caswell, John Hunter, Eric Firing, Jens Hedegaard Nielsen, Nelle Varoquaux, Benjamin Root, Phil Elson, Darren Dale, Jae-Joon Lee, … Nikita Kniazev
*Zenodo* (2017-05-10) https://doi.org/10.5281/zenodo.573577