Supplemental Material

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

Venkat Malladi, Anusha Nagari, Hector L. Franco, and W. Lee Kraus

This document contains the following supplementary information. Page

1) Supplemental Figures

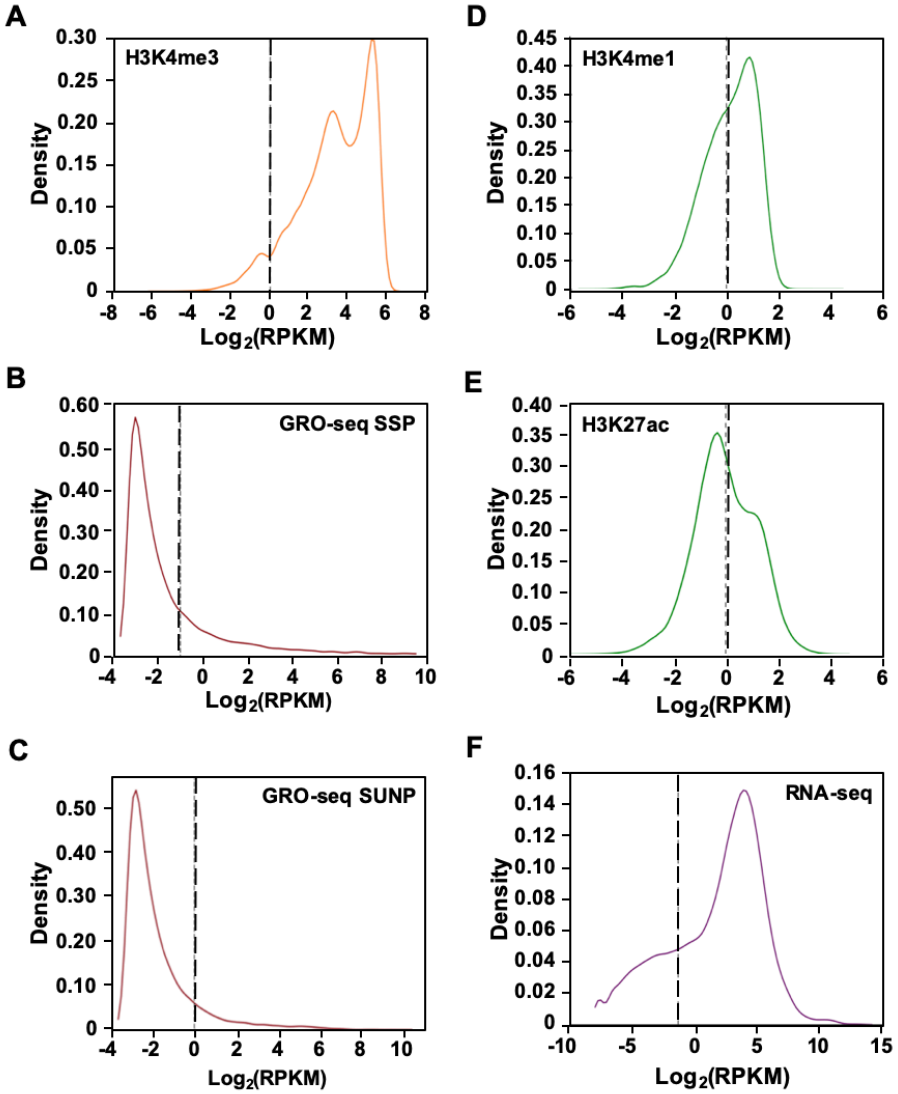


Figure S1. Density plots of enhancer transcription and gene expression levels across all cell types.

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 ≥ 1.

(B) Density plot of short-short paired GRO-seq transcription (SSP) (enhancer mark), cutoff RPKM ≥ 1.

(C) Density plot of short-unpaired GRO-seq transcription (SUNP) (enhancer mark), cutoff RPKM ≥ 0.5.

(D) Density plot of H3K4me1 (enhancer mark), cutoff RPKM ≥ 1.

(E) Density plot of H3K27ac (enhancer mark), cutoff RPKM ≥ 1.

(F) Density plot of RNA-seq (gene expression), cutoff FPKM ≥ 0.4

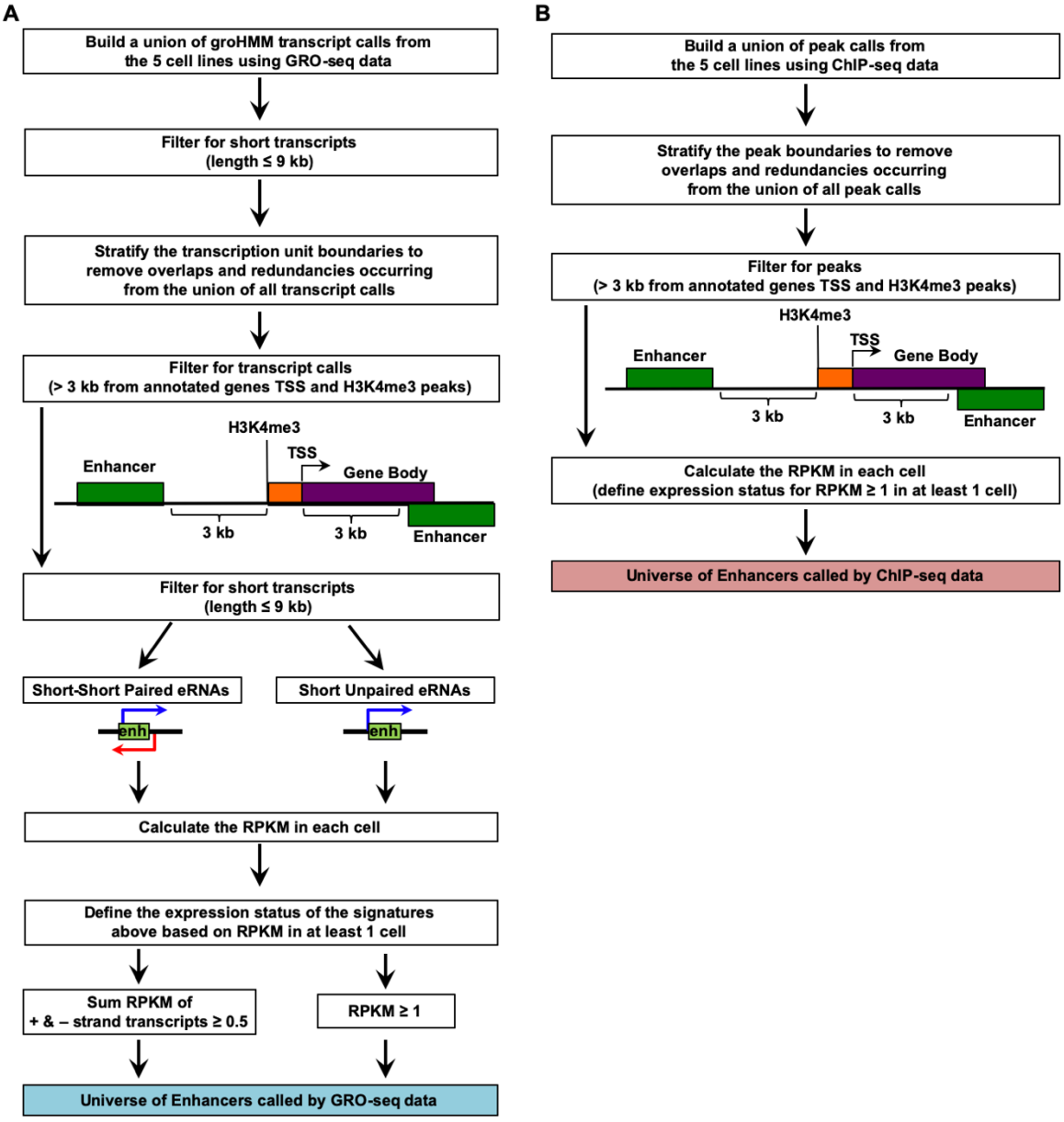
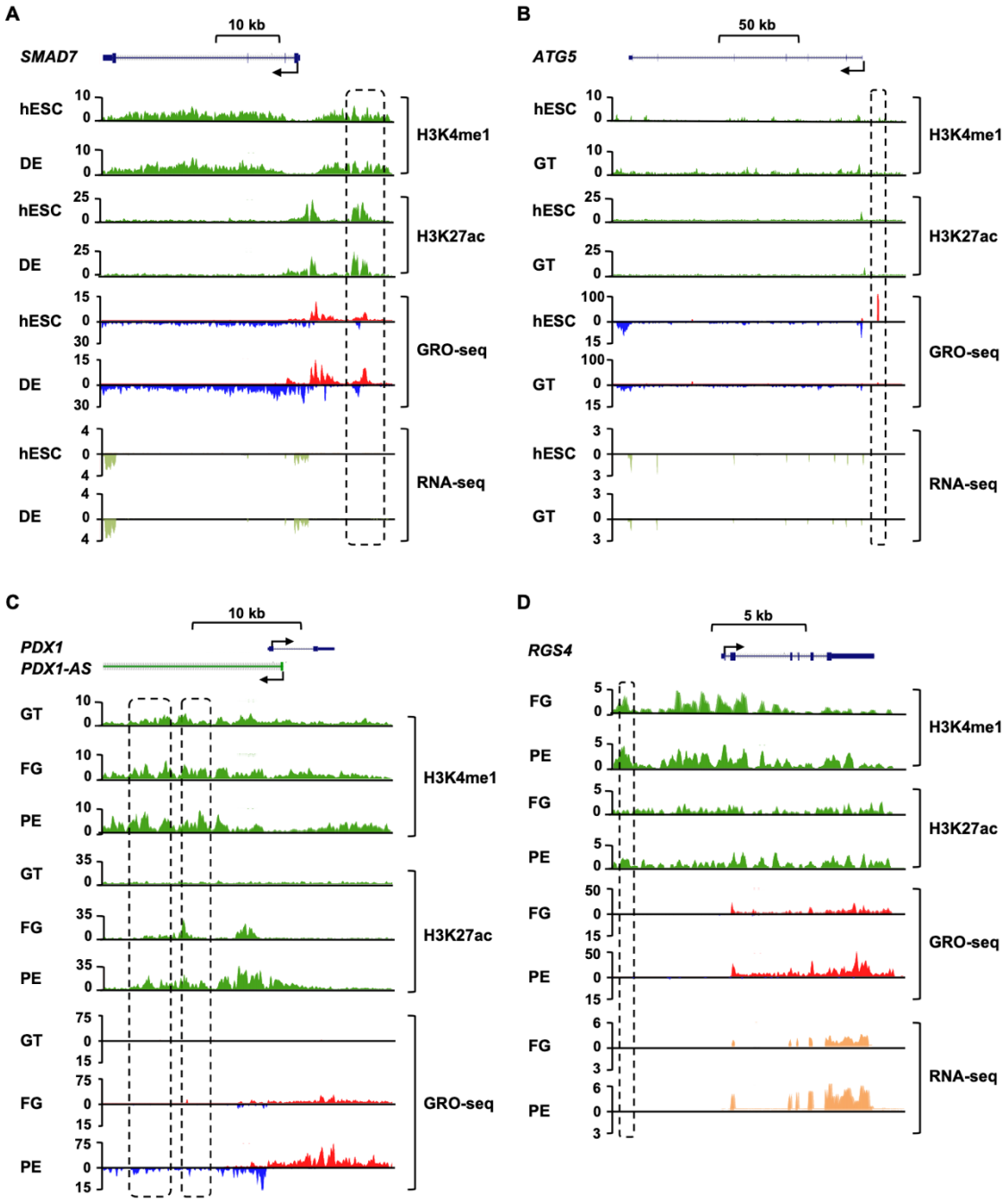


Figure S2. Unbiased, genome-wide prediction of active enhancers.
(A) Overview of the computational pipeline used for the genome-wide annotation of enhancer transcripts (eRNAs) and prediction of active enhancers using GRO-seq data.
(B) Overview of the computational pipeline used for the genome-wide annotation of and prediction of active enhancers using ChIP-seq (H3K4me1 and H3K27ac) data.

[See the figure on the next page]

Figure S3. Enhancer transcription is a more robust predictor of enhancer activity and target gene expression than other features of active chromatin.
UCSC Genome browser views of GRO-seq, histone modification ChIP-seq, and RNA-seq data showing transcribed enhancers (black box with dashed line) and their nearest neighboring genes. 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 highlight histone modifications typically enriched at enhancers (green). The increased transcription determined by GRO-seq (red/blue) for DE correlates with the expression of the nearest gene as determined by RNA-seq (orange/light green).
(B) Browser view showing a transcribed enhancer and its nearest neighboring gene (*ATG5*). The data highlight an enhancer identified by GRO-seq (red/blue) lacking 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 highlight an enhancer identified by histone modifications enriched at enhancers (green). The 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 highlight an enhancer identified by histone modifications enriched at enhancers (green) lacking enhancer transcription identified by GRO-seq (red/blue). The increased enhancer signal determined by histone modifications for PE correlates with the expression of the nearest gene as determined by RNA-seq (orange/light green) and GRO-seq (red/blue).

Figure S3.


A
10 kb
SMAD7
hESC
DE
H3K4me1
H3K27ac
GRO-seq
RNA-seq
B
50 kb
ATG5
hESC
GT
C
10 kb
PDX1
PDX1-AS
GT
FG
PE
D
5 kb
RGS4

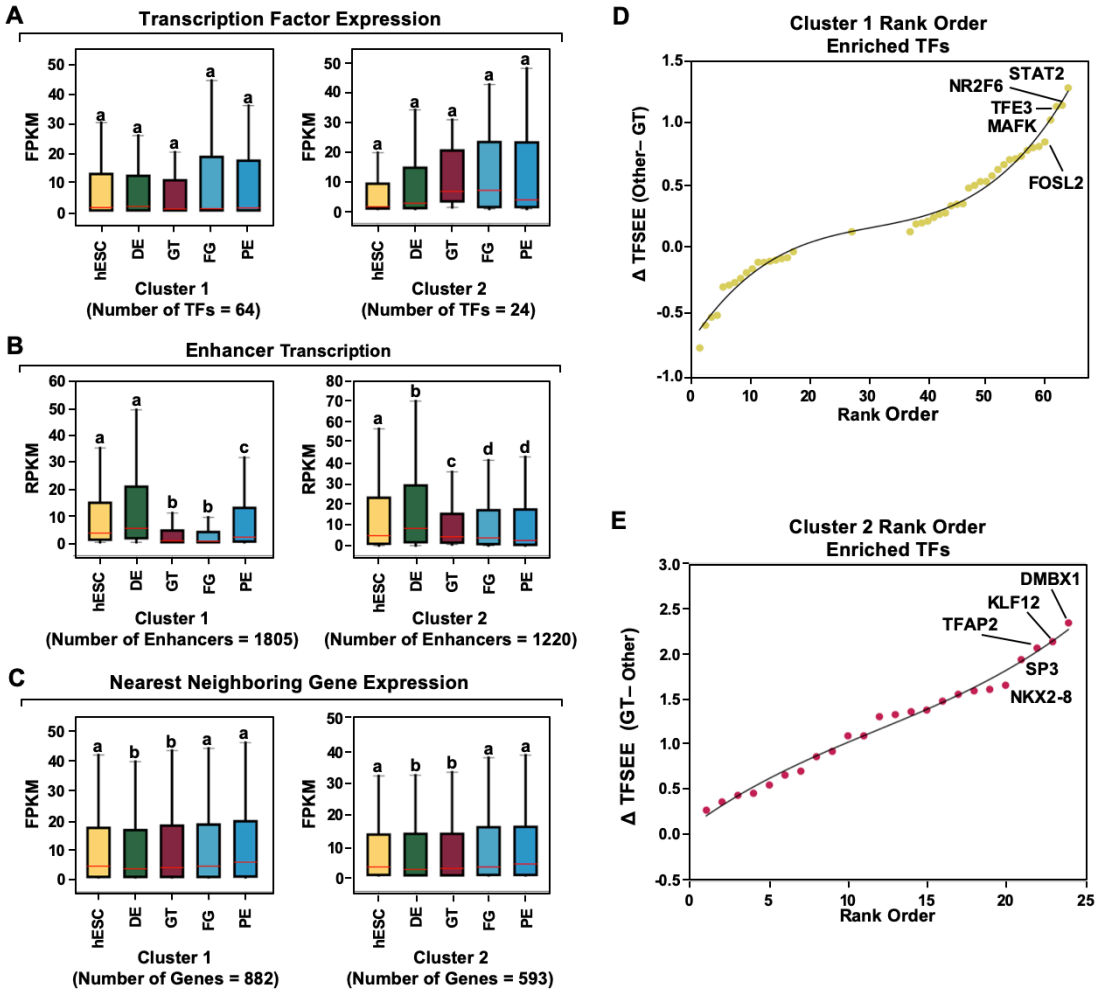


Figure S4. TFSEE-predicted TFs are enriched or depleted in the primitive gut tube during pancreatic differentiation.

(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. The number of TFs in each cluster are in parentheses ($p < 1 \times 10^{-2}$). *(B)* Enhancer transcription as measured by GRO-seq. The number of enhancers in each cluster are in parentheses ($p < 1 \times 10^{-4}$). *(C)* Gene expression as measured by RNA-seq. The number of genes in each cluster are in parentheses ($p < 0.05$).

(D and E) Rank order of TFs enriched in Cluster 1 and Cluster 2 identified using TFSEE. The top five TFs in each cluster are noted.

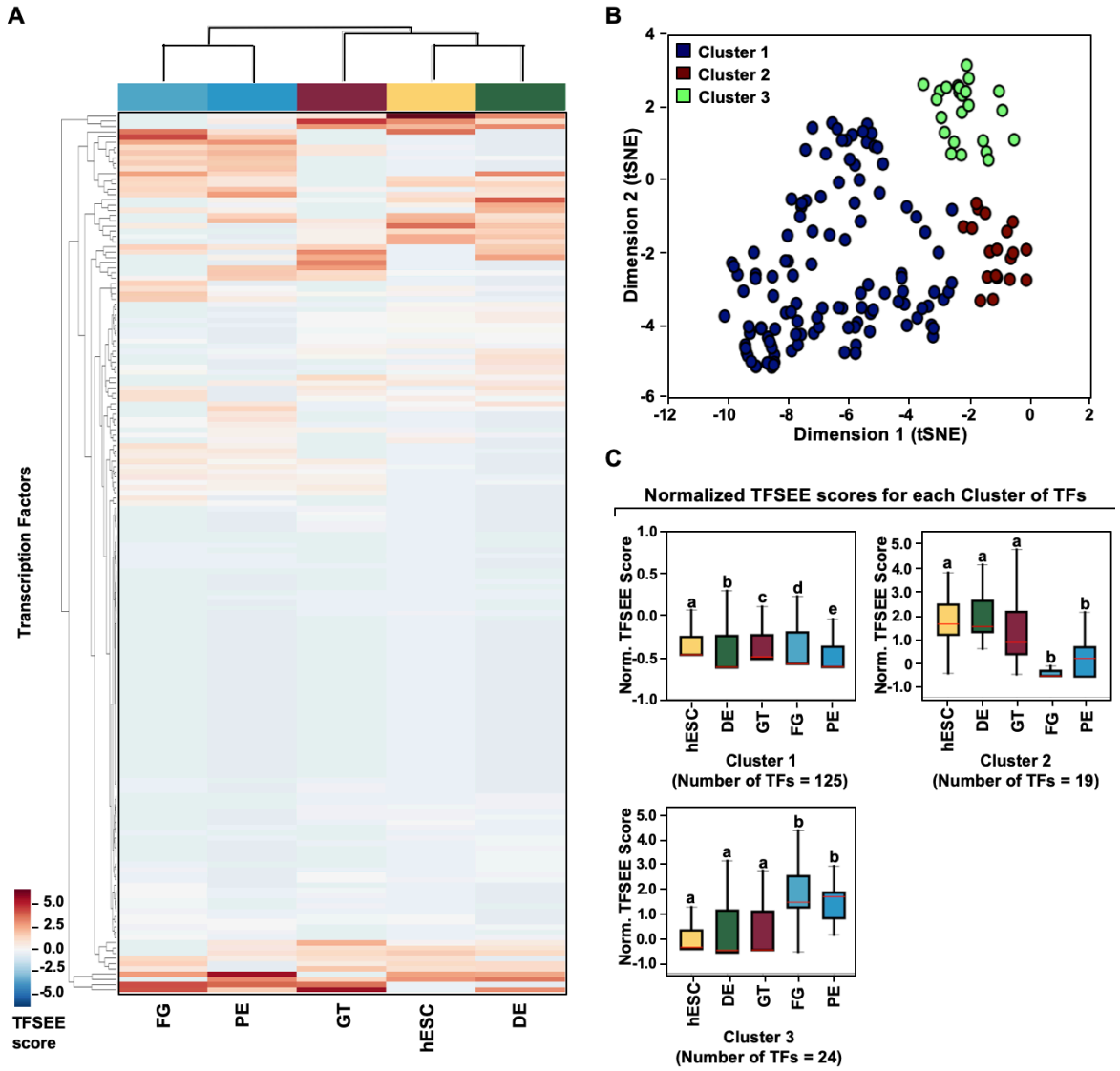


Figure S5. TFSEE using enhancers called by histone modifications identifies cell type-specific enhancers and their cognate TFs that drive gene expression during pancreatic differentiation.

(A) Unsupervised hierarchical clustering of cell line normalized TFSEE scores shown in a heatmap representation.

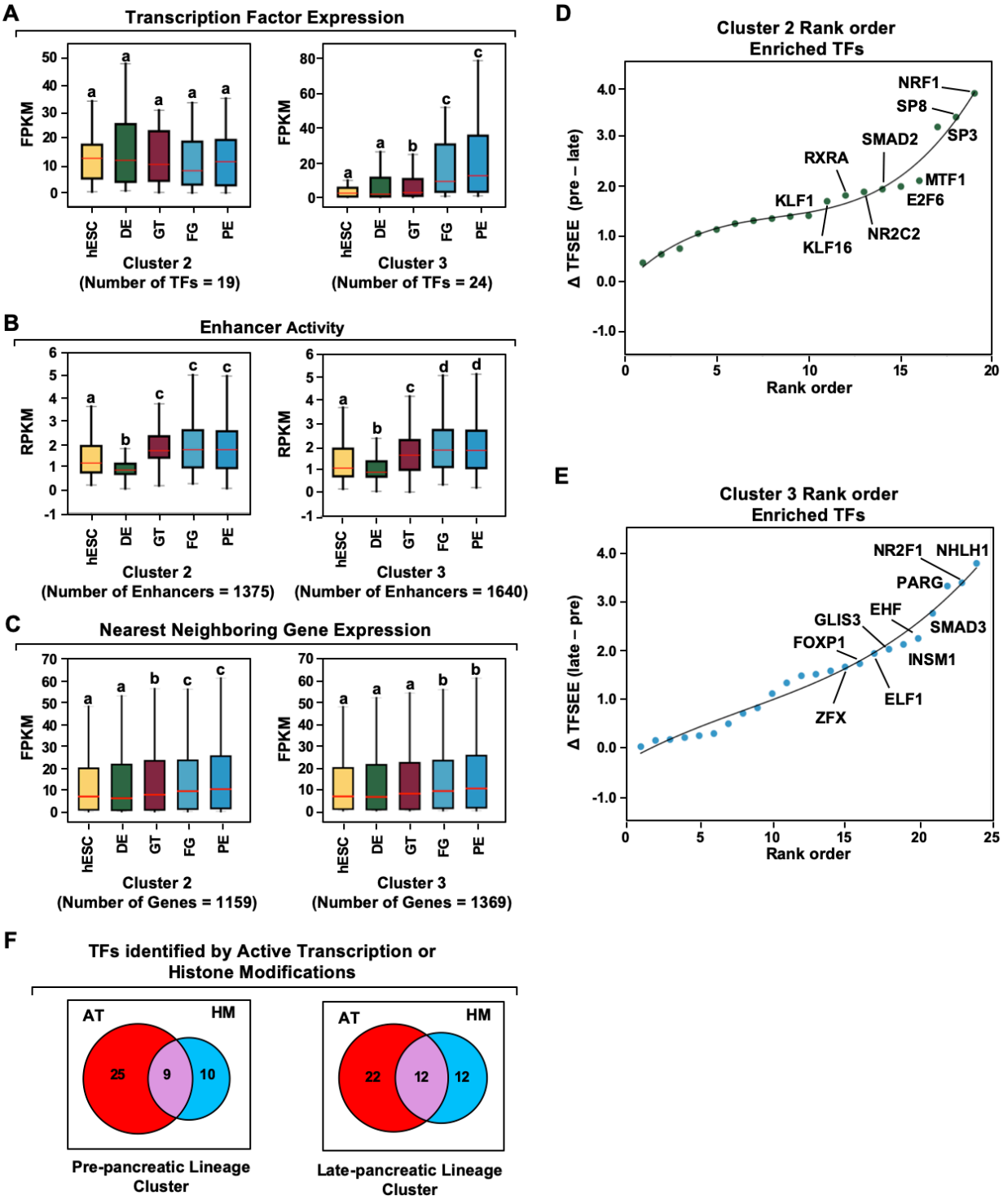
(B) Biaxial t-SNE clustering plot of cell type-normalized TFSEE scores showing evidence of three distinct clusters. Each point represents an individual TF.

(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, $p < 1 \times 10^{-2}$). The number of TFs in each cluster are in parentheses. 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).

[See the figure on the next page]

Figure S6. TFs predicted by TFSEE using enhancers called 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. Cluster 3 highlights TFs highly expressed in FG and PE. TF expression as measured by RNA-seq. The number of TFs in each cluster are in parentheses ($p < 1 \times 10^{-4}$). *(B)* Enhancer activity as determined by ChIP-seq (H3K27ac enrichment). The number of enhancers in each cluster are in parentheses ($p < 1 \times 10^{-4}$). *(C)* Gene expression as measured by RNA-seq. The number of genes in each cluster are in parentheses ($p < 0.05$).
(D and E) Rank order of TFs enriched in Cluster 2 and Cluster 3 identified using TFSEE. The top ten TFs in each cluster are noted.
(F) Venn diagram comparing TF identification by active transcription (AT) or by histone modification (HM) in pre- and late-pancreatic clusters.

Figure S6.


A
Transcription Factor Expression
FPKM
hESC
DE
GT
FG
PE
Cluster 2
(Number of TFs = 19)
Cluster 3
(Number of TFs = 24)
B
Enhancer Activity
RPKM
Cluster 2
(Number of Enhancers = 1375)
Cluster 3
(Number of Enhancers = 1640)
C
Nearest Neighboring Gene Expression
Cluster 2
(Number of Genes = 1159)
Cluster 3
(Number of Genes = 1369)
D
Cluster 2 Rank order
Enriched TFs
Δ TFSEE (pre – late)
Rank order
NRF1
SP8
SP3
SMAD2
RXRA
MTF1
E2F6
KLF1
KLF16
NR2C2
E
Cluster 3 Rank order
Enriched TFs
Δ TFSEE (late – pre)
NR2F1
NHLH1
PARG
EHF
GLIS3
SMAD3
FOXP1
INSM1
ELF1
ZFX
F
TFs identified by Active Transcription or
Histone Modifications
AT
HM
25
9
10
22
12
12
Pre-pancreatic Lineage
Cluster
Late-pancreatic Lineage
Cluster

2) Supplemental Tables

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

Assay	Accession Numbers
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

Table of assays and accession numbers of raw data for the time course of differentiation of human embryonic stem cells (hESC) to pancreatic endoderm (PE) used in this study. All data sets are available from NCBI's Gene Expression Omnibus or EMBL-EBI's ArrayExpress.

Table S2. groHMM tuning parameters.

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

Table summarizing the shape setting parameters and -log transition probabilities used to predict the transcription units for each cell lines using GRO-seq data with groHMM method.