Exploring Lineage-Specific Enhancers by Integrating Enhancer Transcription, Epigenomic Features, Sequence Motifs, and Transcription Factor Expression

This manuscript was automatically generated from vsmalladi/tfsee-manuscript@fde1a6b on November 8, 2017.

Authors

- **Venkat Malladi**
  0000-0002-0144-0564 · vsmalladi · katatonikkat

Abstract

The identification of transcription factors (TF) driving the formation of active enhancers that regulate the expression of target genes remains an open problem. We have developed a computational framework that identifies cell type-specific enhancers and their cognate TFs by integrating multiple genomic assays that probe the transcriptomes (GRO-seq and RNA-seq) and epigenomes (ChIP-seq) of various samples. Our method, called Total Functional Score of Enhancer Elements (TFSEE), integrates the magnitude of enhancer transcription (GRO-seq), enrichment of marks associated with enhancers (H3K4me1 and H3K27ac ChIP-seq), TF mRNA expression levels (RNA-seq), and TF motif p-values (MEME). This method has allowed us to explore the enhancer landscape in different cell types that share common origins or are biologically related, including distinct molecular subtypes of breast cancer, and embryonic stem cells (ESCs) and their derived lineages. 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. 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.

Introduction

Results

Overview of TFSEE model

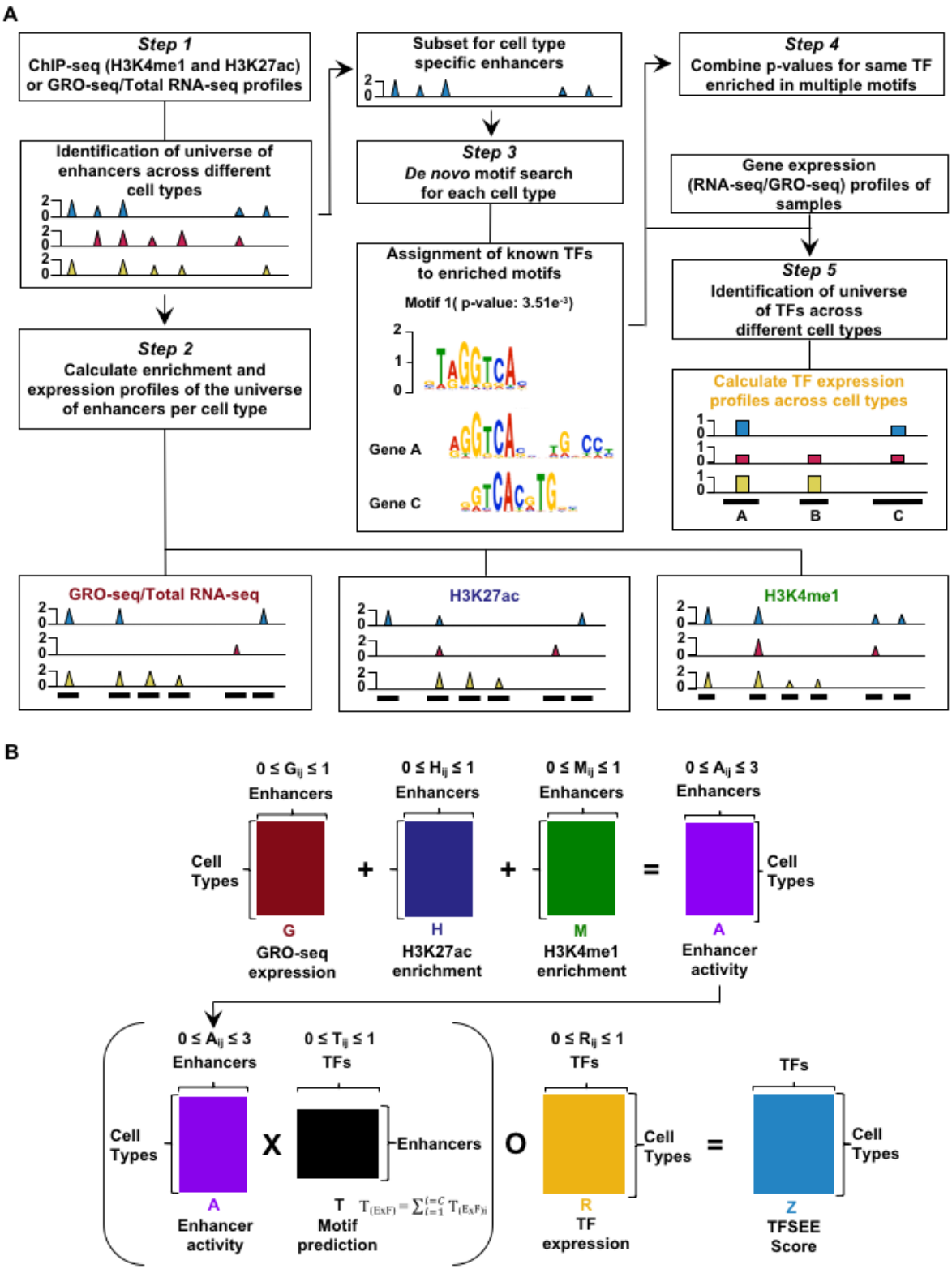

Figure 1: **Overview of Total Functional Score of Enhancer Elements (TFSEE) Method.** The TFSEE method has five steps, followed by an integration stage. TFSEE combines diverse data sets to identify cell type-specific enhancers and their cognate transcription factors (TFs). **(A)** In step 1, epigenomic (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 enriched 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. **(B)** 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.

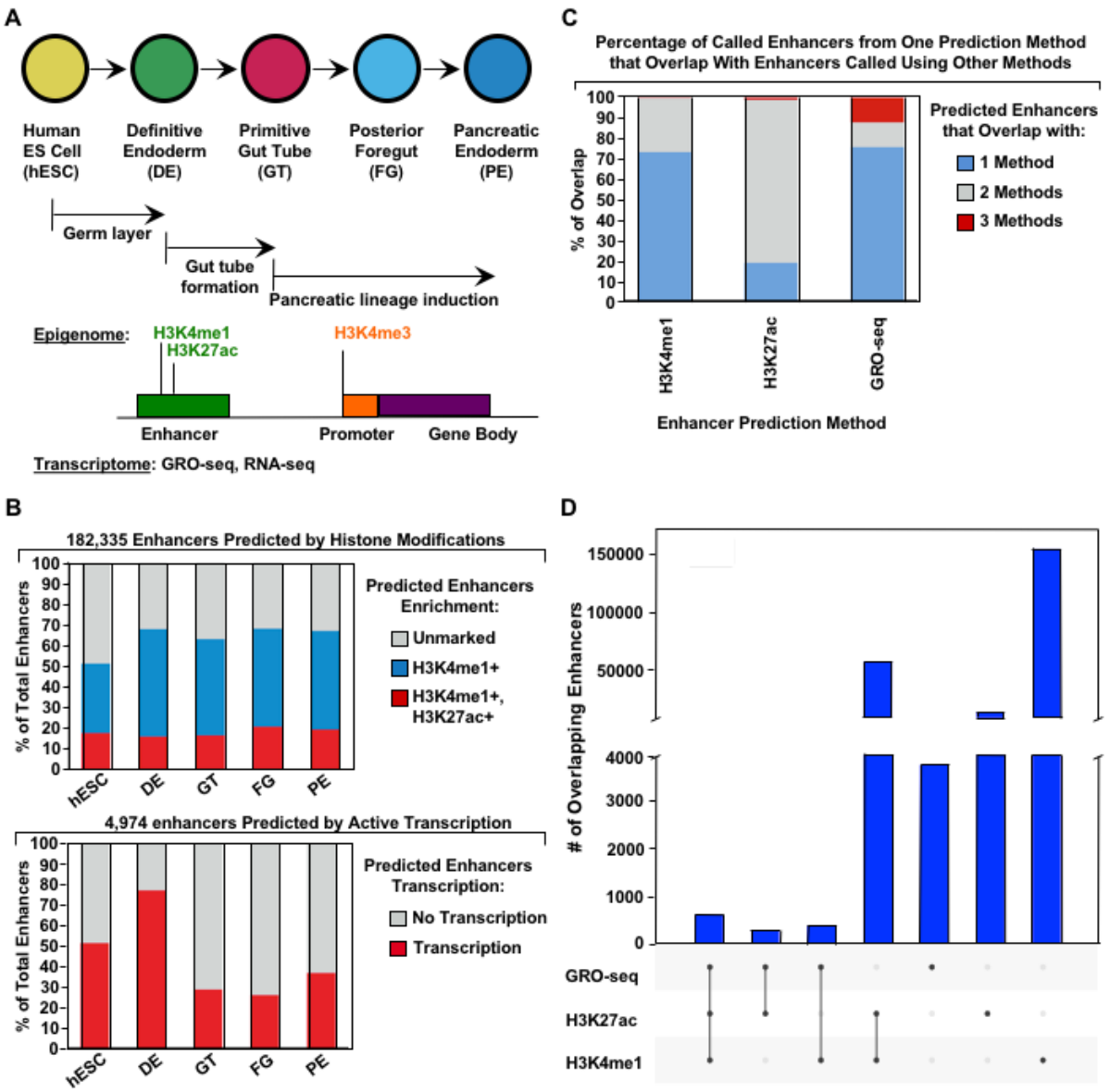

Figure 2: **Comparison of genome-wide prediction of enhancers in pancreatic differentiation.** **(A)** *(Top)* Schematic depiction of pancreatic differentiation starting from Human embryonic stem cells (hESCs) to pancreatic endoderm (PE). *(Bottom)* Depiction of epigenomic (ChIP-seq) and transcriptional (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.

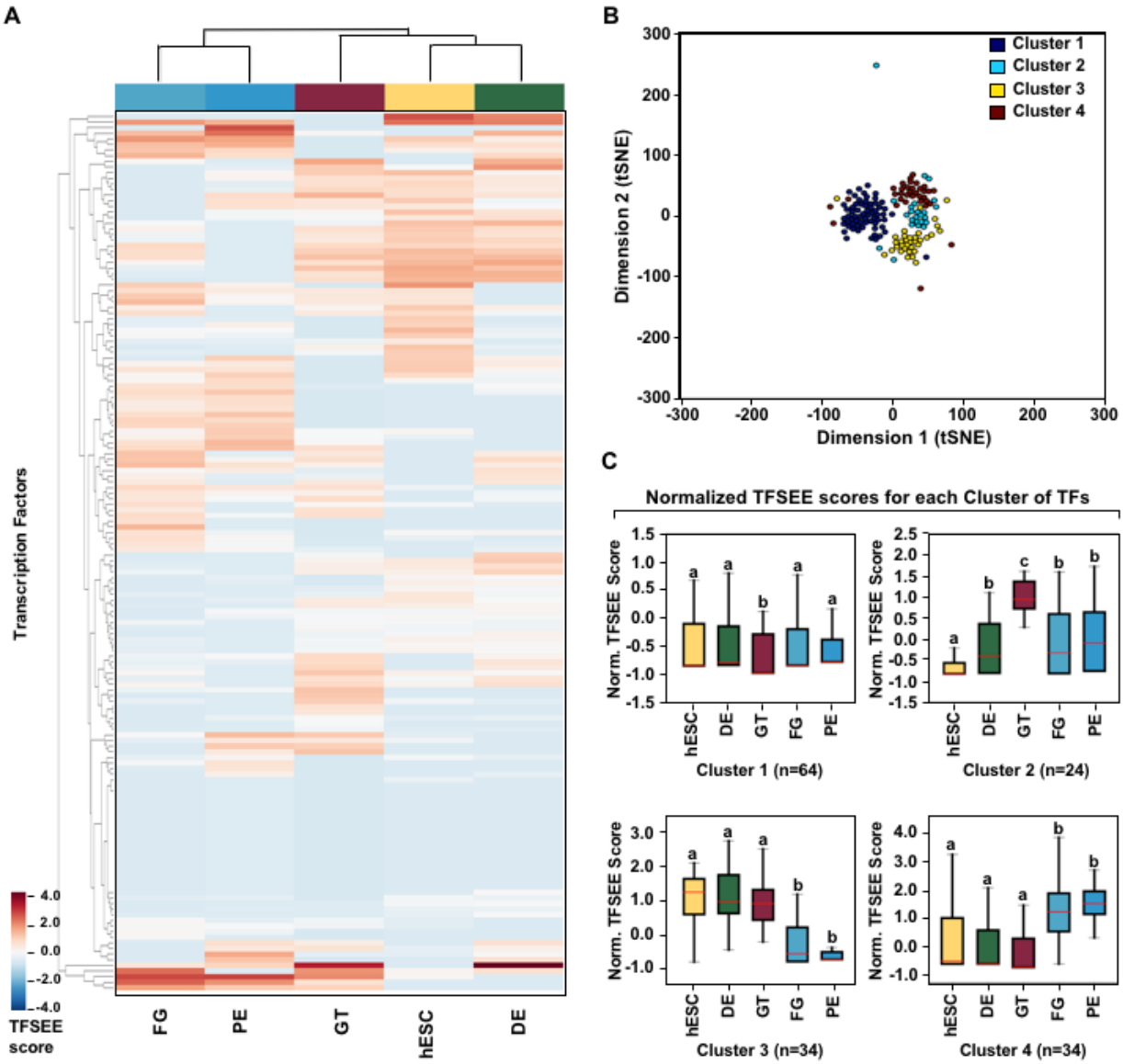

Figure 3: **TFSEE identifies cell type-specific enhancers and their cognate TFs that drive gene expression in pancreatic differentiation. (A)** Unsupervised hierarchical clustering of cell type-normalized TFSEE scores shown in a heatmap representation. hESC (human embryonic stem cell); DE (definitive 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, $p < 1 \times 10^{-4}$). 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).

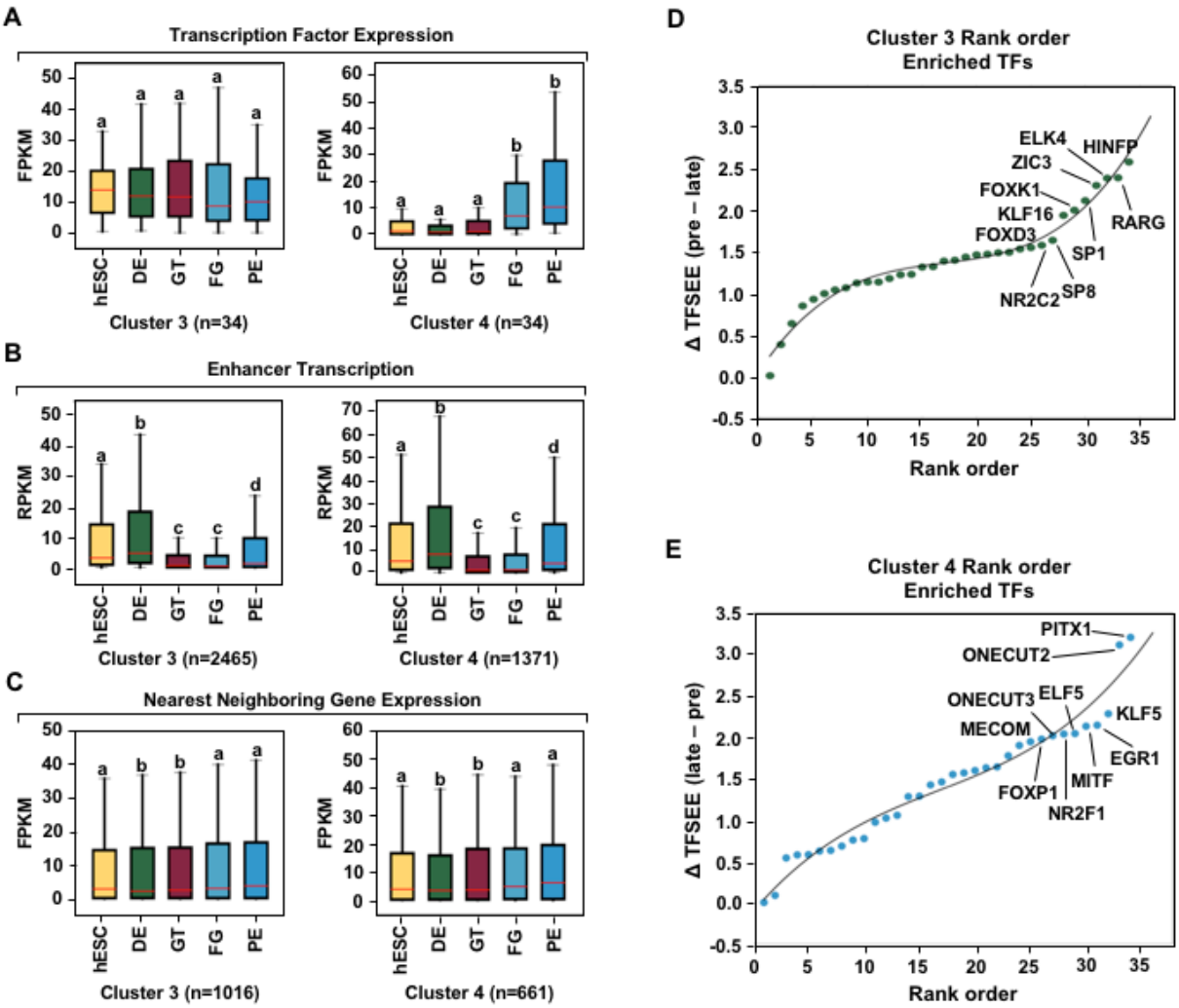

Figure 4: **TFSEE-Predicted TFs are enriched in pre- and late- pancreatic differentiation. (A-C)** Box plots of normalized TF expression (panel A), enhancer transcription (panel B), and gene expression for the nearest neighboring genes to active enhancers (panel C) in pre- (cluster 3) and late-pancreatic (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. ($p < 1 \times 10^{-4}$) **(B)** Enhancer transcription as measured by GRO-seq. Number of enhancers in each cluster are in parenthesis. $p < 1 \times 10^{-4}$). **(C)** Gene expression as measured by RNA-seq. Number of genes in each cluster are in parenthesis. ($p < 0.05$) **(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.

Discussion

Acknowledgments

Material and Methods

Genomic Data Curation

We used previously published GRO-seq, ChIP-seq and RNA-seq data from [1,2] of time course differentiation of human embryonic stem cells (hESC) to pancreatic endoderm (PE). All data sets are available from NCBI's Gene Expression Omnibus [3] or EMBL-EBI's ArrayExpress [4] 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 [5]. The aligned reads are subsequently filtered for quality and uniquely mappable reads using Samtools version 0.1.19 [6] and Picard version 1.127 [7]. Library complexity is measured using BEDTools version 2.17.0 [8] and meet ENCODE data quality standards [9]. Relaxed peaks were called using MACS version 2.1.0 [10] with a p-value of 1×10^{-2} for each replicate, pooled replicates' reads and pseudoreplicates. Peak calls that are replicated from the pooled replicated that are either observed in both replicates, or in both pseudoreplicates are 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 [11]. Quantification of genes against Gencode version 19) [12] annotations was done using default parameters in RSEM version 1.2.31 [13].

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 [14]. The trimmed reads were aligned to the human reference genome (GRCh37/hg19) using default parameters in BWA version 0.7.12 [15].

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 [16] with default parameters.

Defining Transcription Start Sites

We made distinct transcription start sites (TSS) for protein-coding genes from Gencode version 19 [12] annotations using MakeGencodeTSS [17].

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 [18–20] 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 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 [12], and H3K4me3 peaks. These were further classified into (1) short paired eRNAs and (2) short unpaired eRNAs as described previously [21]. For the short paired eRNAs, the sum of the GRO-seq RPKM values for both strands of DNA was used to call an enhancer transcript pair as expressed using a criterion of RPKM $\geq$ 0.5. For the short unpaired eRNAs, an RPKM cutoff of $\geq$ 1 was used to call an enhancer transcript as expressed. The universe of expressed eRNAs (short paired and short unpaired) 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 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 [22]. 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) [23] using TOMTOM [24].

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 are $>$ 3kb from known TSS, protein coding genes from Gencode version 19 annotations [12], and H3K4me3 peaks. A RPKM cutoff of $\geq$ 1 of H3K4me1 and $\geq$ 1 H3K27ac in at least 1 cell line was used to call 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 [22,25]. 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 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) [23] using TOMTOM [24].

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 [16]. For visualization of the multidimensional TFSEE scores, we performed t-distributed stochastic neighbor embedding analysis (t-SNE) [26] 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 [27–29].

Nearest Neighboring Gene Analyses and Box Plots

The universe of expressed genes in each cell line was determined from the RNA-seq data using an FPKM cutoff $>$ 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 [30,31]. Wilcoxon rank sum tests were performed to determine the statistical significance of all comparisons.

References

1. **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

2. **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

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

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

5. **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

6. **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

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

8. **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

9. **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

10. **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

11. **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

12. **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

13. **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

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

15. **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

16. **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

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

18. **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

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

20. **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

21. **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

22. **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

23. **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

24. **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

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

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

27. **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

28. **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

29. **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

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

31. **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