Jeon and Tucker-Kellogg *BMC Genomics* 2020, **21**(Suppl 10):614
https://doi.org/10.1186/s12864-020-07009-8

BMC Genomics

RESEARCH Open Access

Bivalent genes that undergo transcriptional switching identify networks of key regulators of embryonic stem cell differentiation

Ah-Jung Jeon[1] and Greg Tucker-Kellogg[1,2*]

From The 18th Asia Pacific Bioinformatics Conference
Seoul, Korea. 18-20 August 2020

Abstract

Background: Bivalent promoters marked with both H3K27me3 and H3K4me3 histone modifications are characteristic of poised promoters in embryonic stem (ES) cells. The model of poised promoters postulates that bivalent chromatin in ES cells is resolved to monovalency upon differntiation. With the availability of single-cell RNA sequencing (scRNA-seq) data, subsequent switches in transcriptional state at bivalent promoters can be studied more closely.

Results: We develop an approach for capturing genes undergoing transcriptional switching by detecting 'bimodal' gene expression patterns from scRNA-seq data. We integrate the identification of bimodal genes in ES cell differentiation with analysis of chromatin state, and identify clear cell-state dependent patterns of bimodal, bivalent genes. We show that binarization of bimodal genes can be used to identify differentially expressed genes from fractional ON/OFF proportions. In time series data from differentiating cells, we build a pseudotime approximation and use a hidden Markov model to infer gene activity switching pseudotimes, which we use to infer a regulatory network. We identify pathways of switching during differentiation, novel details of those pathway, and transcription factor coordination with downstream targets.

Conclusions: Genes with expression levels too low to be informative in conventional scRNA analysis can be used to infer transcriptional switching networks that connect transcriptional activity to chromatin state. Since chromatin bivalency is a hallmark of gene promoters poised for activity, this approach provides an alternative that complements conventional scRNA-seq analysis while focusing on genes near the ON/OFF boundary of activity. This offers a novel and productive means of inferring regulatory networks from scRNA-seq data.

Keywords: Bivalency, Embryonic stem cells, scRNA-seq, Bimodality, Chromatin state, Pseudotime, Genome regulatory network, Hidden Markov model

*Correspondence: dbsgtk@nus.edu.sg
[1] Department of Biological Sciences, Faculty of Science, National University of Singapore, 16 Science Drive 4, Singapore 117558, Singapore
[2] Computational Biology Programme, Faculty of Science, National University of Singapore, 21 Lower Kent Ridge Road, Singapore 11907, Singapore

Background

All levels of chromatin organization in eukaryotes are associated with covalent histone modifications that expose or constrain the genome for regulation [1, 2]. Because most transcription factors cannot bind nucleosome-bound DNA, nucleosomes must first be mobilized to expose DNA for transcription factor-driven regulation. Thus, chromatin remodeling, histone modifications, and transcriptional regulation are tightly linked.

While nucleosomes at most promoters in most cells carry either activating or repressive histone modifications, a large number of promoters in embryonic stem cells are *bivalent*, marked by both the activating H3K4me3 mark and the repressive H3K27me3 mark [3–5]. Increased bivalency is a hallmark of pluripotency, keeping developmental genes poised for commitment; bivalent promoters in embryonic stem cells are often resolved upon differentiation to either monovalent active (H3K4me3) or monovalent repressive (H3K27me3) states [5, 6].

Our understanding of transcription is being revolutionized by single-cell RNA sequencing (scRNA-seq) [7], which provides new opportunities to investigate the effects of chromatin organization on transcription [8]. However, most of what is known about transcriptional control by chromatin factors is based on population-averaged data, since single-cell technologies to interrogate chromatin state are still limited. In addition, we might expect genes whose expression is associated with changes in chromatin state to toggle between active (ON) and inactive (OFF) expression states in single cells [9], which poses special problems for integrative analysis.

One challenge is that scRNA-seq data commonly yields many genes with measured expression of exactly zero in many cells. Zeros may appear because of limited read depth in single cells, or from technical sources of zero inflation. In either case, the frequent zeros present challenges for both normalization and for detection of differential expression [10]. In most scRNA-seq analyses, genes with a high proportion of zeros are ignored and genes with high variability of non-zero values are kept.

Cell state-dependent genes in scRNA-seq data are often chosen first on the basis of high cell-to-cell variability. These Highly Variable Genes (HVG) may include genes differentially expressed in bulk transcriptomic studies, and may include genes indicative of underlying cell variability. HVG detection methods include exploiting the coefficient of variation [11] or squared coefficient of variation [12]. Some methods like BASiCS [13] require spike-in genes to estimate technical variation. Many HVG detection methods are computationally intensive and require genes to be filtered prior to the main analysis. This filtering often excludes genes with low mean expression.

Motivated by our interest in chromatin bivalency, our working model is that important stochastic expression changes occur right at the transcriptional boundary between active and inactive states [9]. Genes that undergo such transcriptional switching often include a high proportion of zeros in the expression data, or mean expression levels that are not high enough to be detected as highly variable. Therefore, we needed a new method with a conceptually different underlying principle to identify those switching genes.

To identify transcriptional switches at bivalent (including poised) promoters, we focus on genes that are sometimes silent, and other times expressed. We show that a modified Gaussian mixture model can identify genes undergoing transcriptional switches, even in the presence of the abundance of zeros common to scRNA-seq data. We refer to these as "bimodal" genes. We associate these bimodal genes with chromatin state from independent ChIP-seq studies and find that important developmental processes are enriched in bimodal genes carrying bivalent histone marks in pluripotent cells, and that bimodal bivalent genes can be used to characterize cell types.

We employ binarized gene expression for feature selection, and identify known and novel genes activated or silenced upon differentiation of embryonic stem cells to distinct cell types. We show how fractional state proportion of binarized gene expression data is a useful complement to conventional differential expression measures for many genes, and allows us to identify genes that switch activity with cell state change.

We apply a similar approach to time course differentiation data and develop a pseudotime ordering method exploiting probabilistic suffix trees of binarized data. The pseudotime ordering allows us to identify switching events using a hidden Markov model in pseudotime, when genes switch between mostly OFF and mostly ON states. We use the orchestration of these events in pseudotime to infer a regulatory network of gene switching.

Results

Bimodality-based variable gene detection identifies genes in ON/OFF transcriptional states

We define expression variability as the toggle between active and inactive transcriptional states. Variable genes, therefore, are those that are active in some cells and inactive in others in a population. Across the population, we expect the variable genes to show clear separation of expression values, resulting in a bimodal distribution.

For identification of bimodal genes, we define the underlying model of gene expression as a mixed log-normal distribution with possible zero inflation (Eq. 6 in "Methods" section). We allow the Gaussian Mixture Model to have two or three centers. Even though a truly bimodal gene will only have two centers, we observed genes with three centers and decided to include the optional intermediate state, which could arise from imprecise normalization of expression values or from the existence of a true intermediate state. The component with the smallest mean corresponds to the inactive state and we expect it to occur near zero; others will appear anywhere away from – but distinctly away from – zero. Bimodal genes were selected using Hartigan's Dip Test for non-unimodality, followed by fitting a two or three centered Gaussian Mixture Model on appropriately transformed data (see "Methods" section). Genes with $p < 0.01$ and a smallest mean below 1 were selected as bimodal genes.

We applied this approach on two scRNA-seq datasets [14]. The first dataset (H1 data) uses H1 cells and their differentiated derivatives, which were FACS sorted based on known cell types, including definitive endoderm cells (DEC), endothelial cells (EC), neural progenitor cells (NPC), trophoblast-like cells (TB), and human foreskin fibroblasts (HFF). The second dataset (H9 data) uses H9 cells differentiated towards endoderm and collected at various time points (0h, 12h, 24h, 36h, 72h, and 96h) after stimulation.

The two datasets presented two different experimental groupings. In H1 data, cells were grouped based on cell types, while in H9 data, cells were grouped based on experimental time points. We applied our bimodal gene identification method to H1 data to explore differentiation signature genes and genes with poised promoters, and later to H9 data to construct and analyze a pseudotime trajectory based on bimodality.

To include unexpressed or undetected genes – which appear as a single point function at zero – in our mixture model, we added a small amount of normally distributed noise to all transformed data. In the preliminary analysis with H1 data, we varied the amount of added noise and determined the optimal standard deviation of added noise as 1×10^{-5} (Fig. S1 and Table S1, Additional File 1). Too little noise or too much added noise reduced the number of bimodal genes detected. Re-sampling showed that the small amount of added noise generated a stable and consistent identification of bimodal genes (Fig. S1, Additional File 1). After optimization of noise we identified 2610 bimodal genes from H1 data and 4943 bimodal genes from H9 data.

Bivalency distinguishes cell type dependent bimodality in H1 cell differentiation

In embryonic stem cell differentiation, histone bivalency (H3K4me3 and H3K27me3) marks important developmental and differentiation regulatory genes [5]. Bivalently marked genes are kept in a poised state before lineage-specific transcriptional fate is determined. Bivalent genes therefore represent genes that will potentially undergo transcriptional switching upon differentiation. Bivalent genes that show cell type-specific bimodality upon differentiation should therefore include marker genes and signature genes for differentiation.

We integrated the 2610 bimodal genes from scRNA-seq analysis of H1 data with independent ChIP-seq data of the same cell types to focus on those bimodal genes that are bivalently marked in H1 cells. This selection is meant to include those genes poised for transcription while in the undifferentiated state. This integrative analysis resulted in 644 bivalent, bimodal genes from the H1 data.

When cells were clustered based on the expression profiles of bivalent bimodal genes, all of the cells clustered perfectly accordingly to the experimentally determined cell type (Fig. 1a). This analysis shows that the bivalent and bimodal genes indeed show cell type-specific expression patterns. MSigDB analysis provided additional support for this interpretation, as the genes were enriched in developmental and differentiation-specific genes (Table S2, Additional File 1). In the original paper that generated the H1 data set [14], the authors used a curated list of 52 lineage-specific markers for hierarchical clustering of cell. Our method provides an unsupervised discovery of a large number (644 genes) of cell type-specific transcriptional gene switching.

In contrast to bivalent bimodal genes, when we used monovalent H3K4me3-marked bimodal genes in H1 cells for clustering, the cell type-specific clustering was much weaker (Fig. S2, Additional File 1). MSigDB analysis showed that these genes were enriched in cell-cycle regulators. This suggests that the H3K4me3-marked bimodal genes include a large number of genes whose bimodality in scRNA-seq data is not due to changes in cell identity, but due to changes in cellular state. (Table S3, Additional File 1).

Binarization of bimodal gene expression identifies cell type-specific patterns of expression

As active and inactive states of bimodal genes have clearly distinguishable expression values, expression values of bimodal genes can be binarized. For each bimodal gene, the local minimum expression value was taken as the binarization threshold. Cells with expression value less than the threshold were considered inactive (OFF), and cells with expression value more than that were considered active (ON).

We used the clustering obtained above to assign fractional gene activity between 0 and 1 by assigning cells to random windows within each cluster, converting the original 750 cells with binarized expression to 30 windows with fractional activity. Using these fractional expression profiles, we clustered bivalent bimodal genes into 8 groups (Fig. 1b). Visualization of fractional expression values showed much more striking patterns than the

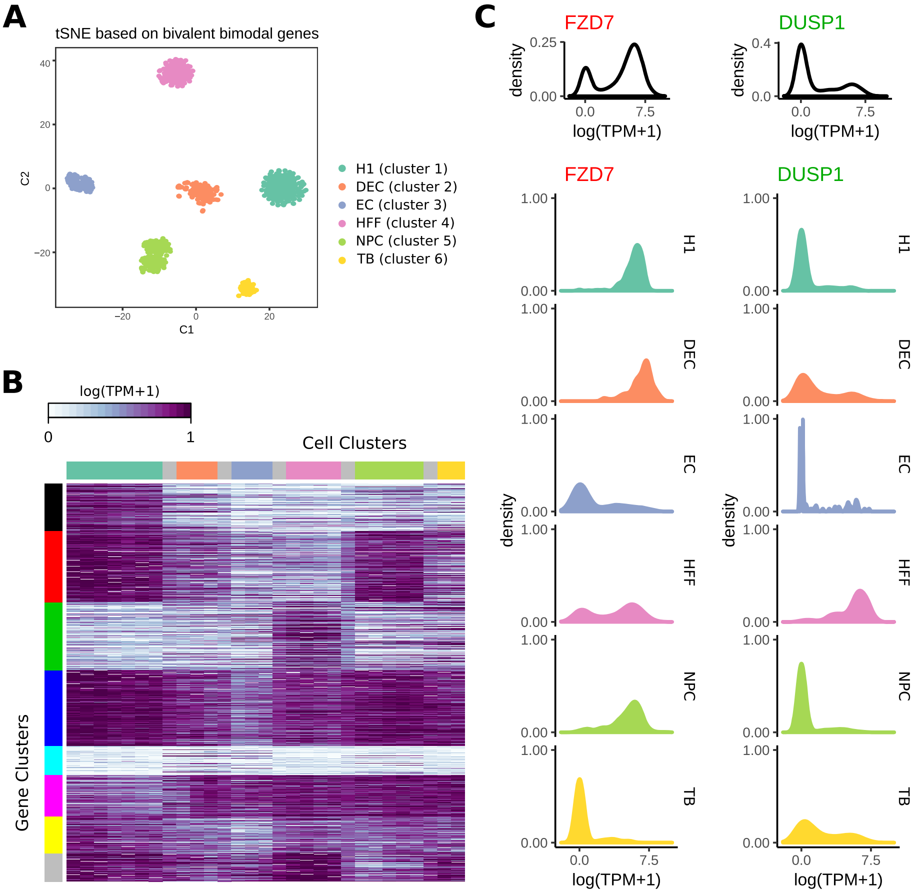

Fig. 1 Bivalent and bimodal genes in H1 differentiation data. (**a**) t-SNE plot of bivalent and bimodal genes in H1 differentiation data. Cells were colored based on cell types, which perfectly matched the clusters. (**b**) Heatmap of fractional expression values obtained from the ratio of active cells per window. Columns represent cells and rows represent genes. Cells were ordered by cell types, split into 30 windows and colored by the cell type (Grey represents the cells in between two cell types). Genes were clustered based on the fractional expression values. (**c**) Example genes from clustering shown in (**b**). FZD7 belongs to the 2nd gene cluster (red) and DUSP1 belongs to the third gene cluster (green). Top panel shows the overall density of log(TPM+1). Bottom panel shows the density of log(TPM+1) for cells in each cell type

individual cells: most groups showed cell type-specific transcriptional states, especially gene clusters 1,2,3, and 4. For example, gene cluster 2 (red) showed mostly active transcription in H1 and NPC cells, and less active transcription in other cell types, while gene cluster 3 (green) showed mostly active transcription in HFF and mixed in others. Example genes and their expression profiles in each cell type are shown in Fig. 1c. FZD7 is from gene cluster 2, and it is mostly in active states in H1, DEC, and NPC cells. It forms a mixed population of both active and inactive states among HFF cells, while is mostly inactive in EC cells. DUSP1 (from gene cluster 3) is mostly active only in HFF cells. Both FZD7 and DUSP1 show overall bimodal expression profiles (top, Fig. 1c) while showing unimodality within each type of cell. These are examples of genes exhibiting cell type-specific bimodality.

Binarization of bimodal genes allows differential expression testing based on bimodality

To identify genes that showed transcriptional switching between cell types, we used the binarized expression values for bimodal genes. For a pair of cell types including H1 and an H1 derivative, we then used Fisher's exact test to identify genes that showed clear changes in the proportions of active cells upon differentiation. The odds ratio was used to infer the direction of differential activity, with genes with odds ratio greater than one being upregulated as less than one being downregulated upon differentiation (Figs. 2 and S9, Additional File 1). "Upregulated" in this usage does not mean expressed at a higher magnitude, but active in a higher proportion of cells. Enrichment analysis showed that this method of differential expression switching was consistent with cell-type specificity. For example, genes that were more often inactive in H1 cells and active

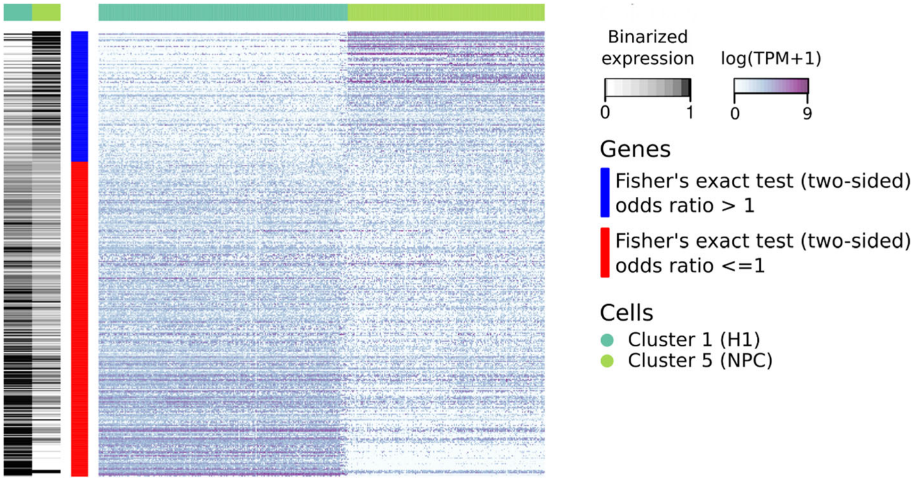

Fig. 2 Genes that undergo transcriptional switching upon differentiating into Neural Progenitor Cells. Genes with adjusted *p*-value less than 0.05 are shown. The heatmap in greyscale (left) shows the fractional expression values across all cells in each cell-type. The heatmap on the right shows the actual log(TPM+1) expression values of the genes

in NPC cells were enriched in developmental, neurogenesis, and neuron differentiation GO categories (Table S4, Additional File 1). Our bimodality-based transcriptional switch identification method successfully identifies signature genes of cell type-specific differentiation.

Variation of information can be used to infer a pseudotime trajectory for binarized bimodal genes in time series data

The analyses described above are valuable for discrete cell states, but many single-cell studies use cells collected over time. With the perspective that gene expression changes represent underlying changes in transcriptional state, we asked whether one could infer the underlying time order of cells in a differentiation model, and use that cell ordering to identify the timing of transcriptional switching. We used data from an endoderm differentiation study of H9 cells (H9 data) [14]. In this study, H9 cells were differentiated towards endoderm and cells were collected and sequenced at different timepoints.

We used the 4943 bimodal genes and asked whether we could use binarized bimodal genes to infer a pseudotime trajectory of cells. Unlike H1 data where a clear distinction between cell types was important to reveal marker genes, an overly clear distinction between time-dependent cell clusters in H9 data would potentially compromise pseudotime trajectory reconstruction. Therefore, we included H3K27me3 monovalent bimodal genes in addition to the H3K4me3/H3K27me3 bivalent bimodal genes. We also kept H3K27me3 monovalent-marked genes to identify genes that showed de-repression upon differentiation. A total of 1371 genes were selected.

The selected genes characterized cells at each timepoint well, with the exception of cells from 72 and 96 hours of treatment (Fig. S3a, Additional File 1), consistent with the similarity at these timepoints noted in the original paper [14]. Cells in cluster 5 and 6 were therefore combined for downstream analysis (Fig. S3b, Additional File 1).

Existing pseudotime inference methods are not designed for genes with a high proportion of zeros, as we see for bimodal genes, and not designed for binarized gene expression. We devised a new algorithm to use binarized bimodal gene expression for cell ordering. In addition, since in this data set cells were collected at known time points throughout differentiation, the order of clusters was known.

Figure 3a graphically summaries our pseudotime trajectory estimation method. Our algorithm orders clusters of cells by known time points, and then generates candidate cell orderings within each cluster. The candidate cell orderings are then used to generate a probabilistic suffix tree of all cells, from which a final pseudotime ordering is inferred. The procedure is described below.

We defined distances between cells as the variation of information (VI, a metric related to mutual information) of binarized gene expression. To order the cells within each cluster, we used the adjacent time points (clusters) as anchors (*connecting cells* in Fig. 3a). VI distances between cell pairs in adjacent clusters were calculated, and the 100 closest pairs were chosen as candidates to connect clusters. For each candidate pair, a candidate ordering was defined by VI distance from the candidate connecting cell, resulting in 100 orderings for the first and last clusters, and 200 orderings for the clusters in between.

Using the obtained set of orderings (*sequences*), we built a probabilistic suffix tree (PST) for each cluster, followed by sequence generation. If no unique sequence could be obtained, we selected the ordering with the highest probability based on the PST. We observed that all

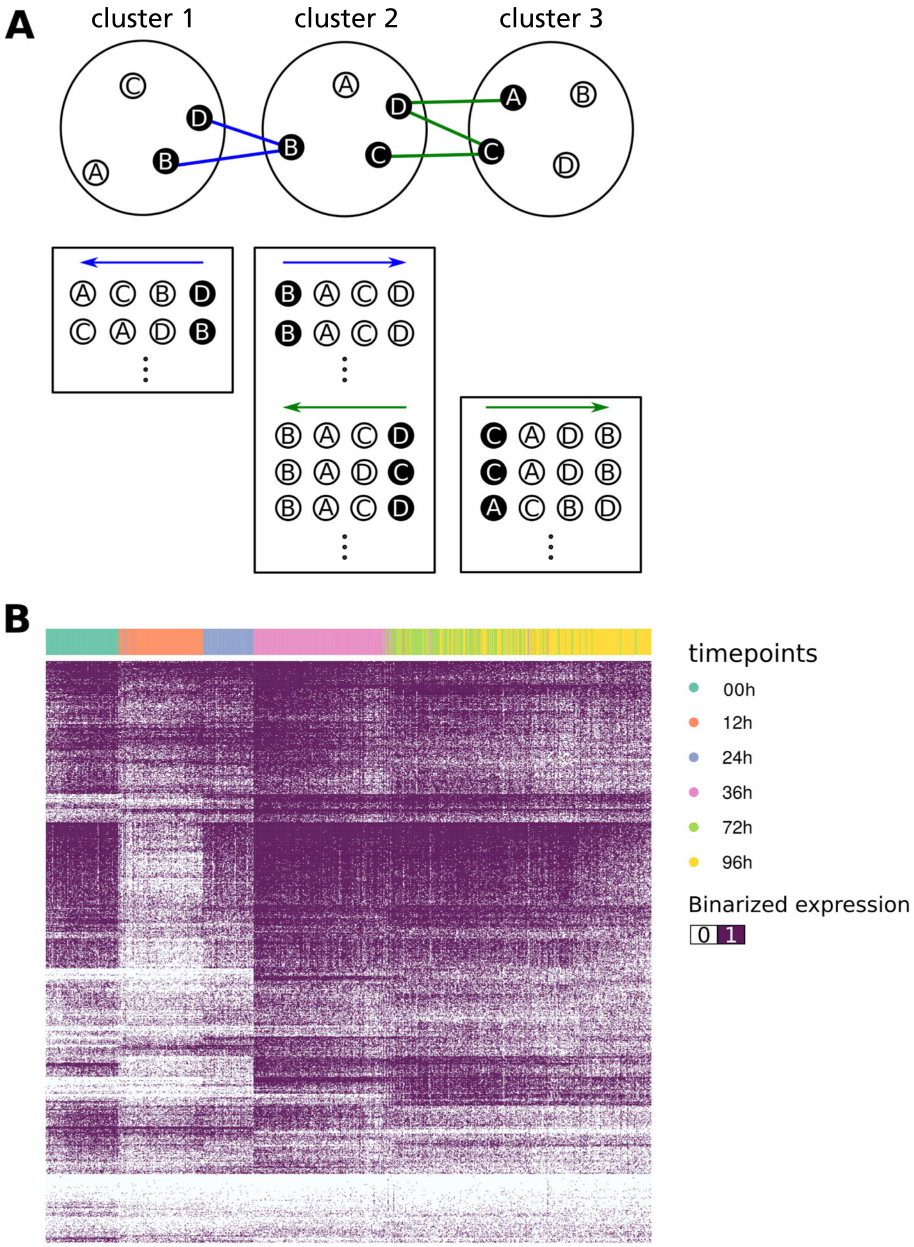

Fig. 3 Pseudotime trajectory estimation (**a**) *Top*. Each circled letter represents a cell in a cluster. Letters highlighted in black represent the *connecting cells* that show low variation of information to cells in the adjacent cluster. Cells in clusters are unordered. *Bottom*. Cells arranged in different orders within rectangles represent the cell orderings (*sequence*) within each cluster based on the variation of information between the *connecting cells* and the rest of the cells in the cluster. Refer to the main text for more explanation. (**b**) Heatmap of log(TPM+1) expression values of bivalent and bimodal genes from H9 time-series data. Cells (columns) are arranged according to the pseudotime ordering and colored according to the experimental timepoints. Genes (rows) are ordered using hierarchical clustering

clusters generated unique sequences except for cluster 1. We believe this is because cluster 1 was formed from untreated control cells, so strictly speaking they were not on the same pseudotime trajectory as the other treated cells. This was also reflected in the t-SNE map (Fig. S3a, Additional File 1), where cluster 1 cells sat noticeably away from the rest of the clusters.

Hierarchical clustering of genes using the final cell order revealed gradual changes in the expression profiles (Fig. 3b) consistent with pseudotime inference.

Identification of transcriptional switches during endoderm differentiation from a hidden Markov model

To identify when transcriptional switching occurs along the reconstructed pseudotime scale, we used the Viterbi algorithm [15] to find the most likely underlying hidden transcriptional states that resulted in the sequence of observed binarized expression values. Out of 1371 genes, 1063 genes showed clear switching patterns with well defined Viterbi paths (*switching genes*). The switching points for each gene were determined based on the transition between active and inactive state paths in pseudotime-ordered cells. We classified genes into 17 switch groups, based on the initial state, final state, and the number of switches (Fig. 4, as well as Table S5 and Fig. S4, Additional File 1). Switch group 1 included genes that were switched on with a single switch from inactive initial to active final state, active final state; switch group 2 included genes that were switched off with a single switch from active initial to inactive final state. The rest of the switch groups included genes with multiple switches through endoderm differentiation.

As cells were collected at time points far enough apart to establish distinct expression profiles, most of the switches occurred near the cell cluster boundaries as shown by the

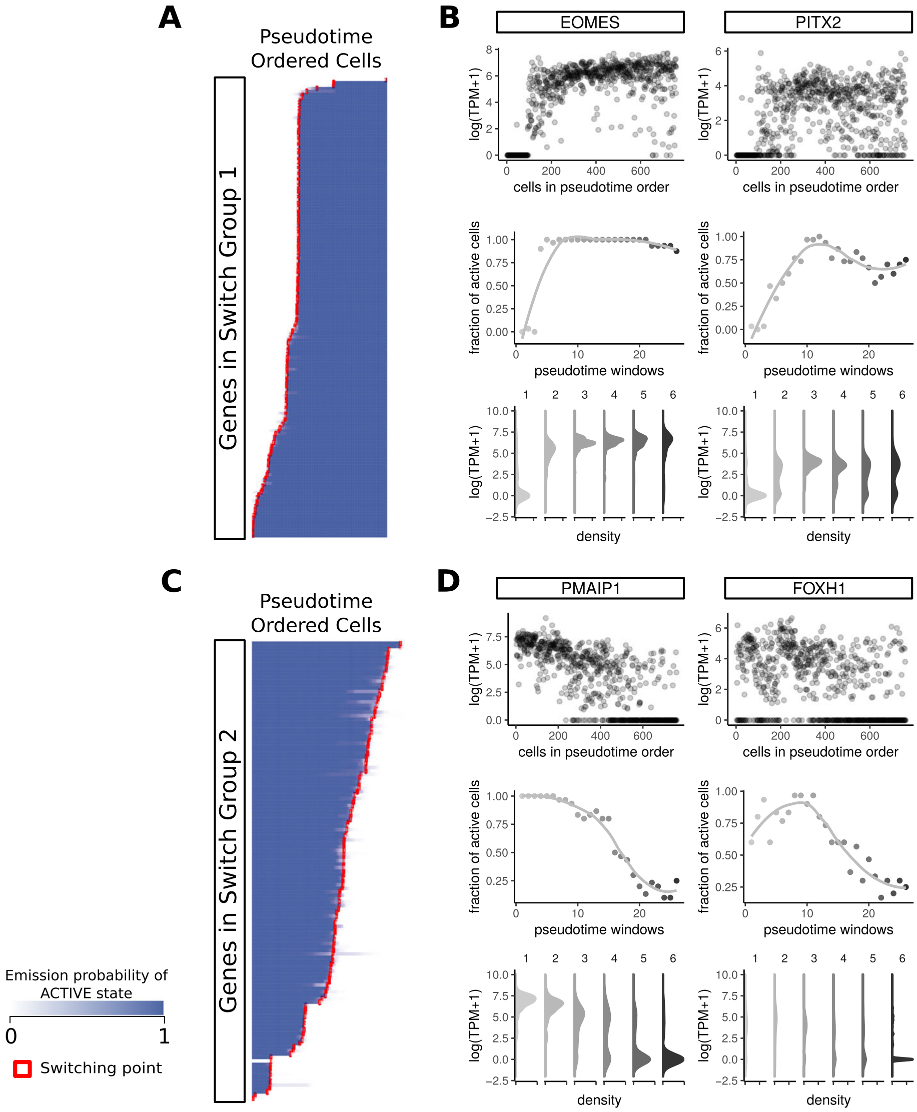

Fig. 4 Switch group 1 and 2 genes with activated and deactivated transcriptional states Genes are grouped according to the switch groups. For switch group 1 and 2, genes are ordered within each group according to the first switching timepoints (denoted in red) in increasing order. (**a,c**) the Viterbi path of active transcriptional state for genes in switch group 1 and 2, (**b,d**) expression profiles of example genes from switch group 1 and 2. Top: log(TPM+1) expression values in pseudotime order. Middle: Fraction of active cells at each pseudotime window. Cells were split into 26 windows. Bottom: Density plot for log(TPM+1) expression values at each time-ordered cell cluster. X-axis scales are 0-0.75, 0-0.6, 0-0.6, and 0-2 respectively

vertical streaky patterns in the heatmap (Fig. S4, Additional File 1). Some genes switched sequentially, as shown by the diagonal switching patterns, especially genes inactivated late in differentiation (Fig. 4, Switch Group 2). Figure 4 shows example genes from the switch groups 1 (EOMES and PITX2) and 2 (PMAIP1 and FOXH1). Fractional expression values (middle) show gradual changes, but the density plots per cell cluster (bottom) show that expression values take distinct bimodal patterns with switch-like activation or deactivation.

Pathway genes undergo switching during endoderm differentiation

To highlight the underlying biological themes behind the switching genes, we computed overlaps with other gene sets in MSigDB [16, 17]. We first focused on gene sets in CP (Canonical pathways), CP:BIOCARTA (BioCarta gene sets), CP:KEGG (KEGG gene sets), and CP:REACTOME (Reactome gene sets) to evaluate the involvement of switching genes in various pathways. The top pathway overlaps included P53 signaling, TGF-beta signaling, SMAD2/3 nuclear signaling, MAPK signaling, and WNT signaling.

The switching genes involved in signaling pathways were mostly specific to one pathway (Fig. S5, Additional File 1, top right), but their enrichment does not tell us if the switching within a pathway is coherent. We visualized pathway-level switching using stairstep graphs of switching genes within each pathway over pseudotime order (Fig. S5, Additional File 1, bottom). Many of the P53 signaling pathway genes were active during the early phase of endoderm differentiation, while the majority of Wnt signaling pathway genes were active during the late stage. Most of the switching genes peaked during the mid-phase before getting switched off, indicating the involvement of these pathways in regulating endoderm differentiation.

The stairstep plot alone makes it appear that many pathways are less active late in differentiation of H9 cells to endoderm. This may be the case, but the emission probabilities for individual genes tell a richer story: many genes that were switched on remained active, while other genes either were transiently activated or switched off.

Transcriptionally switching genes include transcription factors and their targets

To identify potential transcription factor targets among the switching genes, we computed enrichment between the switching genes and the TFT (transcription factor targets) gene set from MSigDB [16] using a Fisher's exact test. The results show striking enrichment for specific transcription factors in this gene set (Table 1). The top 10 hits included SP1, MAZ, E12, LEF1, NFAT, and AP4.

Table 1 Top 10 motif hits for switching genes from MSigDB search. 1340 genes were compared to gene sets in C3 TFT (transcription factor targets) gene sets [16]

Gene Set Name	# Genes in Overlap	FDR q-value
GGGCGGR_SP1_Q6	273	1.13×10^{-64}
GGGAGGRR_MAZ_Q6	231	4.79×10^{-61}
CAGGTG_E12_Q6	221	2.17×10^{-48}
CTTTGT_LEF1_Q2	185	3.19×10^{-43}
TGGAAA_NFAT_Q4_01	170	5.06×10^{-37}
CAGCTG_AP4_Q5	148	6.78×10^{-36}
TTGTTT_FOXO4_01	172	1.09×10^{-33}
AACTTT_UNKNOWN	162	5.16×10^{-33}
GGGTGGRR_PAX4_03	129	2.36×10^{-32}
CTTTAAR_UNKNOWN	94	2.39×10^{-22}

Gene regulation by transcription factors often shows a correlation between a TF and its targets. Some of the transcription factor hits were bimodal genes themselves, so we examined whether the switching pattern of a TF was recapitulated in the switching pattern of predicted targets. We compared the expression of four bimodal TFs (MYC, PITX2, FOXO3, and FOXO1) to the aggregated number of transcriptionally active TF targets that are also transcriptionally switching (Fig. 5). In each case, the aggregated number of active target genes peaked at a similar time point as the expression of the transcription factor. Some motif hits showed overlapping context to some of the earlier pathway hits. LEF1 and E12 (TCF3) came up as the top switching transcription factors, with many downstream targets also undergoing transcriptional switching. These TFs are known to be involved in canonical Wnt/beta-catenin signaling [18], which came up in the pathway search. Beta-catenin nuclear signaling-related genes peak during the mid-phase of endoderm differentiation, before deactivation at the late stage. The majority of E12 and LEF1 targets are active at similar time points as Wnt signaling pathway genes (Fig. S6, Additional File 1). This suggests that the involvement of canonical Wnt/beta-catenin signaling during endoderm differentiation may be mediated through activation of LEF1 and TCF3.

A network of switching genes based on mutual information

Based on the pathway genes identified from the earlier analysis, we wanted to draw a map of key regulatory genes involved in endoderm differentiation. In addition to the pathway genes, we identified *connector genes* that are closely related to the pathway genes based on variation of information. We assumed that the connector genes are the mediating regulators that help to establish a more interconnected map. With 68 switching genes that were

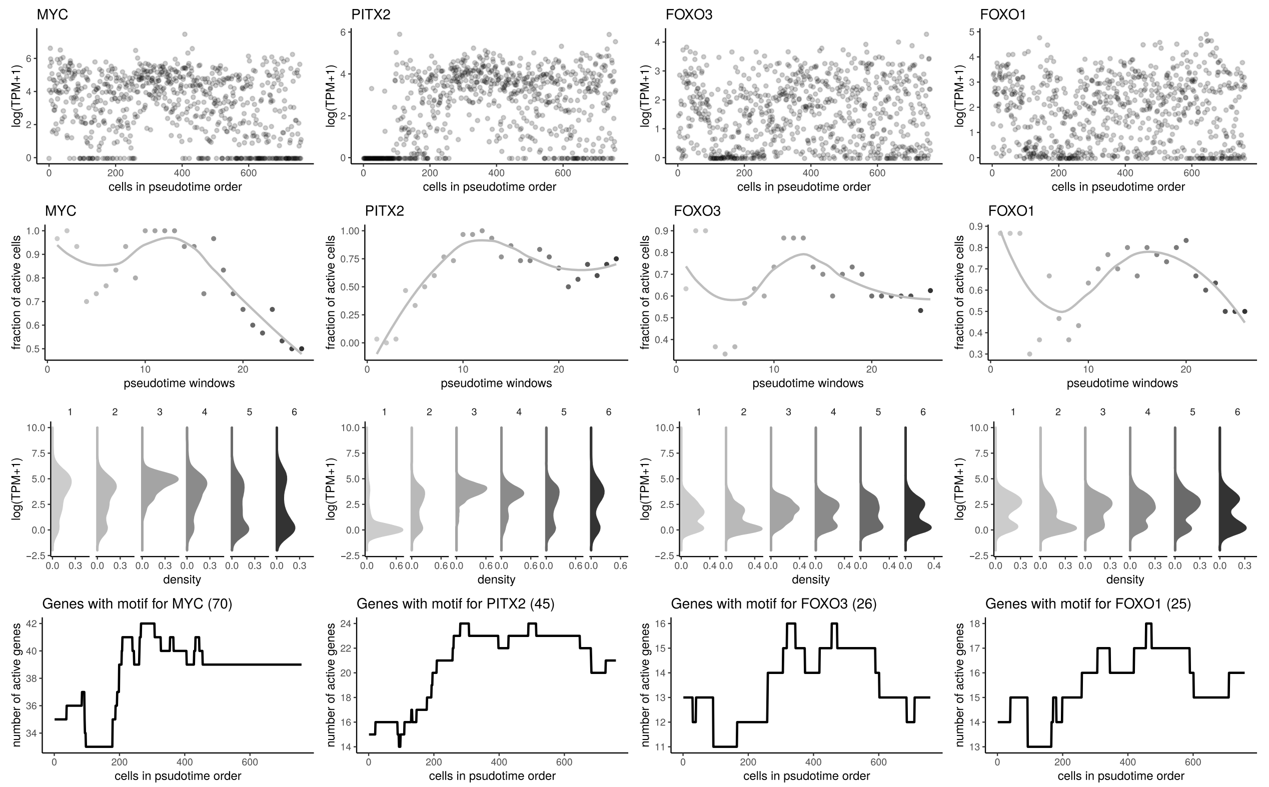

Fig. 5 Transcription factors that undergo switching together with their downstream targets First row: Plots of log(TPM+1) expression values for transcription factors that came up in MSigDB motif search. Second row: Fraction of active cells for transcription factors that came up in MSigDB motif search. Third row: Density plot of log(TPM+1) expression values for transcription factors that came up in MSigDB motif search. Fourth row: Total number of active switching genes with transcription factor motifs

involved in the pathways identified from the earlier pathway search, we selected the top 20 most related connector genes.

Co-expression networks are mapped using a distance metric obtained from expression data [19], to establish significant relationships between pairs of genes. We wanted to build a similar network, but with binarized expression data. We built a network map using Algorithm For The Reconstruction Of Accurate Cellular Networks (ARACNE), which takes the mutual information matrix and returns the inferred network [20] (Fig. 6). We chose ARACNE because binarized expression data lends itself to a mutual information-based network reconstruction, and because the ARACNE algorithm used mutual information to remove uninformative edges that might otherwise complicate the inferred network.

The ARACNE network is undirected, but we assigned direction to some edges based on the switching times of the connected nodes (genes). If gene A had undergone transcriptional switching earlier in pseudotime than gene B, the edge connecting A and B was directed from A to B. This was based on the assumption that a prior event would influence later events. The network may not represent the actual gene regulatory network, but displays the co-switching relationship between key pathway genes during endoderm differentiation.

The network structure shows several interesting properties related to the earlier pathway and transcription factor analyses. Frizzled genes (FZD2, FZD4, FZD5, FZD8) and bone morphogenetic protein (BMP) genes (BMP2, BMP4, BMP7, BMP8A, BMP8B) appeared at different parts of the network, indicating their involvement in different signaling events. In the motif analysis we had found that switching genes were enriched in LEF1 and NFAT targets. From the network map, the edge connecting FZD5 and TCF7 further supports the presence of the canonical Wnt/beta-catenin signaling pathway. On the opposite side of the network map, we saw a hub of genes around HAPLN1. The presence of FZD2 (Frizzled 2), CAMK2D (Calcium/Calmodulin Dependent Protein Kinase II Delta), and DAAM1 (Dishevelled Associated Activator Of Morphogenesis 1) suggests that there is also non-canonical Wnt/Ca2+ signaling pathway present. The time-dependent differential expression of various Wnt

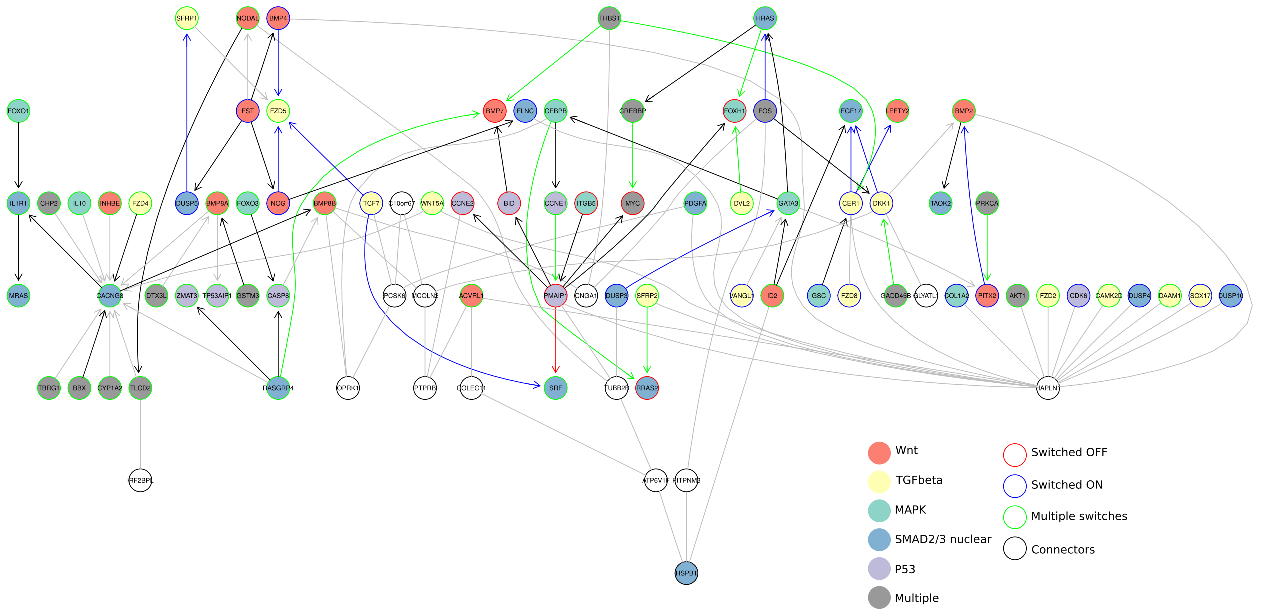

Fig. 6 Co-switching network of genes based on mutual information Undirected edges represent genes that switch at the same time. Directed edges represent the tail gene switchedd earlier than the head gene. Grey edges represent genes that switched at similar time (less than 10 cells apart). The rest of the edge colors followed the color of the tail gene

and frizzled genes further supports the involvement of dynamic Wnt signaling events (Fig. S7, Additional File 1).

Many TGF-beta and Wnt signaling pathway related genes (yellow and red highlighted in Fig. 6) were retained in the network map. Interestingly, we noticed the presence of not only positive regulators of the pathways, but also negative regulators such as the BMP antagonist NOG (Noggin), Wnt antagonists SFRP1 (Secreted Frizzled Related Protein) and DKK1 (Dickkopf WNT Signaling Pathway Inhibitor 1), and TGFb antagonist INHBE (Inhibin Subunit Beta E). These indicate the dynamic regulation of pathways during endoderm differentiation, consistent with the findings from the stairstep plots (Fig. S6, Additional File 1).

We also noticed evidence for cross-talk between TGF-beta, Wnt, and MAPK signaling pathways based on the edges connecting genes of different pathways. Some examples are edges connecting INHBE-CACNG8-FZD4, FST-BMP4-FZD5, FST-DUSP5-SFRP1-FZD5, LEFTY2-CER1-FZD8, and DKK1-BMP2-TAOK2. TGF-beta and Wnt signaling genes also appeared interconnected together at the top left corner of the network map.

Discussion

On variable gene detection and differential expression testing

Differential gene expression testing often includes an implicit assumption that larger gene expression changes are more responsible for phenotypic changes. Consistent with this model, expression counts are first normalized and a mean-variance relationship is often used as an indicator of interest, with highly variable genes (HVG) then selected based on a mean-variance threshold [21]. Variance-based methods focus on detecting changes in the spread of expression data, and are hence excellent in detecting changes in active transcriptional levels.

In this analysis, we instead model differential expression of signature genes happening across the transcriptional ON/OFF boundary, regardless of the actual magnitude of expression values. We choose this approach because the most functionally striking difference in expression is a switch from complete repression to activation, or vice versa. This is related to, but distinct from, the assumption behind variance-based methods.

Crossing the transcriptional on/off boundary, or switching between gene activation and repression, does not ensure bimodally distributed expression values. However, in this study we used data having distinct experimental groupings: either grouped according to cell types (in H1 data) or according to experimental time points (in H9 data). We assumed that signature genes distinguishing the cell groups must have formed stable transcriptional states within each cell group distinct from the originating cells. Such shifts in transcriptional states would result in bimodally distributed expression values across the cell groups, leaving a clear separation between the expression values from the on/off transcriptional states.

We also focused on the toggle between on/off transcriptional states because histone mark-driven transcriptional control mostly involves switching between repression and activation, especially in the poised promoter model of bivalency. This model is relevant for both the H1 and H9 embryonic stem cell differentiation data sets we used.

In a typical population-based transcriptomic analysis, differential expression descriptors (such as "increased/decreased expression") describe gene expression as a process on a continuous scale of change. Differential expression analysis then takes the quantified or normalized read counts and performs statistical tests to discover quantitative changes in expression levels between experimental groups [21]. Genes identified as differentially expressed may show changes of transcription at any level, but bimodal genes take either of the ON/OFF transcriptional states. The magnitude of expression for each bimodal gene in a given cellular state is proportion of active and inactive cells, and gene expression changes are modeled as the dynamic shifts in transcriptional states. Therefore, differential expression testing is done differently for bimodal genes. Since expression values of bimodal genes can be binarized, we build contingency tables of genes in on/off transcriptional states between two cell types, and test to determine whether the proportions of on/off states are different depending on the cell types. We describe gene expression changes as dynamic shifts between active and inactive states. If a gene is increased in expression, in our analysis it is shown by increase in the proportion of cells in which the gene is active.

Bimodality-based identification of variable genes, therefore, presents a complementary method of detecting gene expression changes in scRNA-seq by focusing on the on/off transcriptional states. We are not advocating any superiority of one method over the other, but rather presenting a different method of identifying variable genes. If a choice were to be made between variance-based and bimodality-based methods, it must depend on the purpose of the experiment.

Advantages of binarizing expression data

Since bimodal genes present clearly distinguishable on/off transcriptional states, their expression values could be appropriately binarized, which presented several advantages for later analysis. In addition to using Fisher's exact test for differential gene expression testing, binarized expression makes the data less susceptible to normalization accuracy and less biased towards genes with high expression counts. Binarized counts also enabled the use of mutual information-based analysis to detect patterns of co-regulatory genes. This is in contrast to many of the existing co-expression network methods based on expression correlation [19], and presents an alternative way to select co-regulated genes based on similar transcriptional switching events. Co-switching networks could reveal the genetic relationships that are possibly missed out by co-expression networks. Binarizing the expression data may appear to be prematurely simplifying the data, but the in-depth network and pathway analyses we had done shows that there are more things one could learn in terms of the global relationship between genes through binarization.

Integrating chromatin state information into transcriptomic analysis

We have seen that the presence of H3K4me3/H3K27me3 bivalency successfully identified cell type-specific bimodality in the differentiation datasets. We also saw that H3K4me3 monovalent bimodal genes were enriched in cell state-specific pathways such as cell cycle regulation. Our results suggest that chromatin states indeed offer valuable insight to further interrogate the transcriptomic data.

Our approach may have work especially well in differentiation studies because of the poised promoters of bivalent genes. Several studies have shown that bivalently marked genes are differentiation signature genes [5, 22]. Since bimodal genes are those that undergo transcriptional switching, bivalent and bimodal genes are the differentiation signature genes that indeed show clear cell-type dependent de-repression or de-activation. These are therefore the genes that are highly linked to lineage commitment and differentiation, and were indeed enriched with differentiation markers. A study done in 2017 investigated bulk RNA-seq data for bivalent genes for 34 tissues and primary cell lines, and also made a similar observation that bivalent genes in ES cells tend to show tissue-specific expression pattern [23].

Further studies linking between chromatin state and gene expression are currently limited by the availability of experimental methods. Various challenges include simultaneously measuring both quantities and achieving single-cell level for chromatin state data.

On the use of binarized data for regulatory network inference

Some of the earliest gene regulatory network reconstruction methods based on microarray expression data used Boolean networks. Technologies to infer networks have advanced tremendously since then, as have technologies to measure gene expression. However, single-cell RNA-seq data offers the opportunity to observe – zero inflation notwithstanding – if a gene is transcriptionally on or off in a given cell, and the binary state of its activity is informative. As we show here, genes that might otherwise be missed because they are not expressed at high enough *levels* may be useful to infer regulatory networks and reveal hidden components of pathway activation.

Conclusions

Histone mark modifications are strongly associated with transcriptional regulation, but direct analysis in single cells is limited by current technology. By adopting a strategy to integrate population-level measurements of histone modifications and single-cell transcriptomics, we identified transcriptional switching patterns of bivalent genes in pluripotent cells associated with cell type upon differentiation. The technical strategy, using a mixture model-based method to identify genes with multimodal on/off expression patterns across cell types, is amenable to both pairwise comparisons and time course scRNA-seq experiments.

Binarizing transcript expression levels facilitates analysis of transcriptional state, and supports novel analysis analogous to conventional differential expression as well as novel pseudotime inference, estimate of switching points in pseudotime, and gene regulatory network reconstruction.

Methods

All analysis was performed in R (version 3.5.1, [24]) and most plots were generated using the R package *ggplot2* [25].

Data

ChIP-seq files for H1 cells and H1 cells differentiated by treatment with BMP4 – into mesendoderm, trophoblast and neural progenitor cells were downloaded from the UCSD Human Reference Epigenome Mapping Project (GSE16256, [26])

scRNA-seq files for baseline and differentiated H1 and H9 cells were downloaded from GEO Series GSE75748 [14].

Normalization

For normalization, we used Transcripts per kilobase million (TPM) to scale both datasets. Transcript lengths were based on the GRCh37 genome build from the Ensembl database through biomaRt [27]. We compared different methods of normalization, and found that unnormalized data and data normalized using SCnorm both showed clearly distinct populations of zeros (Fig. S8, Additional File 1). SCnorm normalized data also inferred many more bimodal genes than TPM normalized data. While most bimodal genes (92%) from TPM normalized data were also identified using SCnorm normalized data, read count-based data, regardless of normalization, showed a distinct gap between zeros and the lowest non-zero expression values (Fig. S8, Additional File 1). This is unsurprising, but the gap increased the apparent detection of bimodal genes. TPM-normalized data is effectively continuous on the non-negative real line, and hence allowed detection of bimodality without this artifact. Normalized data were then transformed by log(TPM + 1), and referred to as logTPM.

The two methods of normalization were compared based on the number of bimodal genes detected in H9 data, with the threshold of dip test $p < 0.01$.

Model description

Let Y_g denote the vector of observed gene expression values for a gene in a population of cells. Given a population of cells, the gene may be in active or inactive states in different cells. We define inactive (OFF) and active (ON) transcriptional states, X_{OFF} and X_{ON}, to denote the gene expression model of one gene based on the bimodal nature of transcription. We assume that the log of normalized gene expression values in state i follows a Gaussian distribution with a mean of μ_i and standard deviation of σ_i.

$$Y_g = X_{OFF} + X_{ON} \tag{1}$$

$$ln(X_i) \sim N\left(\mu_i, \sigma_i^2\right) \tag{2}$$

We take the threshold of 1 for mean gene expression values to distinguish the two states.

$$0 \leq \mu_{OFF} < 1 \leq \mu_{ON} \tag{3}$$

Due to the nature of scRNA-seq, we add in in zero inflation component to the model, yielding a sum of two lognormal distributions for the data plus the zero inflation.

$$Y_g = \delta(0) + X_{OFF} + X_{ON} \tag{4}$$

$$ln(Y_g + 1) = \delta(0) + N\left(\mu'_{OFF}, \sigma_{OFF}^2\right) + N\left(\mu'_{ON}, \sigma_{ON}^2\right) \tag{5}$$

We add a small amount of normally distributed noise after log transformation and model the result as a Gaussian mixture.

$$ln\left(Y_g + 1\right) + N\left(0, \epsilon^2\right) = N\left(0, \epsilon^2\right) + N\left(\mu'_{OFF}, \sigma_{OFF}^2 + \epsilon^2\right) + N\left(\mu'_{ON}, \sigma_{ON}^2 + \epsilon^2\right) \tag{6}$$

The amount of noise $\epsilon = 1 \times 10^{-5}$ was optimized over a wide range of values choosing the value that resulted in clear single maximum number of bimodal genes for the given data set (see "Results" section).

Bimodal gene identification and estimation of parameters

We first used Hartigan's dip test [28] to test for unimodality of logTPM data using the R package diptest (version 0.75-7, [29]). Next, we fit Gaussian mixture models with 2 or 3 centers using the R package Mixtools (version 1.1.0, [30]). During optimization of the method, we considered bimodal genes as those with $p < 0.01$ from the Dip test, and with $\mu_1 < 1$, and $\lambda_1 < 0.6$ from mixture model fitting.

To determine the optimal level of noise to be added, normally distributed added noise was tested over the range $\sigma \in [0, 1]$. Each run was evaluated based on the number of bimodal genes detected in H1 data, using the criteria described above. Bimodal gene selection criteria was adjusted in the final analysis to use only $p < 0.01$ for the dip test.

All scripts are available from https://github.com/GTK-lab/modalv.

Cell clustering

For clustering of cells, we used t-Distributed Stochastic Neighbor Embedding (t-SNE) [31, 32] implemented in R (Rtsne, [33]).

Signature analysis

Web-based signature analysis was performed using the MSigDB "Investigate Gene Sets" tool ([17], available from http://software.broadinstitute.org/gsea/msigdb/annotate.jsp) at the Broad Institute, with default parameters. Genes were compared to C2 (curated gene sets) category, C3 (motif gene sets), and C5 (Gene Ontology gene sets) gene sets [16].

Pseudotime trajectory estimation

Matrix of Mutual Information and variation of information between cells were calculated using the R package *infotheo* [34]. The *PST* package was used to build probabilistic suffix trees [35].

Network analysis

Co-switching gene networks were generated using the R package *minet* [36] followed by directed edge setting and annotation as described above.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12864-020-07009-8.

Additional file 1: Additional File 1: Supplementary Figures S1–S9 and Supplementary Tables S1–S5.

Acknowledgments

The authors thank the National University of Singapore for continued support and members of the Tucker-Kellogg lab for helpful comments and feedback.

About this supplement

This article has been published as part of BMC Genomics Volume 21 Supplement 10, 2020: Selected articles from the 18th Asia Pacific Bioinformatics Conference (APBC 2020): genomics. The full contents of the supplement are available online at https://bmcgenomics.biomedcentral.com/articles/supplements/volume-21-supplement-10

Authors' contributions

AJJ and GTK conceived of the idea for the project. AJJ performed the bioinformatics analysis. GTK supervised the project. AJJ and GTK wrote the paper. Both authors read and approved the final manuscript.

Funding

This work was supported by the Singapore Ministry of Education Academic Research Fund, through AcRF Tier 1 grant R154-000-693-112 at AcRF Tier 2 grant R154-000-A24-112 to Greg Tucker-Kellogg. The funding body had no role the design of the study and collection, analysis, interpretation of data or writing of the manuscript.

Availability of data and materials

ChIP-seq data was obtained from GEO at GSE16256, [26]. scRNA-seq data was from GSE75748 [14]. Code for this manuscript is at https://github.com/GTK-lab/modalv.

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests.

Published: 18 November 2020

References

1. Li B, Carey M, Workman JL. The role of chromatin during transcription. Cell. 2007;128(4):707–19. https://doi.org/10.1016/j.cell.2007.01.015.
2. Bannister AJ, Kouzarides T. Regulation of chromatin by histone modifications. Cell Res. 2011;21(3):381–95. https://doi.org/10.1038/cr.2011.22.
3. Azuara V, Perry P, Sauer S, Spivakov M, Jørgensen HF, John RM, Gouti M, Casanova M, Warnes G, Merkenschlager M, Fisher AG. Chromatin signatures of pluripotent cell lines. Nat Cell Biol. 2006;8(5):532–8. http://arxiv.org/abs/16570078.
4. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim T-K, Koche RP, Lee W, Mendenhall E, O'Donovan A, Presser A, Russ C, Xie X, Meissner A, Wernig M, Jaenisch R, Nusbaum C, Lander ES, Bernstein BE. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature. 2007;448(7153):553–60. https://doi.org/10.1038/nature06008.
5. Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, Fry B, Meissner A, Wernig M, Plath K, Jaenisch R, Wagschal A, Feil R, Schreiber SL, Lander ES. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell. 2006;125(2):315–26. https://doi.org/10.1016/j.cell.2006.02.041.
6. Mohn F, Weber M, Rebhan M, Roloff TC, Richter J, Stadler MB, Bibel M, Schubeler D. Lineage-specific polycomb targets and de novo DNA methylation define restriction and potential of neuronal progenitors. Mol Cell. 2008;30(6):755–66. https://doi.org/10.1016/j.molcel.2008.05.007.
7. Griffiths JA, Scialdone A, Marioni JC. Using single-cell genomics to understand developmental processes and cell fate decisions. Mol Syst Biol. 2018;14(4):8046. https://doi.org/10.15252/msb.20178046.
8. Stuart T, Satija R. Integrative single-cell analysis. Nat Rev Genet. 2019;20(5):257–72. https://doi.org/10.1038/s41576-019-0093-7.
9. Raj A, van Oudenaarden A. Nature, nurture, or chance: stochastic gene expression and its consequences. Cell. 2008;135(2):216–26. https://doi.org/10.1016/j.cell.2008.09.050.
10. Lun ATL, Bach K, Marioni JC. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. Genome Biol. 2016;17(1):75. https://doi.org/10.1186/s13059-016-0947-7.
11. Chen H-IH, Jin Y, Huang Y, Chen Y. Detection of high variability in gene expression from single-cell RNA-seq profiling. BMC Genomics. 2016;17(7):508. https://doi.org/10.1186/s12864-016-2897-6.
12. Brennecke P, Anders S, Kim JK, Kołodziejczyk AA, Zhang X, Proserpio V, Baying B, Benes V, Teichmann SA, Marioni JC, Heisler MG. Accounting for technical noise in single-cell RNA-seq experiments. Nat Methods. 2013;10(11):1093–5. https://doi.org/10.1038/nmeth.2645.
13. Vallejos CA, Marioni JC, Richardson S. BASiCS: Bayesian analysis of single-cell sequencing data. PLoS Comput Biol. 2015;11(6):1004333. https://doi.org/10.1371/journal.pcbi.1004333.
14. Chu L-F, Leng N, Zhang J, Hou Z, Mamott D, Vereide DT, Choi J, Kendziorski C, Stewart R, Thomson JA. Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. Genome Biol. 2016;17(1):173. https://doi.org/10.1186/s13059-016-1033-x.

15. Forney GD. The viterbi algorithm. Proc IEEE. 1973;61(3):268–78. https://doi.org/10.1109/PROC.1973.9030.
16. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinformatics. 2011;27(12):1739–40.
17. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102(43): 15545–50.
18. Cadigan KM, Waterman ML. TCF/LEFs and Wnt signaling in the nucleus. Cold Spring Harb Perspect Biol. 2012;4(11):007906. https://doi.org/10.1101/cshperspect.a007906.
19. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9(1):559. https://doi.org/10.1186/1471-2105-9-559.
20. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Favera RD, Califano A. ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics. 2006;7(1):7. https://doi.org/10.1186/1471-2105-7-S1-S7.
21. Wang T, Li B, Nelson CE, Nabavi S. Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. BMC Bioinformatics. 2019;20(1):40. https://doi.org/10.1186/s12859-019-2599-6.
22. Sachs M, Onodera C, Blaschke K, Ebata KT, Song JS, Ramalho-Santos M. Bivalent chromatin marks developmental regulatory genes in the mouse embryonic germline in vivo. Cell Rep. 2013;3(6):1777–84. https://doi.org/10.1016/J.CELREP.2013.04.032.
23. Court F, Arnaud P. An annotated list of bivalent chromatin regions in human ES cells: A new tool for cancer epigenetic research. Oncotarget. 2017;8(3):4110. http://arxiv.org/abs/1811.03954.
24. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2013.
25. Wickham H. Ggplot2: Elegant graphics for data analysis, 3rd edn. New York: Springer-Verlag; 2016.
26. Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, Kellis M, Marra MA, Beaudet AL, Ecker JR, Farnham PJ, Hirst M, Lander ES, Mikkelsen TS, Thomson JA. The NIH roadmap epigenomics mapping consortium. Nat Biotechnol. 2010;28:1045–8.
27. Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat Protoc. 2009;4(8):1184–91. https://doi.org/10.1038/nprot.2009.97.
28. Hartigan JA, Hartigan PM. The dip test of unimodality. Ann Stat. 1985;13(1):70–84. https://doi.org/10.1214/aos/1176348654.
29. Maechler M. diptest: Hartigan's Dip Test Statistic for Unimodality - Corrected; 2016. https://CRAN.R-project.org/package=diptest.
30. Benaglia T, Chauveau D, Hunter DR, Young D. Mixtools: an R package for analyzing finite mixture models. J Stat Softw. 2009;32(6):1–29.
31. van der Maaten LJP, Hinton GE. Visualizing data using t-SNE. J Mach Learn Res. 2008;9:2579–605.
32. van der Maaten LJP. Accelerating t-SNE using tree-based algorithms. J Mach Learn Res. 2014;15:3221–45.
33. Krijthe JH. Rtsne: T-Distributed Stochastic Neighbor Embedding using a Barnes-Hut Implementation; 2015. https://github.com/jkrijthe/Rtsne.
34. Meyer PE. infotheo: Information-Theoretic Measures. R package version 1.2.0. 2014. https://CRAN.R-project.org/package=infotheo.
35. Gabadinho A, Ritschard G. Analyzing state sequences with probabilistic suffix trees: The PST R package. J Stat Softw. 2016;72(3):1–39. https://doi.org/10.18637/jss.v072.i03.
36. Meyer PE, Lafitte F, Bontempi G. minet: AR/Bioconductor package for inferring large transcriptional networks using mutual information. BMC Bioinformatics. 2008;9(1):461. https://doi.org/10.1186/1471-2105-9-461.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.