Briefings in Bioinformatics, 2022, **23(2)**, 1–16

https://doi.org/10.1093/bib/bbac017

Problem Solving Protocol

Visualization, benchmarking and characterization of nested single-cell heterogeneity as dynamic forest mixtures

Benedict Anchang, Raul Mendez-Giraldez, Xiaojiang Xu, Trevor K. Archer, Qing Chen, Guang Hu, Sylvia K. Plevritis, Alison Anne Motsinger-Reif and Jian-Liang Li

Corresponding author: Benedict Anchang, Biostatistics and Computational Biology Branch, National Institute of Environmental Health Sciences. 111 T W Alexander Dr, Research Triangle Park, NC 27709, USA and Center for Cancer Research, National Cancer Institute, Bethesda, MD 20892, USA. Tel +1 984-287-3350; E-mail: benedict.anchang@nih.gov

Abstract

A major topic of debate in developmental biology centers on whether development is continuous, discontinuous, or a mixture of both. Pseudo-time trajectory models, optimal for visualizing cellular progression, model cell transitions as continuous state manifolds and do not explicitly model real-time, complex, heterogeneous systems and are challenging for benchmarking with temporal models. We present a data-driven framework that addresses these limitations with temporal single-cell data collected at discrete time points as inputs and a mixture of dependent minimum spanning trees (MSTs) as outputs, denoted as dynamic spanning forest mixtures (DSFMix). DSFMix uses decision-tree models to select genes that account for variations in multimodality, skewness and time. The genes are subsequently used to build the forest using tree agglomerative hierarchical clustering and dynamic branch cutting. We first motivate the use of forest-based algorithms compared to single-tree approaches for visualizing and characterizing developmental processes. We next benchmark DSFMix to pseudo-time and temporal approaches in terms of feature selection, time correlation, and network similarity. Finally, we demonstrate how DSFMix can be used to visualize, compare and characterize complex relationships during biological processes such as epithelial–mesenchymal transition, spermatogenesis, stem cell pluripotency, early transcriptional response from hormones and immune response to coronavirus disease. Our results indicate that the expression of genes during normal development exhibits a high proportion of non-uniformly distributed profiles that are mostly right-skewed and multimodal; the latter being a characteristic of major steady states during development. Our study also identifies and validates gene signatures driving complex dynamic processes during somatic or germline differentiation.

Keywords: forest mixtures, multimodality, minimum spanning tree, nested models, single-cell trajectory analysis, cell differentiation

Introduction

A major topic of controversy in developmental biology revolves around whether cell types and cell states are continuous, discrete, or a mixture of both during a given developmental process. Single-cell time-course analysis is key to addressing this important question. Most trajectory models are either based on pseudo-time progression including Monocle [1, 2], PHATE [3, 4], tSpace [5], Wanderlust [6], PAGA [7], Slingshot [8], Spanning-tree progression analysis of density-normalized events (SPADE) [9], Palantir [10, 11] and CellRank [12], Scuba [13]. Some are mechanistically driven including RNA velocity [14] and others are time dependent like TRACER [15], Waddington-OT [16] Tempora [17] and CStreet [18]. These models are optimized to visualize either a disjointed or continuous transition of states, but not both. For example, a developmental process like spermatogenesis has been shown to comprise a mixture of discrete and continuous states [19]

Benedict Anchang is a Stadtman principal investigator at National Institute of Environmental Health Sciences and National Cancer Institute leading a computational and systems biology group on multi-scale modeling,visualization and integration of biological processes using single-cell analysis.

Raúl Méndez-Giráldez is a bioinformatics scientist at the National Institute of Environmental Health Sciences. He is developing computational methods to study dynamic trajectories of cell differentiation.

Xiaojiang Xu is senior bioinformatics scientist at National Institute of Environmental of health sciences. He has expertise in data analyses of NGS and single-cell sequencing.

Dr. Trevor K. Archer is the Deputy Director of National Institute of Environmental Health Sciences. His expertise and research efforts are in the field of epigenetics and stem cell biology.

Qing Chen is a postdoc fellow in Stem Cell Biology Group at National Institute of Environmental Health Sciences. She is working on molecular mechanisms underlying pluripotency maintenance and spermatogenesis.

Guang Hu is a Senior Investigator at National Institute of Environmental Health Sciences. He is studying the regulation of pluripotency in mouse and human stem cells.

Sylvia K. Plevritis is a Professor and Chair in the Department of Biomedical Data Science in the Stanford School of Medicine. She develops mathematical models of cancer progression and evaluates the effectiveness of mammography and MRI in screening for breast cancer and CT in screening for lung cancer.

Alison Motsinger-Reif is a Senior Investigator at the National Institute of Environmental Health Sciences. Her work focuses on the development and application of methods in statistical genetics and genomics.

Jian-Liang Li is the Director of Integrative Bioinformatics Support group at National Institute of Environmental Health Sciences. His research interests include single cell and spatial omics data analysis, integration and visualization.

Received: September 2, 2021. **Revised:** November 19, 2021. **Accepted:** January 13, 2022

Published by Oxford University Press 2022.

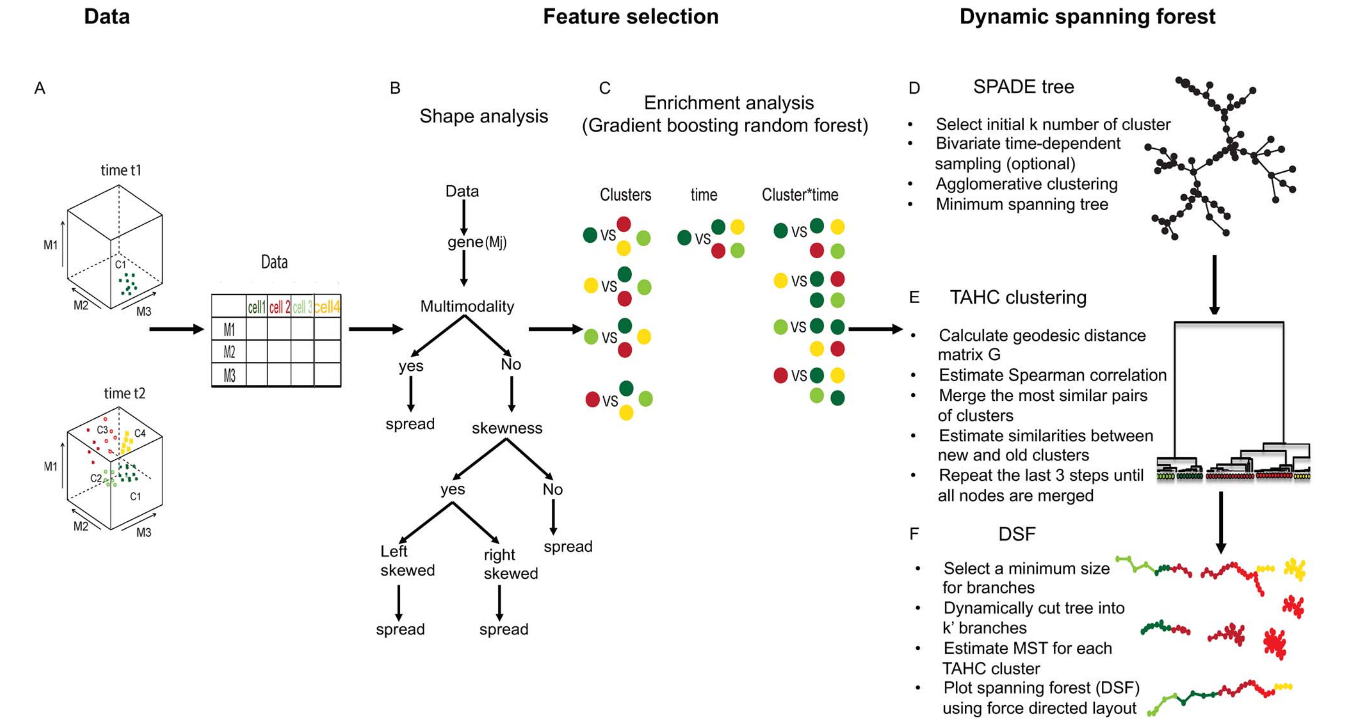

Data
Feature selection
Dynamic spanning forest
A
time t1
M1
C1
M2
M3
time t2
C3
C4
C2
C1
M1
M2
M3
Data
cell1
cell 2
cell 3
cell4
M1
M2
M3
B
Shape analysis
Data
gene(Mj)
Multimodality
yes
No
spread
skewness
yes
No
Left skewed
right skewed
spread
spread
spread
C
Enrichment analysis
(Gradient boosting random forest)
Clusters
time
Cluster*time
vs
D
SPADE tree
• Select initial k number of cluster
• Bivariate time-dependent sampling (optional)
• Agglomerative clustering
• Minimum spanning tree
E
TAHC clustering
• Calculate geodesic distance matrix G
• Estimate Spearman correlation
• Merge the most similar pairs of clusters
• Estimate similarities between new and old clusters
• Repeat the last 3 steps until all nodes are merged
F
DSF
• Select a minimum size for branches
• Dynamically cut tree into k' branches
• Estimate MST for each TAHC cluster
• Plot spanning forest (DSF) using force directed layout

Figure 1. DSFMix takes as input a time course or staging single-cell data and outputs a dynamic spanning forest (DSF). **(A)** A three-dimensional single-cell time-course input data for DSFMix comprising two time points, or stages of development. The green cells in the first time point (t1) occupy a different region in the second time point (t2) and differentiate into light green, red, and yellow cells. **(B)** A binary decision tree and feature selection process representing a shape analysis to generate an optimal lineage marker set for enrichment analysis. The shape analysis step uses a predefined FDR to select variable markers based on the shapes of marginal distribution. The step produces markers whose expression across cells is multimodal, unimodal but symmetrical, left-skewed, or right-skewed. **(C)** The enrichment analysis feature step selects markers for cluster specificity, time specificity and cluster-time interaction specificity using a boosted random forest regression model with binary and multinomial outcome. **(D)** Minimum spanning tree derived using SPADE. **(E)** Tree agglomerative hierarchical clustering uses geodesic distances and spearman correlation between all node pairs to produce a sorted dendrogram that represents the merging process of all nodes from the input tree. **(F)** A dynamic minimum spanning forest produced from minimizing distances between and within clusters in the TAHC clusters. The clusters are derived from a dynamic and iterative branch-cutting method based on the structure of the underlying dendrogram.

which requires new computational models that account for mixtures of discrete and continuous lineages that may result from heterogeneous progenitor cells. Furthermore, most trajectory models require prior knowledge of progenitor cells for optimal lineage reconstruction and do not account for variations in distributions, competing or nested model structures due to the potential heterogeneity of these progenitor cells. Minimum spanning trees (MSTs) have been used successfully to reconstruct developmental processes such as hematopoiesis [20] and intestinal stem-cell differentiation [21]. They are optimal for visualizing very large, high-dimensional datasets at a lower computational complexity cost compared to traditional methods [22]. However, if a tree was applied to model the same sample without major stem or progenitor cells, which might occur during data collection, the resulting tree would be misleading because different lineages would be connected in a single tree. This study outlines dynamic spanning forest mixtures (DSFMix), a computational framework (Figure 1) that has temporal or staged single-cell data collected at discrete time points as input and a mixture of discrete and/or continuous lineages as the output. We subsequently address three key features or limitations of current trajectory-based algorithms that may impact their usage and performance. We particularly focus on tree-based graphical methods.

First, many methods throw away many genes and rely on an initial set of genes or markers that are manually selected or preselected as having highly variant expression [23, 24] or as being highly expressed across cells [25]. Alternative recent approaches may retain highly dropout genes [26] or highly deviant ones [27, 28]. Thus, they may produce biased or incomplete results when suboptimal lineage genes are chosen to build a tree. Median absolute deviation (MAD), a robust measure of variability [29] has been used for gene selection in single-cell analysis [21]. Because MAD is aimed at symmetric distributions and requires a median to estimate parameters, we show that most significant genes expressed during development are multimodal and asymmetric. This motivates the use of location-free scale estimators such as Q_n [29] which is the first quartile of ordered pairwise interpoint absolute distances. Furthermore, use of location-free measures of spread, combined with the optimization of shape features of marginal distributions, can potentially circumvent some of the technical challenges inherent in scRNA-seq data and reduce bias in the marker selection process for downstream analysis. Figures 1A–C summarize the process of selecting an optimal marker set (feature selection) for building a forest. Figure 1A depicts 3D single-cell time-course data comprising two time points, or stages of development. The green cells in the first time point occupy a different region in the second time point and are differentiated into light green, red, and yellow cells. The goal is to classify the expression profile of each gene M_j into a shape class (i.e. multimodal, unimodal but symmetrical, left-skewed, or right-skewed) (Figure 1B). Using a decision tree controlled by a predefined Benjamini-Hochberg false discovery rate (FDR) [30] we optimize a sequential two-step feature gene set. The first step (shape analysis) produces a disjointed set of shape genes. We derive *P*-values corresponding to the associated tests statistics by controlling one- or two-sided tests using a dip test for multimodality [31], a symmetry test [32] for skewness and a novel non-uniform test of dispersion around an unknown median (**Supplementary Information section Methods**). The second step (enrichment analysis, Figure 1C), which is described in more detail later in the paper, generates markers enriched for individual clusters and either time or time–cluster interactions.

A second limitation is that current trajectory models are not generalizable to all dynamic biological structures making it a challenge for benchmarking. As highlighted in [16], some models cannot handle branching trajectories [6] and others are not applicable to developmental time-course datasets or do not incorporate time information [1, 8]. A continuous method like Waddington-OT [16] can leverage on time information as well as model cell growth rates while other tree-based or network models (Scuba [13], TRACER [15], Tempora [17], and CStreet [18]) marginalize over time. However, since tracking individual single cells over time in high-dimensional space is still a major experimental challenge, discrete dynamic models [33] offer a complimentary solution by accounting for uncertainty in time and state space transitions. DSFMix accounts for variations due to time, clusters (i.e. cell types) and variations due to their joint interactions, and can additionally represent complex dynamic structures including simultaneous and nested biological processes as forests. For example, an adaptive response like phenotypic plasticity involves reversible processes occurring simultaneously that are traditionally studied by sequential and staging experiments [15]. Epithelial–mesenchymal transition (EMT) is a developmental process that has been shown to exhibit a complex phenotype characterized by reversible biological processes [15]. A time-independent state transition Markov model TRACER identified distinct EMT and mesenchymal-to-epithelial transition (MET) trajectories from the time course study involving both treatment and withdrawal of TGFβ [15]. In the latter study, an independent visualization tool (PHENOSTAMP) based on t-distributed stochastic neighbor embedding (t-SNE) was used. We show that DSFMix visualizes differentiation between simultaneous complex lineages such as those of EMT and MET, and further breaks down each of these major transition networks into recurrent and statistically significant dynamic subgraphs or patterns called motifs. Furthermore, we also demonstrate how DSFMix forest output can be combined with time ordering to generate a directed tree (minimum cost arborescence [34, 35]) to allow for benchmarking with other time-dependent trajectory models. In practice, given that most available time-course datasets are collected at a few discrete time points, with observed over-representation of highly correlated genes in scRNA-seq datasets, DSFMix uses a boosted random forest regression with binary response to further select top 'shape' genes that are enriched at specific time points, states, and time-state interactions. This gene set is the input for the dynamic spanning forest (DSF) step of the DSFMix algorithm (Figure 1D–F).

The third limitation is the manual annotation of lineages after tree construction by some tree-based methods. For instance, SPADE which produces a MST output combines all cells into a single tree, thereby forcing a continuous relationship among all subpopulations. This introduces uncertainty regarding the boundaries of the derived dynamic subtrees. Tools such as partition-based graph abstraction (PAGA) [7], based on *k*-nearest neighbors (KNN) algorithm, provide an interpretable graph-like map of the resulting data manifold based on estimating the connectivity of manifold partitions. Keeping in mind that a large tree can be broken down into a disconnected forest, DSFMix first constructs a large MST (Figure 1D) based on the genes from the feature selection step, and then separates the tree into a minimum spanning forest (MSF). Spanning forests have been used to capture cellular lineages from static data [36] using *t*-SNE-based clustering. DSFMix accounts for additional time-dependent interactions and provides a complementary solution to the problem of creating a forest of dependent trees. It also provides an unbiased test for comparing and ordering trees within a forest which can be useful for many downstream analyses. Moreover, it provides a flexible framework to identify or study unknown intermediate processes in larger trajectories using sub-trajectory analysis. In this work, we automate the dynamic population identification step by implementing an extension of the tree agglomerative hierarchical clustering algorithm (TAHC) (Figure 1E) introduced by Yu *et al.* [37] that can be used to detect optimal cluster tree motifs in any MST using weighted shortest-path (geodesic) distances. The length of the shortest path between two nodes indicates the distance or degree of connectivity between pairs of vertices in a tree. Finding communities in complex networks has recently become an area of active research across several disciplines. Most approaches like PAGA or Monocle3 use a hierarchical clustering framework such as the well-known divisive Girvan Newman algorithm [38] or agglomerative Louvain method [39]. However, these approaches are not optimal for cluster detection in tree-like structures. Traditionally, clusters are defined as a group of nodes with greater within-cluster density than between-cluster density. These algorithms are not optimal for clustering tree-like topologies with several uniformly dense, connecting paths [37]. A

DSF is derived from an adaptive, iterative process of cluster decomposition and composition [40] on the structure of the underlying TAHC dendrogram from clustering a MST. This process produces both labeled and unlabeled classes of clusters, in contrast to traditional approaches that assign each node to a cluster grouping. Using a predefined minimum cluster node size, DSFMix produces a set of dynamic tree motifs characterized by multiple disjointed and/or nested lineages that are ultimately optimized into a MSF. Since we control time-dependent events, DSFMix output (Figure 1F) is a dynamic MSF with multiple lineages occurring sequentially or simultaneously over time.

In this study, we first demonstrate the importance of forest-based algorithms compared to single tree approaches for visualizing and characterizing several developmental processes. Next, we benchmark DSFMix to traditional pseudo-time and temporal graphical approaches as well as assess the impact of its novel feature selection step compared with classical approaches. In addition, we compare the performance of TAHC with other graphical clustering algorithms, including the GN algorithm, uniform manifold approximation and projection [41] (UMAP) and *t*-SNE [42]. Finally, we demonstrate how DSFMix can be used to visualize, characterize, and compare similar intermediate cell lineages during dynamic biological processes. Specifically, we identify gene signatures driving complex dynamic processes during somatic or germline differentiation using dispersion analysis instead of averages. We illustrate how subtrees can be compared or ordered within a forest and how overfitting can be reduced using a non-parametric test that is robust to unbalanced network sizes. We also demonstrate how DSFMix can be combined with genomic correlation search engines for new scientific discoveries and validation. This study provides a complete framework for and insights on visualizing and investigating single-cell or multicellular heterogeneity as dynamic forest mixtures.

Confounding of relationships by single-tree approaches assuming a continuous lineage

In this section, we provide motivation for the use of forest-based algorithms by highlighting some of the limitations of using single tree approaches like SPADE for visualizing intestinal development [21] and hematopoiesis [20]. Supplementary Figure S1A (see Supplementary Data available online at http://bib.oxfordjournals.org/) shows the result of SPADE using normalized single-cell RNA-seq data from Yan *et al.* [21] to reconstruct the lineage hierarchy during development. The data span all major cell types during mouse intestinal development and thus serve as a good benchmark to understand the interrelatedness of clusters. A large single subpopulation of non-cyclic intestinal stem cells (blue cells) differentiates into secretory Paneth (red) and chemosensory tuft (yellow) progenitor cells. Supplementary Figure S1B (see Supplementary Data available online at http://bib.oxfordjournals.org/) shows the results if most of these non-cyclic stem cells are not included or missing during data generation. In this scenario, SPADE connects all cells into a tree, distorting the lineage orientation by locating the expected central position of cyclic stem cells on one end of the tree and the distal differentiated cells on the other. Thus, to ensure the best performance of a single tree, the cell of origin, as well as major intermediate lineages, must be present in the data.

Hematopoietic stem cells (HSCs) have been shown to differentiate phenotypically and functionally distinct cell types whose lineages are better captured by a forest instead of a tree. The differentiation process is usually defined by a set of known lineage markers. For example, unstimulated human bone marrow cells from a healthy donor were analyzed using 13 surface markers, namely CD45, CD45RA, CD19, CD11b, CD4, CD8, CD34, CD20, CD33, CD123, CD38, CD90 and CD3 [6, 36]. SPADE was applied to recapitulate the continuous hematopoietic hierarchy and 24 subpopulations were manually isolated from the tree, spanning HSC, myeloid and lymphoid lineages. Supplementary Figure S1C (see Supplementary Data available online at http://bib.oxfordjournals.org/) shows a static forest output, highlighting seven major lineage subpopulations characterized by different tree structures: cytotoxic T and helper T cells, pre-B and B cells, natural killers and myeloid cells, namely macrophages and monocytes. These tree motifs can help identify specific combinations of signaling and lineage markers that define major individual lineages.

DSFMix applied to several time course datasets to study cellular progression

We introduce a combination of mass cytometry (CyTOF) and scRNA-seq time course data to demonstrate the broad applicability of DSFMix. The first CyTOF time-course dataset (emtdata) revealed significant phenotypic differences when cells undergo EMT and reversible MET processes [15]. In this study, single-cell data were collected for 20 days after in-vitro treatment of lung cancer cell lines with TGFβ for 10 days, followed by TGFβ withdrawal for the subsequent 10 days. Given that EMT-MET is a well-studied and validated dynamic process, we use the emtdata as the main benchmarking data for comparing DSFMix features with other trajectory models. We also make use of four additional scRNA-seq time-course single-cell datasets to test the DSFMix framework. The first scRNA-seq dataset (spermdata) comprises pooled normalized and batch-corrected (see **Supplementary Information section Methods** for further details) mouse testes data from four published [43–46] and unpublished studies. The dataset consists of approximately 100,000 cells that span 16 postnatal stages during spermatogenesis, associated with days 0, 3, 5, 6, 7, 8, 10, 14, 15, 18, 20, 25, 30, 35, 67 and 80. The data characterize heterogeneous and spatial spermatogonia, spermatocyte, spermatid and mature spermatozoa developmental stages and are confounded by the asynchronous stages of cells collected at different cycles of spermatogenesis. We use DSFMix to capture major developmental processes and characterize the underlying heterogeneity surrounding the transition from spermatogonia to spermatocyte to spermatozoa lineages. The second dataset (ipscdata) is from Zhao *et al.* [47], who investigated pluripotency using chemically induced cellular reprogramming of somatic mouse embryonic fibroblasts (MEFs) to embryonic stem cells (ESCs). The dataset comprises 50,000 cells collected from 12 time points over 21 days. We use DSFMix to capture lineages with heterogeneous intermediate extraembryonic endoderm (XEN)-like transitions. The third dataset (hormonedata) is from Hoffman *et al.* [48], who used it to study the early transcriptional hormonal response of A1–2 breast cancer cell lines treated with glucocorticoid dexamethasone (Dex). The data comprise cells sampled at six time points spanning 18 h post-treatment. We apply DSFMix to the data to better characterize the underlying dynamic drug-response landscape. To further demonstrate how DSFMix can be used to analyze data from two time point (case–control) studies we also make use of a fourth scRNA-seq data to study the immune response to coronavirus disease (COVID-19) [49].

Results

We next investigate how various steps of the DSFMix algorithm, beginning with feature selection compares with other traditional approaches and affects the dynamic forest output.

DSFMix feature selection step output is genes with a very high proportion of nonuniformly, marginally distributed shapes skewed to the right

Typical scRNA-seq data are sparse, and all genes tend to show at least a bi-modal structure, with most having the highest density at zero (Supplementary Figure S2A, see Supplementary Data available online at http://bib.oxfordjournals.org/). During shape analysis (first step), we condition the feature selection on non-zero expression values to mitigate the dropping of numerous low-variable but significant genes with $Q_n = 0$ values. For spermdata, this process results in more than 20,000 genes with $Q_n > 0$ compared to about 1,000 genes if the zeros are included (Supplementary Figure S2A, see Supplementary Data available online at http://bib.oxfordjournals.org/). To score a given gene, we use statistical testing of scale and shape measures of marginal distributions around the unknown median. To control for false positives, we use a predefined FDR threshold. For large datasets (>100,000 cells), we set the cut-off to a recommended <1e−3. Following the binary decision tree in Figure 1B, we determine if each gene M_j is multimodal, unimodal but symmetrical, left-skewed, or right-skewed (**Supplementary Information section Methods**). In scenarios with several *P*-values with a zero value, we rank the genes using the effect sizes, or Q_n values. Figure 2A shows histograms (columns) of distributions of selected genes that are associated with multimodality, unimodal symmetry, left skewness, and right skewness for three of the above scRNA-seq datasets (rows). In summary, the output of the DSFMix shape analysis step is genes that exhibit a very high proportion of non-uniformly, marginally distributed shapes that are mostly skewed to the right. In addition, after Dex treatment, the hormone response data interestingly support a lack of cell multimodality (Figure 2B).

Enrichment analysis using boosted random forest regression shows that multimodal genes are strongly associated with major steady states during development

To select optimal input markers for a forest output, DSFMix requires each cell to be associated with an initial cluster (k_0) and time label (t). These clusters can be easily derived using rapid algorithms such as k-means [50] or from independent standard preprocessing pipelines, e.g. Seurat [24], Monocle [1, 2] and Cytobank [51]. We compare each cluster, time and cluster–time interaction level (Figure 1C) with all alternatives using a logistic boosted random forest framework [52] implemented in the XGBoost R package (**Supplementary Information section Methods**). This approach is fast, scalable and enables the ranking of important genes based on gain, cover and frequency of occurrence.

To construct an optimal DSFMix forest, we select the most differentially expressed gene or marker for each comparison. We also include the most differentially expressed gene across all cluster, time and cluster-time interaction variables using a multinomial model. The results in Supplementary Figure S3A and B (see Supplementary Data available online at http://bib.oxfordjournals.org/) based on the emtdata indicate that even without prior knowledge of initial clusters, gene enrichment at discrete time points, combined with shape analysis, produces very similar SPADE trees (P-value = 0.99). To model EMT, Karacosta *et al.* [15] selected canonical markers: vimentin, ecadherin, MUC1, TWIST, CD24, and CD44, through a combination of principal component analysis and tree-based regression, with the top three principal components as the response. This procedure is also used by many pseudo-trajectory models in scRNA-seq analysis. The DSFMix feature selection framework also controls for additional time-dependent cell–cell variation, resulting in additional markers (e.g. phosphorylated retinoblastoma (pRb), FAP, TROP2, keratin-7, CD45). We also observe a more heterogeneous branching process for the DSFMix 9-marker single-tree model (Supplementary Figure S3C, see Supplementary Data available online at http://bib.oxfordjournals.org/) compared to the 6-marker tree in Karacosta *et al.* We also assess the performance of DSFMix when no feature selection is performed comparing the use of its shape

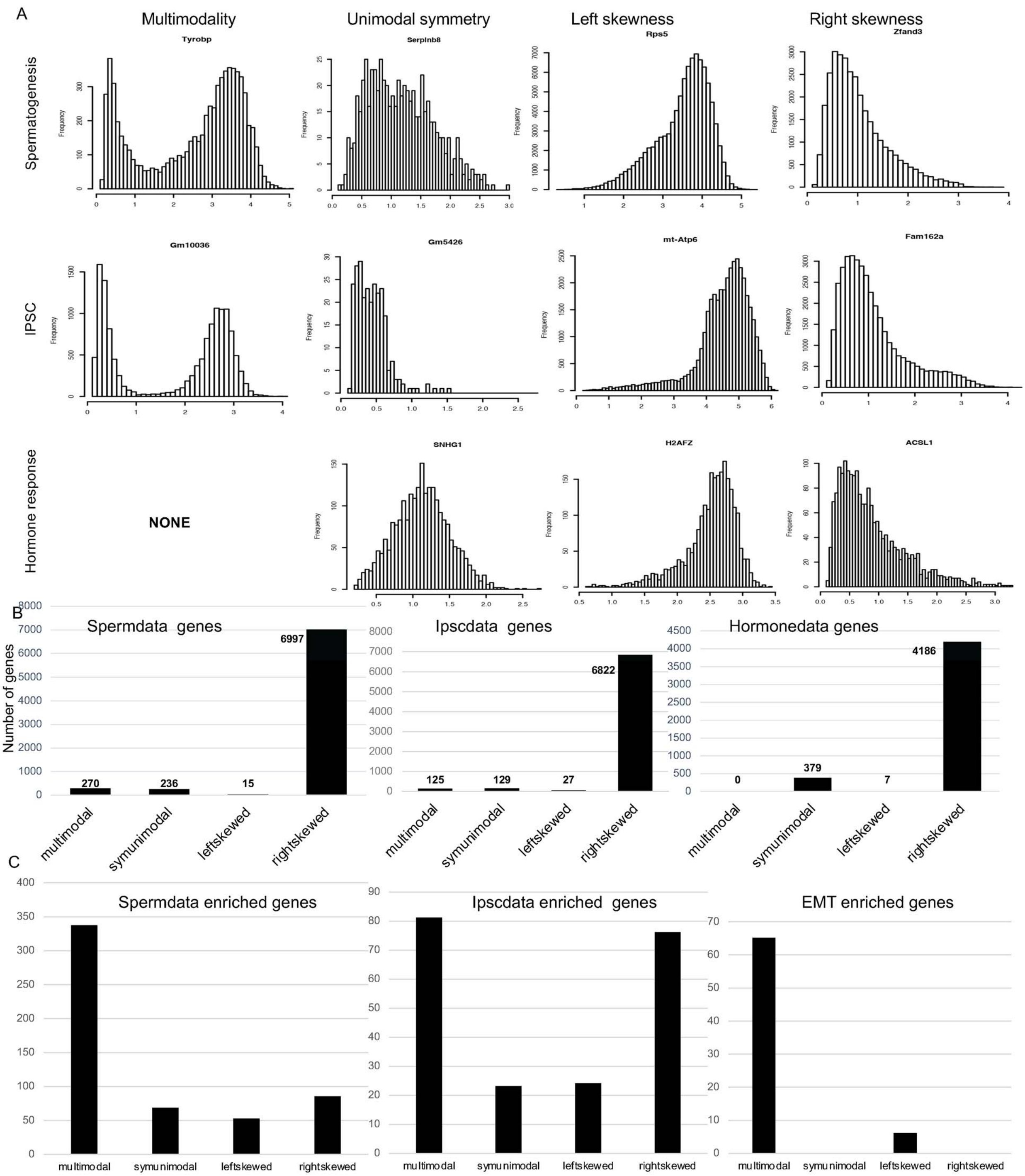


Figure 2. Analysis of DSFMix feature selection step applied to various developmental processes. **(A)** Histograms (excluding zeros) showing examples of genes associated with maximum marginal spread and shape for spermdata (top), ipscdata (middle), and hormonedata (bottom). **(B)** Barplots showing the normalized marginal expression of genes during biological process. A very high proportion of non-uniform distribution of shapes which are mostly skewed to the right is observed. **(C)** Distribution of enrichment genes associated with multimodality, unimodal and symmetrical, left skewness and right skewness for spermdata, ipscdata and emtdata. The multimodal genes are enriched the most during development.

driven feature framework to using all markers within a PCA projection. Supplementary Figure S3D (see Supplementary Data available online at http://bib.oxfordjournals.org/) shows the two SPADE trees under both scenarios. In addition, a weighted edge count test shows no significant difference at 0.05 level (P-value = 1) indicating that the feature selection step of DSFMix is robust to variations in the feature space.

Figure 2C shows the frequency of association between genes that are highly enriched across specific time-dependent clusters and their shape characteristics during spermatogenesis, iPSC reprogramming, and EMT.

Table 1. Comparison of major DSFMix features with pseudotrajectory or temporal trajectory models

Method	Time	Transformed feature space	Trajectory structure	Nested structures	Visualization output
Monocle3	No (pseudotime)	Yes	Continuous	No	*t*-SNE or UMAP
PHATE	No (diffusion pseudotime)	Yes	Continuous	No	PHATE visualization
tSpace	No (pseudotime)	Yes	Continuous	No	tPCA projection
Slingshot	No (pseudotime)	Yes	Continuous	No	PCA projection
PAGA	No (diffusion pseudotime)	Yes	Continuous	No	UMAP
Waddington-OT	Yes (cellular time dependent growth)	Yes	Continuous	No	Force directed graph layout
TRACER	Yes (Time independent markov model)	Yes	Discrete	No	Directed network
Scuba	Stochastic binary tree	Yes	Bifurcation	No	Binary tree
Tempora	Yes (time ordering)	Yes	Discrete	No	Directed network
CStreet	Yes (probabilistic time ordering)	Yes	Discrete	No	Time ordering network
DSFMix	Yes (time ordering)	No	Discrete and/or Continuous	Yes	Spanning tree/forest

Multimodal genes are enriched the most during normal development.

DSFMix output is a minimum spanning forest from dynamic tree-based clustering

This section presents a brief overview of the final step of the DSFMix algorithm. Figure 1D–F summarizes the three key steps for generating a DSFMix forest. Each step is associated with an output figure and text box outlining the underlying sub-steps.

In the first step (Figure 1D), SPADE is implemented, generating a k-node MST derived from the means of the m selected feature markers. The $k >> k_0$ nodes correspond to k-means of clusters generated from a fast, outlier-resistant clustering algorithm based on the Gini index [53] (**Supplementary Information section Methods**). This approach circumvents the need for the down-sampling and up-sampling steps of SPADE and can accommodate most sample sizes from published scRNA-seq datasets. However, for reproducibility, we implement optional bivariate discrete-distribution time-dependent sampling for large datasets (>1 million cells), which guarantees a similar relative proportion of cells across discrete time points and sampled cell states. Each node (size) in the SPADE output represents a median expression (number of cells) within that node.

In the second step (Figure 1E), TAHC is performed on the MST or SPADE tree (**Supplementary Information section Methods**). Instead of traditional distance measures used with agglomerative clustering approaches, which tend to neglect outer nodes, TAHC uses geodesic distances between all node pairs to represent the number of links in the shortest path between two directly linked nodes. The output of TAHC is a sorted dendrogram (Figure 1E) that represents the merging process of all nodes from the input tree. The leaf nodes of the dendrogram are linearly ordered by the smallest distances between merged clusters [54].

In the third step (Figure 1F), DSFMix performs unsupervised clustering of the TAHC dendrogram. One approach to determine the size of the forest k^l is to cut the dendrogram at an optimal height by visual inspection. However, such a predefined static cut-off can lead to singletons as well as non-nested clusters. To determine the unknown number of clusters, DSFMix uses a dynamic branch-cutting top down approach which relies solely on the structure of the dendrogram. DSFMix adapts a heuristic algorithm implemented in the dynamicTreeCut R package [40] and applies an iterative process of cluster decomposition and reconstruction stopping when the number of clusters becomes stable (**Supplementary Information section Methods**). DSFMix uses a circular or Kamada–Kawai force-directed layout (Kamada and Kawai [55]) to view the forest. Each node of the forest can be color-coded to display for each gene, the entire expression distribution across all the cell clusters with node sizes representing their relative sample sizes. Note that we can further convert the DSFMix spanning forest output to a directed forest by conditioning on time using the Edmonds' algorithm [34, 35] (**Supplementary Information section Methods**). This allows potential comparisons with other directed temporal models.

Benchmarking DSFMix to traditional pseudo-time approaches with respect to time, network similarity, and interpretability

Single-cell pseudo-time trajectory models do not explicitly represent real time. Table 1 summarizes the major differences between real time-based models, e.g. DSFMix and Waddington-OT [16] and several pseudo-time algorithms such as Monocle3, PHATE, tSpace, PAGA and Slingshot in terms of variations in time or pseudo-time, feature selection, trajectory structure, nested structure identification and visualization output. The tabulated differences demonstrate the challenges involved in benchmarking these algorithms in addition to lack of an ideal ground truth data due to difficulty in tracking high-dimensional features of individual live cells over time. We then use time correlation analysis (**Supplementary Information section Benchmarking**

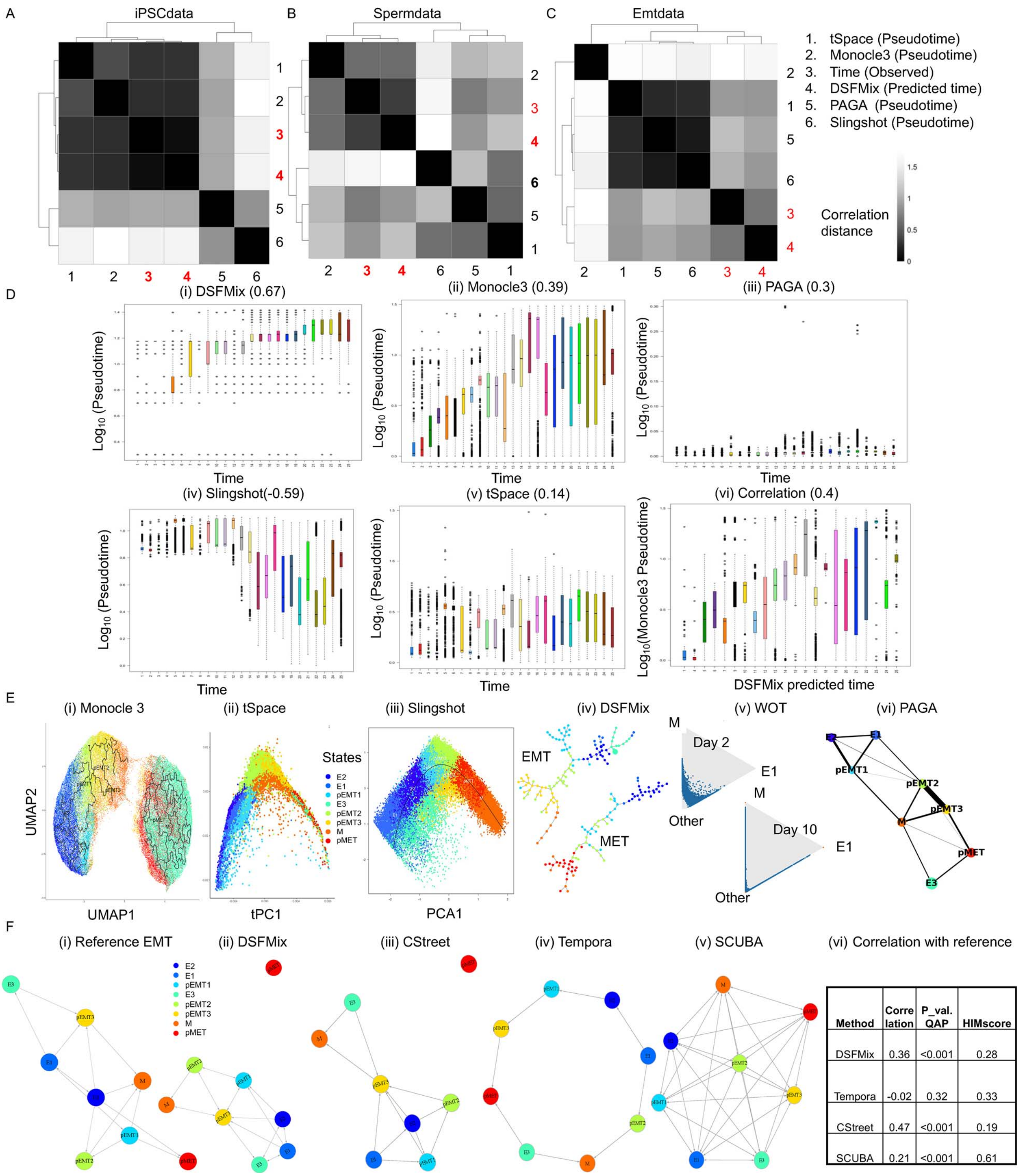
time correlation analysis) to show that predictions of DSFMix strongly correlates more with observed time trends compared to several pseudo- time algorithms such as tSpace, Monocle3, PAGA and Slingshot. Recent studies [3] have used non-linear correlation analysis as an implicit benchmarking measure for these algorithms with the expectation that the progression of a given biological process should also correlate with sampled staged time measurements even in the presence of noise, redundancy and complex branching structures. Figure 3A and B represents the hierarchical clustered heatmaps of the distance matrices based on Spearman's rank correlation coefficient between pseudotime models (1:tSpace, 2:Monocle3, 5:PAGA, 6:Slingshot), observed time (3) and DSFMix forest predicted time ordering (4) for ipscdata, spermdata and emtdata respectively. The similarity scores show that DSFMix is more correlated to observed time trends compared to Monocle3, Slingshot, PAGA and tSpace for all three datasets. This observation is supported by correlation trends from boxplots in Figure 3E (i–v) ordered by time for DSFMix, Monocle3, PAGA, slingshot and tSpace. DSFMix is associated with a high correlation of 0.92 (ipscdata, Supplementary Figure S4A (i), see Supplementary Data available online at http://bib.oxfordjournals.org/), 0.67 (spermdata, Figure 3D (i)) and 0.57 (emtdata, Supplementary Figure S4B (i), see Supplementary Data available online at http://bib.oxfordjournals.org/) respectively followed by Monocle3 with a correlation of 0.74, 0.39 and −0.1 respectively. Figure 3Evi shows a correlation of 0.4 between DSFMix time trend prediction and Monocle3. The high discrepancy in correlations demonstrate strong cell-to-cell and time-dependent variations introduced by the various underlying method assumptions. Figure 3F shows visualizations of underlying trajectories on projected data by all six methods related to the EMT–MET plasticity captured clearly by DSFMix as two separate trees (Figure 3E (iv)).

We use similarity network analysis based on a non-parametric Quadratic Assignment Procedure (QAP) correlation test [56] and the Hamming-Ipsen-Mikhailov (HIM) score [57], to benchmark DSFMix and recently published time dependent trajectory models, with the transition state validated network model during EMT derived from the study by Karacosta *et al.* [15]. We focus on the EMT data (emtdata) after TGFβ stimulation associated with baseline, 2, 6 and 10 days, respectively. By conditioning on time, we reduced the directed DSFMix EMT tree (Supplementary Figure S5B, see Supplementary Data available online at http://bib.oxfordjournals.org/) to 8 states (Figure 3F (ii)) for benchmarking with other methods. Figure 3F shows network comparisons of the EMT reference network (i) with DSFMix network (ii) and recently developed discrete temporal methods; CStreet (iii), Tempora(iv) and Scuba (v). *P*-values related to the correlation measures in Table (vi) indicate that all methods except for Tempora demonstrate a significant correlation structure at 0.05 level with the reference network and the highest correlation is associated with CStreet (0.47).

Furthermore, Supplementary Figure S6A and B (see Supplementary Data available online at http://bib.oxfordjournals.org/) show a better clustering of cells along the main tree axis captured by tree-based clustering (TAHC) algorithm used by DSFMix as compared to the popular Girvan-Newman algorithm on the MST derived from Yan *et al.*'s [21] scRNA-seq dataset. In addition, using meta-analysis we observe a greater cophenetic correlation of TAHC dendrogram intercluster relatedness associated with the UMAP global structure (0.6) than with *t*-SNE (0.5) (Supplementary Figures S6C and S7A, see Supplementary Data available online at http://bib.oxfordjournals.org/). A detailed DSFMix analysis on the labeled subset data covering the full EMT spectrum (Supplementary Figure S8A (i–iv), see Supplementary Data available online at http://bib.oxfordjournals.org/) also shows that DSFMix forest captures several independent and nested mixtures of full and partial EMT lineages which correlate strongly with canonical markers like vimentin.

DSFMix identification of a complex microenvironment of cell types associated with diverse nested differentiation trajectories during spermatogenesis

Spermatogenesis is an essential process for sexual reproduction and has several clinical implications. In this well-studied and characterized developmental process, haploid sperm cells, or spermatids, are produced from germ or spermatogonia cells in the seminiferous tubules of the testes. Recently, several scRNA-seq studies have sought to understand and characterize the heterogeneous lineages governing spermatogonia, spermatocyte, spermatid, and mature spermatozoa formation [43–46]. Germ cell development includes a mixture of discrete initial states followed by continuous trajectories [19]. Furthermore, because of the heterogeneous nature of the testes microenvironment, we expect somatic cells such as sertoli, myoid, leydig, and possibly immune cells to communicate with differentiating cells during spermatogenesis. Our goal is to use DSFMix to visualize this complex molecular dynamic microenvironment as well as to identify key genes associated with intermediate lineages during mouse spermatogenesis. Supplementary Figure S9 (see Supplementary Data available online at http://bib.oxfordjournals.org/) provides heatmaps of expression signatures of genes that are enriched across various stages when DSFMix is applied to spermdata. Gene encoding cell cycle proteins (Cdkn2a) and iron-transporting proteins (Ftl) are overexpressed during early spermatogonia differentiation (days 0–3) while isozyme-encoding genes such as lactate dehydrogenase (Ldhc) are associated with late spermatocyte stages (day 14). Prm encodes for major DNA–binding proteins during the haploid phase of spermatogenesis, and spermatid nuclear transition



Method	Corre lation	P_val. QAP	HIMscore
DSFMix	0.36	<0.001	0.28
Tempora	-0.02	0.32	0.33
CStreet	0.47	<0.001	0.19
SCUBA	0.21	<0.001	0.61

Figure 3. DSFMix Benchmarking analysis with respect to time and network similarity. **(A–C)** Heatmaps representing hierarchical clustering of the pairwise distance matrices derived from spearman correlations between pseudo-time models; tSpace (1), Monocle3(2), PAGA (5) and slingshot (6) observed time (3) and DSFMix (4) predicted time ordering for ipscdata, spermdata and emtdata respectively. Highest similarity between DSFMix (4) predicted time and the observed time trend (3) is observed. **(D)** Box plots representing correlation trends with estimates (top) between the observed time trend and pseudo-time or predicted time for each method applied to the spermdata. Large variations in terms of correlations are shown with the highest correlation (0.67) associated with DSFMix. **(E)** Visualizations of underlying EMT trajectories on projected data by all 6 methods. Evidence of EMT-MET plasticity in terms of 2 independent trees is captured clearly by DSFMix. In panel **E** (iv), the size of the nodes within each subtree is proportional to the number of cells in that node, whereas the length of the edges reflects the Euclidean distance of the median expression. **(F)** Correlation network analysis comparing recently developed discrete temporal methods: (iii) Cstreet, (iv) Tempora, (v) Scuba, and derived directed DSFMix network (ii) with reference EMT (i) network. Quantitative correlation measures in Table (vi) show that all methods except for Tempora demonstrate a significant correlation structure at 0.05 level with the reference using Quadratic Assignment Procedure (QAP), with CStreet capturing the highest correlation.

proteins (Tnp1) are overexpressed in later stages (>day 18). The feature selection step at FDR of 0.0001 resulted in 174 genes (Supplementary Table S1, see Supplementary Data available online at http://bib.oxfordjournals.org/). Supplementary Figure S8B (first row, see Supplementary Data available online at http://bib.oxfordjournals.org/) (ii) shows a SPADE MST colored by stages, with multiple progenitor lineages (blue). Figure 4A shows a spanning forest comprising six MSTs and portrays a clear progression of early (left, blue) to late (right, red) spermatogenesis. Tree (1) illustrates a complex interplay of spermatogenesis (left branch) and somatic cells (right branch). The heterogeneous branch (right) is associated with genes expressed in sertoli, leydig, myoid, interstitial, and immune lineages (see Figure 4B–D and Supplementary Table S2, see Supplementary Data available online at http://bib.oxfordjournals.org/, for the list of genes involved). Trees (2) and (3) represent spermatogonia and spermatocyte differentiation (Supplementary Tables S3 and S4, see Supplementary Data available online at http://bib.oxfordjournals.org/), respectively, while trees (4–6) show overrepresentations of differentiated spermatocytes and spermatids (Supplementary Tables S5–S7, see Supplementary Data available online at http://bib.oxfordjournals.org/). Note that cells (in blue) associated with the early spermatogenesis stage are more closely related in the DSFMix forest than in the SPADE tree.

DSFMix identification of lineage genes during spermatogenesis and quantification of pairwise similarities between forest lineages

To identify and rank significant genes associated with each dynamic forest tree, we formulate a dispersion test using double MAD instead of mean differences to control for asymmetry (**Supplementary Information section Methods**). Significant dynamic genes such as Apoe (1), Selenop (2), Ftl (3), Mlf1 (5), Rps5 (4), and Eef1a1 (6) are associated with spermatogenesis [58–63]. To test for similar or nested structures, we compare all pairwise forest trees using a non-parametric, two-sample, MST-based multivariate exact test [64] based on permutations. This test is robust to unbalanced network sizes and uses weighted edge counts to determine the statistics under the null (see Supplementary Information section Methods). Under the null, we expect that edges in the similarity graph are more likely to connect nodes from the same sample. We reject the null hypothesis of equal distribution in favor of the alternative if the number of between-sample edges is significantly less than what is expected under the null. Maa *et al.* [65] showed that the edge-count test, based on the MST derived from Euclidean distances, is consistent with all alternatives for multivariate data. We condition the test on a single MST derived from each tree pair. Figure 4E presents a heatmap that shows the pairwise edge-count statistics among all trees. Tree (1) is associated with the smallest values (blue), providing evidence of nested relationships with all other trees. Most subtrees are significantly different from each other at the 0.05 level. The similarity *P*-values can be used as distance measures to order subtrees within the forest environment for better visualization.

Validation of tissue specificity, pathway enrichment and identification of master regulators associated with spermatogenesis through correlation analysis on DSFMix signature genes

Next-generation sequencing correlation engines are powerful tools for examining gene expression relationship across species, studies, and technologies. In this work, we use the Illumina correlation engine managed by BaseSpace [66] to functionally validate our gene signatures during spermatogenesis and identify key transcriptional factors whose dysregulation may negatively affect normal spermatogenesis. Supplementary Figure S10 (see Supplementary Data available online at http://bib.oxfordjournals.org/) confirms that gene expression changes in testes have the strongest correlations with our work (*P*-value = 2.16e−36). Furthermore, expression of most signature genes in this study is associated with single-organism processes, cell differentiation, and the cellular processes involved in reproduction. Interestingly, mutation of the RNF17 gene, which encodes for proteins in the testes [67], is negatively correlated with the DSFMix gene signature, including the knockdown signature of ubiquitin, an essential protein found in all eukaryotic cells [68]. Further, an additional gene, MAMLD1, a causative gene for 46 XY disorders of sex development associated with abnormal development of the testes [69], is negatively correlated with the expression of our gene signature. These compelling findings demonstrate DSFMix's potential as a powerful hypothesis generator when combined with correlation engines.

DSFMix identifies nested and branching processes including switching gene signatures during intermediate and terminal embryonic stem cell reprogramming

Chemical reprogramming of one somatic cell to another is a powerful tool for studying the molecular heterogeneity of pluripotency. Accordingly, we subsequently applied DSFMix to the ipscdata. The results are summarized in Supplementary Information section 'DSFMix identifies complete, early, intermediate, and late dynamic lineages during induced embryonic chemical reprogramming'. We detected a switch-like gene signature program associated with intermediate (XEN) and terminal stem-cell formation (Figure 5A). Supplementary Figure S8B (first row, see Supplementary Data available online at http://bib.oxfordjournals.org/) (iii) identifies a global dynamic branching trajectory from MEFs (blue) to ESCs (red) with three branching lineages instead of two as captured by Monocle in Zhao *et al.* [47]. The additional branch (Figure 5B) is associated with the regulation of genes such as CrxOS which has been shown to maintain the self-renewal capacity of murine ESCs [70].

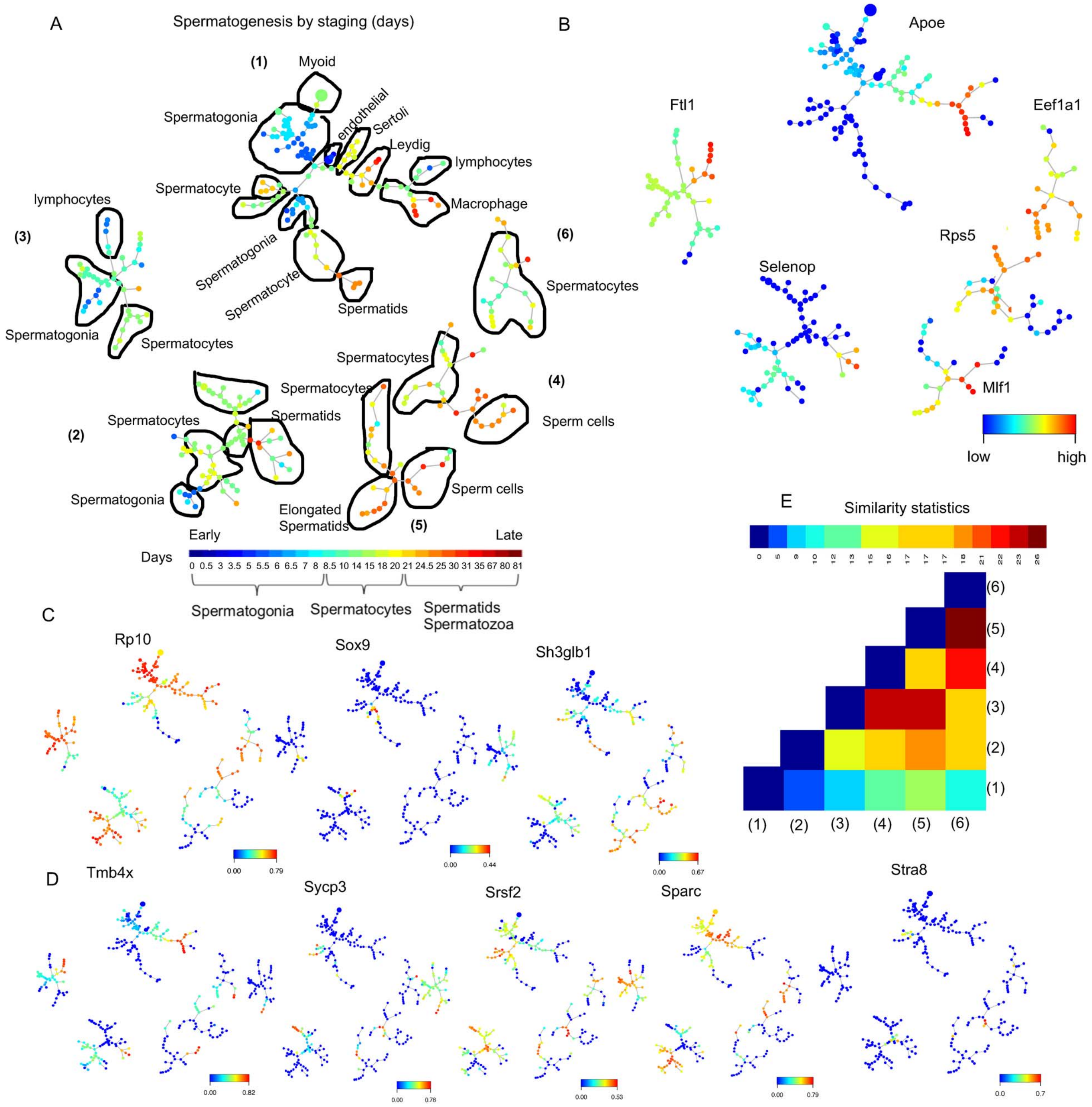


Figure 4. DSFMix Analysis on spermatogenesis. **(A)** Forest comprising six MSTs colored by median staging times spanning approximately 80 days. Cells undergoing early spermatogenesis are colored in blue while late sperm formation is colored in red. **(B)** DSF plots highlighting the most dynamic genes that span individual subtrees during spermatogenesis. **(C and D)** Forest plots highlighting key genes that are regulated during branching stages of spermatogenesis. **(E)** Heatmap of similarity test statistics between all pairwise trees (1–6) in **A** based on two-sample multivariate weighted edge count test. Blue represents identical trees, and red represents statistically significant dissimilar trees. Trees (2) and (3) have greater similarity with the major tree (1) compared to the rest. In all the figure panels, size of each node within each subtree is proportional to the number of cells in that node, whereas the length of the edges reflects the Euclidean distance of the mean expression between the 2 connected nodes.

Figure 5C shows a forest of stem cell pluripotency with seven trees (Supplementary Table S8, see Supplementary Data available online at http://bib.oxfordjournals.org/ for the complete list of genes in all trees by DSFMix). MEFs are mainly connected to tree (2) (Supplementary Table S10, see Supplementary Data available online at http://bib.oxfordjournals.org/ for the complete list of genes in tree 2), representing linear trajectories to ESCs. The switch-like signatures are associated with tree (5) (Supplementary Table S13, see Supplementary Data available online at http://bib.oxfordjournals.org/ for the complete list of genes in tree 5). Figures 5D–F shows tree-specific dynamic genes, including several transcription factors such as Sox17, Twist1, Pou5f1, and Sall4 (Supplementary Figure S11, see Supplementary Data available online at http://bib.oxfordjournals.org/, and Supplementary Tables S9–S15, see Supplementary Data available online at http://bib.oxfordjournals.org/, for the lists of genes selected by DSFMix per tree) during iPSC reprogramming. Figure 5G indicates a strong

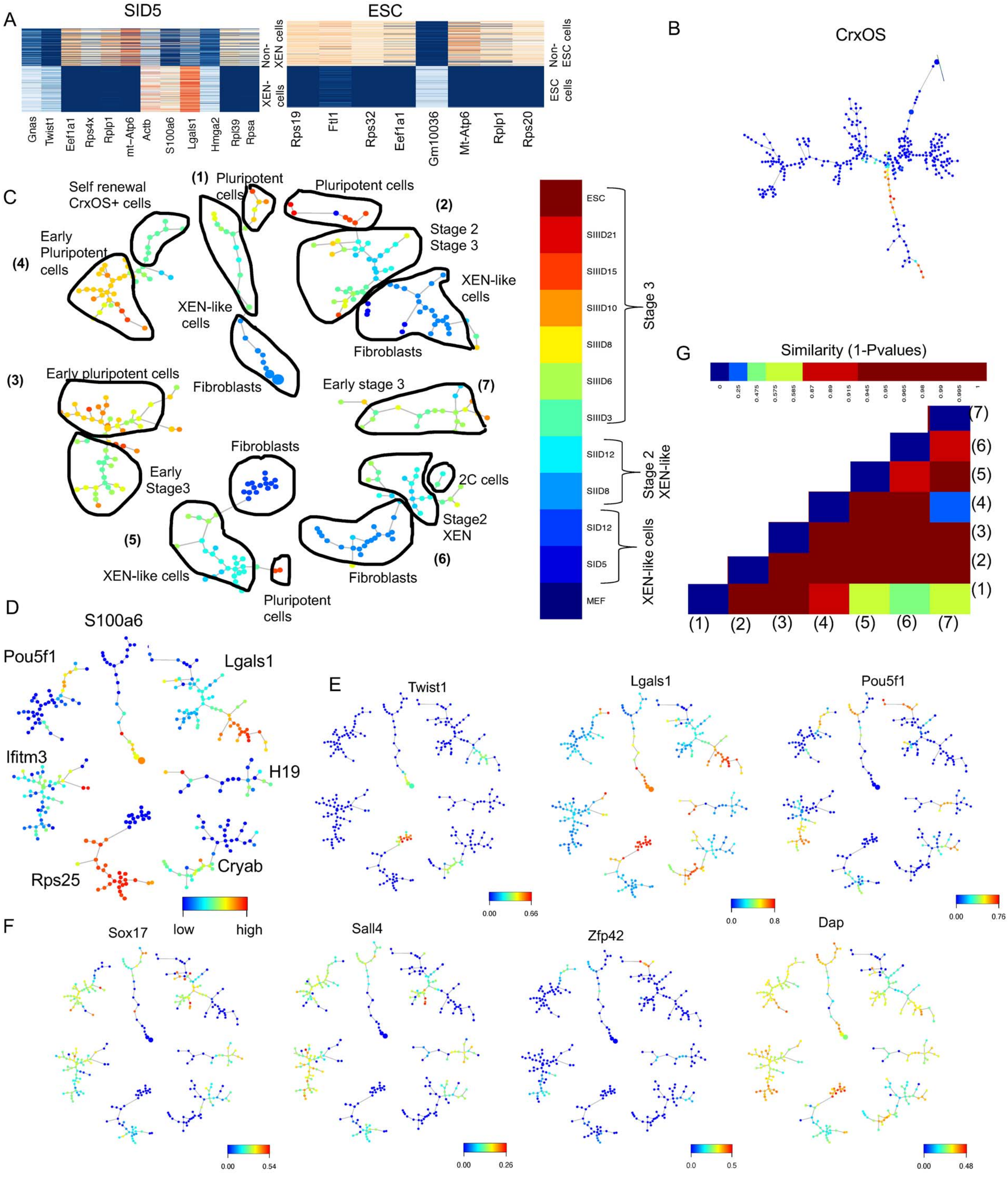


Figure 5. DSFMix analysis on chemically induced pluripotent stem cell (CiPSC) reprogramming. **(A)** Heatmaps of gene signatures showing switch-like programs associated with the transition of MEFs to intermediate extraembryonic endoderm (XEN-like) cells at day 5 as well as embryonic stem cell (ESC) formation. **(B)** SPADE single tree highlighting uniqueness of CrxOS expression in one of its three terminal branches during iPSC reprogramming. **(C)** DSFMix forest comprising seven trees colored by the 12 timepoints corresponding to 3 stages, spanning ~21 days. Cells from stage I after induction were collected at days 5 and 12; cells at stage II were collected at days 8 and 12; and cells at stage III were collected at days 3, 6, 8, 10, 15, and 21. DSF subtrees capturing several linear dynamic lineages spanning complete Fibroblast-XEN-ESC lineages (1,2,5), Fibroblasts including two-cell (2C) embryonic-like cells (6), intermediate XEN-like cells (7), and early pluripotency (3,4) lineages. **(D)** DSF plot highlighting significant dynamic genes spanning individual subtrees during iPSC. (**E and F)** DSF plots highlighting key genes whose expression changes dynamically over time during CiPSC. **(G)** Heatmap of similarity related *P*-values between tree pairs in **C**. Trees (2) and (3) are significantly different from the other trees while trees (4) and (7) show strong similarity.

similarity between trees (4) and (7) (Supplementary Tables S12 and S15, see Supplementary Data available online at http://bib.oxfordjournals.org/, for the lists of genes selected by DSFMix in tree 4 and 7 respectively).

DSFMix identifies a strong transcriptional response after 8 h as well as synergistic drug combinations with Dex treatment for prostate and breast cancers

Intratumor heterogeneity is a major cause of drug resistance. Increased glucocorticoid receptor activities have been observed in several metastatic cancers, including breast cancer [71]. In this study, we used DSFMix to capture the underlying dynamic transcriptional response of the T47D A1–2 breast cancer cell line to Dex treatment. A detailed analysis (**Supplementary Information section 'DSFMix identifies a dynamic gene signature associated with continuous hormonal transcriptional response with DEX' and Supplementary Figures S12 and S13,** see Supplementary Data available online at http://bib.oxfordjournals.org/) highlights time-dependent genes (Supplementary Figure S12, see Supplementary Data available online at http://bib.oxfordjournals.org/), DSF construction genes (Supplementary Figure S13A, see Supplementary Data available online at http://bib.oxford journals.org/), output forest with four trees (Supplementary S13, see Supplementary Data available online at http://bib.oxfordjournals.org/), and profiles of enzyme coding genes (e.g. DLG5, ACSL1, and HSD17B1) associated with transitions after 8 h (tree 3, Supplementary Figure S13D–G, see Supplementary Data available online at http://bib.oxfordjournals.org/). Supplementary Figure S13, see Supplementary Data available online at http://bib.oxfordjournals.org/, and Supplementary Tables S16–S19, see Supplementary Data available online at http://bib.oxfordjournals.org/, illustrate other dynamic tree-specific genes associated with hormone response. Similarity (*P*-values) between trees (Supplementary Figure S13H, see Supplementary Data available online at http://bib.oxfordjournals.org/) suggests overfitting at the 0.05 level, implying that combining cells from trees (2) and (4) into a single tree produces a more parsimonious forest. Hence, DSFMix can be used to order trees within a forest as well as minimize overfitting. Correlation analysis further confirms that changes in gene expression associated with daidzein, estradiol, and genisten (Supplementary Figure S14, see Supplementary Data available online at http://bib.oxfordjournals.org/) are strongly positively correlated with changes due to Dex treatment, indicating potential synergistic drug combinations at the single-cell level for prostate and breast cancers.

Discussion

Current knowledge of developmental biology is restricted by our ability to determine whether cell types and cell states during a given developmental process are continuous, discrete, or a mixture of the two. In this study, we address this major challenge and provide a computational model framework that uses temporal and stage single-cell data collected at discrete time points as input and a mixture of dependent trees, denoted here DSFMix, as output. DSFMix uses binary decision-tree models to select significant time-dependent markers associated with marginal distributions of multimodality and skewness. The selected markers are then used to connect all cells with an MST, and the tree is then broken up into a DSF. The trees of the forest are derived by combining tree-based clustering with a dynamic branch-cutting method based on the shape of the underlying dendrogram.

In this paper, we benchmark DFSMix using feature selection, correlation time course analysis and network similarity as well as highlight several advantages of our method as compared to pseudo-time single-graphical-driven approaches for visualizing and characterizing major cell types and states involved in developmental processes. Subsequently, we assess the impact of our feature selection step compared with traditional approaches, which are mainly focused on location-based variables and do not control for time or shape variation using curated single cell datasets. We identify time-dependent additional markers not included in Karacosta *et al.* [15] to build an EMT map. In addition, we show that DSFMix input genes exhibit a very high proportion of non-uniform, marginally distributed shapes that are mostly skewed to the right. We also show that multimodal genes are strongly associated with major steady states during development, a finding that challenges current downstream statistical methods that are optimized for averages and bimodality. Further, focusing on multimodal genes could potentially maximize cluster relatedness and data interpretability. Comparing the performance of TAHC on an MST with the GN algorithm reveals that traditional graphical clustering methods break branches into several equal-sized clusters and performs poorly when identifying clusters along the main axis of trees. On the other hand, TAHC tends to capture unique clusters along the main axis while efficiently clustering cells on branches. In addition, comparing a TAHC dendrogram with dendrograms from UMAP and *t*-SNE clustering reveals that inter-cluster relatedness in the TAHC tree has a stronger correlation with the UMAP global structure than *t*-SNE, although the correlation is weak (0.6). This may be problematic for clustering *t*-SNE and UMAP outputs because, for nearest neighbor algorithms, there is no guarantee that inter-cluster low-dimensional distances are correctly preserved in high-dimensional space. Most recently developed trajectory methods depend on KNN graphs derived from pooled time course data while time-dependent models like Waddington-OT or TRACER [15] use stochastic Markov processes to infer cellular progression as latent trajectories. Our EMT benchmarking analysis demonstrates the challenges of comparing latent and non-latent space temporal models without ground truth temporal data at the single cell level. Instead of using single scores we suggest the use of distributions of similarity network to account for structure uncertainty. The weighted edge count test is optimal for comparing MSTs and the QAP test is optimal for comparing directed networks. In addition, most temporal trajectory models do not provide optimal interpretable trajectories with granularity at the single-cell level. Note that by increasing the number of nodes, DSFMix can capture mixtures of both single-cell and cluster trajectories [36].

Finally, we demonstrate how DSFMix can be used on scRNA-seq data to visualize, compare, and characterize nested, intermediate, and complex dynamic features of spermatogenesis, induced iPSC reprogramming, and hormonal transcriptional response. We identify a complex microenvironment of germ and somatic cell types associated with diverse differentiation trajectories during spermatogenesis. A method like SPADE forces several asynchronous cell types into a single tree challenging the interpretability of early and intermediate lineages. Furthermore, SPADE could not resolve the nested structures that exist in temporal data. In contrast, using DSFMix, we were able to identify several dependent tree structures, dynamic genes during spermatogenesis and quantify the pairwise similarities between lineages during spermatogenesis. We were also able to identify switch-like signature programs of genes associated with the early transition of MEFs to intermediate XEN-like cells and during terminal ESC formation. DSFMix further confirmed that early hormonal transcriptional response from treatment of breast cancer cells with Dex produced no significant multimodal genes. This contrasts with the EMT data, which exhibited strong multimodality. In addition, major transcriptional changes occurred immediately after treatment and again after 8 h (Supplementary Figure S13B and C, see Supplementary Data available online at http://bib.oxfordjournals.org/). Quantitative similarity analysis of subtrees using a weighted edge-count test statistic suggests overfitting demonstrating that DSFMix can be combined with similarity testing to order trees within a forest as well as to reduce overfitting. DSFMix can also be applied in a supervised setting to investigate relationships between the components of larger biological systems such as the human body. For example, **Supplementary Information section Visualization of immune response due to coronavirus disease (COVID-19)** and Supplementary Figure S15 (see Supplementary Data available online at http://bib.oxfordjournals.org/) summarize a detailed DSFMix analysis on immune response to COVID-19. The patient-specific features associated with disease progression are clearly visualized as individual trees in the forest.

In summary, DSFMix is a powerful tool to integrate discrete cell lineages and continuous molecular data. It provides a novel framework and insights for visualizing, benchmarking, and investigating relationships between network motifs and underlying clusters in weighted dynamic networks. It is a powerful hypothesis-generating tool when combined with correlation search engines. This study motivates the need to develop better tree-related clustering algorithms to improve our mechanistic understanding of developmental biology at single-cell resolution. DSFMix is implemented using R available on GitHub (https://github.com/NIEHS/DSFMix).

Key Points

- Forest-based methods such as dynamic spanning forest mixtures (DSFMix) are optimal for modeling and visualizing complex heterogeneous processes and interactions including simultaneous processes.
- The marginal expression of genes during normal development exhibits a high proportion of non-uniformly distributed patterns that are mostly skewed to the right and multimodal, the latter being a characteristic of major steady states during development.
- Traditional graphical clustering methods are suboptimal when identifying clusters in data structures connected by different tree-like structures.
- Benchmarking with respect to time ordering and network similarity shows a higher correlation of DSFMix ordering with real time compared to pseudo-time trajectory approaches resulting in real time-dependent visualization and characterization of cell transitions.
- Several gene signatures at the single-cell level were identified and validated driving several complex dynamic processes such as epithelial–mesenchymal transition, spermatogenesis, stem cell pluripotency, and early transcriptional response from hormones and COVID-19.

Supplementary data

Supplementary data are available online at https://academic.oup.com/bib.

Data availability

We downloaded published raw scRNA-seq datasets for mouse spermatogenesis from Gene Expression Omnibus (GEO) under accession codes GSE121904 [44], GSE124904 [45] and GSE117707 [46] and from the ArrayExpress database under accession code EMTAB-6946 [43]. We also included unpublished WT postnatal scRNA-seq data at stage 8 with accession number GSE162642 in the analysis. We pooled raw sequencing datasets from 25 samples from different stages of spermatogenesis to generate a data of about 11,000 cells. We downloaded raw scRNA-seq datasets for mouse MEFs and chemically induced pluripotent stem cells from GEO under accession code GSE114952 [47] to generate a data of 50 000 individual cells from 12 time points throughout the reprogramming process. We also downloaded normalized scRNA-seq data used to study early transcriptional hormonal response from GEO under accession code GSE141834 [48]. We downloaded arcsinh-transformed CyTOF time-course data for the EMT analysis from [15] and COVID-19 immune response scRNA-seq normalized data from the COVID-19 Cell Atlas website portal (https://www.covid19cellatlas.org/). Raw sequencing data are also available in GEO under accession number GSE150728 [49].

Authors' contributions

B.A. contributed to the concept and algorithm of DSFMix. B.A. and S.K.P. were involved in the design of the static SPADE Forest concept. B.A. and R.M.G. performed the benchmarking analysis. B.A., A.A.M. and R.M.G. were involved in writing the initial manuscript. B.A., Q.C., X.X., J.L., T.K.A. and G.H. were involved in data generation and pre-processing. B.A., R.M.G, X.X., A.A.M. and J.L. performed sensitivity analyses. All authors were involved in interpreting the results.

Acknowledgements

This research was supported by the Intramural Research Program of the National Institutes of Health (NIH), National Institute of Environmental Health Sciences (NIEHS).

Funding

This work is supported by funding from the National Institute of Environmental Health Sciences (Grant/Award Number: 1ZIAES103350-01/02).

References

1. Trapnell C, Cacchiarelli D, Grimsby J, *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* 2014;**32**:381–6.
2. Cao J, Spielmann M, Qiu X, *et al.* The single-cell transcriptional landscape of mammalian organogenesis. *Nature* 2019;**566**: 496–502.
3. Moon KR, van Dijk D, Wang Z, *et al.* Visualizing structure and transitions in high-dimensional biological data. *Nat Biotechnol* 2019;**37**:1482–92.
4. Moon KR, van Dijk D, Wang Z, *et al.* Author correction: visualizing structure and transitions in high-dimensional biological data. *Nat Biotechnol* 2020;**38**:108.
5. Dermadi D, Bscheider M, Bjegovic K, *et al.* Exploration of cell development pathways through high-dimensional single cell analysis in trajectory space. *iScience* 2020;**23**:100842.
6. Bendall SC, Davis KL, Amir el AD, *et al.* Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell* 2014;**157**:714–25.
7. Wolf FA, Hamey FK, Plass M, *et al.* PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol* 2019;**20**:59.
8. Street K, Risso D, Fletcher RB, *et al.* Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* 2018;**19**:477.
9. Qiu P, Simonds EF, Bendall SC, *et al.* Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. *Nat Biotechnol* 2011;**29**:886–91.
10. Setty M, Kiseliovas V, Levine J, *et al.* Author correction: characterization of cell fate probabilities in single-cell data with Palantir. *Nat Biotechnol* 2019;**37**:1237.
11. Setty M, Kiseliovas V, Levine J, *et al.* Characterization of cell fate probabilities in single-cell data with Palantir. *Nat Biotechnol* 2019;**37**:451–60.
12. Lange M, Bergen V, Klein M, *et al.* CellRank for directed single-cell fate mapping. *Nature Methods* 2022. https://doi.org/10.1038/s41592-021-01346-6.
13. Marco E, Karp RL, Guo G, *et al.* Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. *Proc Natl Acad Sci U S A* 2014;**111**:E5643–50.
14. Bergen V, Lange M, Peidli S, *et al.* Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat Biotechnol* 2020;**38**:1408–14.
15. Karacosta LG, Anchang B, Ignatiadis N, *et al.* Mapping lung cancer epithelial-mesenchymal transition states and trajectories with single-cell resolution. *Nat Commun* 2019; **10**:5587.
16. Schiebinger G, Shu J, Tabaka M, *et al.* Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell* 2019;**176**:928–943.e22.
17. Tran TN, Bader GD. Tempora: cell trajectory inference using time-series single-cell RNA sequencing data. *PLoS Comput Biol* 2020;**16**:e1008205.
18. Zhao C, Xiu W, Hua Y, *et al.* CStreet: a computed cell state trajectory inference method for time-series single-cell RNA sequencing data. *Bioinformatics* 2021;**37**:3774–80.
19. Green CD, Ma Q, Manske GL, *et al.* A comprehensive roadmap of murine spermatogenesis defined by single-cell RNA-Seq. *Dev Cell* 2018;**46**:651–667.e10.
20. Bendall SC, Simonds EF, Qiu P, *et al.* Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science* 2011;**332**:687–96.
21. Yan KS, Gevaert O, Zheng GXY, *et al.* Intestinal enteroendocrine lineage cells possess homeostatic and injury-inducible stem cell activity. *Cell Stem Cell* 2017;**21**:78–90.e6.
22. Probst D, Reymond JL. Visualization of very large high-dimensional data sets as minimum spanning trees. *J Chem* 2020;**12**:12.
23. Brennecke P, Anders S, Kim JK, *et al.* Accounting for technical noise in single-cell RNA-seq experiments. *Nat Methods* 2013;**10**: 1093–5.
24. Butler A, Hoffman P, Smibert P, *et al.* Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* 2018;**36**:411–20.
25. Duo A, Robinson MD, Soneson C. A systematic performance evaluation of clustering methods for single-cell RNA-seq data. *F1000Res* 2018;**7**:1141.
26. Andrews TS, Hemberg M. M3Drop: dropout-based feature selection for scRNASeq. *Bioinformatics* 2019;**35**:2865–7.
27. Townes FW, Hicks SC, Aryee MJ, *et al.* Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. *Genome Biol* 2019;**20**:295.
28. Townes FW, Hicks SC, Aryee MJ, *et al.* Author correction: feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. *Genome Biol* 2020;**21**: 179.
29. Rousseeuw PJ, Croux C. Alternatives to the median absolute deviation. *J Am Stat Assoc* 1993;**88**:1273–83.
30. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics* 2001;**29**:1165–88 1124.

31. Hartigan JA, Hartigan PM. The dip test of Unimodality. *The Annals of Statistics* 1985;**13**:70–84.
32. Miao W, Gel YR, Gastwirth JL. A new test of symmetry about an unknown. In: *Random Walk, Sequential Analysis and Related Topics*. Fudan University, Shanghai, China: WORLD SCIENTIFIC, 2006, 199–214.
33. Devaney RLL. *A First Course In Chaotic Dynamical Systems: Theory And Experiment*. Chapman and Hall/CRC, 20201 online resource, 318.
34. Chu Y. On the shortest arborescence of a directed graph. *Sci Sin* 1965;**14**:1396–400.
35. Edmonds J. *J Res Natl Bur Stand Sec B* 1967;**71B**:233–40.
36. Anchang B, Hart TD, Bendall SC, *et al.* Visualization and cellular hierarchy inference of single-cell data using SPADE. *Nat Protoc* 2016;**11**:1264–79.
37. Yu M, Hillebrand A, Tewarie P, *et al.* Hierarchical clustering in minimum spanning trees. *Chaos* 2015;**25**:023107.
38. Newman ME, Girvan M. Finding and evaluating community structure in networks. *Phys Rev E Stat Nonlin Soft Matter Phys* 2004;**69**:026113.
39. Blondel VD, Guillaume J-L, Lambiotte R, *et al.* Fast unfolding of communities in large networks. *J Stat Mech Theory Exp* 2008;**2008**:P10008.
40. Langfelder P, Zhang B, Horvath S. Defining clusters from a hierarchical cluster tree: the dynamic tree cut package for R. *Bioinformatics* 2008;**24**:719–20.
41. Becht E, McInnes L, Healy J, *et al.* Dimensionality reduction for visualizing single-cell data using UMAP. *Nat Biotechnol* 2019;**37**: 38–44.
42. van der Maaten L, Hinton G. Viualizing data using t-SNE. *J Mach Learn Res* 2008;**9**:2579–605.
43. Ernst C, Eling N, Martinez-Jimenez CP, *et al.* Staged developmental mapping and X chromosome transcriptional dynamics during mouse spermatogenesis. *Nat Commun* 2019;**10**:1251.
44. Grive KJ, Hu Y, Shu E, *et al.* Dynamic transcriptome profiles within spermatogonial and spermatocyte populations during postnatal testis maturation revealed by single-cell sequencing. *PLoS Genet* 2019;**15**:e1007810.
45. Law NC, Oatley MJ, Oatley JM. Developmental kinetics and transcriptome dynamics of stem cell specification in the spermatogenic lineage. *Nat Commun* 2019;**10**:2787.
46. Wang Z, Xu X, Li JL, *et al.* Sertoli cell-only phenotype and scRNA-seq define PRAMEF12 as a factor essential for spermatogenesis in mice. *Nat Commun* 2019;**10**:5196.
47. Zhao T, Fu Y, Zhu J, *et al.* Single-cell RNA-Seq reveals dynamic early embryonic-like programs during chemical reprogramming. *Cell Stem Cell* 2018;**23**:31–45.e7.
48. Hoffman JA, Papas BN, Trotter KW, *et al.* Single-cell RNA sequencing reveals a heterogeneous response to glucocorticoids in breast cancer cells. *Commun Biol* 2020;**3**:126.
49. Wilk AJ, Rustagi A, Zhao NQ, *et al.* A single-cell atlas of the peripheral immune response in patients with severe COVID-19. *Nat Med* 2020;**26**:1070–6.
50. Lloyd S. Least squares quantization in PCM. *IEEE Trans Inf Theory* 1982;**28**:129–37.
51. Kotecha N, Krutzik PO, Irish JM. Web-based analysis and publication of flow cytometry experiments. *Curr Protoc Cytom* 2010;**53**:Unit10.17.
52. James G, Witten D, Hastie T, *et al. An Introduction to Statistical Learning: With Applications in R*. Vol. **103**. Springer Texts in Statistics. New York, NY: Springer New York, 2013.
53. Gagolewski M, Bartoszuk M, Cena A. Genie: a new, fast, and outlier-resistant hierarchical clustering algorithm. *Inform Sci* 2016;**363**:8–23.
54. Sakai R, Winand R, Verbeiren T, *et al.* Dendsort: modular leaf ordering methods for dendrogram representations in R. *F1000Res* 2014;**3**:177.
55. Kamada T, Kawai S. An algorithm for drawing general undirected graphs. *Inf Process Lett* 1989;**31**:7–15.
56. Krackardt D. QAP partialling as a test of spuriousness. *Social Networks* 1987;**9**:171–86.
57. Jurman G, Visintainer R, Filosi M *et al.* The HIM glocal metric and kernel for network comparison and classification. In: *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. 2015, pp. 1–10.
58. Hanissian SH, Akbar U, Teng B, *et al.* cDNA cloning and characterization of a novel gene encoding the MLF1-interacting protein MLF1IP. *Oncogene* 2004;**23**:3700–7.
59. Boitani C, Puglisi R. Selenium, a key element in spermatogenesis and male fertility. In: Cheng CY (ed). *Molecular Mechanisms in Spermatogenesis*. New York, NY: Springer New York, 2008, 65–73.
60. Leichtmann-Bardoogo Y, Cohen LA, Weiss A, *et al.* Compartmentalization and regulation of iron metabolism proteins protect male germ cells from iron overload. *Am J Physiol Endocrinol Metab* 2012;**302**:E1519–30.
61. Paoli D, Zedda S, Grassetti D, *et al.* Are apolipoprotein E alleles correlated with semen quality? *Int J Androl* 2012;**35**:714–9.
62. Chen J, Jiang D, Tan D, *et al.* Heterozygous mutation of eEF1A1b resulted in spermatogenesis arrest and infertility in male tilapia, *Oreochromis niloticus*. *Sci Rep* 2017;**7**:43733.
63. Kong J, Han H, Bergalet J, *et al.* A ribosomal protein S5 isoform is essential for oogenesis and interacts with distinct RNAs in Drosophila melanogaster. *Sci Rep* 2019;**9**:13779.
64. Chen H, Chen X, Su Y. A weighted edge-count two-sample test for multivariate and object data. *J Am Stat Assoc* 2018;**113**: 1146–55.
65. Maa J-F, Pearl DK, Bartoszyński R. Reducing multidimensional two-sample data to one-dimensional interpoint comparisons. *Ann Stat* 1996;**24**:1069–74 1066.
66. Alexa A. *BaseSpaceR: R SDK for BaseSpace RESTful API.*, 2020, R package version 1.34.30.
67. Pan J, Goodheart M, Chuma S, *et al.* RNF17, a component of the mammalian germ cell nuage, is essential for spermiogenesis. *Development* 2005;**132**:4029–39.
68. Bose R, Manku G, Culty M, *et al.* Ubiquitin-proteasome system in spermatogenesis. *Adv Exp Med Biol* 2014;**759**:181–213.
69. Miyado M, Yoshida K, Miyado K, *et al.* Knockout of murine Mamld1 impairs testicular growth and daily sperm production but permits normal postnatal androgen production and fertility. *Int J Mol Sci* 2017;**18**:1300.
70. Saito R, Yamasaki T, Nagai Y, *et al.* CrxOS maintains the self-renewal capacity of murine embryonic stem cells. *Biochem Biophys Res Commun* 2009;**390**:1129–35.
71. Lin KT, Wang LH. New dimension of glucocorticoids in cancer treatment. *Steroids* 2016;**111**:84–8.