frontiers in Bioengineering and Biotechnology

METHODS
published: 04 February 2020
doi: 10.3389/fbioe.2020.00018

CALISTA: Clustering and LINEAGE Inference in Single-Cell Transcriptional Analysis

Nan Papili Gao[1,2], *Thomas Hartmann*[1], *Tao Fang*[1] *and Rudiyanto Gunawan*[1,3]*

[1] Institute for Chemical and Bioengineering, ETH Zurich, Zurich, Switzerland, [2] Swiss Institute of Bioinformatics, Lausanne, Switzerland, [3] Department of Chemical and Biological Engineering, University at Buffalo, Buffalo, NY, United States

We present Clustering and Lineage Inference in Single-Cell Transcriptional Analysis (CALISTA), a numerically efficient and highly scalable toolbox for an end-to-end analysis of single-cell transcriptomic profiles. CALISTA includes four essential single-cell analyses for cell differentiation studies, including single-cell clustering, reconstruction of cell lineage specification, transition gene identification, and cell pseudotime ordering, which can be applied individually or in a pipeline. In these analyses, we employ a likelihood-based approach where single-cell mRNA counts are described by a probabilistic distribution function associated with stochastic gene transcriptional bursts and random technical dropout events. We illustrate the efficacy of CALISTA using single-cell gene expression datasets from different single-cell transcriptional profiling technologies and from a few hundreds to tens of thousands of cells. CALISTA is freely available on https://www.cabselab.com/calista.

Keywords: single-cell, gene expression, transcriptional bursts, random dropouts, cell clustering, lineage progression, pseudotime, cell differentiation

OPEN ACCESS

Edited by:
Xianwen Ren,
Peking University, China

Reviewed by:
Xiaojian Shao,
National Research Council Canada (NRC-CNRC), Canada
Yong Wang,
Academy of Mathematics and Systems Science (CAS), China

***Correspondence:**
Rudiyanto Gunawan
rgunawan@buffalo.edu

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *25 September 2019*
Accepted: *10 January 2020*
Published: *04 February 2020*

Citation:
Papili Gao N, Hartmann T, Fang T and Gunawan R (2020) CALISTA: Clustering and LINEAGE Inference in Single-Cell Transcriptional Analysis. Front. Bioeng. Biotechnol. 8:18. doi: 10.3389/fbioe.2020.00018

INTRODUCTION

The differentiation of stem cells into multiple cell types relies on the dynamic regulation of gene expression (Ralston and Shaw, 2008). In this regard, advances in single-cell gene transcriptional profiling technology have given a tremendous boost in elucidating the decision making process governing stem cell commitment to different cell fates (Kalisky et al., 2018). The applications of single-cell transcriptional analysis have led to new insights on the functional role of cell-to-cell gene expression heterogeneity in the physiological cell differentiation process (Guo et al., 2010; Kumar et al., 2014; Cacchiarelli et al., 2015; Reinius et al., 2016; Richard et al., 2016). Along with the surge in single-cell transcriptional profiling studies, algorithms for analyzing single-cell transcriptomics data have received increasing attention. In comparison to measurements from aggregate or bulk samples of cell population, single-cell gene expression profiles display much higher variability, not only due to technical reasons, but also because of the intrinsic stochastic (bursty) dynamics of the gene transcriptional process (Kærn et al., 2005). In particular, the stochastic gene transcription has been shown to generate highly non-Gaussian mRNA count distributions (Raj et al., 2006), which complicate data analysis using established methods that rely on a standard noise distribution model (e.g., Gaussian or Student's t-distribution).

Numerous algorithms have recently been developed specifically for the analysis of single-cell gene expression data. A class of these computational algorithms is geared toward identifying cell groups or clusters within a heterogeneous cell population. Traditional clustering algorithms

such as *k*-means and hierarchical clustering have been applied for such a purpose (Grün et al., 2015; Treutlein et al., 2016; Stumpf et al., 2017). Several other single-cell clustering strategies, such as CIDR (Lin et al., 2017), pcaREDUCE (Žurauskiene and Yau, 2016) and SNN-cliq (Xu and Su, 2015), adapt more advanced algorithms such as nearest neighbors search. Time-variant clustering strategies have also been implemented to elucidate the appearance of multiple cell lineages (Huang et al., 2014; Marco et al., 2014). In addition, consensus clustering methods, such as SC3 (Kiselev et al., 2017), have received much interest thanks to their superior stability and robustness. Finally, a likelihood-based method called Simulated Annealing for Bursty Expression Clustering (SABEC) (Ezer et al., 2016) employs a mechanistic model of the bursty stochastic dynamics of gene transcriptional process to cluster cells.

Another important class of algorithms deals with the reconstruction of lineage progression during cell differentiation process and the pseudotemporal ordering of single cells along the cell developmental path(s). The lineage progression describes the transition of stem cells through one or several developmental stages during the cell differentiation. This progression may comprise a single developmental path from the progenitor cells to one final cell fate, as well as bifurcating paths leading to multiple cell fates. In this class of algorithms, the reconstruction of the lineage progression and developmental paths is commonly implemented for the purpose of pseudotemporal cell ordering. The pseudotime of a cell represents the relative position of the cell along the developmental path and is typically normalized to be between 0 and 1. By plotting the gene expression against the pseudotime of the cells along a developmental path, one obtains a dynamic trajectory of the gene expression based on which the gene regulations driving the cell fate decision-making can be inferred. Numerous algorithms are available for single-cell transcriptional data analysis for cell lineage inference and cell ordering, notably DPT (Haghverdi et al., 2016), MONOCLE 2 (Trapnell et al., 2014; Qiu et al., 2017), and PAGA ((Wolf et al., 2019); for a more complete list, see a recent review by Cannoodt et al. (2016).

In this work, we developed Clustering and Lineage Inference in Single Cell Transcriptional Analysis (CALISTA), a numerically efficient and highly scalable toolbox for an end-to-end analysis of single-cell transcriptomics data. CALISTA is capable of and has been tested for analyzing datasets from major single-cell transcriptional profiling technologies, including scRT-qPCR and scRNA-sequencing with both plate-based (e.g., SMART-seq) and droplet-based platforms (scDrop-seq). CALISTA enables four essential analyses of single-cell transcriptomics in stem cell differentiation studies, namely single-cell clustering, reconstruction of lineage progression, transition gene identification and cell pseudotime ordering. In existing literature, these analyses are typically carried out by stringing several task-specific tools together in a bioinformatics pipeline. But, the basic assumptions behind different tools (e.g., regarding the distribution of data noise) maybe incompatible, an issue that has not been delved into more carefully in the literature. In contrast, the different analyses in CALISTA are fully compatible with each other as they are based on the same likelihood-based approach using probabilistic models of gene transcriptional bursts and random dropout events.

In the next section, we describe the algorithmic aspects and functionalities of CALISTA. The single-cell clustering of CALISTA is adapted from a previous method SABEC with a significant improvement in computational times, while the remaining CALISTA analyses represent novel contributions. For this reason, we focus the performance evaluation of CALISTA on lineage inference and cell pseudotime ordering, and compare CALISTA with widely-used bioinformatics packages including MONOCLE2 (Trapnell et al., 2014; Qiu et al., 2017) and SCANPY (Wolf et al., 2018). Subsequently, we illustrate CALISTA's end-to-end analysis using single-cell transcriptional profiles from the differentiation of human induced pluripotent stem cells (iPSCs) into mesodermal (M) or undesired endodermal (En) cells (Bargaje et al., 2017). Finally, we demonstrate the scalability of CALISTA in analyzing large datasets from scDrop-seq studies.

RESULTS

Single-Cell Transcriptional Analysis Using CALISTA

Figure 1 summarizes the four analyses of single-cell transcriptional profiles in CALISTA, including: (1) clustering of cells, (2) reconstruction of cell lineage progression, (3) identification of key transition genes, and (4) pseudotemporal ordering of cells. In CALISTA, we adopt a likelihood-based approach where the likelihood of a cell is computed using a probability distribution of mRNA defined according to the two-state model of gene transcriptional process (Peccoud and Ycart, 1995) and when appropriate, a random dropout event model (see section Methods). A random dropout occurs when mRNA molecules of a gene are not detected even though the true mRNA count is non-zero. The single-cell clustering in CALISTA is an adaptation of the algorithm SABEC (Ezer et al., 2016), where the single-cell clustering is carried out in two steps as illustrated in **Figure 1b**: (1) independent runs of maximum likelihood clustering, and (2) consensus clustering. SABEC has a high computational requirement due to the implementation of simulated annealing in the maximum likelihood step. This requirement prohibits the application of SABEC to large scRNA datasets with >10 K cells, such as from scDrop-Seq. CALISTA offers a substantial numerical speed-up over SABEC thanks to the implementation of a greedy algorithm and the reduction in the model parameter space (see **Supplementary Note S1** and **Supplementary Table S1**). CALISTA offers a parallel computing option which enables running the analysis over multiple computing cores for further speed-up.

The rest of the single-cell analyses in CALISTA represent new contributions of this work. For reconstructing cell lineage progression, we treat single-cell clusters as cell states through which stem cells transition during the cell differentiation. Here, CALISTA allows the calculation of distances between any pair of cell clusters. The cluster distance—defined as the maximum difference in the cumulative likelihood value upon reassigning the cells from the original cluster to the other cluster (see

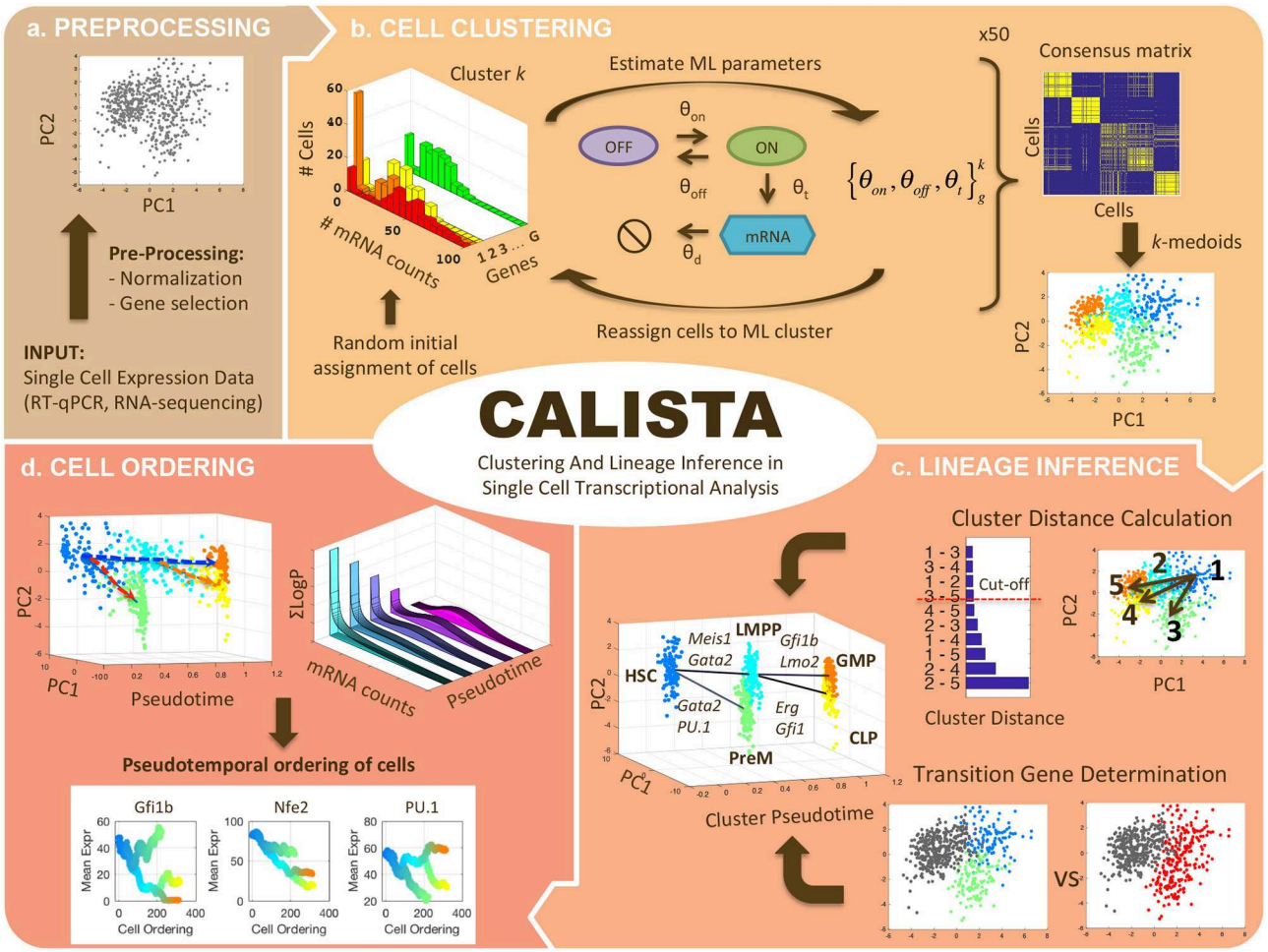

FIGURE 1 | CALISTA single-cell analysis workflow. **(a)** Single-cell expression input data are first pre-processed. **(b)** Single-cell clustering in CALISTA combines maximum likelihood and consensus clustering. In the maximum likelihood step, CALISTA relies on the two-state model of the gene transcription process in combination with a model of random dropouts to describe the distribution of single-cell gene expression. The maximum likelihood clustering is implemented multiple times, each starting from random cell assignment into clusters, using a greedy algorithm, the results from which are then used to generate a consensus matrix. The final step implements *k*-medoids clustering algorithm using the consensus matrix. **(c)** CALISTA uses cluster distances—a measure of dissimilarity in the gene expression distribution between any two clusters—to reconstruct the lineage progression graph. The state transition edges are added in increasing magnitude of cluster distances. In addition, CALISTA provides transition genes between any two connected clusters in the lineage graph based on the gene-wise likelihood differences between having the cells separately and having them in one cluster. **(d)** For pseudotemporal ordering of the cells, CALISTA reassigns each cell to a transition edge and computes its pseudotime by maximizing the cell likelihood. CALISTA uses a linear interpolation for evaluating the cell likelihood between two connected clusters. Given a developmental path in the lineage progression, CALISTA generates a moving averaged gene expression trajectory using the pseudotemporally ordered cells in the path.

section Methods)—gives a measure of dissimilarity in their gene expression distributions between any two clusters. CALISTA generates the lineage progression graph by sequentially adding state transition edges connecting closely distanced clusters until every cell cluster is connected to at least another cluster (**Figure 1c**). CALISTA also provides an interface for users to edit the lineage progression graph, i.e., adding or removing state transition edges, based on the cluster distances and other available information about the cell differentiation. For assigning directionality to the edges, CALISTA relies on user-provided information, for example information on the cell stage or sampling time, the starter/progenitor cells or the expected temporal profiles of the expression of marker genes. CALISTA also allows multiple initial clusters.

For any two connected clusters in the lineage progression, one can further use CALISTA to obtain the set of transition genes. The transition genes are determined based on the differences of the likelihood between having the cells in separate clusters and having them together in a single cluster (see section Methods). Here, the likelihood difference corresponding to a gene reflects the informative power of that gene for segregating cells into two clusters. The transition genes may point to candidate gene markers and genes regulating the state transition during differentiation. SABEC also allows

the determination of transition genes, but using a different strategy, called Estimation of Pairwise changes in Kinetics (EPiK), that is based on the statistical significance of the difference in the two-state model parameters between any two clusters.

The final component of CALISTA concerns with the pseudotemporal ordering of cells along a developmental path—defined as a sequence of connected clusters—in the lineage progression graph (**Figure 1d**). More specifically, given a developmental path in the reconstructed lineage progression, CALISTA produces a list of the cells ordered in increasing pseudotimes. For this purpose, we first assign a pseudotime to each cluster, which is normalized such that the starting cluster in the lineage progression graph has a pseudotime of 0 and the final cell cluster (or clusters) has a pseudotime of 1. Subsequently, we assign each cell to a transition edge that is pointing to or emanating from the cluster to which the cell belongs, again by adopting the maximum likelihood principle (see section Methods). Here, we assume that the distribution of the single-cell gene expression varies monotonically between cell states (clusters). For simplicity, the likelihood of a cell along a transition edge is computed using a linear interpolation of the cell likelihood values from the connected clusters. Each cell is then assigned to the transition edge that maximizes its likelihood value. Analogously, the cell pseudotime is computed by a linear interpolation of the cluster pseudotimes and set to the corresponding maximum point of the cell likelihood value.

Comparison of CALISTA Performance With Other Methods

We compared the performance of CALISTA with two widely-used single-cell bioinformatics packages for lineage inference and pseudotime cell ordering: MONOCLE 2 (Qiu et al., 2017) and SCANPY (Wolf et al., 2018). More specifically, in SCANPY package, we used Partition-based Graph Abstraction (PAGA) for lineage progression inference (Wolf et al., 2019) and Diffusion Pseudotime (DPT) for pseudotime cell ordering (Haghverdi et al., 2016).

In the first comparison, we generated *in silico* single-cell expression data of the cell differentiation of central nervous system (CNS) using a stochastic differential equation (SDE) model proposed by Qiu et al. (2012). We simulated single-cell data for 9 time points and 200 cells per time point, totaling 1,800 cells (see section Methods). As shown in **Figure 2A**, the simulated single-cell data clearly display two cell lineage bifurcations, as expected in this cell differentiation system (Qiu et al., 2012, 2018): (1) CNS precursors (pCNSs) differentiating into neurons and glia cells; (2) glia cells differentiating into astrocytes and oligodendrocytes (ODCs). **Figures 2B–D** show the reconstructed lineage progressions produced by MONOCLE 2, PAGA, and CALISTA, respectively. PAGA produced the most inaccurate lineage, deviating significantly from the expected lineage (**Figure 2C** vs. **Figure 2A**). MONOCLE 2 performed better than PAGA, producing a lineage progression that is in general agreement with the *in silico* lineage graph. But, looking at MONOCLE 2's lineage more carefully, the method identified many more bifurcation or branching points than expected (13 vs. 2). CALISTA outperformed both MONOCLE 2 and PAGA, generating a lineage progression that agrees very well with the *in silico* lineage.

Figures 2E,F depict the pseudotemporal cell ordering for the simulated CNS single-cell expression produced by MONOCLE2, DPT, and CALISTA, respectively. Besides visual comparisons of the pseudotemporal ordering, we also computed the correlations between the pseudotimes from each of the methods and the *in silico* times of the cells, i.e., the simulation times at which the single-cell mRNA data were sampled (see **Supplementary Table S2**). Among the three algorithms compared, CALISTA's pseudotimes have the highest correlation with the *in silico* cell times (correlation ρ of 0.856), followed by DPT ($\rho = 0.769$) and lastly MONOCLE 2 ($\rho = 0.571$). The cell orderings in **Figures 2E–G** further confirm the advantage of CALISTA over the other methods.

We further evaluated CALISTA's performance using four single-cell gene transcriptional datasets from cell differentiation systems with a variety of lineage topologies, including Bargaje et al. study on the differentiation of human induced pluripotent stem cells (iPSC) into cardiomyocytes (Bargaje et al., 2017), Chu et al. study on the differentiation of human embryonic stem cells (hESC) into endodermal cells (Chu et al., 2016), Moignard et al. study on hematopoietic stem cell (HSC) differentiation (Moignard et al., 2013), and Treutlein et al. study on mouse embryonic fibroblast (mEF) differentiation into neurons (Treutlein et al., 2016). We again compared CALISTA with the same single-cell analyses as in the *in silico* case study above. **Figures 3** summarizes the reconstructed lineage progression of the cell differentiation using MONOCLE 2, PAGA, and CALISTA. The cell differentiation in these cell systems follows the lineage progression drawn in **Figure 4A**. As in the *in silico* case study above, CALISTA generated the most accurate lineage progressions, followed by MONOCLE 2 and lastly PAGA. **Figures 4B–D** show the pseudotemporal ordering of cells produced by MONOCLE 2, DPT, and CALISTA, respectively. In assessing the accuracy of the pseudotimes, we relied on the known lineage progression and cell capture times, since the true cell differentiation times are not known in these datasets. All three methods performed equally well for the HSC differentiation dataset by Moignard et al. (2013). For iPSC dataset (Bargaje et al., 2017), CALISTA gave the most accurate pseudotimes, while MONOCLE2 and DPT had difficulties in assigning pseudotimes for one of the final cell type due to the close similarity of the gene expression with the progenitor iPSCs (see next section for more detail). For mEF dataset (Treutlein et al., 2016), CALISTA produced pseudotimes that are most consistent with the known lineage and capture times, followed by DPT and then MONOCLE 2. Finally, for hESC dataset (Chu et al., 2016), CALISTA again outperformed DPT and MONOCLE2, but here MONOCLE 2 performed better than DPT. In summary, for both simulated and real life single-cell gene expression datasets, CALISTA is able to reconstruct lineage progression and single-cell pseudotimes much better than widely-used single-cell gene expression analyses.

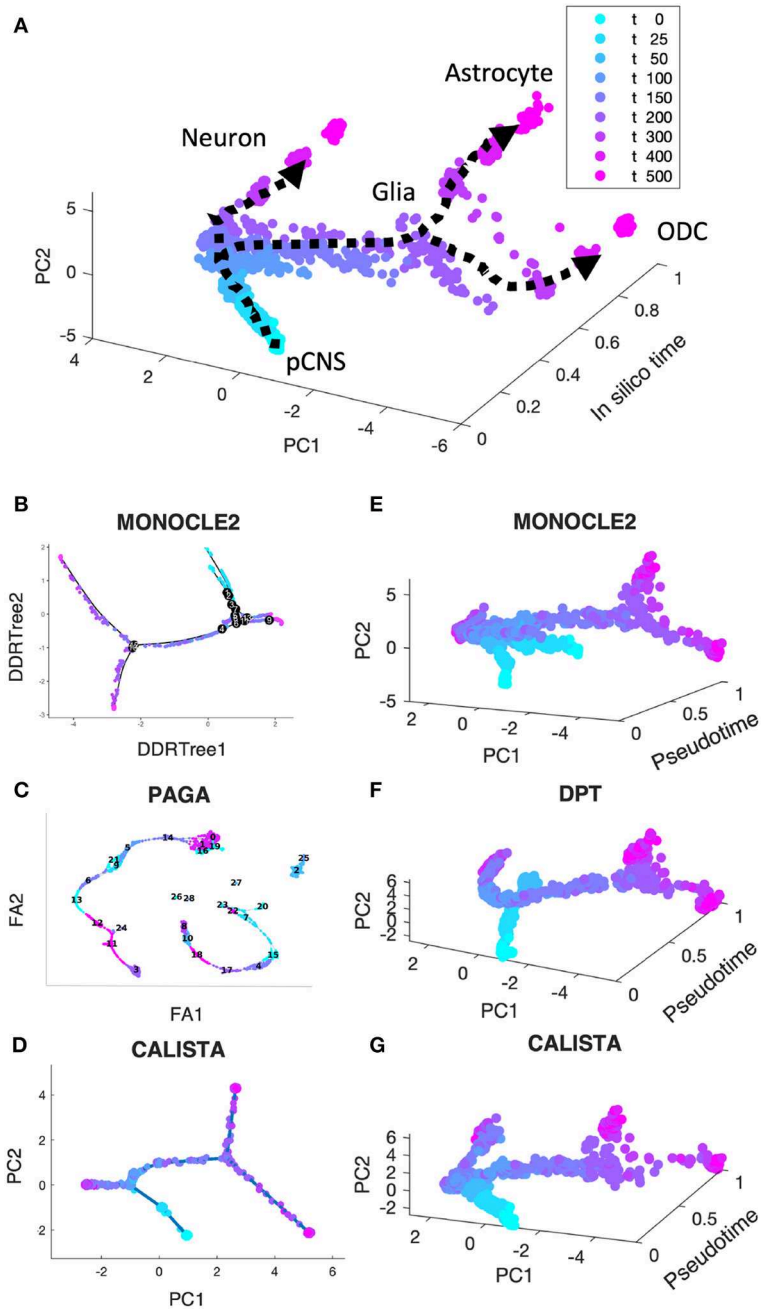

FIGURE 2 | Performance comparison of CALISTA, MONOCLE 2 and SCANPY (PAGA and DPT) using *in silico* single-cell gene expression data of cell differentiation in the central nervous system (CNS). **(A)** Single-cell gene expression data of CNS differentiation simulated using a model proposed by Qiu et al. show two branching/bifurcation points (Qiu et al., 2012): (1) Progenitor CNSs forming neurons and glia cells; (2) Glia cells forming astrocytes and oligodendrocytes (ODCs). **(B–D)** Reconstructed lineage progression by MONOCLE 2, PAGA (via SCANPY) and CALISTA, respectively. DDRTree: discriminative dimensionality reduction via learning tree (Mao et al., 2015), FA, ForceAtlas2 (Hua et al., 2018), PC: principal component. **(E–G)** Pseudotemporal ordering of cells by MONOCLE 2, DPT, and CALISTA, respectively.

Application to the Differentiation of Induced Pluripotent Stem Cells to Cardiomyocytes

In the following, we demonstrated an end-to-end analysis of single-cell gene expression data using CALISTA. Here, we used the single-cell gene expression dataset from the differentiation of human iPSCs into cardiomyocytes in Bargaje et al. study (Bargaje et al., 2017). The dataset includes single-cell expression of 96 genes measured by RT-qPCR for 1896 cells collected across 8 time points (day 0, 1, 1.5, 2, 2.5, 3, 4, 5) after induction to differentiate.

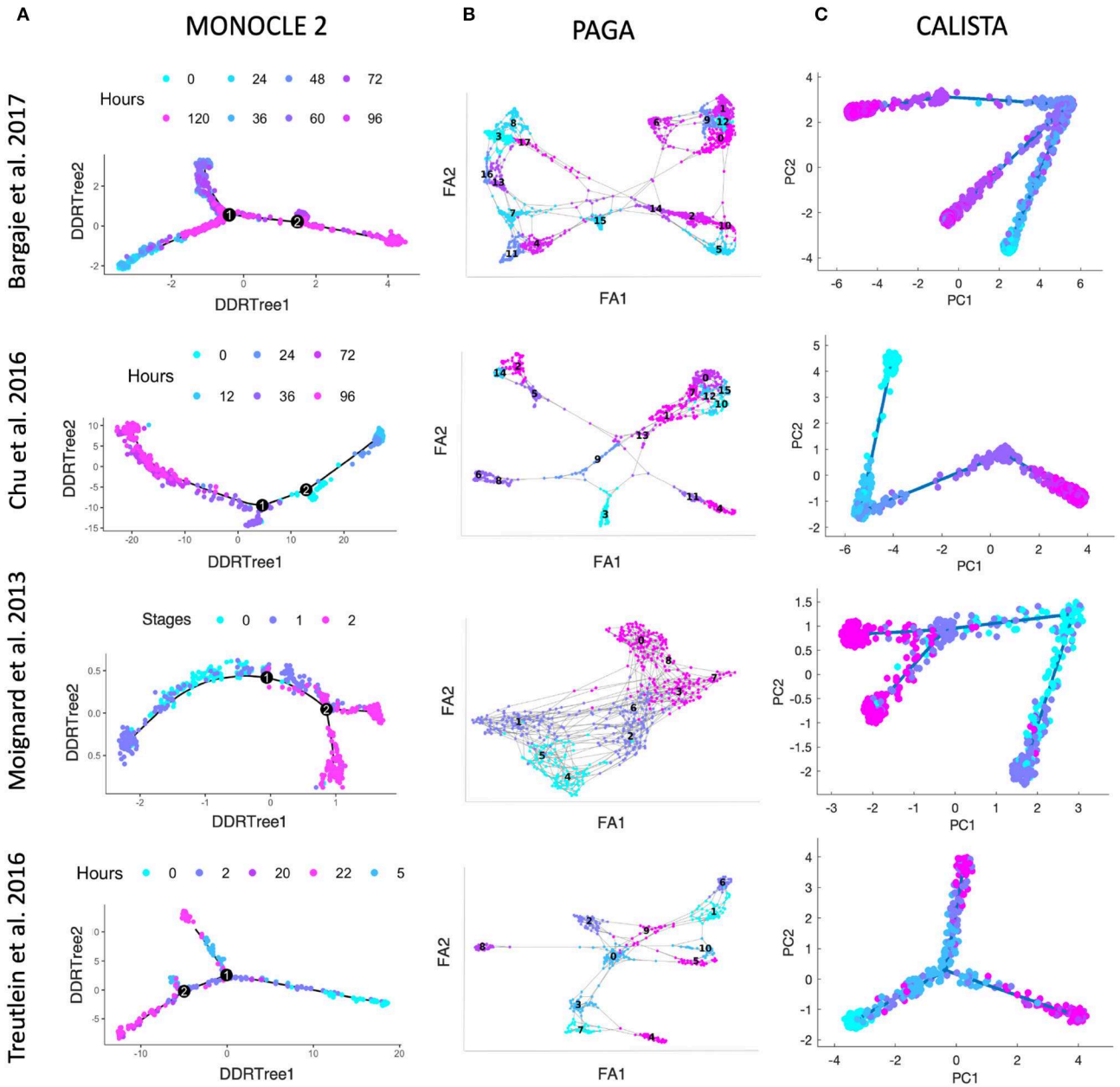

FIGURE 3 | Comparison of lineage progressions reconstructed from single-cell transcriptional profiles by MONOCLE 2, PAGA, and CALISTA. (Top row) Induced pluripotent stem cell (iPSC) differentiation into cardiomyocytes in Bargaje et al. study (Bargaje et al., 2017). (Second row) Human embryonic stem cell differentiation into endodermal cells in Chu et al. study (Chu et al., 2016). (Third row) Hematopoietic stem cell differentiation in Moignard et al. study (Moignard et al., 2013). (Bottom row) Mouse embryonic fibroblast differentiation into neurons in Treutlein et al. study (Treutlein et al., 2016). **(A)** MONOCLE 2. **(B)** PAGA. **(C)** CALISTA. DDRTree: discriminative dimensionality reduction via learning tree (Mao et al., 2015), FA, ForceAtlas2 (Hua et al., 2018), PC: principal component. The colors indicate the cell sampling times.

Figure 5A shows the developmental states identified in the original study: epiblast cells (E) in the early stage (day 0, 1, 1.5), primitive streak (PS)-like progenitor cells in the intermediate stage (day 2, 2.5), and a lineage bifurcation into either the desired mesodermal (M) or undesired endodermal (En) cell fate in the late stage (day 3, 4, 5).

First, we clustered the cells by using CALISTA. The optimal number of clusters was chosen to be five based on the eigengap plot (see **Supplementary Figure S1**). The single-cell clustering of CALISTA, as shown in **Figure 5B**, recapitulates the previously identified developmental states. Here, clusters 1, 2, and 5 contain mostly E, PS, and En cells, respectively, while M cells are split

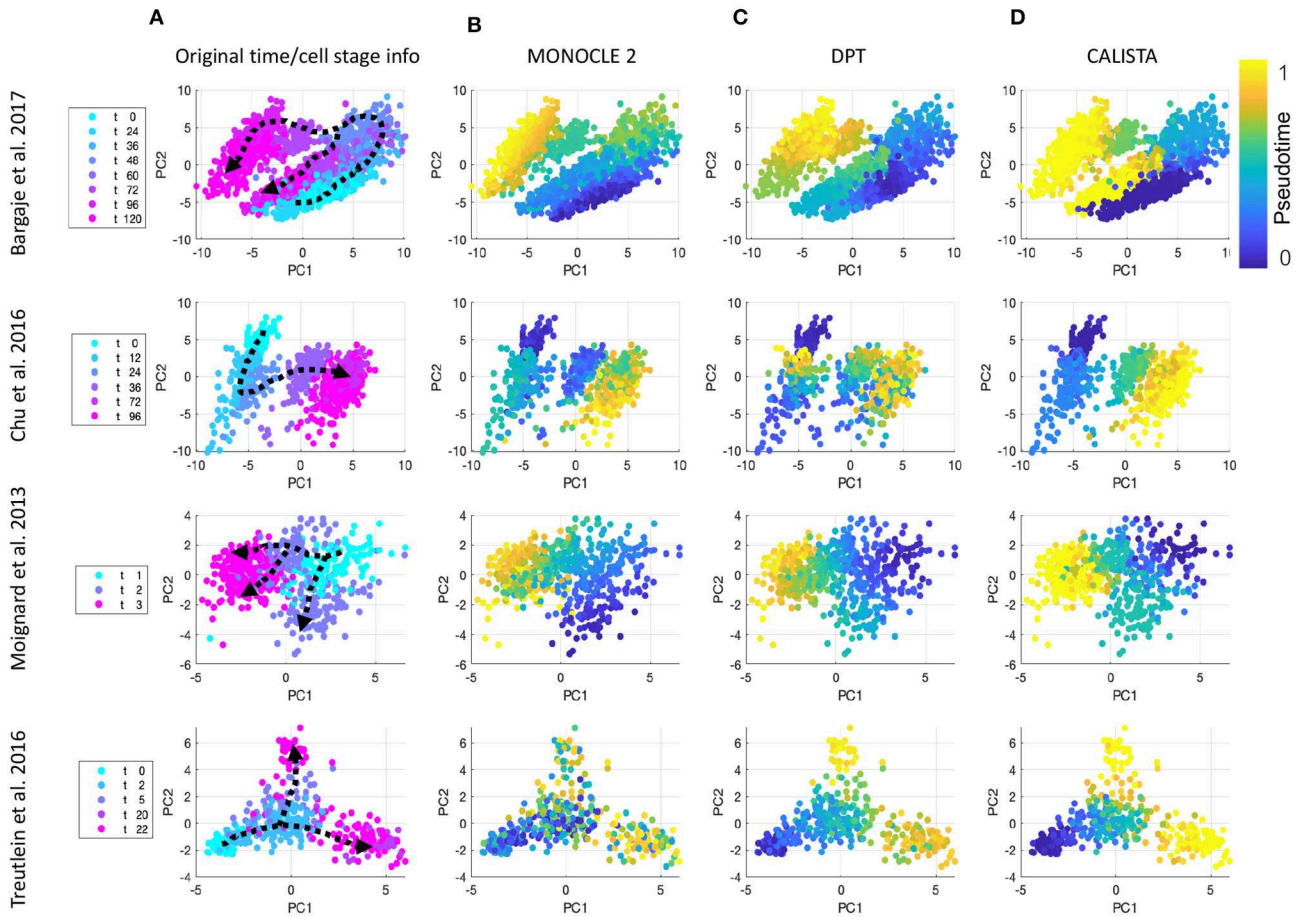

FIGURE 4 | Comparison of pseudotemporal cell ordering using single-cell transcriptional profiles by MONOCLE 2, DPT, and CALISTA. (Top row) Induced pluripotent stem cell (iPSC) differentiation into cardiomyocytes in Bargaje et al. study (Bargaje et al., 2017). (Second row) Human embryonic stem cell differentiation into endodermal cells in Chu et al. study (Chu et al., 2016). (Third row) Hematopoietic stem cell differentiation in Moignard et al. study (Moignard et al., 2013). (Bottom row) Mouse embryonic fibroblast differentiation into neurons in Treutlein et al. study (Treutlein et al., 2016). **(A)** Single-cell gene expression data. Pseudotemporal cell ordering from **(B)** MONOCLE 2, **(C)** DPT, and **(D)** CALISTA. PC: principal component. The colors in the first column indicate the cell sampling times, and those in the **(B–D)** indicate the pseudotimes.

between clusters 3 and 4 that are demarcated by different capture times (see cluster compositions in **Supplementary Figure S2**).

After single-cell clustering, we employed CALISTA to infer the lineage progression graph. The cluster pseudotimes were set to the modes (most frequent values) of the cell capture times in the clusters divided by the maximum cell capture time (see **Supplementary Figure S3**). The directionality of the state transition edges was set according to the cluster pseudotimes, pointing from a cluster with a lower pseudotime to that with a higher pseudotime. As shown in **Figure 5C**, the lineage progression graph reconstructed by CALISTA reproduces the lineage bifurcation event as the cells transition from PS-like cells to take on either M or En cell fates (Bargaje et al., 2017). Subsequently, for each state transition edge, we identified the set of transition genes (see **Supplementary Figure S4**). Among the identified transition genes in the lineage progression graph (25 genes in total, see **Supplementary Table S3**), many are known lineage specific transcriptional regulators involved in the iPSC cell differentiation, such as EOMES, GATA4, GSC, HAND1, KIT, MESP1, SOX17, and T (T-Box transcription factor T) (Bargaje et al., 2017).

Finally, we employed CALISTA to generate the pseudotemporal ordering of cells along the two distinct developmental paths in the lineage progression: (1) the M path forming mesodermal cells (cluster 1–2–3–4; see red dashed path in **Figure 5D**) and (2) the En path forming endodermal cells (cluster 1–2–5; see green dashed path in **Figure 5D**). After assigning cells to the state transition edges and prescribing the cell pseudotimes, we ordered cells belonging to each developmental trajectory in increasing pseudotimes with a total of 1,408 cells in the M path and 1,215 cells in the En path.

The likelihood value of each cell computed during the pseudotemporal cell ordering can further be visualized as a landscape plot. **Figure 5E** depicts the negative log-likelihood surface of the cells over the first two principal components. A higher value on the surface indicates a cell state with broader

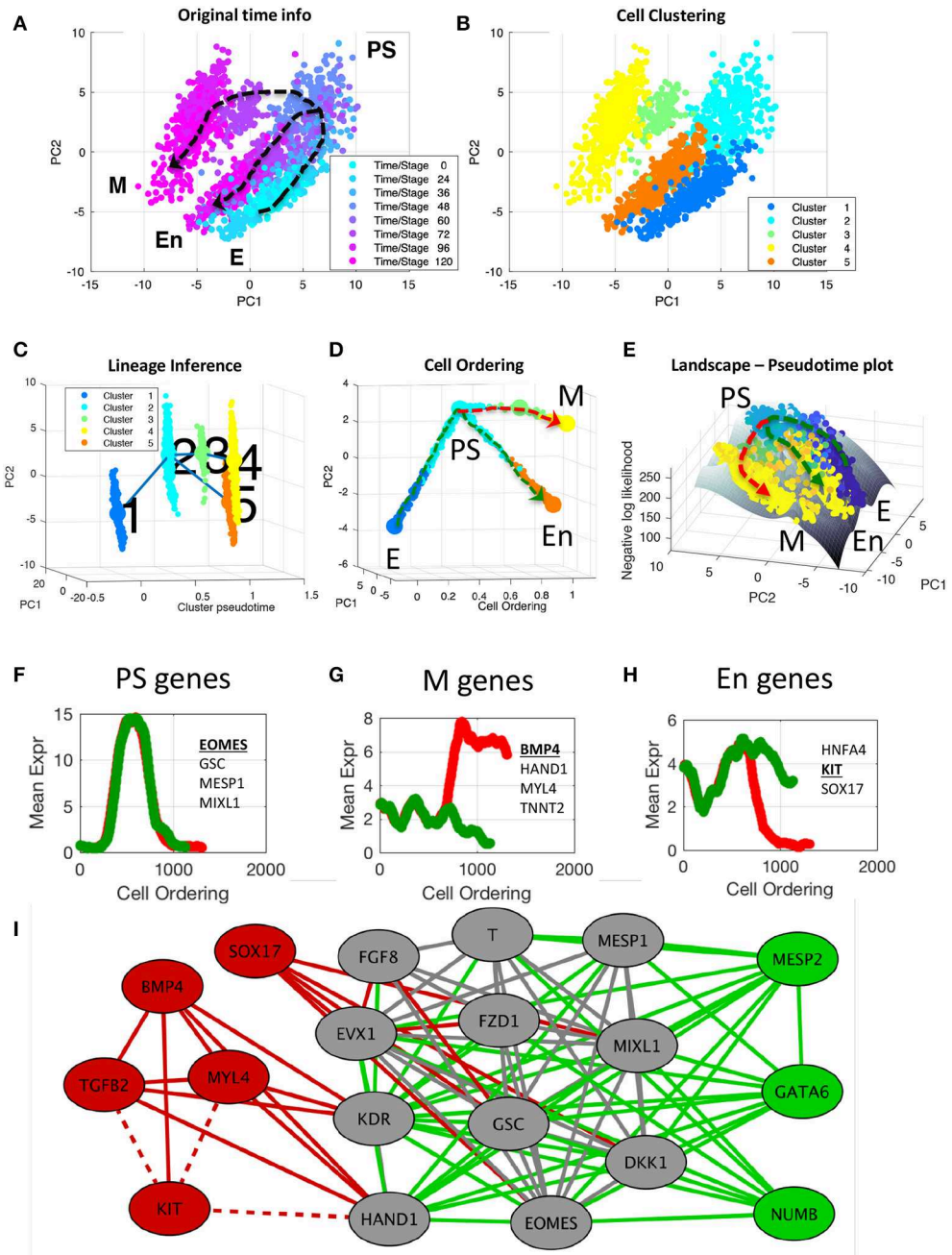

FIGURE 5 | End-to-end analysis of single-cell transcriptional profiles during iPSC differentiation into cardiomyocytes. The single-cell gene expression dataset was taken from the study of Bargaje et al. (2017). **(A,B)** PCA plots of single-cell dataset along the first two principal components. Each dot represents a cell, where the colors indicate **(A)** the capture time info and **(B)** CALISTA single-cell clusters. **(C)** Lineage progression graph reconstructed by CALISTA. The graph was generated by adding transition edges in increasing order of cluster distances until each cluster is connected by an edge. Following this procedure, an edge connecting cluster 1 and 5 was originally added in the graph. We manually removed this edge as the edge bypassed intermediate capture times. **(D)** Pseudotemporal ordering of cells. Cells were first assigned to the state transition edges in the lineage progression, and then pseudotemporally ordered along two developmental paths—mesodermal (M) path (red) and endodermal (En) path (green). **(E)** Cell likelihood landscape. The landscape shows the surface of the negative log-likelihood of the cells. A higher value thus indicates a state of increased gene expression uncertainty. **(F–H)** Moving window average of gene expression along M and En developmental paths for representative **(G)** PS, **(H)** M, and **(I)** En genes (underlined and boldfaced). **(I)** Gene modules of M (red and gray) and En paths (green and gray) [drawn using Cytoscape (Smoot et al., 2011)]. The gray nodes and edges are common between the M and En modules. Solid (dashed) lines indicate positive (negative) gene correlations.

mRNA distributions, i.e., a state of higher uncertainty in the gene expression. As shown in **Figure 5E**, iPSC cells start their journey from a valley in this surface, implying that the progenitor cells are at a low uncertainty state. As the cell differentiation progresses, cells pass through an intermediate state with higher uncertainty, where a peak uncertainty is reached at or around

the cell lineage bifurcation. After the bifurcation, cells follow two paths toward lower uncertainty, leading to two valleys corresponding to distinct cell fates (M and En fates). The rise-and-fall in gene expression uncertainty have also been reported in other cell differentiation systems, suggesting that stem cells go through a transition state of high uncertainty before committing to their final cell fate(s) (Richard et al., 2016; Stumpf et al., 2017).

To visualize the gene expression trajectories along the two cell differentiation paths, we calculated the moving average expression values of transition genes for the pseudotemporally ordered cells using a moving window comprising 10% of the total cells in each path (see **Figures 5F–H** and **Supplementary Figure S5**). A number of transition genes follow highly similar expression trajectories along the M and En paths, with an increase in expression from E to PS-like state, followed by a decrease in expression from PS-like to M or En state (see **Figure 5F**). The majority of genes with the aforementioned trajectory are known PS-like markers [for example EOMES, GSC, MESP1, and MIXL1; (Ng et al., 2005; Bargaje et al., 2017; Tiyaboonchai et al., 2017)], and thus, we refer to these genes as PS-genes (see **Supplementary Table S3**). Another group of transition genes show differential profiles between the M and En paths after the lineage bifurcation. Here, we define M-genes as the genes with higher expressions along the M path than the En path (see **Figure 5G**). Correspondingly, we define En-genes as genes with higher expressions along the En path than the M path (see **Figure 5H**). Notably, many of the known M marker genes [e.g., BMP4, HAND1, MYL4 and TNNT2; (Jagtap et al., 2011; Bargaje et al., 2017)] are among the M-genes, and several of the known En marker genes [e.g., HNFA4, KIT, and SOX17; (Thomas et al., 2006; Ng et al., 2010; Bargaje et al., 2017)] are among the En-genes (see **Supplementary Table S3**). In general, after the lineage bifurcation, the expressions of M genes increase along the M path, but are either suppressed (BMP4, HAND1, KDR) or unchanged (MYL4, TGFB2, TNNT2) along the En path. Among the known En markers, the expression of HNFA4 increases along the En path after the lineage bifurcation but remains relatively unchanged along the M path. The expression of KIT follows the opposite profile, where the gene expression is downregulated along the M path and is relatively unchanged along the En path. While the expression of SOX17, like KIT, decreases along the M path, the gene is upregulated along the En path immediately after the lineage bifurcation before being downregulated toward the end of the developmental trajectory.

Finally, we constructed gene co-expression networks for the M and En developmental paths based on the pseudotemporal profiles of the gene expression (pairwise Pearson correlation, $p \leq 0.01$ and correlation value ≥ 0.8, see **Supplementary Figure S6**). We identified cliques in the M and En gene co-expression networks, i.e., a subset of genes (at least 5) that are connected to each other, using a maximal clique analysis by the Bron-Kerbosch algorithm (Bron and Kerbosch, 1973). **Figure 5I** depicts the cliques from the M and En gene co-expression networks, showing three gene regulatory modules: one module specific to the M path (red), another specific to the En path (green), and a shared module (gray). Most of the PS genes, including known PS marker genes, belong to the shared module as expected. Among the genes in the shared module, DKK1, FZD1, and T are involved in Wnt signaling pathway, which is known to promote cell differentiation (Davidson et al., 2012). GATA6, which plays an important role in the endoderm commitment (Tiyaboonchai et al., 2017), belongs to the En module. On the other hand, the M module shows activating relationships among mesoderm gene markers (e.g., BMP4, MYL4, and HAND1), and antagonistic relationships between several M genes and KIT, an En gene marker.

Application of CALISTA to Massively Parallel Drop-Seq Datasets

To demonstrate the scalability of CALISTA, we analyzed single-cell expression datasets from droplet-based assays. Single-cell Drop-seq is a massively parallel genome-wide expression profiling technology capable of analyzing thousands of cells in a single experiment. However, the bioinformatic analysis of large single-cell transcriptomics datasets poses a significant computational challenge (Angerer et al., 2017). To address this challenge, CALISTA offers a parallelization option for handling large datasets by splitting the greedy optimization runs among multiple computing cores (see **Supplementary Note S2** for more details on CALISTA implementation).

We first tested the single-cell clustering performance of CALISTA in analyzing Drop-seq datasets using the single-cell study of mouse spinal cord neurons by Sathyamurthy et al. (2018) (∼18 K nuclei) and the study of peripherical blood mononuclear cells by Zheng et al. (2017) (∼68 K cells). We noted that *k*-medoid clustering for large consensus matrices is computationally prohibitive. Thus, we bypassed the consensus clustering and took among the independent runs of the greedy algorithm, the cell clustering corresponding to the highest likelihood value. For Sathyamurthy et al. dataset, CALISTA clustering identified nine single-cell clusters based on the eigengap plot, which agrees well with the original study [by comparing **Supplementary Figure S7** with **Figure 1D** in the original publication; (Sathyamurthy et al., 2018)]. For the largest cluster (52% of total data) containing neurons, CALISTA further split the cells into two subpopulations—one comprising cells with high expressions of Snap25, Syp, and Rbiox3, and the other comprising cells with medium expression of these genes. Notably, CALISTA was able to distinguish clearly Schwann and Meningeal cells, which in the original study [using SC3; (Kiselev et al., 2017)], were placed into the same cluster (Sathyamurthy et al., 2018). Meanwhile, for Zheng et al. scDrop-seq dataset, CALISTA generated cell clusters that show a general agreement with the original study (see **Supplementary Figure S8**).

We then tested CALISTA's lineage progression reconstruction on scDrop-seq data of ∼38K cells, taken from 12 developmental stages of zebrafish embryogenesis (Farrell et al., 2018). The results of CALISTA are summarized in **Figure 6**. For this large dataset, we employed a modified clustering procedure (see **Supplementary Note S2**) and identified 82 clusters, 23 of which comprise cells at the final developmental stage. **Figure 6A** depicts the reconstructed lineage progression by CALISTA (see also **Supplementary Video 1**), while **Figure 6B** shows the cell type labels for the cell clusters at the final developmental

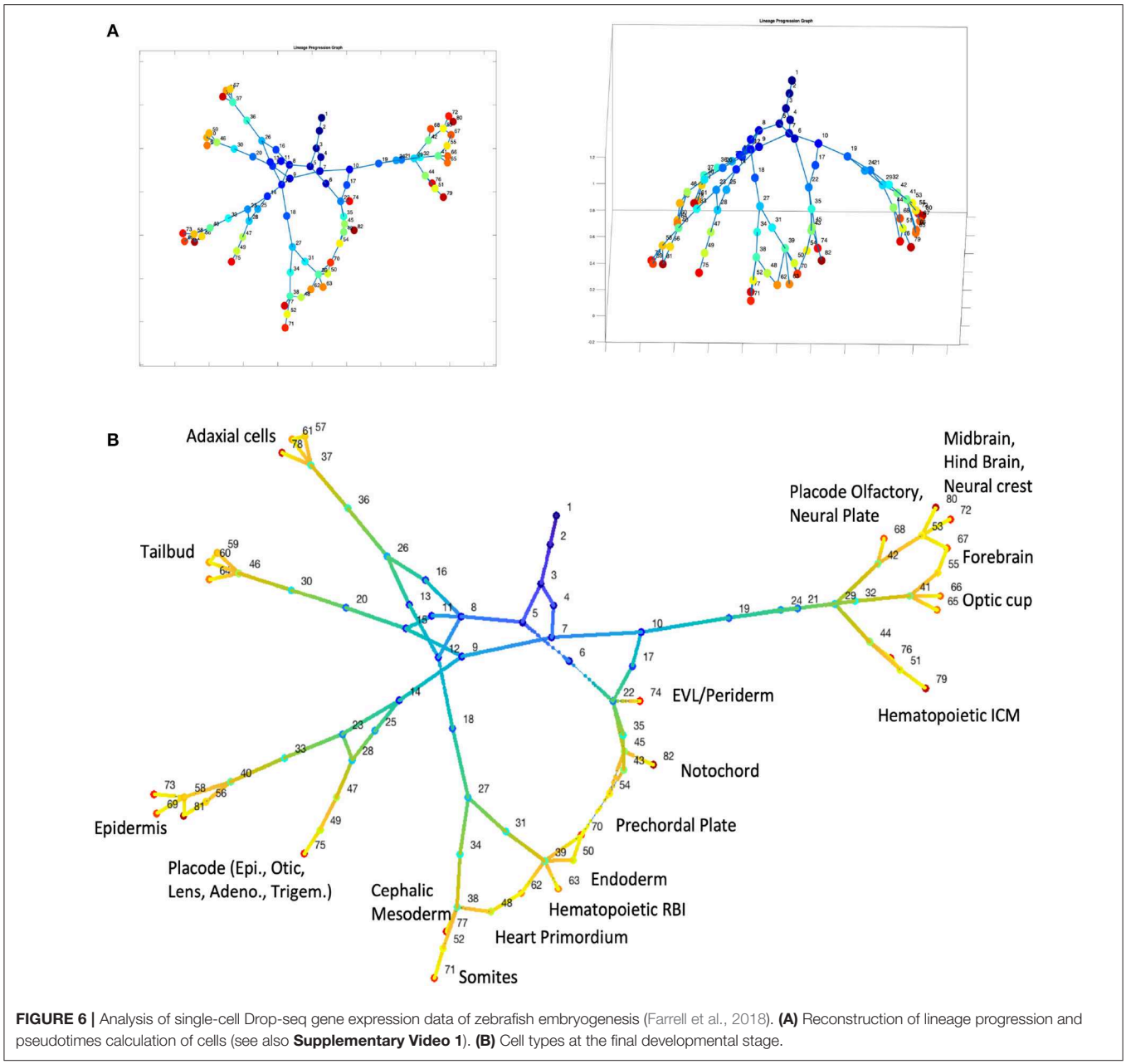

FIGURE 6 | Analysis of single-cell Drop-seq gene expression data of zebrafish embryogenesis (Farrell et al., 2018). **(A)** Reconstruction of lineage progression and pseudotimes calculation of cells (see also **Supplementary Video 1**). **(B)** Cell types at the final developmental stage.

stage based on the expression levels of 26 key gene markers (see **Supplementary Figure S9**). The results are again in good agreement with the known cell types (Farrell et al., 2018).

DISCUSSION

CALISTA provides four bioinformatic analyses for single-cell expression data that are essential in studies of stem cell differentiation. The analyses can be implemented either sequentially in a data analytics pipeline or separately in a standalone application. Throughout the development of the analyses in CALISTA, we used the same likelihood approach based on the two-state stochastic gene transcriptional model. Thus, the analyses are fully compatible with each other. The use of a mechanistic model in CALISTA brings an additional advantage, because the model parameters have relevance to the mechanism of gene transcription process and thus can provide insights into the gene regulations during stem cell differentiation. To the best of our knowledge, CALISTA is the first mechanistic model-based toolbox that allows an end-to-end analysis of single-cell transcriptional profiles. Despite the focus on stem cells in our work, each of the tools in CALISTA is agnostic to the source of the single-cell transcriptomic data analysis and thus can be used for other studies.

CALISTA's single-cell clustering analysis is an adaptation of another method called SABEC (Ezer et al., 2016), with a

significant improvement in computational efficiency, thanks to the implementation of greedy algorithm. This improvement is critical and necessary for the analysis of extremely large single-cell datasets with tens to hundreds of thousands of cells, which will become more common in the near future. In addition to computational efficiency, CALISTA is able to account for random technical dropout events, a feature that is missing from SABEC. With the exception of the single-cell clustering analysis, CALISTA includes novel algorithms for transition gene identification, lineage progression inference and pseudotime calculations. As demonstrated in numerous case studies above, when compared to popular single-cell analysis methods for lineage inference and pseudotemporal cell ordering, including MONOCLE 2, PAGA, and DPT, CALISTA performed better than these methods.

30696980example in setting the number of cell clusters or during the curation of the lineage progression graph (see CALISTA tutorials on https://www.cabselab.com/calista). The user interface enables incorporating existing biological knowledge of the cell differentiation system, which is often difficult—if not impossible—to codify. Although such prior knowledge is not necessary for using CALISTA, the ability to incorporate this knowledge, whenever available, is a useful and important feature in the analysis of single-cell transcriptional profiles.

Obviously, the transcriptome only provides one aspect of the cell phenotype. Besides single-cell transcriptomic profiling, single-cell sequencing technology have enabled other types of molecular profiling at the single cell level, such as the genome and epigenome. Integrating gene expression with other single-cell profiling data is an area of active interest for algorithmic development (Stuart and Satija, 2019). A few notable examples of recent algorithms include those for the integration of scRNA-seq data and single-cell data of DNA accessibility from ATAC-seq (Duren et al., 2018; Zeng et al., 2019).

METHODS

Input Data and Data Preprocessing

The single-cell gene expression matrix should be formatted into a $N \times G$ matrix, where G denotes the number of genes, N denotes the number of cells, and the matrix element $m_{n,g}$ is the transcriptional expression value of gene g in the n-th cell. CALISTA accepts expression values from RT-qPCR (2^{C_t} value) and scRNA-Seq including both plate-based [e.g., log(RPKM) or log(TPM)] and droplet-based measurements (e.g., gene UMI counts). For UMI data from scDrop-seq, CALISTA further scales the expression matrix by dividing each gene UMI count with the total UMI count in the corresponding cell and then multiplying the value with the median of the total UMI counts among cells (Zheng et al., 2017).

Before performing single-cell analysis, we first preprocess the single-cell expression matrix by removing the genes and cells (i.e., columns and rows of the expression matrix, respectively) with a large fraction of zero expression values, exceeding a user-defined threshold (default threshold: 100% for genes and 100% for cells). For scRNA-seq datasets, CALISTA further selects a number of informative genes Y for the single-cell analysis following a previously described procedure (Macosko et al., 2015), with Y is to the minimum among the following: 200, the number of genes G, and a user-defined percentage of the number of cells N.

Next, CALISTA scales the single-cell expression values such that the maximum value of any gene is 200. The scaling is carried out as follows:

$$\widehat{m}_{n,g} = 2^{\frac{log_2(m_{n,g}+1)}{log_2(m_{max,g})} log_2(200)} \quad (1)$$

where $m_{n,g}$ and $\widehat{m}_{n,g}$ represent the original and scaled expression value of gene g in the n-th cell respectively, and $m_{max,g}$ is the maximum expression value in gene g (i.e., the maximum of $m_{n,g}$ over all cells). The scaling above allows CALISTA to use a pre-computed table for the maximum likelihood in the clustering analysis, thereby reducing the computational cost significantly. We tested using *in silico* single-cell gene expression datasets generated using the two-state gene transcriptional model, and confirmed that the scaling does not affect the clustering accuracy (see **Supplementary Note S3**). After scaling, for scDrop-Seq data, we re-rank the top Y genes in increasing value of the gene-wise likelihood v^g computed following Equation (10) below, with the random dropout model in Equation (4) incorporated in the calculation of likelihood. A lower gene-wise likelihood value indicates a broader distribution of single-cell expression. Genes with likelihood values exceeding a given threshold (by default set at the elbow of the curve of likelihood vs. gene rank) is removed from further analysis.

Stochastic Two-State Gene Transcriptional Model

For describing the mRNA distribution of a gene, CALISTA relies on the two-state model developed by Peccoud and Ycart (1995). The model characterizes the stochastic bursty gene transcriptional process at the single-cell level. More specifically, the model describes a promoter that switches stochastically between an OFF (inactive or non-permissive) state, where the gene transcription cannot start, and an ON (active or permissive) state, where the gene transcription can proceed. The set of reactions describing the stochastic gene transcription in the two-state model are as follows:

$$\begin{aligned} OFF &\xrightarrow{\theta_{on}} ON \\ ON &\xrightarrow{\theta_{off}} OFF \\ ON &\xrightarrow{\theta_t} ON + mRNA \\ mRNA &\xrightarrow{\theta_d^m} \oslash \end{aligned}$$

where θ_{on} is the rate of the promoter activation, θ_{off} is the rate of the promoter inactivation, $\theta_t$is the rate of mRNA production when the promoter is active, θ_d is the rate constant of mRNA degradation, and m denotes the number of mRNA molecules. At steady state, the probability distribution of mRNA count m can be

approximated by the following density function (Raj et al., 2006):

$$P_2\left(m;\ \theta_{on},\ \theta_{off},\ \theta_t\right) = \frac{\Gamma\left(\theta_{on}+m\right)\Gamma\left(\theta_{on}+\theta_{off}\right)\theta_t^m}{\Gamma\left(m+1\right)\Gamma\left(\theta_{on}+\theta_{off}+m\right)\Gamma\left(\theta_{on}\right)}\ {}_1F_1\left(\theta_{on}+m,\ \theta_{on}+\theta_{off}+m,\ -\theta_t\right) \quad (2)$$

where ${}_1F_1$ represents the confluent hypergeometric function of the first kind.

Random Dropout Event Model

If desired and when appropriate, users can account for random dropout events in CALISTA. The inclusion of random dropouts is particularly suitable when dealing with single-cell transcriptional profiles from Drop-Seq technology. In CALISTA, the dropout probability is modeled by a negative exponential function with an optimal decay constant of λ, as follows:

$$P_d\left(0|m\right) = e^{-\lambda m} \quad (3)$$

where $P\left(0|m\right)$ denotes the probability that the measured single-cell mRNA count is zero when the true number of mRNA molecules is m. A decay constant $\lambda = 0$ gives a dropout probability of 1, i.e., dropout occurs regardless of the true mRNA count. The parameter λ is estimated from the plot of the fraction of zeros against the mean expression across all measured genes, following the procedure described in Pierson et al. (Pierson and Yau, 2015) (see examples in **Supplementary Figure S10**). Combining the dropout event model with the two-state gene transcription model above give the following distribution function of measured single-cell mRNA readouts $\widehat{m}$:

$$P\left(\widehat{m};\ \theta_{on},\ \theta_{off},\ \theta_t\right) = \begin{cases} \sum_m P_d\left(0|m\right) P_2\left(m;\ \theta_{on},\ \theta_{off},\ \theta_t\right) & for\ \widehat{m}=0 \\ \left(1-e^{-\lambda\widehat{m}}\right) P_2\left(\widehat{m};\ \theta_{on},\ \theta_{off},\ \theta_t\right) & for\ \widehat{m}>0 \end{cases} \quad (4)$$

Cell Clustering

As illustrated in **Figure 1B**, CALISTA combines maximum likelihood and consensus clustering algorithms for single-cell clustering analysis. Following a previous method SABEC (Ezer et al., 2016), CALISTA uses the stochastic two-state model of gene transcription to derive the steady-state probability distribution of gene expression values. The probability distribution is used to prescribe a likelihood value to a cell based on the measured single-cell expression values of its genes (Ezer et al., 2016). In contrast to SABEC that employs a simulated annealing algorithm, CALISTA implements a greedy algorithm to solve the likelihood maximization problem (see **Supplementary Note S1**). The greedy algorithm is implemented repeatedly for a specified number of times, each starting from a different random initial cell assignment, to generate a consensus matrix. The (i, j)-th element of the consensus matrix records the number of times that the i-th and j-th cells are assigned to a cluster. The final cell clustering is obtained by applying k-medoids on the consensus matrix (Bhat, 2014) (see **Supplementary Note S1** for more details).

Lineage Inference

The first novel algorithm in CALISTA is the reconstruction of cell lineage graph, which reflects the lineage progression in the differentiation process (**Figure 1C**). Based on the view that cell clusters represent cell states, the nodes of the lineage graph comprise cell clusters, while the edges represent state transitions in the lineage progression. For inferring the lineage graph, CALISTA computes cluster distances based on dissimilarities in the gene expressions among cells from two clusters. Again, CALISTA adopts a likelihood-based strategy using the probability distribution of mRNA from the stochastic two-state gene transcription model to define the cluster distances.

Cluster Distance

Given K clusters from CALISTA single-cell clustering analysis above or the clustering provided by the user, CALISTA evaluates a $K \times K$ dissimilarity matrix S where the element s_{kj} gives the likelihood of cells from cluster k to be assigned to cluster j, computed as follows:

$$s_{kj} = \frac{1}{N_k}\frac{1}{G}\left(\sum_{n_k=1}^{N_k}\sum_{g=1}^{G}\ln P\left(\widehat{m}_{n_k,g};\ \theta^*\left(g,\ j\right)\right)\right) \quad (5)$$

where N_k is the set of cell indices for the cells in cluster k and $\theta^*\left(g,\ j\right) = \left\{\theta^*_{on}, \theta^*_{off}, \theta^*_t\right\}_j^k$ is the parameter vector of two-state gene transcription that maximizes the joint probability of obtaining the gene expression values for gene g of the cells in the cluster j, as follows:

$$\theta^*\left(g,\ j\right) = \arg\max_{\Theta}\sum_{n\in N_j}\ln P\left(\widehat{m}_{ng};\ \theta\right). \quad (6)$$

The probability $P\left(\widehat{m};\ \theta\right)$ is computed using the steady state probability distribution from the two-state gene transcription model as given in Equation (2) (assuming that random dropout events are insignificant) or using the distribution function defined in Equation (4).

Note that the diagonal element s_{kk} is the sum of the cell likelihood in the original cluster assignment, i.e., $s_{kk} = \sum_{n\in N_k}\sum_{g\in G}\ln P\left(\widehat{m}_{n,g};\ \theta^*\left(g,\ k\right)\right)$, and thus are larger than other elements s_{kj}, $j \neq k$. The dissimilarity coefficients in the matrix S is subsequently normalized by subtracting each element with the diagonal element of the corresponding rows, as follows:

$$\widehat{s}_{kj} = s_{kk} - s_{kj} \quad (7)$$

Since the likelihood always takes on negative values, the normalized dissimilarity coefficient assumes a positive value. A larger $\widehat{s}_{kj}$ reflects a higher degree of dissimilarity between the two clusters. The distance between the j-th and k-th cluster, denoted by d_{kj}, is defined as:

$$d_{kj} = max\left(\widehat{s}_{kj}, \widehat{s}_{jk}\right) \quad (8)$$

Lineage Graph Construction

CALISTA generates a lineage graph by connecting single-cell clusters (states) based on the cluster distances. The lineage graph describes the state transition of stem cells during the cell differentiation process, under the assumption that the transitions occur between closely related cell states (i.e., between clusters with low distances). Briefly, CALISTA starts with a fully disconnected graph of single cell clusters, and sequentially adds one transition edge at a time in increasing magnitude of cluster distance until each cluster is connected by at least one edge. Once the lineage graph has been established, CALISTA assigns directionalities to the edges in the lineage graph according to user-provided information, e.g., starting cells/clusters or expected gene expression profiles. Given information of the starting cells, CALISTA defines the cell cluster(s) containing these cells as the starting cluster(s). On the other hand, given the expected trajectory of some marker genes, CALISTA uses the mean expression of the gene marker(s) in each cell cluster to determine the starting cell cluster.

When the stage or time information is provided for the cells, CALISTA implements the following lineage reconstruction procedure. First, large outliers in the cluster distances (i.e., cluster distances that are larger than the median value by 3 scaled median absolute deviation) are removed from further consideration. Single-cell clusters are then labeled by their most frequent (mode) cell stage or time. Clusters with the lowest stage/time label are the starting clusters. CALISTA constructs a connected graph by assigning one (and only one) incoming edge for each cluster, except for the starting cluster(s), from the cluster with the lowest cluster distance among the set of feasible parent cluster(s). Here, the feasible parents of a given cluster j are any clusters with time/stage labels that are the nearest to but do not exceed a cutoff cell times/stages in cluster j. The default cutoff is set to the 5th percentile of the cell times/stages in cluster j. This cutoff is used to ensure that there exists sufficient difference in the cell times between the parent and daughter cell clusters.

Besides the automated lineage reconstruction above, CALISTA further allows the users to manually add or remove transition edges between pairs of clusters/nodes through a user-friendly GUI (see CALISTA user manual).

Transition Genes

Another novel contribution in CALISTA is an algorithm to extract the set of transition genes between any two connected clusters in the lineage graph. Here, transition genes are defined as genes whose single-cell expressions are highly informative in grouping cells into the two clusters. More specifically, CALISTA evaluates the likelihood difference between having cells assigned to two separate clusters and having the cells together in one cluster, again using the steady-state distribution of mRNA from the two-state gene transcription model. Given two clusters j and k, we compute the following:

$$v_{jk}^{g} = v_{j}^{g} + v_{k}^{g} - v_{j+k}^{g} \tag{9}$$

with

$$v_{j}^{g} = \sum_{n \in N_j} \ln P\left(\widehat{m}_{n,g}; \theta^{*}\left(g, j\right)\right) \tag{10}$$

$$v_{k}^{g} = \sum_{n \in N_k} \ln P\left(\widehat{m}_{n,g}; \theta^{*}\left(g, k\right)\right) \tag{11}$$

$$v_{j+k}^{g} = \sum_{n \in N_k \cup N_j} \ln P\left(\widehat{m}_{n,g}; \theta^{*}\left(g, j+k\right)\right) \tag{12}$$

where the optimal parameter vector $\theta^{*}\left(g, j+k\right)$ is obtained according to Equation (6) for all cells from clusters j and k together. The value of v_{jk}^{g} reflects the informativeness of single-cell gene expression of gene g for grouping cells into two clusters j and k by the maximum likelihood principle in CALISTA. For each edge in the lineage graph, CALISTA generates a rank list of genes in decreasing values of v_{jk}^{g}. The transition genes correspond to the set of top genes in the list such that the ratio between the sum of v_{jk}^{g} among these genes and the total sum of v_{jk}^{g} among all genes exceeds a given threshold (default threshold: 50%).

Pseudotemporal Ordering of Cells

Given a lineage progression graph among cell clusters, the third and last novel algorithm in CALISTA concerns with the pseudotemporal ordering of cells. For this purpose, we first assign a pseudotime to each cluster. If the time or stage information of the cells is provided, then the pseudotime of a cluster is set to the mode of the time/stage of the cells in the cluster divided by the largest time/stage. When the time/stage information is not available, but the starting cluster is known (e.g., from knowledge of starting cells or marker genes), we assign a pseudotime of 0 for the starting cluster. We then evaluate the sum of the cluster distances along each path in the lineage progression and identify the maximum cumulative cluster distance. The pseudotime of a cluster is given by its cumulative cluster distance to the starting cluster divided by the maximum cumulative cluster distance. Once the cluster pseudotimes have been set, we assign each cell to one of the state transition edges and compute the cell pseudotime using the maximum likelihood principle (see *Cell assignment to transition edges* below). Finally, given a developmental path in the linage progression, CALISTA provides a pseudotemporal ordering of cells that have been assigned to the edges belonging to the path.

Cell Assignment to Transition Edges

For pseudotemporal ordering of the cells, CALISTA first assigns cells to the edges in the lineage graph. In the following illustration, let us consider a cell n in cluster k. CALISTA allocates the cell n to one of the edges that are incident to cluster k, again following the maximum likelihood principle. For this purpose, we define the likelihood value of a cell n to belong to an edge pointing from any cluster j to cluster k as follows:

$$\Lambda_{j \to k}\left(n\right) = \max_{t} \Lambda_{j}\left(n\right) + \frac{\left(t - t_j\right)}{\left(t_k - t_j\right)}\left(\Lambda_{k}\left(n\right) - \Lambda_{j}\left(n\right)\right) \tag{13}$$

where t_k denotes the cluster pseudotime label and $\Lambda_k(n)$ defines the likelihood value of the n-th cell to be in cluster k, i.e., $\Lambda_k(n) = \sum_{g\in G} \ln P\left(\widehat{m}_{n,g}; \theta^*(g, k)\right)$. Similarly, we define the likelihood value of the same cell to belong to an edge pointing from cluster k to any cluster l by the following:

$$\Lambda_{k\to l}(n) = \max_t \Lambda_k(n) + \frac{(t-t_k)}{(t_l-t_k)}\left(\Lambda_l(n) - \Lambda_k(n)\right) \quad (14)$$

CALISTA computes all possible $\Lambda_{j\to k}(n)$ and $\Lambda_{k\to l}(n)$ lineage graph, and assigns the cell to the edge that gives the maximum of all $\Lambda_{j\to k}(n)$ and $\Lambda_{k\to l}(n)$ values. The pseudotime of the cell $t(n)$ is set to t that gives the maximum likelihood value, as follows:

$$\begin{gathered} t(n) = \arg\max_t \Lambda_j(n) + \frac{(t-t_j)}{(t_k-t_j)}\left(\Lambda_k(n) - \Lambda_j(n)\right) \\ or \\ t(n) = \arg\max_t \Lambda_k(n) + \frac{(t-t_k)}{(t_l-t_k)}\left(\Lambda_l(n) - \Lambda_k(n)\right) \end{gathered} \quad (15)$$

depending on the cell assignment to edges above.

Cell Ordering Along a Developmental Path

Given a lineage progression graph, users can identify one or several developmental paths. A developmental path is defined as the sequence of connected clusters in the lineage progression graph with transition edges pointing from one cluster to the next in the sequence. CALISTA generates a pseudotemporal ordering along a given developmental path by first identifying cells belonging to the state transition edges in the path and order these cells according to their pseudotimes. Note that in defining the likelihood function for assigning cells to edges, we have assumed that the steady state probability distributions of gene expressions vary linearly between two connected clusters or states. But the result of the cell ordering does not change if we replace the linear interpolation function with any monotonic function.

In silico Single-Cell Time-Stamped Expression Data Generation

For testing the performance of CALISTA, we simulated synthetic single-cell gene expression data using the stochastic differentiation equation (SDE) model of the gene network (12 genes) governing the differentiation of central nervous system (CNS) proposed by Qiu et al. (2012). We simulated single-cell gene expression data for 9 time points (t_{sim} = 0, 1, 2, 4, 6, 8, 12, 16, 20) and a total of 1,800 cells (i.e., 200 cells per time point) using a time step of 0.04. Prior to the sampling, we simulated the model until steady state. In order to simulate asynchronous cell differentiation, for each time point $t_{sim,i}$, we took cells from random simulation times assuming a Gaussian distribution with a mean of $t_{sim,i}$ and a standard deviation of 0.2. The *in silico* data are included in CALISTA package (http://www.cabselab.com/calista).

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

AUTHOR CONTRIBUTIONS

NP and RG designed the computational framework and workflow and wrote the manuscript. NP developed open-course tool and performed all data analyses. NP, TH, and TF collected all necessary data and performed the preliminary analysis. All authors read and approved the final manuscript.

FUNDING

This work was supported by the Swiss National Science Foundation (grant number 157154 and 176279 - SinCity).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2020.00018/full#supplementary-material

Supplementary Table S1 | Comparison of computational runtimes for single-cell clustering: SABEC vs. CALISTA.

Supplementary Table S2 | Comparison of pseudotimes: *in silico* (ground truth), Monocle 2, DPT, and CALISTA.

Supplementary Table S3 | Transition genes in Bargaje et al. study (Bargaje et al., 2017) of iPSC differentiation into cardiomyocytes.

Supplementary Table S4 | Single-cell clustering evaluation: cell cluster assignments, adjusted Rand index, and *in silico* datasets.

Supplementary Video S1 | Three-dimensional projection of single-cell gene expression and lineage progression of zebrafish embryogenesis (Farrell et al., 2018).

REFERENCES

Angerer, P., Simon, L., Tritschler, S., Wolf, F. A., Fischer, D., and Theis, F. J. (2017). Single cells make big data: new challenges and opportunities in transcriptomics. *Curr. Opin. Syst. Biol.* 4, 85–91. doi: 10.1016/j.coisb.2017.07.004

Bargaje, R., Trachana, K., Shelton, M. N., McGinnis, C. S., Zhou, J. X., Chadick, C., et al. (2017). Cell population structure prior to bifurcation predicts efficiency of directed differentiation in human induced pluripotent cells. *Proc. Natl. Acad. Sci. U.S.A.* 114, 2271–2276. doi: 10.1073/pnas.1621412114

Bhat, A. (2014). K-medoids clustering using partitioning around medoids for performing face recognition. *Int. J. Soft Comput. Math. Control* 3, 1–12. doi: 10.14810/ijscmc.2014.3301

Bron, C., and Kerbosch, J. (1973). Algorithm 457: finding all cliques of an undirected graph. *Commun. ACM* 16, 575–577. doi: 10.1145/362342.362367

Cacchiarelli, D., Trapnell, C., Ziller, M. J., Soumillon, M., Cesana, M., Karnik, R., et al. (2015). Integrative analyses of human reprogramming reveal dynamic nature of induced pluripotency. *Cell* 162, 412–424. doi: 10.1016/j.cell.2015.06.016

Cannoodt, R., Saelens, W., and Saeys, Y. (2016). Computational methods for trajectory inference from single-cell transcriptomics. *Eur. J. Immunol.* 46, 2496–2506. doi: 10.1002/eji.201646347

Chu, L.-F., Leng, N., Zhang, J., Hou, Z., Mamott, D., Vereide, D. T., et al. (2016). Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. *Genome Biol.* 17:173. doi: 10.1186/s13059-016-1033-x

Davidson, K. C., Adams, A. M., Goodson, J. M., McDonald, C. E., Potter, J. C., Berndt, J. D., et al. (2012). Wnt/β-catenin signaling promotes differentiation, not self-renewal, of human embryonic stem cells and is repressed by Oct4. *Proc. Natl. Acad. Sci. U.S.A.* 109, 4485–4490. doi: 10.1073/pnas.1118777109

Duren, Z., Chen, X., Zamanighomi, M., Zeng, W., Satpathy, A. T., Chang, H. Y., et al. (2018). Integrative analysis of single-cell genomics data by coupled nonnegative matrix factorizations. *Proc. Natl. Acad. Sci. U.S.A.* 115, 7723–7728. doi: 10.1073/pnas.1805681115

Ezer, D., Moignard, V., Göttgens, B., and Adryan, B. (2016). Determining physical mechanisms of gene expression regulation from single cell gene expression data. *PLoS Comput. Biol.* 12:e1005072. doi: 10.1371/journal.pcbi.1005072

Farrell, J. A., Wang, Y., Riesenfeld, S. J., Shekhar, K., Regev, A., and Schier, A. F. (2018). Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science* 360:eaar3131. doi: 10.1126/science.aar3131

Grün, D., Lyubimova, A., Kester, L., Wiebrands, K., Basak, O., Sasaki, N., et al. (2015). Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature* 525, 251–255. doi: 10.1038/nature14966

Guo, G., Huss, M., Tong, G. Q., Wang, C., Li Sun, L., Clarke, N. D., et al. (2010). Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. *Dev. Cell* 18, 675–685. doi: 10.1016/j.devcel.2010.02.012

Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F., and Theis, F. J. (2016). Diffusion pseudotime robustly reconstructs lineage branching. *Nat. Methods* 13, 845–848. doi: 10.1038/nmeth.3971

Hua, J., Lin Huang, M., and Wang, G. (2018). Graph layout performance comparisons of force-directed algorithms. *Int. J. Performabil. Eng.* 14, 67–76. doi: 10.23940/ijpe.18.01.p8.6776

Huang, W., Cao, X., Biase, F. H., Yu, P., and Zhong, S. (2014). Time-variant clustering model for understanding cell fate decisions. *Proc. Natl. Acad. Sci. U.S.A.* 111, E4797–E4806. doi: 10.1073/pnas.1407388111

Jagtap, S., Meganathan, K., Gaspar, J., Wagh, V., Winkler, J., Hescheler, J., et al. (2011). Cytosine arabinoside induces ectoderm and inhibits mesoderm expression in human embryonic stem cells during multilineage differentiation. *Br. J. Pharmacol.* 162, 1743–1756. doi: 10.1111/j.1476-5381.2010.01197.x

Kærn, M., Elston, T. C., Blake, W. J., and Collins, J. J. (2005). Stochasticity in gene expression: from theories to phenotypes. *Nat. Rev. Genet.* 6, 451–464. doi: 10.1038/nrg1615

Kalisky, T., Oriel, S., Bar-Lev, T. H., Ben-Haim, N., Trink, A., Wineberg, Y., et al. (2018). A brief review of single-cell transcriptomic technologies. *Brief. Funct. Genomics* 17, 64–76. doi: 10.1093/bfgp/elx019

Kiselev, V. Y., Kirschner, K., Schaub, M. T., Andrews, T., Yiu, A., Chandra, T., et al. (2017). SC3: consensus clustering of single-cell RNA-seq data. *Nat. Methods* 14, 483–486. doi: 10.1038/nmeth.4236

Kumar, R. M., Cahan, P., Shalek, A. K., Satija, R., DaleyKeyser, A. J., Li, H., et al. (2014). Deconstructing transcriptional heterogeneity in pluripotent stem cells. *Nature* 516, 56–61. doi: 10.1038/nature13920

Lin, P., Troup, M., and Ho, J. W. K. (2017). CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data. *Genome Biol.* 18:59. doi: 10.1186/s13059-017-1188-0

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., et al. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* 161, 1202–1214. doi: 10.1016/j.cell.2015.05.002

Mao, Q., Wang, L., Goodison, S., and Sun, Y. (2015). "Dimensionality reduction via graph structure learning," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD'15* (New York, NY: ACM Press), 765–774.

Marco, E., Karp, R. L., Guo, G., Robson, P., Hart, A. H., Trippa, L., et al. (2014). Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. *Proc. Natl. Acad. Sci. U.S.A.* 111, 5643–5650. doi: 10.1073/pnas.1408993111

Moignard, V., Macaulay, I. C., Swiers, G., Buettner, F., Schütte, J., Calero-Nieto, F. J., et al. (2013). Characterization of transcriptional networks in blood stem and progenitor cells using high-throughput single-cell gene expression analysis. *Nat. Cell Biol.* 15, 363–372. doi: 10.1038/ncb2709

Ng, E. S., Azzola, L., Sourris, K., Robb, L., Stanley, E. G., and Elefanty, A. G. (2005). The primitive streak gene Mixl1 is required for efficient haematopoiesis and BMP4-induced ventral mesoderm patterning in differentiating ES cells. *Development* 132, 873–874. doi: 10.1242/dev.01657

Ng, V. Y., Ang, S. N., Chan, J. X., and Choo, A. B. H. (2010). Characterization of Epithelial cell adhesion molecule as a surface marker on undifferentiated human embryonic stem cells. *Stem Cells* 28, 29–35. doi: 10.1002/stem.221

Peccoud, J., and Ycart, B. (1995). Markovian modeling of gene-product synthesis. *Theor. Popul. Biol.* 48, 222–234. doi: 10.1006/tpbi.1995.1027

Pierson, E., and Yau, C. (2015). ZIFA: dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biol.* 16:241. doi: 10.1186/s13059-015-0805-z

Qiu, X., Ding, S., and Shi, T. (2012). From understanding the development landscape of the canonical fate-switch pair to constructing a dynamic landscape for two-step neural differentiation. *PLoS ONE* 7:e49271. doi: 10.1371/journal.pone.0049271

Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H. A., et al. (2017). Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods* 14, 979–982. doi: 10.1038/nmeth.4402

Qiu, X., Rahimzamani, A., Wang, L., Mao, Q., Durham, T., Mcfaline-Figueroa, J. L., et al. (2018). Towards inferring causal gene regulatory networks from single cell expression measurements. *bioRxiv* 426981. doi: 10.1101/426981

Raj, A., Peskin, C. S., Tranchina, D., Vargas, D. Y., and Tyagi, S. (2006). Stochastic mRNA synthesis in mammalian cells. *PLoS Biol.* 4:e309. doi: 10.1371/journal.pbio.0040309

Ralston, A., and Shaw, K. (2008). Gene expression regulates cell differentiation. *Nat. Educ.* 1, 127. Available online at: https://www.nature.com/scitable/topicpage/gene-expression-regulates-cell-differentiation-931/

Reinius, B., Plaza Reyes, A., Edsgärd, D., Codeluppi, S., Petropoulos, S., Deng, Q., et al. (2016). Single-Cell RNA-seq reveals lineage and X chromosome dynamics in human preimplantation embryos. *Cell* 165, 1012–1026. doi: 10.1016/j.cell.2016.03.023

Richard, A., Boullu, L., Herbach, U., Bonnafoux, A., Morin, V., Vallin, E., et al. (2016). Single-cell-based analysis highlights a surge in cell-to-cell molecular variability preceding irreversible commitment in a differentiation process. *PLoS Biol.* 14:e1002585. doi: 10.1371/journal.pbio.1002585

Sathyamurthy, A., Johnson, K. R., Matson, K. J. E., Dobrott, C. I., Li, L., Ryba, A. R., et al. (2018). Massively parallel single nucleus transcriptional profiling defines spinal cord neurons and their activity during behavior. *Cell Rep.* 22, 2216–2225. doi: 10.1016/j.celrep.2018.02.003

Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P.-L., and Ideker, T. (2011). Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* 27, 431–432. doi: 10.1093/bioinformatics/btq675

Stuart, T., and Satija, R. (2019). Integrative single-cell analysis. *Nat. Rev. Genet.* 20, 257–272. doi: 10.1038/s41576-019-0093-7

Stumpf, P. S., Smith, R. C. G., Lenz, M., Schuppert, A., Müller, F.-J., Babtie, A., et al. (2017). Stem cell differentiation as a non-markov stochastic process. *Cell Syst.* 5, 268–282.e7. doi: 10.1016/j.cels.2017.08.009

Thomas, T., Nowka, K., Lan, L., and Derwahl, M. (2006). Expression of endoderm stem cell markers: evidence for the presence of adult stem cells in human thyroid glands. *Thyroid* 16, 537–544. doi: 10.1089/thy.2006.16.537

Tiyaboonchai, A., Cardenas-Diaz, F. L., Ying, L., Maguire, J. A., Sim, X., Jobaliya, C., et al. (2017). GATA6 Plays an important role in the induction of human definitive endoderm, development of the pancreas, and functionality of pancreatic β cells. *Stem Cell Rep.* 8, 589–604. doi: 10.1016/j.stemcr.2016.12.026

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., et al. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* 32, 381–386. doi: 10.1038/nbt.2859

Treutlein, B., Lee, Q. Y., Camp, J. G., Mall, M., Koh, W., Shariati, S. A. M., et al. (2016). Dissecting direct reprogramming from fibroblast to neuron using single-cell RNA-seq. *Nature* 534, 391–395. doi: 10.1038/nature18323

Wolf, F. A., Angerer, P., and Theis, F. J. (2018). SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* 19:15. doi: 10.1186/s13059-017-1382-0

Wolf, F. A., Hamey, F. K., Plass, M., Solana, J., Dahlin, J. S., Göttgens, B., et al. (2019). PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol.* 20:59. doi: 10.1186/s13059-019-1663-x

Xu, C., and Su, Z. (2015). Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics* 31, 1974–1980. doi: 10.1093/bioinformatics/btv088

Zeng, W., Chen, X., Duren, Z., Wang, Y., Jiang, R., and Wong, W. H. (2019). DC3 is a method for deconvolution and coupled clustering from bulk and single-cell genomics data. *Nat. Commun.* 10:4613. doi: 10.1038/s41467-019-12547-1

Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., et al. (2017). Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* 8:14049. doi: 10.1038/ncomms14049

Žurauskiene, J., and Yau, C. (2016). pcaReduce: hierarchical clustering of single cell transcriptional profiles. *BMC Bioinform.* 17:140. doi: 10.1186/s12859-016-0984-y

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Papili Gao, Hartmann, Fang and Gunawan. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.