BIRDccNEST: Interpretable single cell characterization with inferred directed cell networks

Gizem Cicekli[*]
Department of Computer Science
Tufts University
Medford, MA 02155
Gizem.Cicekli@tufts.edu

Adrita Samanta
Department of Computer Science
Tufts University
Medford, MA 02155
Adrita.Samanta@tufts.edu

Hao Zhu [†]
Department of Computer Science
Tufts University
Medford, MA 02155
Hao.Zhu@tufts.edu

Donna K. Slonim
Department of Computer Science
Tufts University
Medford, MA 02155
Donna.Slonim@tufts.edu

Abstract

We introduce BIRDccNEST (pronounced "bird's nest"), an efficient unsupervised framework for characterizing cells and defining trajectories in single cell RNA-sequencing data by inferring directed cell-cell relationship networks. These networks are then transformed into cluster flow networks describing directed relationships between cell-cell communities, naturally capturing an interpretable trajectory and characterizing subgroups of cells. We demonstrate that this approach finds interpretable and more coherent cell communities and trajectories on several data sets. Code is available at: https://bcb.cs.tufts.edu/BIRDccNEST.html

1 Introduction

Single-cell RNA sequencing (scRNA-seq) has emerged as a powerful tool to profile transcription at the resolution of individual cells, allowing researchers to discover heterogeneity within tissues, decypher developmental processes, and reveal cellular responses to changing conditions [18, 14]. Single cell datasets are inherently high-dimensional, sparse, and noisy, presenting unique challenges and opportunities.

Pre-processing steps, such as filtering out low-quality cells and genes, normalizing for sequencing depth, and correcting for technical variation such as batch effects, are critical to ensure data quality and interpretability [23]. After pre-processing, to make the data more tractable for downstream analyses, dimensionality reduction techniques such as principal component analysis (PCA) or t-stochastic neighbor embedding (t-SNE) are almost always employed [10, 1, 12] as the first step of the analysis pipeline. At its best, dimensionality reduction can help capture the most informative genes and reduce noise while maintaining the transcriptional structure of each cell and cell type [25].

ScRNA-seq data analysis extends well beyond the identification of cell types and includes the exploration of dynamic biological processes such as differentiation, activation, or cell cycle progression, a process referred to as *trajectory inference* [28]. Trajectory inference aims to elucidate continuous biological processes by ordering cells

[*]corresponding author
[†]Current address: Center for Virology and Vaccine Research, Harvard Medical School, Boston, MA 02115

20th Conference on Machine Learning in Computational Biology (MLCB 2025).

along a pseudo-temporal trajectory. Mapping groups of cells to such an ordering enables characterizing expression changes across cell states. In non-linear processes such as differentiation, trajectories may help identify the genes responsible for determining cell fate [21].

Trajectory inference also begins with applying dimensionality reduction techniques such as PCA to the normalized counts matrix. Most trajectory inference methods then cluster cells in the reduced-dimensional space, compute cluster centroids, and commonly compute a minimum spanning tree (MST) for those centroids. The nodes of the MST are then ordered, often using some domain knowledge for orientation by specifying a "root". Established trajectory inference algorithms such as Monocle [28], Slingshot [24], or PAGA [29] use the tree or an inferred graph structure to fit principal curves and smooth paths or account for branches by averaging across the curves for consistency. While there are nominally-unsupervised trajectory inference algorithms, many still require the user to specify a root node.

Here, we introduce BIRDccNEST (pronounced "bird's nest"), which stands for "BI-directional RegDiffusion Cell-Cell Networks EStimating Trajectories." BIRDccNEST is an efficient alternative framework for characterizing cells and defining trajectories in scRNA-seq data by inferring directed cell-cell relationship networks.

We demonstrate on several published datasets that BIRDccNEST finds new cell subtypes more coherent than the originally published cell states identified in the corresponding data, while automatically capturing the underlying temporal or developmental trajectories by the inferred directed edges between cells. We previously introduced RegDiffusion, a fast, stable method for inference of directed regulatory relationships between genes in scRNA-seq data sets [30]. RegDiffusion gets its speed from a computational advance proposed for its underlying probabilistic diffusion model learning method [13], dramatically accelerating the inference process. Our idea here is to apply the same approach to the transposed expression matrix, identifying an ordering between the cells instead while capturing cell similarity. Since BIRDccNEST retains the speed of RegDiffusion and infers relationships among *cells*, it naturally learns trajectories represented in the data without requiring extensive computational resources or data filtering that would lead to information loss.

2 Methods

Broadly, BIRDccNEST infers a cell-cell network by applying the RegDiffusion algorithm to the transposed normalized expression counts matrix. The output of this step is an asymmetric adjacency matrix of cell-cell edge weights. We then use community detection methods to infer clusters within the network. Next we reduce these networks to reflect simplified relationships between the clusters by creating what we call *cluster flow networks*. Finally, we find oriented maximum spanning trees in the cluster flow networks. Details are specified below.

2.1 RegDiffusion cell-cell network inference

Given scRNA-seq normalized expression matrix $X^T \in \mathbb{R}^{m \times n}$, where n is the number of cells and m is the number of genes measured, BIRDccNEST learns a weighted directed adjacency matrix, $A \in \mathbb{R}^{n \times n}$, representing a network whose nodes correspond to the individual cells in the data set. In this network an edge weight thus captures the similarity of the expression distributions in the two cells over all genes. We expect network nodes corresponding to cells of similar types to be connected by highly-weighted bi-directional edges, whereas nodes corresponding to less similar cells might be connected by edges in only one direction reflecting the cells' underlying lineage.

This intuition is most consistently reflected by edges with the highest inferred edge weights. This may reflect a peculiarity of RegDiffusion, that it can learn edge weights near zero, reflecting very low confidence in the inferred relationships. Accordingly, we keep only those edges with edge weights in the top quartile, so long as the resulting network is strongly connected. This is true for all data sets here except mDC (see Section 2.4), where we kept the top 40% of the edge weights to reconnect the network.

2.2 Cell communities and trajectory inference algorithm

We next apply a standard network community detection method to group together cells of similar cell type/state. Here we used Louvain clustering [2] because it is a simple and widely understood greedy algorithm. Specifically, the method optimizes a modularity score that captures the density of weighted edges within a community compared to edges between communities. Although, here we only show results for Louvain, any other method such as Leiden [27] or DSD with Spectral clustering [5] can be adopted that may yield more coherent clusters for a given dataset. For reproducibility, we listed selected resolution values and any other setting in Table 3 in the Appendix.

We expect to see more directed edges connecting clusters in a manner reflecting the underlying lineage. That is, if cell states in community A precede cell states in community B, we expect more high-confidence directed edges from cells in A to cells in B than in the opposite direction. Capturing this intuition, we define *cluster flow networks* by creating graphs where each community is represented by a single node. A directed edge then connects two communities' nodes if the pruned cell-cell network includes *any* inter-cluster edges in the same direction, and the number of such edges in the cell-cell network becomes the weight of the flow-network edge.

Oriented maximum spanning trees (OMSTs, aka maximum spanning arborescences) are rooted maximum-weight spanning trees such that there is a directed walk from root node r to all other nodes [15]. We used a variant of Edmonds' algorithm [7, 26], implemented by the NetworkX package, to find the maximum spanning arborescence in our cluster flow networks. This gives us a directed trajectory in the cluster flow network that can be assessed for its ability to reflect cell lineage.

2.3 Evaluating cell communities

Because results depend on how cells are grouped in the cell-cell network, we care about how well the clustering algorithm identifies the inherent underlying groups of cells. We therefore assessed clusters from two perspectives: a) how well the clusters capture consistent expression patterns within clusters (i.e. coherence), and b) whether clusters show better distinguishable biological groupings based on known cell types/state markers (i.e. do they represent clearly interpretable and annotateble cell types).

We first confirmed the correctness of clusterings simply by how well they recapitulated known cell types or time points in the data. To further assess whether the transcriptional profiles characterized different yet valid cell groupings, we evaluated how well clusters captured consistent expression patterns of variation in the underlying data (i.e. cluster coherence). We used the Davies–Bouldin Index (DBI) [6], a standard metric quantifying clusters' internal consistency and between-cluster separation, and the Silhouette Score [19], which quantifies how well individual points fit their assigned cluster compared to others. Both coherence metrics were computed using their scikit-learn implementations: for DBI, *lower* values indicate more coherent and separated clusters, while for the Silhouette Score, *higher* values indicate the same. Additionally, we tested whether improvements in coherence scores across all datasets were significant using the Wilcoxon signed-rank paired test.

In cases where our DBI is low and Silhoutte score is high but had clusters that did not reflect expected cell states, we pursued further biological exploration that may yield novel understanding. Thus, for some datasets, we explored how well the expected, or our newly-discovered, clusters capture expected expression patterns of known cell type/state markers. This analysis verifies that the dimension-reduced transcription space does a good job of defining and distinguishing interpretable cell types/states.

2.4 Datasets

To evaluate our approach, we used the BEELINE datasets [17]. BEELINE is a collection of datasets comprising a standardized benchmark to study gene regulation and cell differentiation across species. It includes single-cell RNA-seq datasets from both human and mouse samples. Each dataset is pre-processed for quality control and includes published cell type or state labels, making these a good choice for evaluating our framework. Table 1 summarizes the datasets used.

Data set	Reference	Species	Cells	Genes
hESC	Chu et al. [4]	human	758	4,406
hHep	Camp et al. [3]	human	425	4,336
mESC	Hayashi et al. [11]	mouse	421	1,120
mDC	Shalek et al. [22]	mouse	383	3,755
mHSC-L	Nestorowa et al. [16]	mouse	847	692
mHSC-GM		mouse	888	1,595
mHSC-E		mouse	1071	704

Table 1: Characterization of BEELINE datasets by species and size

3 Results and Discussion

Table 2 presents an overview of the coherence of the cell communities discovered in the cell-cell networks. The Davies-Bouldin indices and Silhoutte scores are shown for all of the datasets considered here using the cell types

Data set	DBI of published cell types↓	DBI of BIRDccNEST communities ↓	Silhouette score of published cell types↑	Silhouette score of BIRDccNEST communities ↑
hESC	3.82	**3.72**	0.072	**0.079**
hHep	5.01	**3.24**	0.085	**0.108**
mESC	2.44	**2.31**	0.141	**0.154**
mDC	7.33	**3.08**	0.004	**0.096**
mHSC-L	6.41	**4.82**	-0.005	**0.025**
mHSC-GM	5.49	**4.72**	0.027	**0.043**
mHSC-E	7.85	**3.07**	0.052	**0.127**
	Significance of improvements (both metrics): Wilcoxon signed-rank p-value = 0.015			

Table 2: Comparing cluster quality by Davies-Bouldin Index (DBI) for which lower scores correspond to more coherent and separated clusters and Silhouette Score for which higher scores correspond to more coherent and separated clusters. The metrics take original published cell types of each dataset as clusters. The paired Wilcoxon signed-rank test shows improvements in coherence scores are *significant*.

identified by their original publications, and using the communities identified by our approach. Lower DBI values indicate more separate and coherent clusters. Higher Silhoutte scores indicate more separate and coherent clusters. For all datasets, our approach is more effective at finding coherent clusters; in some cases, dramatically so.

The rest of this section presents results for the BEELINE hESC, hHep, and mESC datasets. We chose to focus on these because they have clear underlying trajectories. Our results for the other datasets appear in the Appendix. We also provide a comparison to other TI algorithms in Section 5.4 in the Appendix.

3.1 Human embryonic stem cell differentiation

The hESC dataset characterizes the differentiation of H1 human embryonic stem cells into definitive endoderm (DE) cells by RNA-sequencing at six time points over a period of 96 hours [4]. Endoderm is one of the three germ layers formed during early embryonic development; DE cells eventually give rise to a number of organs and tissues, including gut, lungs, and liver.

In Figure 1a, we see the raw cell-cell network inferred for this dataset by RegDiffusion, with nodes colored by their corresponding cells' known time points. (While the network was inferred using all the data, only half the cells, chosen at random, are shown, to reduce clutter.) Although the inferred network structure is broadly consistent with the known time points, RegDiffusion infers many near zero weight edges. Figure 1b shows the same network pruned to include only the high-confidence edges inferred by RegDiffusion; the cells are now more clearly separated into their major clusters, though the 72 and 96 hour time points are clustered together, and the 12 and 24 hour timepoints are barely separated. Pruning also confirms that the high-confidence edges tend to connect cells in similar cell states, with low-confidence edges linking more distant cell states.

Figure 1c shows a heatmap of cell counts for each time point in each of the six inferred communities, with cell counts in each square. Here we begin to see how our unsupervised approach may find more correlated expression patterns, compared to those in cells sorted by time point alone. We see that the 12 and 24 hour cells are clustered together, while communities 3, 4, and 5 all contain both 72- and 96-hour cells. This is consistent with the PCA shown in Chu, et al. [4], recreated in Figure 1d, where the 12 and 24 hour samples somewhat overlap, and the 72 and 96 hour samples are almost inseparable. The authors note that by 72 hours the differentiation process has largely stabilized, and that there may be more subtle transformations causing other structure within these groups. This observation is consistent with the hypothesis that our cell-network communities better capture structure in the data, a hypothesis supported by the lower DBI found for these communities. We see a scatter plot of these communities, without edges, in Figure 1e; this is analogous to a PCA plot of all cells, but this different projection seems to better capture transcriptional patterns. By re-introducing network edges, Figure 1f highlights how these inferred communities are connected within the network, supporting the community structure (Figure 1f).

In Figure 1g, we see the corresponding cluster flow network constructed using these inferred communities along with the edge weights between all pairs of cluster nodes. The oriented maximum spanning tree (Figure 1h) extracted from the cluster flow network produces the community ordering 0, 1, 2, 3, 5, 4, which makes the most sense given the composition of communities 3-5 shown in Figure 1c. We hypothesize that community 4 is the most characteristically definitive-endoderm (DE) group, while community 3 has slightly fewer fully-differentiated cells, and community 5 cells are in an intermediate state.

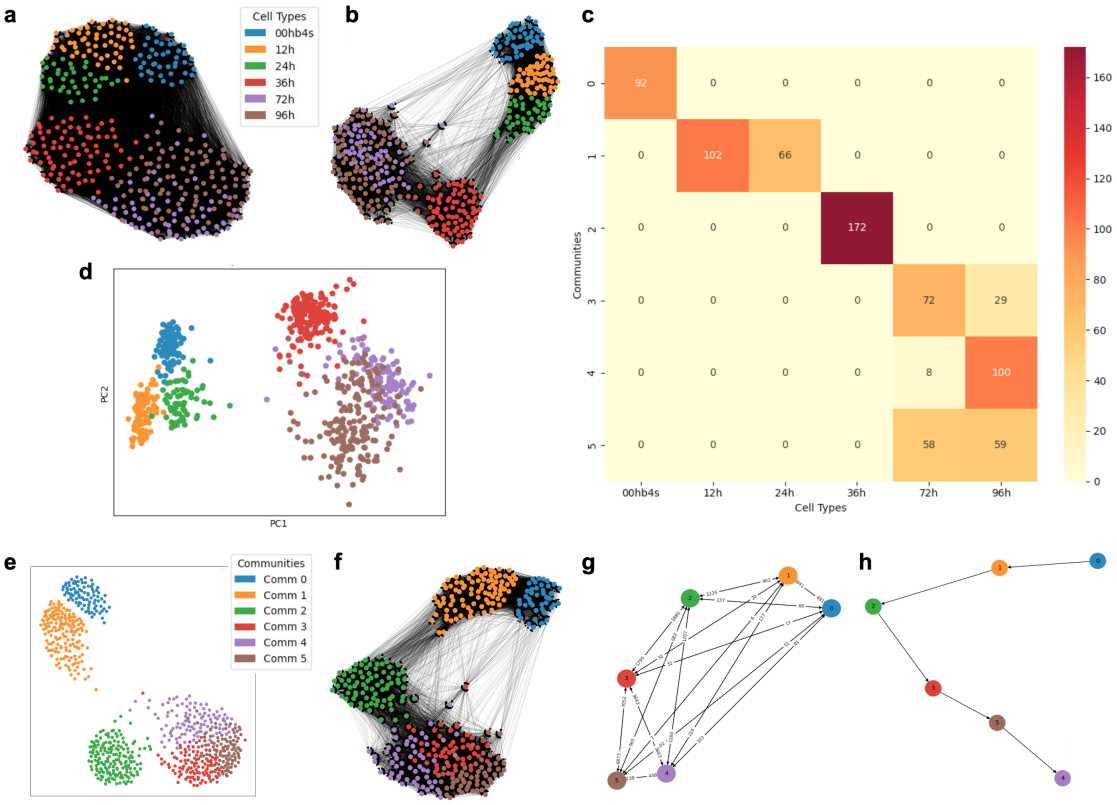

Figure 1: **Results for the hESC dataset.** Cell type legend applies to subfigures a, b, and d; Communities legend applies to subfigures e-h. a) Inferred cell-cell network, showing a sample of 50% of nodes. b) Pruned cell-cell network with top quartile edge weights, showing the same sample of nodes. c) Heatmap of overlap between known time points and inferred communities. d) PCA of dataset for first two principal components. e) All cell-cell network nodes colored by inferred communities, displayed without edges. f) Pruned cell-cell network of inferred communities, showing the same sample of nodes. g) Cluster flow network with edge counts. h) Oriented maximum spanning tree showing inferred trajectory.

This observation is further supported by our examination of DE marker genes highlighted by Chu, et al. [4]. Figure 2 shows this comparison. The violin plots on the left side reflect the established time point cell states; the violin plots on the right side come from the inferred communities. We observe that our inferred communities follow approximately the same trend in gene expression over the cell stage specific markers. For example, for the marker gene *POU5F1*, we see on the left that its expression drops over time during stem cell differentiation. On the right, we see approximately a similar trend, with higher expression in early communities (0 and 1) and lower expression in later ones (3, 5, 4).

3.2 Mouse embryonic stem cell differentiation

The mESC dataset examines a similar differentiation trajectory, following the progress of mouse embryonic stem cells (mESCs) differentiating into primitive endoderm (PrE) cells, with cells profiled at 0, 12, 24, 48, and 72 hours [11]. Figure 3 shows the same plots as in Figure 1 for this new data set. The heatmap (Figure 3c) confirms that in this case, the inferred communities follow a simple and direct trajectory, almost consistent with the known time points. Still, the 48 and 72 hour time points are clustered into two communities, one with more 48 hour than 72 hour samples, and the other the reverse, suggesting that the PrE cells have separate transcriptional characteristics that don't fully correspond to the number of hours of differentiation.

We also observe that the overlap between these two cell states is visible in PCA space (Figure 3d). The lower DBI for the BIRDccNEST communities again suggests that they reflect more coherent expression patterns than strictly grouping cells by time point. The inferred trajectory shown in Figure 3h is completely consistent with our

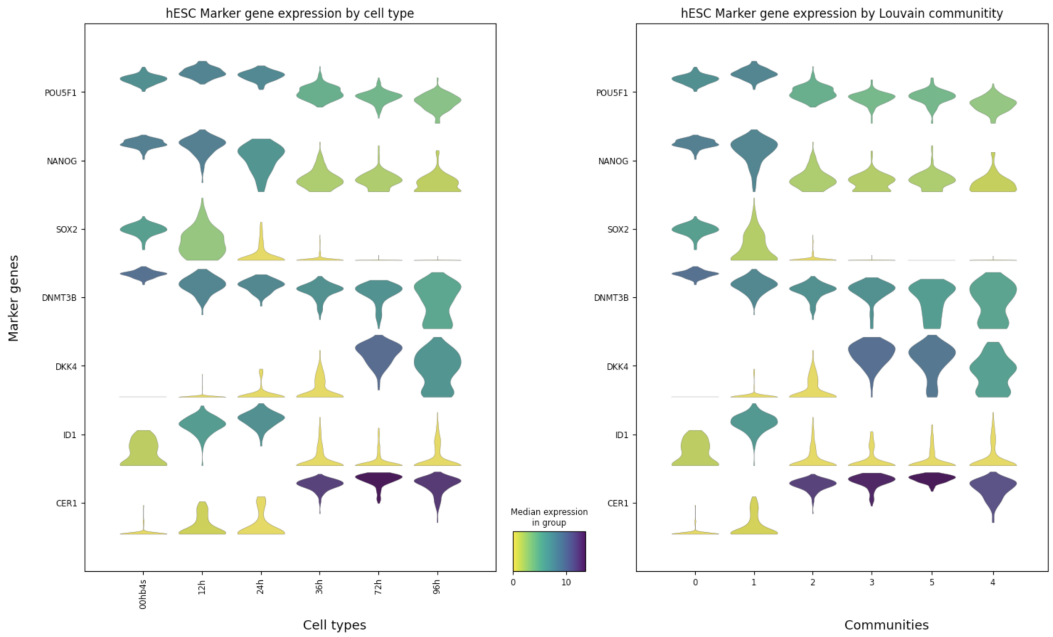

Figure 2: **Gene expression patterns of stage-specific markers during hESC differentiation.** Communities are ordered to reflect their inferred trajectory.

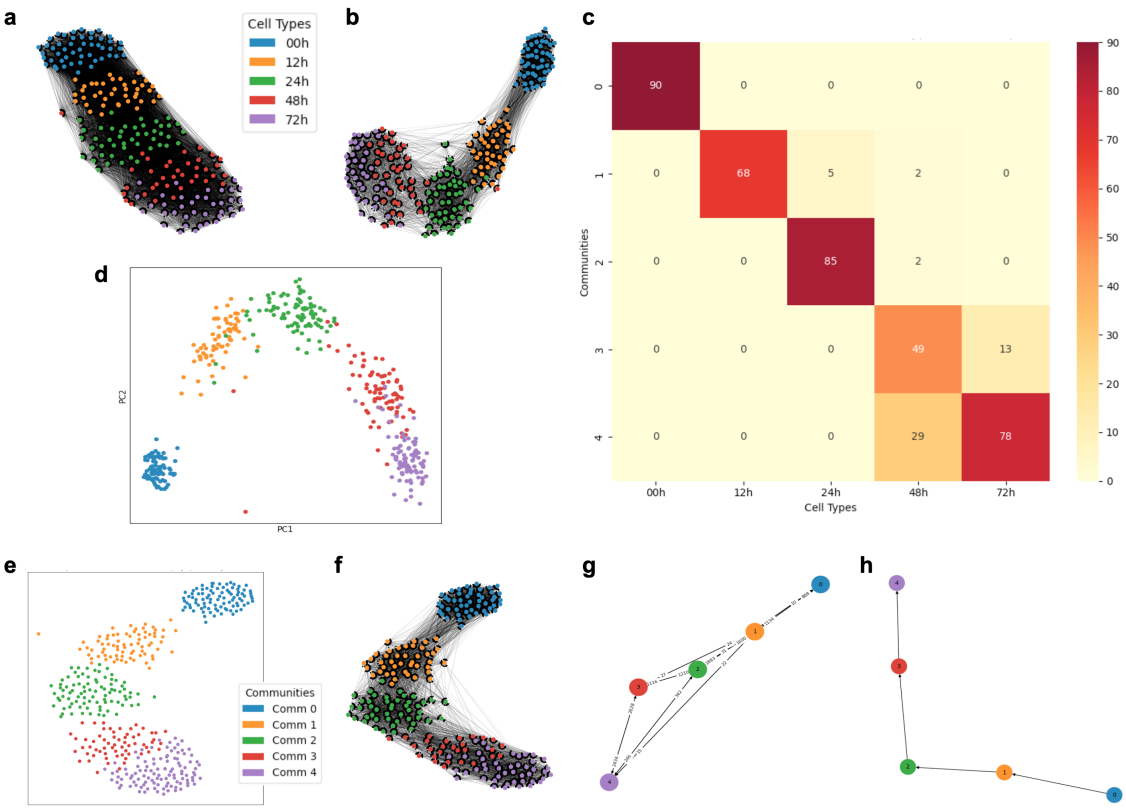

Figure 3: **Results for the mESC dataset.** The panels in this figure are the same as those in Figure 1.

expectations. These expectations were further confirmed by examining marker gene expression trends, seen in the associated violin plots (Figure 6 in the Appendix).

3.3 Human hepatocyte differentiation from iPSCs

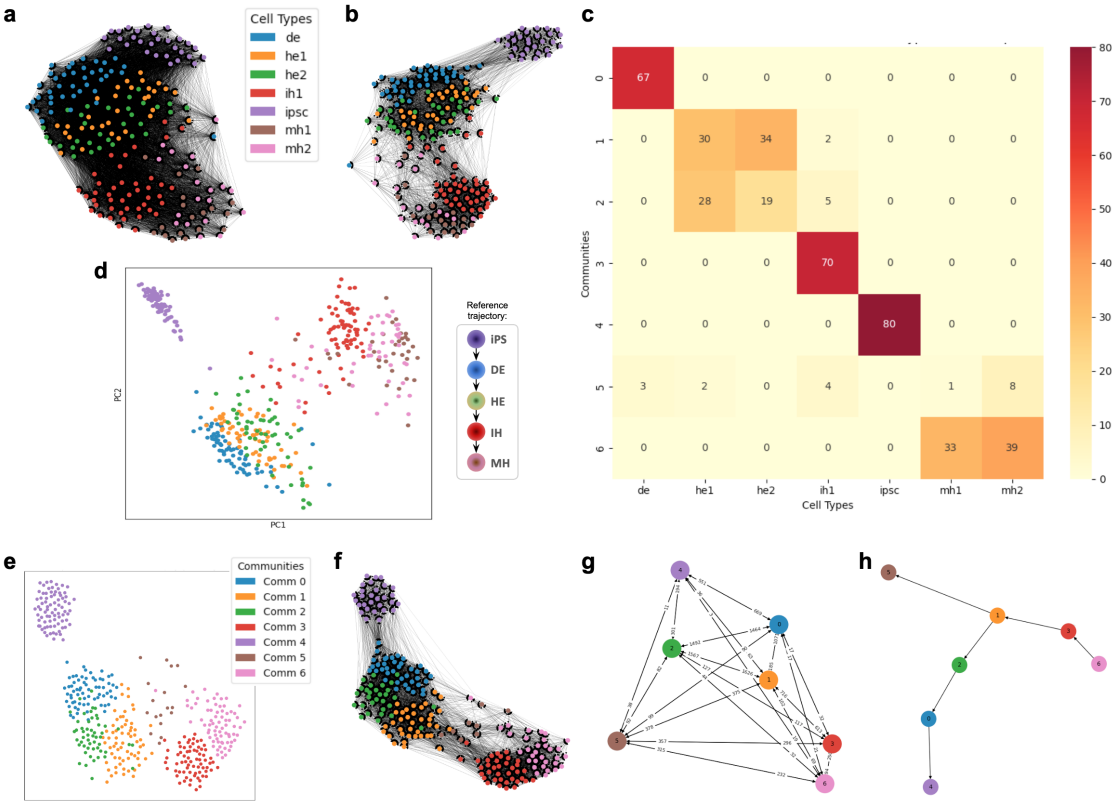

Figure 4: **Results for the hHep dataset.** The panels in this figure are the same as those in Figure 1, except that next to panel d) we include the reference trajectory pattern ordering the cell types established by Camp, et al. [3]

In Figure 4 we see the results for the human hepatocyte differentiation data set originally published by Camp, et al. [3]. This is a longer differentiation experiment that starts with pluripotent stem cells and goes beyond primitive or definitive endoderm to mature hepatocytes (liver cells). Specifically, the time points sampled include induced pluripotent stem cells (iPSCs), corresponding to day 0; definitive endoderm (DE) cells at day 6; day 8 hepatic endoderm (HE), a developmental layer already committed to the hepatic lineage; day 14 immature hepatoblast-like cells (IH), and mature hepatocyte-like cells (MH) after 21 days. The authors further highlighted two numbered subgroups for the HE and MH classes, which are implied to represent more or less differentiated versions of the named cell types (see their Extended Figure 1 in [3]).

The iPSCs appear distinct from most other cell types under any data projection. Beyond those, we see some similar cluster heterogeneity using PCA on the original data (Figure 4d) and with the inferred cell-cell network (Figures 4a, b, and e). The distinctions between DE and the two HE clusters are blurry in both, but they are somewhat more separable with BIRDccNEST. The same can be said for IH and the two MH clusters - the hepatocyte-like cells are, in practice, separated more by expression patterns than by known cell label. Figure 4c shows that BIRDccNEST breaks the HE cells into two communities but in a way orthogonal to the HE grouping from the original publication. In contrast, the two MH groups are almost all combined into one community.

The most interesting community is number 5, which contains a small number of cells from almost all cell types except iPSCs, and which appears less coherent than the others in Figure 4e. We hypothesize that these are unhealthy or dying cells that resemble each other more than they resemble their expected state in the hepatic lineage. Figure 5 shows hepatic marker gene distributions for all communities. Given that more than half the cells in cluster 5 are MH cells, the distributions of markers like *ALB, FABP1, APOA4*, and *GSTA1* for this cluster should at least partially reflect the pattern seen in the MH1 and MH2 clusters, but they patently fail to do so.

We further see that this community causes a problem in the flow network (Figure 4g), where low weights connecting to its node imply that it is distanced from the rest of the nodes using the Fruchterman-Reingold spring

layout for visualization [8]. The OMST ordering is driven by a surprisingly highly weighted edge from cluster 6 to cluster 3, which reverses the ordering of the entire trajectory. Edmonds' algorithm naturally roots the tree at such nodes with high weight outgoing edges and lower weight incoming ones. So the inferred trajectory goes backwards from MH to IH, the two HE clusters, DE, and then iPSCs. Thus, it still establishes the linear relationship correctly but reverses the order, and also identifies cells that should be removed from the analysis.

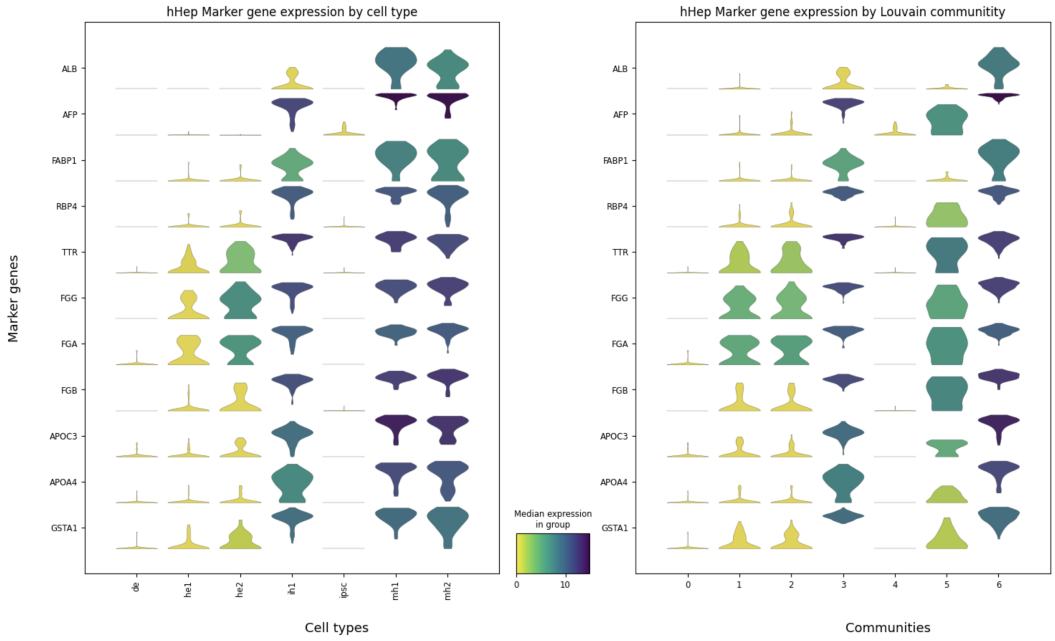

Figure 5: **Gene expression patterns of stage-specific markers during hHep differentiation.**

4 Conclusions

Here we have introduced BIRDccNEST, a simple and rapid approach for visualizing relationships between cell types in single-cell data. We observe that the inference of edges weighted by the similarity of expression distributions of the cells corresponds to the role that dimensionality reduction techniques play in conventional post-PCA analyses. The added benefit of automatically inferring a cell-cell network is that the inherent structure of the network already draws cells of the same type or state together. Additionally, there is already an inherent direction associated with these relationships, making trajectory inference more intuitive and requiring less prior knowledge or any supervision. While the community orderings from the OMST are not always perfect, BIRDccNEST also identifies outliers in the data whose removal may improve the resulting analyses. We note that popular trajectory algorithms such as Slingshot that utilize spanning trees are also prone to making similar mistakes due to the deficiency of cluster based spanning trees [1].

Our intuition in finding oriented maximum spanning trees in the cluster flow networks is partly based on how other trajectory inference algorithms use spanning trees for lineage discovery. However, we wanted to find maximal directed flows between all cluster nodes, as we expected these to best reflect an underlying lineage trajectory in our framework setting. We are considering methods to reduce the root node selection bias or improve its robustness, so that orderings are not rooted by atypical edge weights due to community size imbalances.

Another concern might be scalability. While the algorithms here all run in a few seconds on the BEELINE data sets, these are benchmarks that are all quite small. We have also run the algorithm on datasets with thousands of both cells and genes; it runs in tens of seconds at most. However, we have also chosen to investigate datasets with more complex trajectories, when larger sizes allow for greater community size imbalances, thus further exacerbating the orientation issues we saw with hHep. We therefore envision future work exploring cluster size normalization techniques and consensus building techniques as in Slingshot for the final orientation process.

Finally, in developing BIRDccNEST, we observed that the community detection process in the inferred cell-cell network may influence the results. We therefore would expect that strategically comparing multiple community detection algorithms would be another fruitful direction for future work.

Acknowledgements

We wish to thank Lenore Cowen and members of the Tufts BCB Group and CS Department for their feedback on earlier versions of this content, and colleagues Vicky Yang, Rebecca Batorsky, and AKC award 03279 for student support (GC).

References

[1] Robert A. Amezquita, Aaron T. Lun, Etienne Becht, Vince J. Carey, Lindsay N. Carpp, Ludwig Geistlinger, Federico Marini, Kevin Rue-Albrecht, Davide Risso, Charlotte Soneson, and et al. Orchestrating single-cell analysis with bioconductor. *Nature Methods*, 17(2):137–145, Dec 2019.

[2] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

[3] J Gray Camp, Keisuke Sekine, Tobias Gerber, Henry Loeffler-Wirth, Hans Binder, Malgorzata Gac, Sabina Kanton, Jorge Kageyama, Georg Damm, Daniel Seehofer, et al. Multilineage communication regulates human liver bud development from pluripotency. *Nature*, 546(7659):533–538, 2017.

[4] Li-Fang Chu, Ning Leng, Jue Zhang, Zhonggang Hou, Daniel Mamott, David T Vereide, Jeea Choi, Christina Kendziorski, Ron Stewart, and James A Thomson. Single-cell rna-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. *Genome biology*, 17:1–20, 2016.

[5] The DREAM Module Identification Challenge Consortium, Sarvenaz Choobdar, Mehmet E. Ahsen, Jake Crawford, Mattia Tomasoni, Tao Fang, David Lamparter, Junyuan Lin, Benjamin Hescott, Xiaozhe Hu, Johnathan Mercer, Ted Natoli, Rajiv Narayan, Aravind Subramanian, Jitao D. Zhang, Gustavo Stolovitzky, Zoltán Kutalik, Kasper Lage, Donna K. Slonim, Julio Saez-Rodriguez, Lenore J. Cowen, Sven Bergmann, and Daniel Marbach. Assessment of network module identification across complex diseases. *Nature Methods*, 16(9):843–852, September 2019.

[6] David L Davies and Donald W Bouldin. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2):224–227, 2009.

[7] Jack Edmonds et al. Optimum branchings. *Journal of Research of the national Bureau of Standards B*, 71(4):233–240, 1967.

[8] Thomas MJ Fruchterman and Edward M Reingold. Graph drawing by force-directed placement. *Software: Practice and experience*, 21(11):1129–1164, 1991.

[9] Laleh Haghverdi, Florian Buettner, and Fabian J. Theis. Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics*, 31(18):2989–2998, 05 2015.

[10] Ashraful Haque, Jessica Engel, Sarah A. Teichmann, and Tapio Lönnberg. A practical guide to single-cell rna-sequencing for biomedical research and clinical applications. *Genome Medicine*, 9(1), Aug 2017.

[11] Tetsutaro Hayashi, Haruka Ozaki, Yohei Sasagawa, Mana Umeda, Hiroki Danno, and Itoshi Nikaido. Single-cell full-length total rna sequencing uncovers dynamics of recursive splicing and enhancer rnas. *Nature communications*, 9(1):619, 2018.

[12] Lukas Heumos, Anna C. Schaar, Christopher Lance, Anastasia Litinetskaya, Felix Drost, Luke Zappia, Malte D. Lücken, Daniel C. Strobl, Juan Henao, Fabiola Curion, and et al. Best practices for single-cell analysis across modalities. *Nature Reviews Genetics*, 24(8):550–572, Mar 2023.

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[14] Saiful Islam, Una Kjällquist, Annalena Moliner, Pawel Zajac, Jian-Bing Fan, Peter Lönnerberg, and Sten Linnarsson. Characterization of the single-cell transcriptional landscape by highly multiplex rna-seq. *Genome research*, 21(7):1160–1167, 2011.

[15] Bernhard Korte and Jens Vygen. *Combinatorial Optimization. Algorithms and Combinatorics*, volume 21, chapter Spanning trees and arborescences. Springer Berlin Heidelberg, 2018.

[16] Sonia Nestorowa, Fiona K Hamey, Blanca Pijuan Sala, Evangelia Diamanti, Mairi Shepherd, Elisa Laurenti, Nicola K Wilson, David G Kent, and Berthold Göttgens. A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation. *Blood, The Journal of the American Society of Hematology*, 128(8):e20–e31, 2016.

[17] Aditya Pratapa, Amogh P Jalihal, Jeffrey N Law, Aditya Bharadwaj, and TM Murali. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nature methods*, 17(2):147–154, 2020.

[18] Daniel Ramsköld, Shujun Luo, Yu-Chieh Wang, Robin Li, Qiaolin Deng, Omid R Faridani, Gregory A Daniels, Irina Khrebtukova, Jeanne F Loring, Louise C Laurent, et al. Full-length mrna-seq from single-cell levels of rna and individual circulating tumor cells. *Nature biotechnology*, 30(8):777–782, 2012.

[19] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

[20] Wouter Saelens, Robrecht Cannoodt, Helena Todorov, and Yvan Saeys. A comparison of single-cell trajectory inference methods. *Nature Biotechnology*, 37(5):547–554, May 2019.

[21] Manu Setty, Vaidotas Kiseliovas, Jacob Levine, Adam Gayoso, Linas Mazutis, and Dana Pe'Er. Characterization of cell fate probabilities in single-cell data with palantir. *Nature biotechnology*, 37(4):451–460, 2019.

[22] Alex K Shalek, Rahul Satija, Joe Shuga, John J Trombetta, Dave Gennert, Diana Lu, Peilin Chen, Rona S Gertner, Jellert T Gaublomme, Nir Yosef, et al. Single-cell rna-seq reveals dynamic paracrine control of cellular variation. *Nature*, 510(7505):363–369, 2014.

[23] Oliver Stegle, Sarah A Teichmann, and John C Marioni. Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics*, 16(3):133–145, 2015.

[24] Kelly Street, Davide Risso, Russell B Fletcher, Diya Das, John Ngai, Nir Yosef, Elizabeth Purdom, and Sandrine Dudoit. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC genomics*, 19:1–16, 2018.

[25] Shiquan Sun, Jiaqiang Zhu, Ying Ma, and Xiang Zhou. Accuracy, robustness and scalability of dimensionality reduction methods for single-cell rna-seq analysis. *Genome biology*, 20:1–21, 2019.

[26] Robert Endre Tarjan. Finding optimum branchings. *Networks*, 7(1):25–35, 1977.

[27] V. A. Traag, L. Waltman, and N. J. Van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific Reports*, 9(1):5233, March 2019.

[28] Cole Trapnell, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J Lennon, Kenneth J Livak, Tarjei S Mikkelsen, and John L Rinn. Pseudo-temporal ordering of individual cells reveals dynamics and regulators of cell fate decisions. *Nature biotechnology*, 32(4):381, 2014.

[29] F Alexander Wolf, Fiona K Hamey, Mireya Plass, Jordi Solana, Joakim S Dahlin, Berthold Göttgens, Nikolaus Rajewsky, Lukas Simon, and Fabian J Theis. Paga: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome biology*, 20:1–9, 2019.

[30] Hao Zhu and Donna Slonim. From noise to knowledge: Diffusion probabilistic model-based neural inference of gene regulatory networks. *Journal of Computational Biology*, 31(11):1087–1103, November 2024.