Robust lineage reconstruction from high-dimensional single-cell data

Published online 20 May 2016

Nucleic Acids Research, 2016, Vol. 44, No. 14 ***e122***
doi: 10.1093/nar/gkw452

Robust lineage reconstruction from high-dimensional single-cell data

Gregory Giecold[1,2], Eugenio Marco[1,2], Sara P. Garcia[1,2], Lorenzo Trippa[1,2] and Guo-Cheng Yuan[1,2,3,*]

[1]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, MA 02215, USA, [2]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA and [3]Harvard Stem Cell Institute, Cambridge, MA 02138, USA

Received January 4, 2016; Revised April 12, 2016; Accepted May 9, 2016

ABSTRACT

Single-cell gene expression data provide invaluable resources for systematic characterization of cellular hierarchy in multi-cellular organisms. However, cell lineage reconstruction is still often associated with significant uncertainty due to technological constraints. Such uncertainties have not been taken into account in current methods. We present ECLAIR (Ensemble Cell Lineage Analysis with Improved Robustness), a novel computational method for the statistical inference of cell lineage relationships from single-cell gene expression data. ECLAIR uses an ensemble approach to improve the robustness of lineage predictions, and provides a quantitative estimate of the uncertainty of lineage branchings. We show that the application of ECLAIR to published datasets successfully reconstructs known lineage relationships and significantly improves the robustness of predictions. ECLAIR is a powerful bioinformatics tool for single-cell data analysis. It can be used for robust lineage reconstruction with quantitative estimate of prediction accuracy.

INTRODUCTION

Over the past few years, high-throughput sequencing, flow and mass cytometry, microfluidics along with other technologies have evolved to the point that the measurements of gene expression and protein levels are now possible at the single-cell resolution (1), providing an unprecedented opportunity to systematically characterize the cellular heterogeneity within a tissue or cell type. The high-resolution information of cell-type composition has also provided new insights into the cellular heterogeneity in cancer and other diseases (2). Single-cell data present new challenges for data analysis, and computational methods for addressing such challenges are still under-developed (3). Here we focus on a common challenge: to infer cell lineage relationships from single-cell gene expression and proteomic data. While several methods have been developed (4–8), one common limitation is that the resulting lineage is often sensitive to various factors including measurement error, sample size and the choice of pre-processing methods. However, such sensitivity has not been systematically evaluated.

Ensemble learning is an effective strategy for enhancing prediction accuracy and robustness that is widely used in science and engineering (9,10). The key idea is to aggregate information from multiple prediction methods or subsamples. This approach has also been applied to unsupervised clustering, where multiple clustering methods are applied to a common dataset and consolidated into a single partition called the consensus clustering (11).

Here we apply such an ensemble strategy to aggregate information from multiple estimates of lineage trees. We call our method ECLAIR, which stands for Ensemble Cell Lineage Analysis with Improved Robustness. We show that ECLAIR improves the overall robustness of lineage estimates and is generally applicable to diverse data-types Moreover, ECLAIR provides a quantitative evaluation of the uncertainty associated with each inferred lineage relationship, providing a guide for further biological validation.

MATERIALS AND METHODS

ECLAIR consists in three steps: 1. ensemble generation; 2. consensus clustering and 3. tree combination. An overview of our method is shown in Figure 1.

Ensemble generation

Given a dataset, we generate an ensemble of partitions out of a population of n cells by subsampling, which can be either uniform or non-uniform. For large sample size, we prefer to use a non-uniform, density-based subsampling strategy in order to enrich for under-represented cell types.

*To whom correspondence should be addressed. Tel: +1 617 582 8532; Fax: +1 617 632 2444; Email: gcyuan@jimmy.harvard.edu
Present address: Eugenio Marco, Editas Medicine, Cambridge, MA 02142, USA.

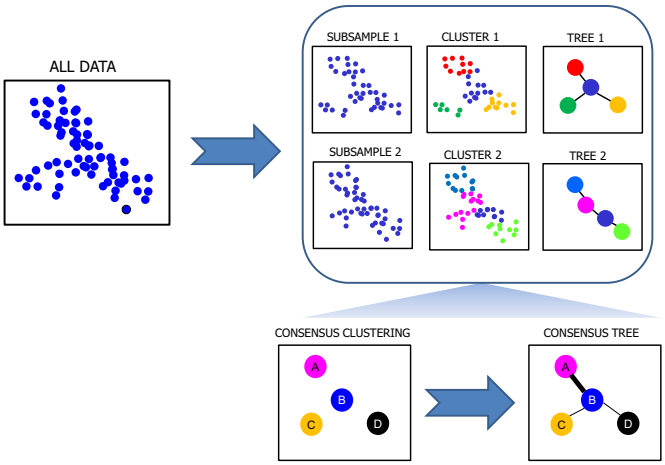

Figure 1. Overview of the ECLAIR method. First, multiple subsamples are randomly drawn from the data. Each subsample is divided into cell clusters with similar gene expression patterns, and a minimum spanning tree is constructed to connect the cell clusters. Next, consensus clustering is constructed by aggregating information from all cell clusters. Finally, a lineage tree connecting the consensus clusters (CC) is constructed by aggregating information from the tree ensemble.

Specifically, a local density at each cell is estimated as the number of cells falling within a neighborhood of fixed size in the gene expression space. If the local density is above a maximum threshold value, a cell is sampled with a probability that is inversely proportional to the local density. If the local density is below a minimum threshold value, the cell is discarded to avoid technical artifacts In other situations, the cell is always included. The resulting subsample exhibits a nearly uniform coverage of the gene expression space while removing outliers in the cell population.

Each subsample is divided into clusters with similar gene expression patterns. The specific clustering algorithm is determined by the user and can be selected from *k*-means (12), affinity propagation (13) or DBSCAN (Density-based Spatial Clustering of Applications with Noise) (14). In practice, we find that *k*-means clustering typically offers a good balance between robustness of the final estimates and computational costs.

For a given clustering solution, a fully connected graph is constructed by connecting every cluster pair, with the edge weight defined as the average Euclidean distance (in the gene expression vector space) between all pairs of cells straddling the two clusters. A minimum spanning tree (MST) is defined as the tree connecting all clusters with minimal total weight. We use Prim's algorithm (15) to identify the MST. Later on, the path length between a pair of cells, x and y, is defined as the minimum number of edges along the a MST path connecting their corresponding clusters, and denoted by L(x,y).

The above clustering and linkage procedure is repeated M times, each corresponding to a random subsample. After each iteration, the resulting clusters are expanded to incorporate every cell in the population: each cell that has not been subsampled is assigned to its closest cluster. In the end, each tree in the ensemble provides a specific estimate of the lineage tree for the entire cell population.

Our goals are to aggregate information from the ensemble and to obtain a robust estimate of the lineage tree.

Consensus clustering

We start by aggregating the clustering information, searching for a consensus clustering that is on average the most consistent with the different M partitions in the ensemble, using a strategy proposed by Strehl and Ghosh (11). For a population of n cells, the similarity between a pair of clusterings, $\lambda^{(a)}$ and $\lambda^{(b)}$, which contains $k^{(a)}$ and $k^{(b)}$ clusters respectively, is quantified by the normalized mutual information (NMI), defined as:

$$\phi^{(NMI)}(\lambda^{(a)}, \lambda^{(b)}) = \frac{\sum_{h=1}^{k^{(a)}} \sum_{l=1}^{k^{(b)}} n_{h,l} \log(\frac{n \cdot n_{h,l}}{n_h^a n_l^b})}{\sqrt{(\sum_{h=1}^{k^{(a)}} n_h^{(a)} \log \frac{n_h^{(a)}}{n})(\sum_{h=1}^{k^{(b)}} n_l^{(b)} \log \frac{n_l^{(b)}}{n})}}$$

here, $n_h^{(a)}$ and $n_l^{(b)}$ denote the numbers of cells in the corresponding clusters, and $n_{h,l}$ stands for the number of cells in their intersection.

For an ensemble of M partitions, $\lambda^{(1)}, \cdots, \lambda^{(M)}$, the consensus clustering $\lambda^* = \{C_1^*, \cdots, C_K^*\}$ is defined as the one that maximizes the average NMI with the M partitions in the ensemble. The solution is computed by combining three approximation algorithms, CSPA, HGPA and MCLA, and selecting the one that performs the best (11).

Tree combination

The final step of ECLAIR amounts to constructing a representative tree connecting the consensus clusters (CC). We first construct a fully connected graph G^*, with the weight of the edge connecting clusters C_i^* and C_j^* given by:

$$W_{ij} = \frac{1}{n_i^* n_j^* M} \sum_{x \in C_i^*} \sum_{y \in C_j^*} \sum_{m=1}^{M} L^{(m)}(x, y)$$

where, n_i^* and n_j^* refer to the number of cells in C_i^* and C_j^*, respectively, and $L^{(m)}(x, y)$ is the path length between cells x and y in the m-th tree $T^{(m)}$. From G^*, again, we extract a MST, T^*, by using Prim's algorithm. This is from now on referred to as the ECLAIR tree. In what follows we show that the ECLAIR tree provides a robust estimate of the lineage relationship.

Tree visualization

We use the igraph Python package (http://igraph.org/python/) to visualize the various trees generated by SPADE (4) or ECLAIR. To facilitate visualization, we encode the overall gene expression pattern associated with a cell cluster in a particular coloring scheme. Specifically, the raw gene expression pattern is subjected to Principal Component Analysis (PCA). The first three components are rescaled to the [0, 1] interval, and together define a unique color in the RGB (red-green-blue) encoding scheme. As such, clusters with similar expression patterns will have similar colors. Besides, the size of each node is scaled according to the number of cells it contains.

Lineage tree comparison and robustness estimation

When we compare two lineage trees, we need to compare not only their edge connections but also their nodes' (i.e. cell clusters) identities, since the variation associated with subsampling results in different cell clusters. Although there exists a body of literature on graph comparison (16), we are not aware of any method that takes into account the differences in the nodes' composition. We have therefore developed new metrics that are suitable for comparing lineage trees.

First, we define a metric to compare the overall similarity between two lineage trees: T_1 and T_2. For each tree, we evaluate the path length between every pair of cells in the population, based on the edge connectivity. The correlation between the two sets of path length values is used as a metric to compare the overall similarity of T_1 and T_2. In particular, when T_1 and T_2 are trained on two non-overlapping subsets obtained by partitioning the dataset, the correlation coefficient becomes a measure of the ECLAIR tree estimate variability.

Second, we define edge-specific dispersion rates to evaluate the robustness of each edge within a lineage tree T^*. Specifically, for each edge E_{ij} connecting a pair of clusters C_i^* and C_j^*, we define the dispersion rate D_{ij} associated with E_{ij} as the standard deviation

$$D_{ij} = \sqrt{E(L^2(x, y)) - E^2(L(x, y))} \quad (1)$$

of the path length $L(x, y)$, where (i) x and y range over all possible pairs of cells selected from C_i^* and C_j^* respectively, (ii) the distance $L(x, y)$ refers to an independent ECLAIR tree, trained by a hypothetical distinct dataset with identical size and identical distribution of the cell composition. The dispersion rate D_{ij} is unknown because we work with a single dataset. Still, $D_{ij} \approx 0$ implies that nearly every pair (x, y) from C_i^* and C_j^* after having been mapped on a hypothetical independent ECLAIR tree would preserve their distance of 1 edge.

We estimate $E(L^2(x, y))$ and $E(L(x, y))$ to evaluate D_{ij}. To this end, we first randomly created 50 ECLAIR trees each obtained from a different dataset obtained by resampling. Then, by averaging over the resulting 50 trees we estimate $E(L^2(x, y))$ and $E(L(x, y))$. The results are denoted by $\check{D}_{ij}$. The edges associated with lower $\check{D}_{ij}$ values are more reproducible and therefore more likely to reflect true lineage relationships.

In addition, we provide an alternative uncertainty metric, denoted by $\hat{D}_{ij}$ from a single ECLAIR tree. Here we consider the variation among the trees within one ECLAIR ensemble, to quantify the variation within the tree ensemble, as the standard deviation of path lengths $L^{(m)}(x, y)$, $m = 1, ..M$ across the individual trees, $T^{(m)}$, within each ensemble. We compute $\hat{D}_{ij}$ by averaging over the M individual trees. To distinguish the two approaches, we call $\check{D}_{ij}$ the inter-ensemble dispersion rate and $\hat{D}_{ij}$ the intra-ensemble dispersion rate. Whereas $\check{D}_{ij}$ is an estimate of D_{ij}, it is computationally costly. On the other hand, $\hat{D}_{ij}$ provides a more computationally variability metric. As shown in the Results section, $\hat{D}_{ij}$ and $\check{D}_{ij}$ are well correlated.

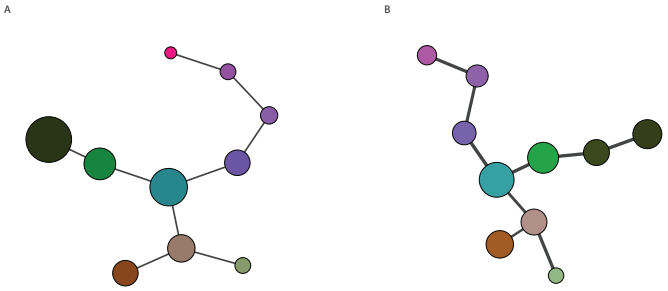

Figure 2. ECLAIR correctly reconstructs the lineage tree in mouse early embryo. (**A**) The lineage tree constructed by SCUBA, based on temporal information in the data. This tree has been experimentally validated. (**B**) The lineage tree constructed by ECLAIR, without using temporal information. The size of each node is proportional to the number of cells in the corresponding cluster. The color of each node indicates the gene expression pattern associated with the corresponding cell cluster. In (B), the edge width is inversely proportional to the estimated dispersion rate.

Numerical implementation and software package

ECLAIR is implemented as an open-source Python package, which can be accessed at https://www.github.com/GGiecold/ECLAIR. In order to manipulate large datasets, we have made a number of efforts to optimize storage and numerical efficiency, including: a scalable Python module utilizing the HDF5 data structure, a sparse-matrix and streamlined implementation of the Strehl-Ghosh approximation algorithms for consensus clustering, along with scalable and efficient implementations of the affinity propagation and DBSCAN algorithms. All packages are accessible from either the aforementioned Github website or the Python Package Index.

RESULTS

Reconstruction of cell lineages in mouse embryos

We used ECLAIR to analyze a public quantitative polymerase chain reaction (qPCR) dataset, which contains gene expression information for 48 genes in 438 cells isolated from early mouse embryos (17). Previously, we developed a method called SCUBA (Single-cell Clustering Using Bifurcation Analysis) to reconstruct and experimentally validate the cell lineages based on the temporal information (8). Here we applied ECLAIR to reanalyze this dataset without using the temporal information. One of the goals is to evaluate if ECLAIR, blinded to any temporal information, estimates a tree consistent with the temporal information in the dataset. The clustering was done by using k-means, with k set to 11 as suggested by a gap statistics analysis (18). We generated an ensemble of 50 trees, each obtained from a subsample containing 75% of the total number of cells. As shown in Figure 2, the SCUBA and ECLAIR trees have strikingly similar overall structures, and the corresponding nodes have similar gene expression patterns. Similar results were obtained by replacing k-means by either DBSCAN (14) or affinity propagation (13) clustering algorithm (Supplementary Figure S1). Taken, together, these results suggest a good accuracy of the ECLAIR algorithm.

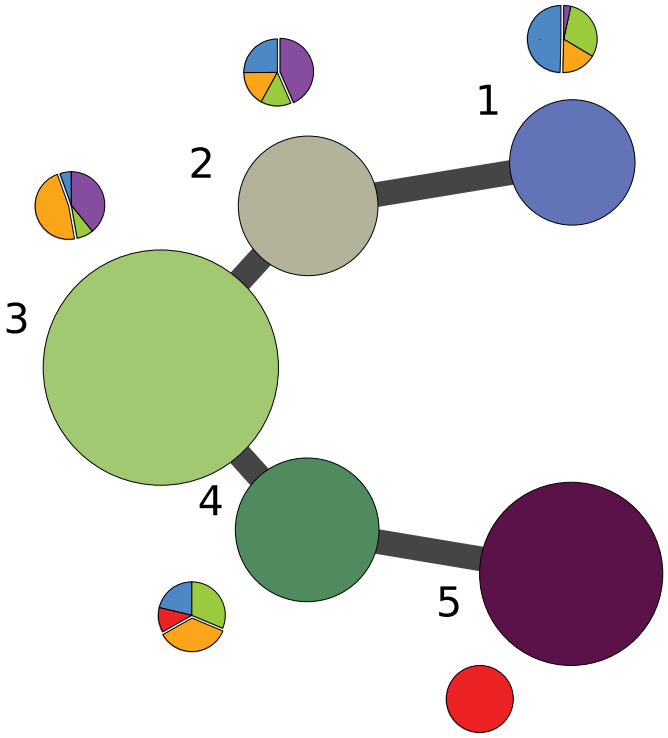

Figure 3. ECLAIR identifies lineage tree in blood-forming potential cells. The pie-charts represent the cell-type composition of each node using the same color scheme as in (19). The numbers indicate the labels for each CC.

ECLAIR is scalable to large dataset analysis

To test the scalability of ECLAIR, we analyzed two larger single-cell gene expression datasets. This first (19) is a qPCR dataset which consists of expression levels of 33 transcription factors and 9 marker genes among 3934 cells with blood-forming potential. These cells can also be divided into five cell-types: primitive streak (PS), neural plate (NP), head fold (HF), GFP+ four somite (4SG) and Flk1+GFP− (4SFG) cells. As in the previous section, we generated an ensemble of 50 lineage trees each obtained from a density-based subsample. Here we used a 50% down-sampling rate due to the large sample size. We set $k = 5$ to be comparable with the original paper (19). The resulting ECLAIR analysis tree contains five nodes along a single branch (Figure 3).

To test whether our ECLAIR analysis is consistent with prior biological knowledge, we analyzed the cell-type composition of each CC identified by ECLAIR and compared with the original analysis. We found that our consensus clustering identified similar but more refined structure than the original results. In particular, CC 1 is mainly composed of PS cells but also contains NP and HF cells, similarly to Cluster I in the original paper (Figure 1E in (19) and Supplementary Figure S2a and c in (19)). CC2 is mainly composed of 4SFG cells but also has contributions from PS, HF and NP cells, similar to the upper part of Cluster II in Supplemental Figure S2c in (19). CC3 is composed HF and 4SFG cells, in accordance with middle part of Cluster II in Supplemental Figure S2c in (19). CC4 has nearly equal contribution from NP cells, well corresponding to the lower part of Cluster II in Supplemental Figure S2c in (19). Finally, CC5 is entirely contributed by 4SG cells, well corresponding to Cluster III in the original paper. As such, our ECLAIR analysis correctly recapitulated the cellular hierarchy.

Recently, single-cell RNAseq technologies have been rapidly developed to profile the whole transcriptome at single-cell resolution (1). To test whether ECLAIR can also be used to analyze single-cell RNAseq data, we analyzed a large dataset consisting of expression levels of 8716 genes in 2730 cells (20). The high-dimensionality of the data prevents direct application of ECLAIR. To circumvent this difficulty, we reduced the data dimensionality by using a computationally efficient variant of randomized PCA (21). We retained the top 50 principal components, which accounted for 74% of the total variance and applied ECLAIR to analyze the reduced dataset using the same parameter setting except for changing k to 19 in order to be comparable with the original study.

ECLAIR analysis identified a lineage tree which can be divided into two major branches (Figure 4A). The upper branch (CC1–CC8) recapitulated the progression from early progenitors CC8 to erythroid lineages, whereas the lower branch (CC8–CC19) recapitulated the progression from early progenitors to the myeloid lineages including monocytes, granulocytes and basophils. Close examination of the expression pattern of 33 lineage marker genes (Figure 4) identified striking similarity with the clustering patterns in the original paper (Figure 1c in (20)). In particular, CC1-3 are associated with Klf1, Gfi1b and Gata1 expression, characteristic of erythroid lineage progenitors, similar to Clusters 1–6 in the original study (Figure 1c in (20)). CC4-7 represent transitional states, whereas CC8 is enriched with the early progenitor TF Gata2, which is similar to the C7 population in (20). CC9-12 are transitional states. CC13-18 are enriched with Cebpa, Mpo and several other myeloid lineage markers and associated with neutrophils, monocytes and eosinophils (similar to Clusters 14–17 in (20)). Taken together, these analyses strongly suggest that our ECLAIR analysis is scalable to analyzing large datasets without losing accuracy.

ECLAIR significantly improves robustness over SPADE

To evaluate the robustness of ECLAIR, we analyzed a publicly-available mass cytometry dataset (22), which contains the expression levels of 9 protein markers for 500 000 cells from the mouse hematopoietic system. This dataset was originally used as the basis to develop SPADE (4), one of the prevalent methods for lineage reconstruction. Despite its wide applications, it has been noted that two independent runs of SPADE often lead to substantially different results. In the latest version, the developers attempted to solve the problem by fixing the value of a random seed for subsampling. However, this simple approach does not resolve the intrinsic problem of variability in the algorithm output.

ECLAIR has a similar model structure compared to SPADE, but uses an ensemble-based approach to improve the estimate robustness. We compared the performance of ECLAIR and SPADE based on the correlation between pairwise path lengths. For each method, we generated 50 trees. Note that each ECLAIR tree itself is obtained by

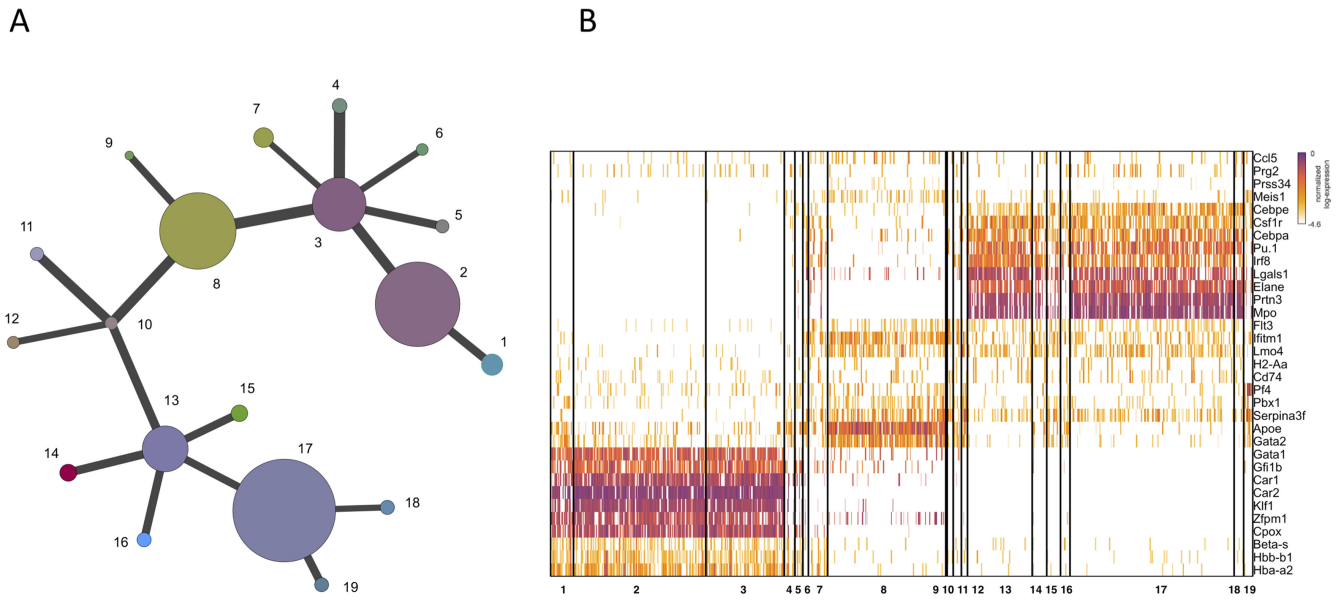

Figure 4. Analysis of the single-cell RNAseq data in (20). (**A**) The lineage tree inferred by ECLAIR. The numbers represent the label for each CC. (**B**) Heatmap showing the expression pattern of 33 lineage markers.

aggregating information from 50 individual lineage trees from subsamples. We set the down-sampling parameter to 50% for both SPADE and ECLAIR. To quantify the reproducibility of each method, we compared the cell-pair path lengths obtained from any pair of the 50 tree estimates (Figure 5). For both methods, SPADE and ECLAIR, we obtain independent trees' estimates from non-overlapping subsets of data and graph cell-pair path lengths (see 'Materials and Methods' section for details). It is clear that the distribution is more densely populated near the diagonal for ECLAIR (Figure 5A) compared to SPADE (Figure 5C). For ECLAIR, the correlation coefficient between different pairs of trees varies between 0.73 and 0.94, with a mean of 0.86 and standard deviation of 0.05. For SPADE, the correlation coefficient between different pairs of trees varies from 0.70 and 0.83, with a mean of 0.75 and a standard deviation of 0.03, indicating ECLAIR significantly improves reproducibility.

In order to evaluate the robustness with respect to training data differences, we randomly divided the whole cell population into three equal-size non-overlapping subsets, labeled S1, S2 and S3, respectively. For each method, we constructed two lineage trees: one using S1 as the training set, whereas the other using S2 for model training. S3 was reserved as the testing set. Again, we compared the cell-pair path lengths obtained from two different trees from each method, based on cells in the S3 subset.). This procedure was repeated 10 times. The distribution is more densely populated near the diagonal for ECLAIR (Figure 5B) compared to SPADE (Figure 5D for ECLAIR, the correlation between different training sets varies from 0.72 to 0.91, with a mean value of 0.82; whereas for SPADE, the correlation varies between 0.62 and 0.82 with an average value of 0.75. Again, ECLAIR is significantly more reproducible compared to SPADE.

Edge-specific dispersion rate

In view of the significant variation across the lineage trees obtained from different methods or, as we have shown here, from different training sets even using the same method, it is important to systematically quantify the robustness of different edges and to identify those edges that are the most robust. To this end, we have developed a quantitative metric, called dispersion rate, to evaluate the robustness of each edge in a lineage tree (see 'Materials and Methods' section for detail).

For an ECLAIR tree obtained from the mass cytometry dataset (Supplementary Figure S2A), we estimated the edge-specific dispersion rates as a way to quantify uncertainty. We started by using the inter-ensemble dispersion rate $\check{D}_{ij}$. To this end, we obtained 50 ECLAIR trees, each from a randomly selected subsample containing 50% of the cells. Comparing the structure of two randomly picked ECLAIR trees (Supplementary Figures S2A and B), we find that the edges associated with lower dispersion rates (thicker edges) are conserved between the trees, indicating that the dispersion rate is an informative metric for edge robustness. For comparison, we also show two randomly picked SPADE trees obtained from the same data (Supplementary Figures S2C and D). We see that the overall structure is more variable.

Next, we compared the intra-ensemble $\hat{D}_{ij}$ and inter-ensemble $\check{D}_{ij}$ estimates (Figure 6 and Supplementary Figure S3). Although the definition and interpretation of these two uncertainty metrics is different, their correlation coefficient estimates is 0.69, indicating substantial agreement of the dispersion rates. In our examples we verified that $\hat{D}_{ij}$ is computationally less demanding than $\check{D}_{ij}$.

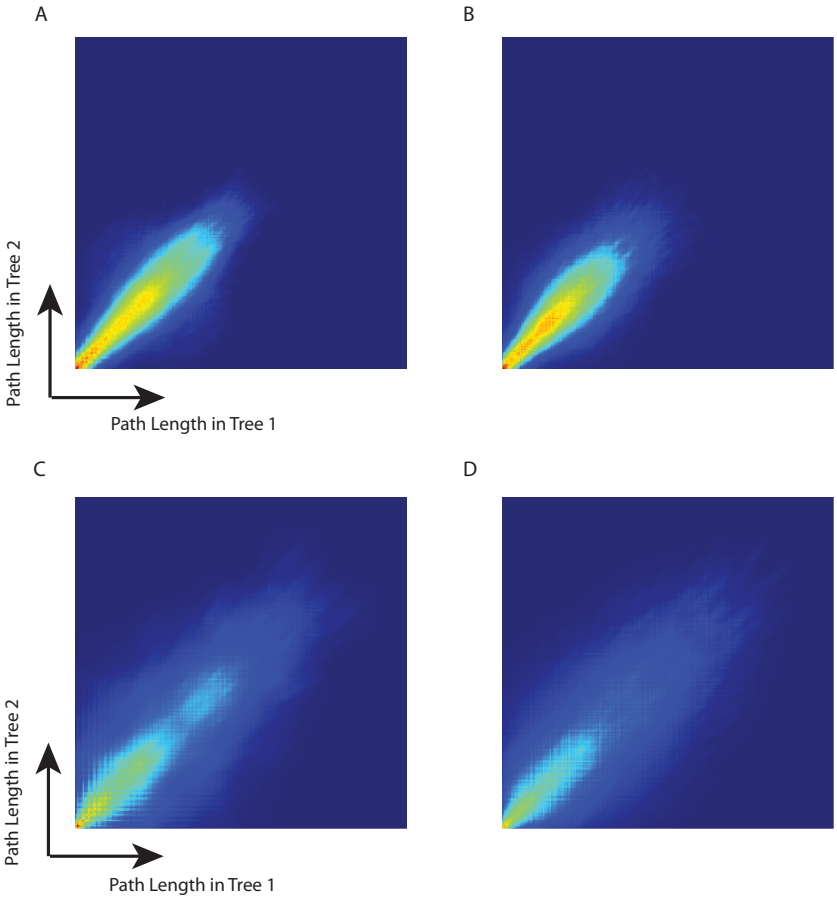

Figure 5. Comparison of the reproducibility between ECLAIR (**A** and **B**) and SPADE (**C** and **D**). Each heatmap shows the probability density of the cell-pair path length estimated using two trees obtained from the same method. (A) Two independent runs of ECLAIR on the same training set. (B) Two independent runs of ECLAIR on different training datasets. (C) Two independent runs of SPADE on the same training set. (D) Two Independent runs of SPADE on different training datasets.

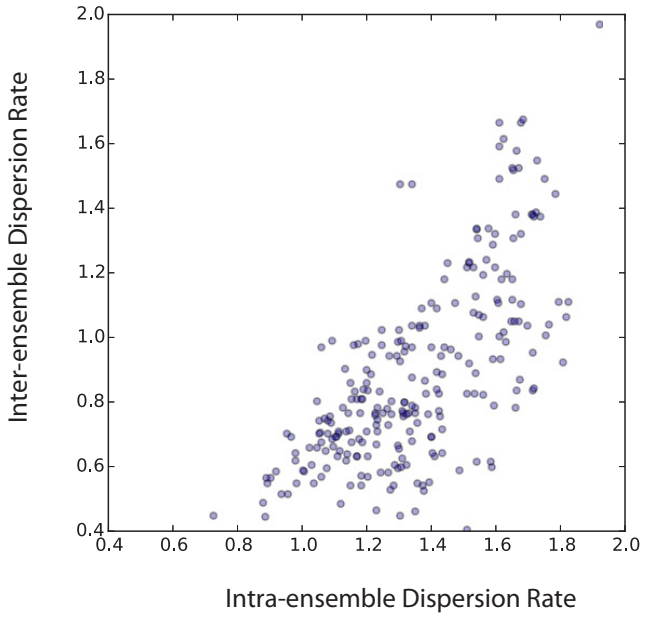

Figure 6. Correlation between the intra-ensemble and inter-ensemble dispersion rates.

DISCUSSIONS AND CONCLUSIONS

One important goal in single-cell analysis is to map the cellular hierarchy within a cell population. Computational based predictions provide important guide for reconstructing lineage relationships which can then be experimentally tested. On the other hand, it must be recognized that such predictions are often associated with a high degree of uncertainty. Our ECLAIR method provides a systematic way to evaluate the uncertainty associated with lineage reconstruction. By comparing with SPADE, a state-of-the-art method for lineage reconstruction, we show that our method has improved the overall robustness and further quantifies the uncertainty for each predicted lineage relationship. The most reliable link may be prioritized for further experimental validation. We recognize that a number of similar methods have been recently developed (6,7,23). However, these methods are intended to reorder cells, without inference of multi-lineage relationships, therefore they are not discussed in our study.

Our ECLAIR analysis still does not entirely remove the variation even when using the same training dataset. However, if the ensemble size were set to a much large size, it follows from Cramer's theorem that the ECLAIR tree con-

verges to a single solution, although the rate of convergence is slow and at odds with most practical purposes (24).

To aid with biological interpretation, we have represented the lineage relationship as a tree, as is commonly done. On the other hand, the full graph contains additional useful information, and it is more reproducible than the MST estimate itself (Supplementary Figure S4). The average correlation of graph weights is 0.96, which is much higher than that for the ECLAIR tree path lengths ($R = 0.86$). This is because the tree structure is sensitive to small perturbations of the underlying data (Supplementary Figure S5). The graph structure may be more suitable for comparison of different studies or subpopulations because of its enhanced reproducibility. Nonetheless, the MST approach still provides an intuitive interpretation as sequential events during cell differentiation. As such, ECLAIR provides both representations as model outputs.

Our ECLAIR method has important limitations. As in other similar methods, we assume that gene expression similarity is regulated by lineage similarity. However, this ignores the contributions of other mechanisms such as spatial organizations. As such, we should only view the predicted cell lineages as one possible hypothesis. More comprehensive experimental data are required to more accurately define cell lineage information.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

Authors' contributions: G.C.Y., E.M., G.G. and L.T. and conceived and designed the computational method. G.G., E.M., S.G. and L.T. implemented the method and analyzed the data. G.G. developed the software package. G.C.Y., L.T. and G.G. wrote the paper. All authors read and approved the final manuscript.

FUNDING

Claudia Barr Award (to G.C.Y.); NIH [R01HL119099 to G.C.Y.]. Funding for open access charge: NIH.
Conflict of interest statement. None declared.

REFERENCES

1. Sandberg,R. (2014) Entering the era of single-cell transcriptomics in biology and medicine. *Nat. Methods*, **11**, 22–24.
2. Saadatpour,A., Lai,S., Guo,G. and Yuan,G.C. (2015) Single-cell analysis in cancer genomics. *Trends Genet.*, **31**, 576–586.
3. Stegle,O., Teichmann,S.A. and Marioni,J.C. (2015) Computational and analytical challenges in single-cell transcriptomics. *Nat. Rev. Genet.*, **16**, 133–145.
4. Qiu,P., Simonds,E.F., Bendall,S.C., Gibbs,K.D. Jr, Bruggner,R.V., Linderman,M.D., Sachs,K., Nolan,G.P. and Plevritis,S.K. (2011) Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. *Nat. Biotechnol.*, **29**, 886–891.
5. Treutlein,B., Brownfield,D.G., Wu,A.R., Neff,N.F., Mantalas,G.L., Espinoza,F.H., Desai,T.J., Krasnow,M.A. and Quake,S.R. (2014) Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature*, **509**, 371–375.
6. Bendall,S.C., Davis,K.L., Amir el,A.D., Tadmor,M.D., Simonds,E.F., Chen,T.J., Shenfeld,D.K., Nolan,G.P. and Pe'er,D. (2014) Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell*, **157**, 714–725.
7. Trapnell,C., Cacchiarelli,D., Grimsby,J., Pokharel,P., Li,S., Morse,M., Lennon,N.J., Livak,K.J., Mikkelsen,T.S. and Rinn,J.L. (2014) The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.*, **32**, 381–386.
8. Marco,E., Karp,R.L., Guo,G., Robson,P., Hart,A.H., Trippa,L. and Yuan,G.C. (2014) Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, E5643–E5650.
9. Wolpert,D.H. and Mecready,W.G. (1997) No free lunch theorems for optimization. *IEEE Trans. Evol. Comput.*, **1**, 67–82.
10. Polikar,R. (2006) Ensemble based systems in decision making. *IEEE Circuits Syst. Mag.*, **6**, 21–45.
11. Strehl,A. and Ghosh,J. (2002) Cluster ensembles { a knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.*, **3**, 583–617.
12. Hartigan,J.A. (1975) *Clustering Algorithms*. Wiley , NY
13. Frey,B.J. and Dueck,D. (2007) Clustering by passing messages between data points. *Science*, **315**, 972–976.
14. Ester,M., Kriegel,H.-P., Sander,J. and Xu,X. (1996) A Density-based algorithm for discovering clusters in large spatial databases with noise. In: Simoudis,E, Han,J and Fayyad,U (eds). *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. AAAI Press, pp. 226–231.
15. Prim,R.C. (1957) Shortest connection networks And some generalizations. *Bell Syst. Tech. J.*, **36**, 1389–1401.
16. Conte,D., Foggia,P., Sansone,C. and Vento,M. (2004) Thirty years of graph matching in pattern recognition. *Int. J. Pattern Recognit. Artif. Intell.*, **18**, 265–298.
17. Guo,G., Huss,M., Tong,G.Q., Wang,C., Li Sun,L., Clarke,N.D. and Robson,P. (2010) Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. *Dev. Cell*, **18**, 675–685.
18. Tibshirani,R., Walther,G. and Hastie,T. (2001) Estimating the number of clusters in a data set via the gap statistic. *J. R. Stat. Soc. B.*, **63**, 411–423.
19. Moignard,V., Woodhouse,S., Haghverdi,L., Lilly,A.J., Tanaka,Y., Wilkinson,A.C., Buettner,F., Macaulay,I.C., Jawaid,W., Diamanti,E. *et al.* (2015) Decoding the regulatory network of early blood development from single-cell gene expression measurements. *Nat. Biotechnol.*, **33**, 269–276.
20. Paul,F., Arkin,Y., Giladi,A., Jaitin,D.A., Kenigsberg,E., Keren-Shaul,H., Winter,D., Lara-Astiaso,D., Gury,M., Weiner,A. *et al.* (2015) Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors. *Cell*, **163**, 1663–1677.
21. Halko,N., Martinsson,P.G. and Tropp,J.A. (2009) Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Rev.*, **53**, 217–288.
22. Bendall,S.C., Simonds,E.F., Qiu,P., Amir el,A.D., Krutzik,P.O., Finck,R., Bruggner,R.V., Melamed,R., Trejo,A., Ornatsky,O.I. *et al.* (2011) Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science*, **332**, 687–696.
23. Shin,J., Berg,D.A., Zhu,Y., Shin,J.Y., Song,J., Bonaguidi,M.A., Enikolopov,G., Nauen,D.W., Christian,K.M., Ming,G.L. and Song,H. (2015) Single-cell RNA-seq with waterfall reveals molecular cascades underlying adult neurogenesis. *Cell Stem Cell*, **17**, 360–372.
24. Topchy,A.P., Law,M.H.C., Jain,A.K. and Fred,A.L. (2004) Analysis of consensus partition in cluster ensemble. In: Rastogi,R, Morik,K, Bramer,M and Wu,X (eds). *Fourth IEEE International Conference on Data Mining*. IEEE, pp. 225–232.