Continuous cell-state density inference and applications for single-cell data

Anonymous Author

Single-cell sequencing continuous to advance our understanding of cell biology, and critical cellular processes such as cell-differentiation. It has been natural to interpret the data as discrete measurements of individual cells and using k-nearest-neighbor graphs to represent the whole population has been a successful computational strategy. However, growing resolution and abundance of single-cell assays and interest to computationally decipher continuous cellular processes[1-4] call for a likewise continuous representations of the cell populations. This encompasses not only the discrete observed states but instead a likelihood of occurrence for all possible cell states enabling even more specialize methods to model this continuity. To this end we have developed scDensity, an algorithm that leverages diffusion-map representation, nearest-neighbor distributions, and Gaussian processes to infer a differentiable function of the cell-state density representing the whole population (Fig.1). scDensity outperforms existing approaches for single-cell density estimations in accuracy, robustness, and resolution for RNA and ATAC modalities. scDensity is computationally efficient and scales to atlas-size single cell datasets. The resulting density function can comprehensibly represent entire cell populations and enable multiple novel downstream applications. This advancement could serve as a new paradigm of single-cell analysis.

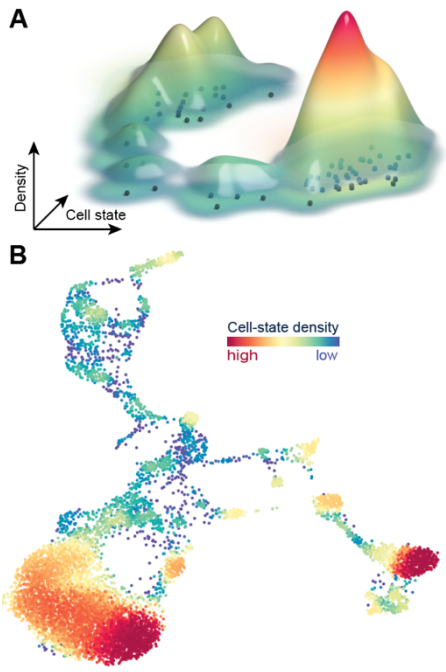

Fig.1 A Continuous cell-state density function over a simulated set of discrete cell states in two dimensions. **B** UMAP of T-cell depleted bone marrow single-cell RNA-seq sample colored by cell-state density.

Algorithm details

We define a model describing the data distribution, so we can use Bayesian inference on its parameters (Fig.2). The central parametrization of the model is a function over the entire cell-state space that describes the logarithm of the cell-state density. To ensure differentiability of this function we employ a Gaussian prior with a Matern covariance function and heuristic to select a fixed length scale. This allows us to sample function values for each observed cell while ensuring that the resulting function is defined even for unobserved cell states, and it links density values between close cells involving multiple samples in the estimation of the local density, increasing its statistical power. A key ingredient is the connection between the density function and the data distribution: Since the cell-state space has high dimensionality, it is intractable to integrate any function over it and to normalize our density function to a probability density. So, instead of treating the measured cell states as samples of the density function, we compute the distance to the nearest neighbor from each cell and use this as a sample from a nearest-neighbor distribution, which we deduced from the Poisson distribution, that connects this distance with the local density. This allows us to fit a density function for higher dimensionalities than previously feasible[5].

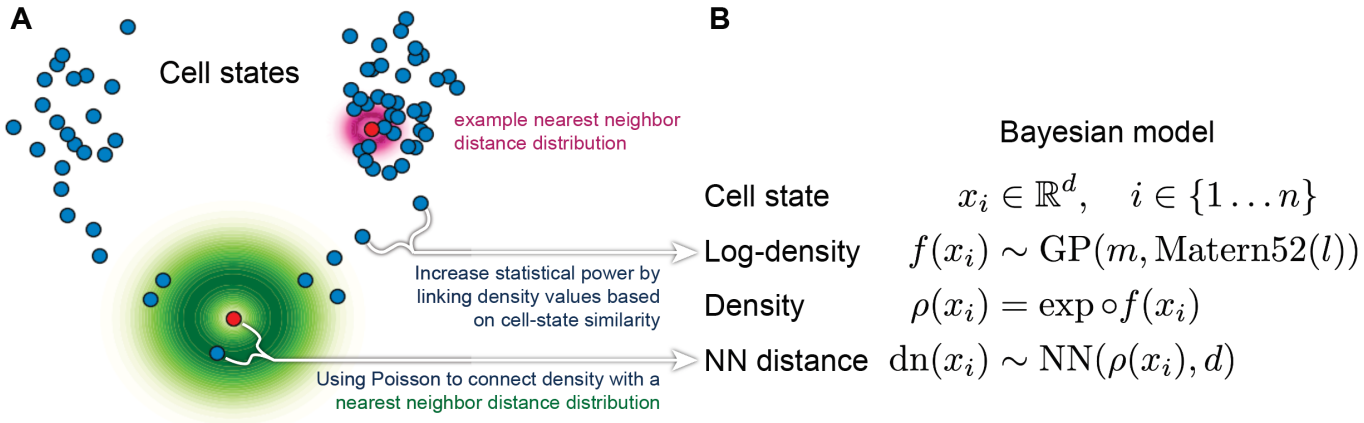

Fig.2 A Cells in 2-dimensional state space with colored nearest neighbor distributions. **B** Bayesian model used to infer the cell-state density.

To improve the scalability of the Gaussian process, we implemented a low rank approximation of the involved covariance matrix. In addition to the rank reduction through k-means inducing points, we apply an improved Nyström approximation[6] to generate density estimates in an extremely efficient manner without sacrificing accuracy. Concretely, scDensity can estimate densities for a dataset of 250k cells in less than 15 minutes. Our implementation using jax[7] allows evaluation of cell-state density for unobserved states, automatic differentiation of the density function, and broad user control including support for other covariance functions.

Modeling relevance

Using a continuous density function as inferred by scDensity overcomes multiple limitations of the traditionally discrete representation for single-cell datasets and cell-differentiation modelling: (i) Instead of only considering cell-state transitions between observed states, any cell-state change can be considered. (ii) Transition rates to different cell states can be informed by density gradients and the Boltzmann equation. (iii) The density quantification can be used to analyses changes in cell prevalence along differentiation trajectories within a dataset or between population of different conditions, allowing conclusions about proliferation and apoptosis rates. (iv) Measurement uncertainty is implicitly encoded in the representation. The smoothness of the density function, which is informed by the data, indicates positive probabilities for the existence of any unobserved cell state. (v) The representation's complexity and size in memory depends on the complexity of the represented cell population and not the number of measured cells. Additional measurements increase the accuracy and reduce uncertainty without directly increasing the cost of specialized downstream applications.

Downstream applications allow automatic detection of rare cell-types and bottlenecks or checkpoints of the cell differentiation process. Furthermore, the differentiable density function enables the application of calculus - We demonstrate this by investigating the effects of the density gradient of cellular state diffusion through the diffusion current that could play an essential role to understand the mechanisms of homeostasis cell populations. Other novel applications include the high-resolution compositional changes of cell populations under different conditions, usage of the Boltzmann equation to compute transition rates between "stable" cell states with high-density, and enrichment of discrete samples through Hamilton Monte Carlo sampling.

Biological relevance

Single-cell studies have indicated that differentiation and disease trajectories are punctuated with regions of variable cell-state density[3,8,9] (Fig.1B). Applied to diverse single-cell datasets, scDensity demonstrates that low-density regions in trajectories are rare intermediate cell-states with critical roles in their biological system. In human hematopoiesis, scDensity revealed that rare cell states are hallmarks of lineage specification and accompanied by upregulation of essential master transcription factors. E.g., common lymphoid progenitors (<0.5% of cells in marrow) emerge from the hematopoietic progenitor pool as a low-density state in the B-cell differentiation trajectory (Fig.3A,C). The emergence is accompanied by localized upregulation of EBF1, the B-cell master regulator (Fig.3B). We further leverage scDensity to identify bottlenecks of iPS differentiation[10], chart the emergence of metastatic cells from mouse models of tumors[11], and detect stem-like populations in regeneration[12], all of which are detected as rare cell states. We demonstrate how cell-state-density enabled single-cell analysis empowers the identification and focus on rare but crucial transitory cell states that are often overlooked in typical single-cell analyses.

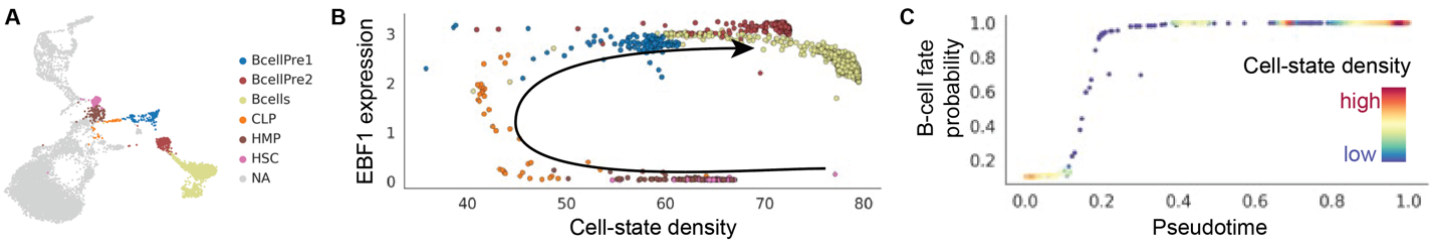

Fig.3 Example data from T-cell depleted bone marrow single-cell RNA-seq experiment (s. Fig.1 B for cell-state density). **A** UMAP embedding of cell states highlighting the B-cell trajectory in color. **B** Cell-state density and EBF1 expression along B-cell trajectory with arrow indicating temporal ordering. **C** Fate probability for cells in B-cell trajectory along pseudotime, colored by cell-state density.

References

1 Qiu, X. *et al.* Reversed graph embedding resolves complex single-cell trajectories. *Nat Methods* **14**, 979-982 (2017). https://doi.org:10.1038/nmeth.4402

2 Setty, M. *et al.* Characterization of cell fate probabilities in single-cell data with Palantir. *Nat Biotechnol* **37**, 451-460 (2019). https://doi.org:10.1038/s41587-019-0068-4

3 Bendall, S. C. *et al.* Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell* **157**, 714-725 (2014). https://doi.org:10.1016/j.cell.2014.04.005

4 Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F. & Theis, F. J. Diffusion pseudotime robustly reconstructs lineage branching. *Nature Methods* **13**, 845-848 (2016). https://doi.org:10.1038/nmeth.3971

5 W, W. & ScottDavid. Nonparametric density estimation for high-dimensional data—Algorithms and applications. (2019). https://doi.org:WICS1461

6 Pourkamali-Anaraki, F. & Becker, S. Improved fixed-rank Nyström approximation via QR decomposition: Practical and theoretical aspects. *Neurocomputing* **363**, 261-272 (2019). https://doi.org:https://doi.org/10.1016/j.neucom.2019.06.070

7 Bradbury, J. *et al.* (2018).

8 Plasschaert, L. W. *et al.* A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte. *Nature* **560**, 377-381 (2018). https://doi.org:10.1038/s41586-018-0394-6

9 Gonzalez-Gonzalez, F. J. *et al.* Single-Cell Transcriptomic Analysis of Human Lung Provides Insights into the Pathobiology of Pulmonary Fibrosis. *American journal of respiratory and critical care medicine* **199** (2019). https://doi.org:10.1164/rccm.201712-2410OC

10 Schiebinger, G. *et al.* Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming. *Cell* **176**, 928-943.e922 (2019). https://doi.org:10.1016/j.cell.2019.01.006

11 LaFave, L. M. *et al.* Epigenomic State Transitions Characterize Tumor Progression in Mouse Lung Adenocarcinoma. *Cancer Cell* **38**, 212-228.e213 (2020). https://doi.org:10.1016/j.ccell.2020.06.006

12 Strunz, M. *et al.* Alveolar regeneration through a Krt8+ transitional stem cell state that persists in human lung fibrosis. *Nat Commun* **11**, 3559 (2020). https://doi.org:10.1038/s41467-020-17358-3