# Revealing lineage-related signals in single-cell gene expression using random matrix theory

Mor Nitzan[a,b,1] and Michael P. Brenner[b]

[a]School of Computer Science and Engineering, Racah Institute of Physics, Faculty of Medicine, The Hebrew University of Jerusalem, 9190401 Jerusalem, Israel; and [b]School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138

Edited by Peter J. Bickel, University of California, Berkeley, CA, and approved January 28, 2021 (received for review August 13, 2019)

**Gene expression profiles of a cellular population, generated by single-cell RNA sequencing, contains rich information about biological state, including cell type, cell cycle phase, gene regulatory patterns, and location within the tissue of origin. A major challenge is to disentangle information about these different biological states from each other, including distinguishing from cell lineage, since the correlation of cellular expression patterns is necessarily contaminated by ancestry. Here, we use a recent advance in random matrix theory, discovered in the context of protein phylogeny, to identify differentiation or ancestry-related processes in single-cell data. Qin and Colwell [C. Qin, L. J. Colwell, *Proc. Natl. Acad. Sci. U.S.A.* 115, 690–695 (2018)] showed that ancestral relationships in protein sequences create a power-law signature in the covariance eigenvalue distribution. We demonstrate the existence of such signatures in scRNA-seq data and that the genes driving them are indeed related to differentiation and developmental pathways. We predict the existence of similar power-law signatures for cells along linear trajectories and demonstrate this for linearly differentiating systems. Furthermore, we generalize to show that the same signatures can arise for cells along tissue-specific spatial trajectories. We illustrate these principles in diverse tissues and organisms, including the mammalian epidermis and lung, *Drosophila* whole-embryo, adult *Hydra*, dendritic cells, the intestinal epithelium, and cells undergoing induced pluripotent stem cells (iPSC) reprogramming. We show how these results can be used to interpret the gradual dynamics of lineage structure along iPSC reprogramming. Together, we provide a framework that can be used to identify signatures of specific biological processes in single-cell data without prior knowledge and identify candidate genes associated with these processes.**

single-cell data | cellular lineage | random matrix theory | spectral analysis

Advances in single-cell RNA sequencing (scRNA-seq) have led to the identification of diverse and heterogeneous cellular populations (1–3). Sampling a large number of nonsynchronized single cells also enables reconstruction of tissue structure (4–6), the cell cycle (7), and differentiation pathways (reviewed in refs. 8 and 9). Trajectory inference in particular, including inferring developmental trajectories or lineage relationships between cells, is a fast-moving research field and includes dozens of methods that have been proposed to approach this challenge [e.g., in refs. 10–17 and recent benchmarking analysis (9)]. In general, these inference approaches rely on the assumption that single-cell expression profiles span a low dimensional manifold in high dimensional expression space, which represents an underlying biological process and the progression, or "phase," of single cells along that process. Therefore, in principle, it is possible to both reconstruct the process and recover the locations of single cells along it. Using rich single-cell data, it has also been possible to infer regulatory interaction networks within single cells and infer their variation across different perturbations and conditions [e.g., refs. 11 and 18 and recent benchmarking study (19)]. However, the biological picture is more complicated: The expression profile of each cell contains multiple signatures, simultaneously encoding information about its location within a tissue and its microenvironment, multiple temporal processes, and differentiation pathways as well as internal patterns of regulatory interactions, along with technical or experimentally driven signals (2). Despite recent advances (e.g., refs. 20–25), disentangling these different biological processes from each other and from signals attributed to the experimental procedure in single-cell expression data is still a major challenge.

The key insight of this paper is that different biological processes may exhibit different geometric, or topological, structures in gene expression space, which induce distinct patterns in the covariance spectrum of single-cell data (which can relate both different cells and different genes to each other). While biological processes such as differentiation, the cell cycle, and spatial relationships induce correlations between different cells, regulatory interactions induce correlations between different genes. This is directly analogous to protein sequence data that exhibit both phylogeny-driven correlations between different sequences as well as correlation between different amino acids in individual sequences arising from physical structural interactions. Qin and Colwell (26) demonstrated that phylogenetic correlations between sequences lead to a characteristic signature in the eigenvalue distribution of the sequence covariance matrix, specifically a power-law tail of large eigenvalues, in which the exponent of the power law is determined by the underlying branching process.

Here, we demonstrate that this same phenomenon occurs in scRNA-seq data, in which lineage or developmental signals give rise to unique covariance structures, reflecting either cell-to-cell or gene-to-gene correlations between the vectors representing the gene expression profiles for each cell or between the vectors representing the expression of each gene across all cells,

## Significance

**Single-cell technologies are rapidly advancing, allowing us to gauge the heterogeneity and structure of cellular communities, tissues, and full organisms. The correlations between genes and between cells within such systems can reveal patterns of regulatory interactions, physical structure, and temporal progression of cells along biological processes. However, it is generally a challenge to identify and tease apart these mixed signals within the noisy, high-dimensional single-cell data. Here, we show it is possible to detect a signature for lineage in the covariance spectrum of single-cell data, predict how it will change with developmental time, and predict how it can be extended to examine the spatial structure of a tissue.**

Author contributions: M.N. and M.P.B. designed research, performed research, analyzed data, and wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.



[1]To whom correspondence may be addressed. Email: mor.nitzan@mail.huji.ac.il.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1913931118/-/DCSupplemental.

Published March 8, 2021.

respectively. Specifically, we find power-law tails of large eigenvalues in the spectra of the gene–gene covariance matrix of the data. We demonstrate this phenomenon for various types of single-cell data, including mammalian lung and epidermis, fly whole embryo, whole *Hydra*, and cells undergoing induced pluripotent stem cells (iPSCs) reprogramming. We show that single yeast cells do not exhibit such signature for lineage. We show that power-law tails of large eigenvalues are also expected to arise for populations of cells along linear trajectories in differentiating systems and demonstrate such signal for subpopulations of dendritic cells (DCs). Furthermore, we generalize these findings to populations of cells whose expression correlations are induced by their spatial organization within a tissue, which may be a result, for example, of external gradients of oxygen, morphogenes or nutrients, patterns of cellular movement, or cell–cell communication. Specifically, we exemplify our findings in the case of enterocytes along the crypt-to-villus axis in the intestinal epithelium. We support these findings by showing that the genes driving the power-law covariance patterns are indeed related to lineage and developmental processes. Finally, we show that we can predict the dynamically changing spectral signatures of differentiating cells sequenced at multiple time points, such as in iPSC reprogramming.

## Results

Single-cell data generated by scRNA-seq is composed of expression profiles of single cells, in which each profile is a vector of expression levels for each gene in the cell. Correlations within this data can arise both from regulatory interactions between different genes and from interactions and structural or dynamic relationships between different cells. These can result from direct cell-to-cell interactions, the relative location of cells within a tissue, the relative progression of cells through temporal processes such as the cell cycle or differentiation pathways, responses to environmental cues, and other diverse biological functions. These different types of correlations, or covariance structures, each contribute to the overall covariance of the single-cell data. Here, we show that differentiation and developmental processes generate a distinct covariance signature due to their hierarchical structure in gene expression space, which we can thus identify in single-cell data.

**Lineage Is Predicted to Generate Power-Law Patterns in Single-Cell Spectra.** To set the stage, we consider a simple model for single-cell lineage progression. Here, each bifurcating trajectory of single cells, originating, for example, from stem cells and ending in terminally differentiated cells, is constructed as follows (*SI Appendix*): a random binary expression profile of length $p$ is chosen as the root (e.g., stem cell) so that the expression level of each of the $p$ genes is approximated as being "ON" or "OFF." The expression profile then goes through a smooth differentiation process. The profile is first randomly changed $m$ times, such that each of the $m$ steps consist of a change in the state of a single gene (from ON to OFF or vice versa). It then goes through a developmental bifurcation, meaning that the profile is duplicated, and each branch goes independently through $m$ additional changes. The cells go through $b$ such bifurcations until they reach their $n$ terminally differentiated states. Note that in principle, $m$ is not necessarily fixed for the whole tree. This model is designed to emulate the differentiation process of a population of single cells, as is reflected by a scRNA-seq dataset of unsynchronized single cells. We are interested in uncovering signals related to lineage in the gene–gene covariance matrix of the single-cell data, $XX^T/n$, where $X$ is a gene-by-cell data matrix ($X_{i,j}$ indicates the expression of gene $i$ in cell $j$).

This model is formally analogous to that of Qin and Colwell (26), who consider genetic mutations in a protein phylogeny, where in that case $m$ is the number of genetic mutations and $p$ is the number of amino acids. They demonstrate that the eigenvalue distribution of the covariance matrix of the "leaves" of the phylogeny (in our case, corresponding to terminally differentiated cells) has a power-law structure, $\lambda \sim r^{-\log(2\alpha)/\log(2)}$, for each eigenvalue $\lambda$, in which $\alpha = \exp(-\frac{4m}{p})$ and $r > 1$ is the eigenvalue rank (*SI Appendix*). We found such power-law patterns in a synthetic single-cell dataset, which we generated according to the differentiation model described above, where the eigenvalues of the covariance matrix as a function of their rank are well fitted by the model's prediction (Fig. 1, *Left*). The power-law tails of large eigenvalues are a general feature of lineage structures and generalize to more realistic models, in which gene expression is not binary and the number and extent of changes in gene expression along each of the lineage branches is not identical (*SI Appendix*, Fig. S1*A*). In fact, the power-law tails can be analytically shown to emerge for inhomogeneous lineages and general tree topologies (26). However, there is a deviation from the predicted power-law tail and its slope can change when different genes have nonuniform probabilities for changing their state along the branches (*SI Appendix*, Fig. S1*B*). Single-cell data many times contain cells not only at terminally differentiated states (as the basic model refers to) but cells that are sampled along the entire differentiation process. In such cases, we empirically find that the eigenvalue distribution of the covariance structure still resembles the expected power-law tail of large eigenvalues, while the deviation from the power-law tail starts at larger eigenvalues relative to the basic model, including only "leaf" (terminally differentiated) cells (*SI Appendix*, Fig. S1*C*).

Finally, the emergence of eigenvalue power laws persists for more realistic dynamic bifurcating trajectories of single cells, driven by changes in functional expression programs, such that the sampling of cells and RNA expression counts and their associated noise statistics resemble those of scRNA-seq data [*SI Appendix*, Fig. S1*D*, simulated using a framework for probabilistic simulation of single-cell RNA-seq tree-like topologies (27)].

We found that the power-law pattern does not arise in the covariance spectra of some basic null (lineage-free) models for single-cell gene expression. The simplest (admittingly too simple) null model for gene expression is that of independent cells and genes. The null covariance structure of $n$ single-cell profiles composed of $p$ genes, that do not exhibit any correlations, is given by a central result in random matrix theory, the Marcenko–Pastur (MP) distribution (28): $f(\lambda) = \frac{\sqrt{(b_+ - \lambda)(\lambda - b_-)}}{2\pi c\lambda}, b_\pm = (1 \pm \sqrt{c})^2$, where $c = n/p$. The universality of this eigenvalue distribution extends beyond the binary distributions we consider in this section and applies to matrices whose entries are independent and identically distributed random variables with zero mean and finite variance. Fig. 1 (*Middle*) shows how the MP law emerges for a simulated population of uncorrelated single-cell expression profiles. In contrast to the differentiation model (with a power-law tail of large eigenvalues), a group of independent expression profiles yields similarly sized large eigenvalues, which results in a flat (zero slope) line of eigenvalues versus their rank.

Going beyond independent gene expression profiles, we note that the effects of batch structure or clustering, as common features of single-cell data, do not generate by themselves power-law tails in the eigenvalue distributions (*SI Appendix*, Fig. S2*A*). Another correlation structure in single-cell data is imposed by gene regulatory interactions. To reason about the effects of such regulatory patterns, we use a simple energy-based model to describe the gene–gene interactions derived from the cellular regulatory network (*SI Appendix*). This family of models was previously used to model and infer the structure of gene regulatory networks (e.g., refs. 29–32) and is analogous to the model used for phenotypic interactions between amino acids in protein

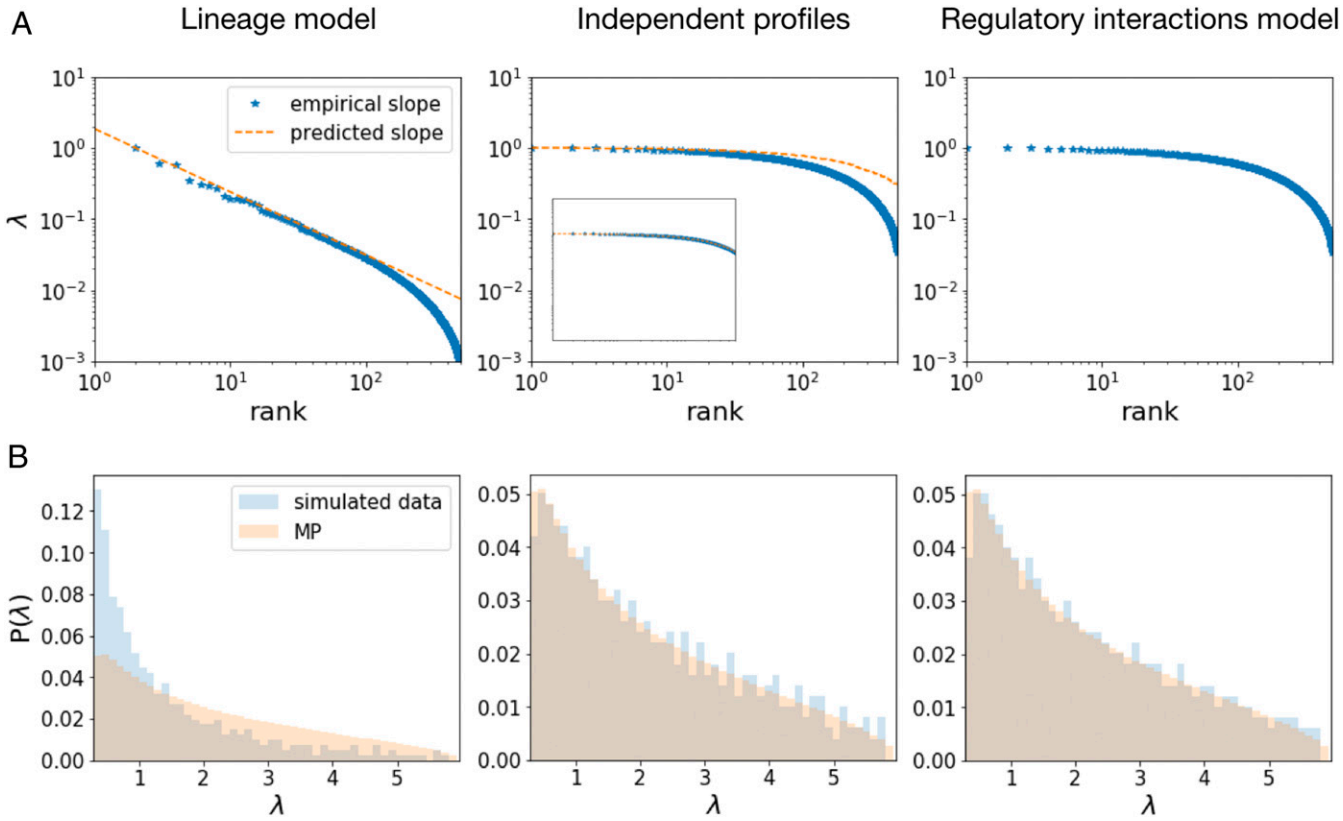


**Fig. 1.** Eigenvalue patterns expected for the covariance matrix of different forms of single-cell data. The eigenvalues of the covariance matrix as a function of their rank (r) (*A*) and the eigenvalue distribution (*B*) generated by the differentiation model (*Left*), a population of independent expression profiles (*Middle*), and cells correlated by regulatory interactions of their genes (*Right*). The differentiation and regulatory interactions models are described in the main text for a population of binary expression profiles. The empirical slope of the eigenvalue versus rank (blue stars) is compared to the predicted slope (dotted orange line) for the lineage model (for $r > 1$) and to the MP-based slope for the independent model. The eigenvalue distributions for the different models are compared to the MP distribution. Parameters for all simulations: number of cells: $2^{10}$; number of genes: 500 (2,000 genes in the independent profiles inset); number of changes in the expression profile along each branch for lineage model: 10; and the gene interaction matrix for the regulatory model is set by assigning a probability of 0.1 for every pair of genes, independently, to interact. The strength of interaction is uniformly sampled from −1,1: The expression profiles for the regulatory interaction model are each evolved 1,000 times according to the Potts model (*SI Appendix*).

sequences (26). Consistent with the results of ref. 26, we show numerically in our context that the eigenvalue distribution of the sample covariance matrix is described well by the MP distribution and the slope of the eigenvalues as a function of their rank is ∼0 (Fig. 1, *Right*). The results are consistent for varying fractions of interactions and different network structures (*SI Appendix*, Fig. S3).

Finally, we generated synthetic scRNA-seq datasets, which capture many features observed in real single-cell data, including high expression outlier genes, differing sequencing depths between cells, and technical dropouts (zero inflation). In short, building on the Splatter statistical framework (33), we generate a synthetic scRNA-seq dataset by a process based on a gamma-Poisson hierarchical model, in which the mean expression level of each gene is sampled from a gamma distribution, and the count that is measured experimentally for each cell is then sampled from a Poisson distribution, in a way that is regularized by expected statistical features of scRNA-seq data. We find that this more complex, realistic scenario as well, without lineage, does not result in power-law spectral patterns (*SI Appendix*, Fig. S2*B*).

**Detection of Power-Law Patterns in Single-Cell Spectra.** Following these analytical and numerical results, the striking covariance structure associated with developmental and differentiation processes of a population of single cells suggests that we should be able to identify it in single-cell sequencing data and further identify genes that are driving these processes. Specifically, we expect the distribution of covariance eigenvalues to substantially deviate from a corresponding MP distribution for single-cell datasets with an underlying lineage signal, with the large eigenvalues following a power-law distribution as a function of their rank. We tested this prediction for a variety of tissues, organisms, and conditions. These included the following:

- The mouse epidermis, forming the outer layer of the mammalian skin. Because of the regenerative capacity of the epidermis, it contains cells along all stages of the differentiation process, from stem cells to terminally differentiated cells. We analyzed a single-cell dataset of the mouse epidermis, which was previously shown to include the whole hierarchy of differentiating cells (34).
- The mouse lung. We chose a single-cell dataset shown to include traces of the cellular hierarchy of the distal mouse lung epithelium (35).
- Reprogramming of mouse embryonic fibroblasts into iPSC. We focused on a dataset which includes single cells sequenced at half-day intervals across 18 d of the reprogramming process (36).
- *Hydra*, a cnidarian polyp which renews all cells of its body throughout its life. We analyzed a single-cell dataset of whole

adult *Hydra*, which includes cellular lineages ranging from stem cells to progenitors, and terminally differentiated cells (37).
- The fly embryo at the onset of gastrulation. We analyzed a dataset including single cells collected from thousands of individual *Drosophila* embryos at this early developmental stage, when the embryo consists of only a few thousands of cells (6).
- Budding yeast. This dataset includes diverse *Saccharomyces cerevisiae* cells sequenced in different environmental conditions, in which we focused on the wild-type strain grown in standard yeast rich media (38).

Further details of the single-cell datasets we analyzed can be found in *SI Appendix*. When analyzing the scRNA-seq datasets (*SI Appendix*) associated with biological systems expected to encode signals of differentiation and developmental pathways, including the epidermis, lung, iPSC reprogramming, *Hydra*, and fly datasets, we find that the large eigenvalues of the gene–gene covariance matrix as a function of their rank approximately follow a power-law tail, as predicted (Fig. 2). In addition, the corresponding eigenvalue distributions are statistically significantly different from their respective MP distributions (Fig. 2, *Insets*, statistical analysis is elaborated in *SI Appendix*). The yeast single cells, on the other hand, generate a distinct spectral signature from the rest, showing no detectable lineage traces, which could potentially point to low mother–daughter expression correlation or to rapidly diminishing correlations along ancestry paths. When the cell–cell correlation structure of the data are destroyed by randomly permuting each of the single-cell datasets over its cells for every gene independently, the resulting eigenvalue distributions resemble the respective null MP distributions, and the power-law patterns in the eigenvalues of the covariance matrices are lost (*SI Appendix*, Figs. S4 and S5), as expected.

**Distinct Spectral Features of Neighboring Cells.** We next aim to contrast groups of cells capturing lineage information with groups of cells that are more homogeneous, while keeping the groups biologically comparable and retaining many of the natural features and correlations in the original single-cell datasets. We do so by subsampling scRNA-seq datasets to examine eigenvalue signatures generated by groups of neighboring cells relative to groups of randomly selected cells (containing the same number of cells). To do this for a given dataset for a group of size $k$, we choose a cell at random and then select its $k$ nearest neighbors based on the pairwise Euclidean distance between all cells in gene expression space, reduced to the top variable genes (*SI Appendix*). We find that the eigenvalues of the covariance matrices of the neighboring cells diverge from the randomly selected cells (Fig. 3 and *SI Appendix*, Fig. S6). The large eigenvalues of the neighboring cells, as expected, follow a curve with smaller absolute slope (Fig. 3, $P$ value $< 0.05$ for all datasets reported). In addition, the largest eigenvalue associated with randomly selected cells is greater than that of the neighboring cells (Fig. 3, *Insets*, $P$ value $< 0.05$ for all datasets reported). This may be expected as the number of effective lineage bifurcations associated with the neighboring cells is smaller than that of the randomly selected cells (which sample the whole tree) and the largest eigenvalue scales exponentially with the number of bifurcations according to the lineage model (*SI Appendix*). For

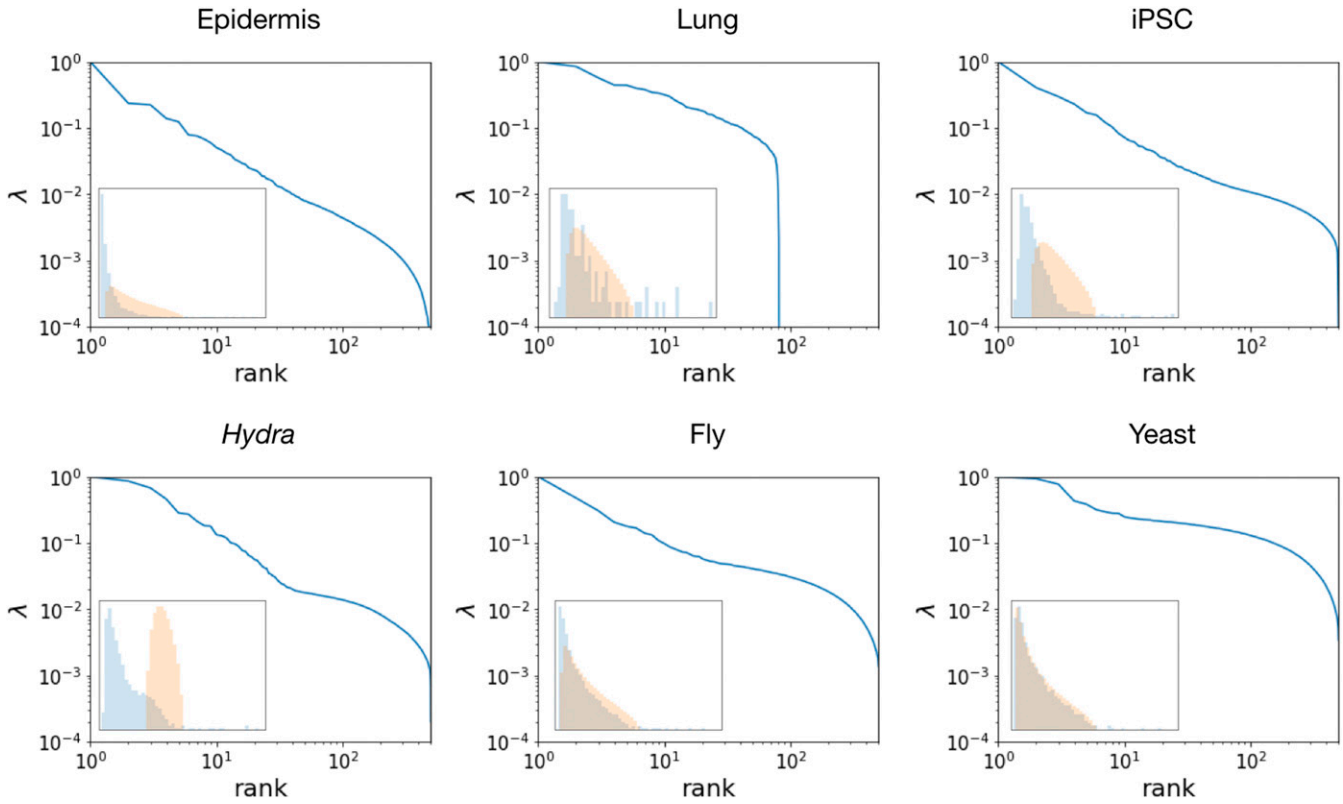


**Fig. 2.** The large eigenvalues as a function of rank of single-cell datasets carrying lineage signal follow a power-law tail. We can detect power-law tails in the large eigenvalues of the sample covariance matrices of single-cell datasets collected from the mouse epidermis and lung, fly embryo, whole *Hydra*, and cells undergoing iPSC reprogramming. Specifically, on a log–log scale, the large eigenvalues fit a line with respect to the rank of the eigenvalues. Yeast colonies do not exhibit such power-law statistics. Each set of eigenvalues is normalized by the largest respective eigenvalue. (*Insets*) The eigenvalue distributions (blue histograms) of all systems but yeast are statistically significantly different from their respective MP distributions (orange histograms). Statistical analysis comparing the single-cell and MP distributions is elaborated in *SI Appendix*.

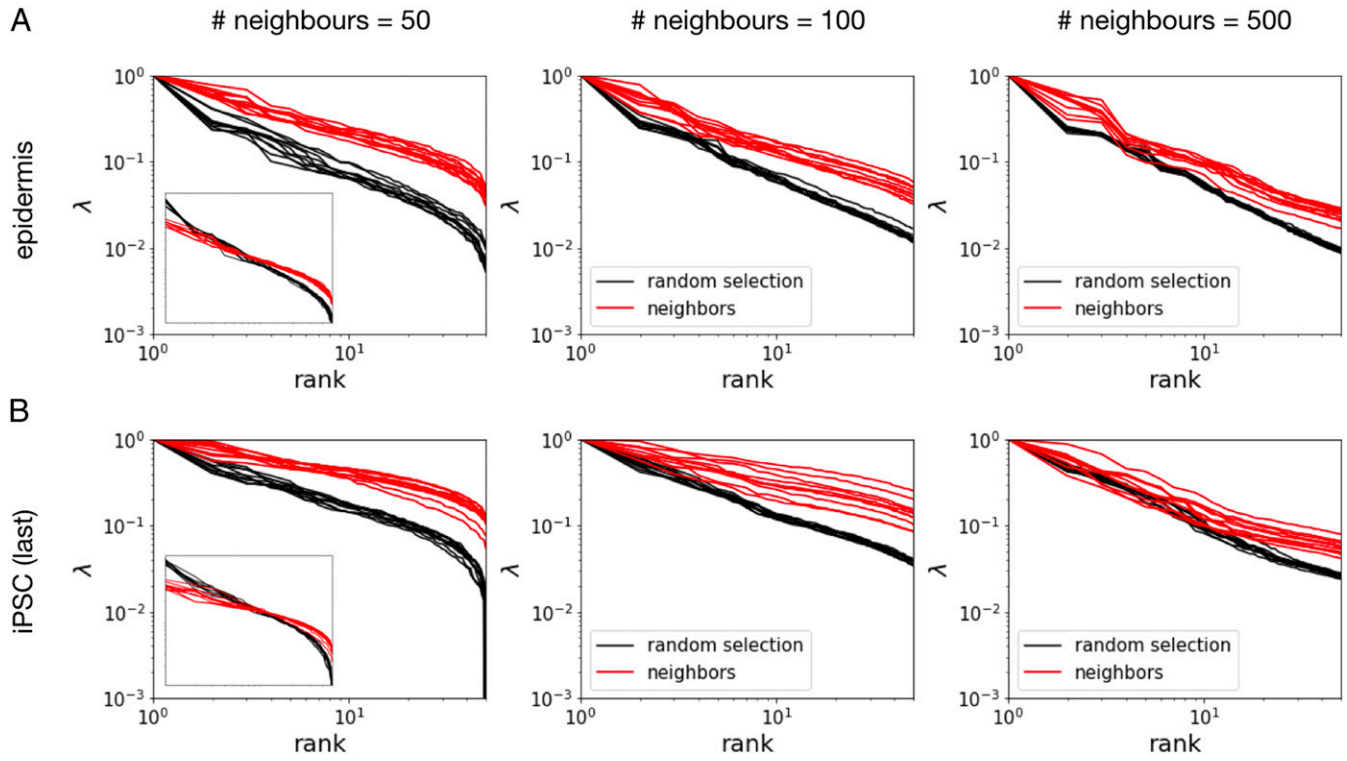


**Fig. 3.** Distinction in behavior of the eigenvalues of single-cell covariance matrices versus their rank between groups of neighboring cells and randomly selected cells. Results are shown for the mouse epidermis (*A*) and the last time point of the iPSC experiment (*B*) for neighborhoods of 50, 100, and 500 cells. Shown are 10 realizations for groups of neighboring cells (red) and randomly selected cells (black). Each set of eigenvalues is normalized by the largest respective eigenvalue. The KS statistic and its associated *P* value for comparing the distributions of slopes (linear fit on a log–log plot) of neighboring versus randomly selected cells are (1.0, 1.5e-45), (1.0, 1.5e-45), and (0.9, 1.9e-39) for the epidermis for 50, 100, and 500 cells; (1.0, 5.3e-42), (1.0, 1.2e-44), and (0.9, 2.3e-35) for the iPSC dataset for 50, 100, and 500 cells (based on 100 realizations). The ratio of slope values of neighboring cells relative to randomly selected cells are (0.63, 0.70, 0.89) for the epidermis and (0.60, 0.64,0.85) for the iPSC dataset for 50, 100, and 500 cells. *Insets*: unnormalized eigenvalues versus rank. The KS statistic and its associated *P* value for comparing the distributions of largest eigenvalues of neighboring versus randomly selected cells are (1.0, 1.2e-44) for the epidermis and (1.0, 1.6e-45) for the iPSC dataset (based on 100 realizations).

example, for the epidermis data, a linear fit on a log–log plot for the ranked normalized eigenvalues has an average slope −0.7 for groups of 50 neighboring cells, relative to an average slope of −1.1 for groups of 50 randomly selected cells (Kolmogorov–Smirnov [KS] statistic: 1.0, *P* value: 1.5e-45 based on 100 realizations). For the same epidermis data, the largest eigenvalue of neighboring cells is 6.48 on average, across realizations, and is 14.38 on average for randomly selected cells (KS statistic: 1.0, *P* value: 1.2e-44 based on 100 realizations). In addition, as the neighborhood size grows and the groups of neighboring cells capture more of the overall structure of the single-cell data (e.g., multiple branches in a lineage trajectory), the spectra of groups of neighboring cells grows to resemble the features of the corresponding randomly selected cell groups; the ratio of slope values of neighboring cells relative to randomly selected cells increase from 0.63 to 0.89 for the epidermis and from 0.60 to 0.85 for the iPSC dataset, going from neighborhoods of 50 to 500 cells (Fig. 3).

**Genes Dominating the Power-Law Domain Are Enriched for Lineage-Related Biological Processes.** To support the hypothesis that the patterns we discover in the large eigenvalues of the covariance structures of the different single-cell datasets are driven by differentiation and developmental processes, we set out to characterize the genes dominating these patterns and, specifically, the high modes of the covariance matrix of the data (i.e., the eigenvectors corresponding to the large eigenvalues). We use gene ontology (GO) enrichment analysis (39, 40) to identify the biological processes that are statistically significantly enriched in the modes corresponding to the power-law regime (*SI Appendix*). We found that the genes dominating the high modes of the datasets we could test for, in which the number of cells is greater than the number of variable genes, were enriched for GO terms related to lineage relationships, including development, morphogenesis, differentiation, and proliferation processes (*SI Appendix*). Specifically, in the iPSC reprogramming dataset, we find enrichment for development processes (top 300 modes); the fly dataset is enriched for differentiation, development, and morphogenesis processes (top 30 modes); and the epidermis dataset is enriched for differentiation, development, and proliferations processes (top 300 modes). While we chose the number of modes approximately corresponding to the power-law regime, the GO enrichment itself is independent of the way the eigenvalues are distributed. We note, however, that for the epidermis and iPSC datasets, the top 10 modes are not enriched for any of the corresponding lineage-related processes listed above, which suggests that the lineage enrichment is not driven by the top few modes (the focus of standard analysis pipelines) and that the broader power-law regime plays an important role. This finding does not only provide support for our analysis but also suggests a path for discovery of genes related to such hierarchical biological processes based on single-cell data.

**Power-Law Patterns Generated by Linear Trajectories.** In this section, we show that a hierarchy of shared covariance structure, and corresponding power-law signatures in the eigenvalue

distribution, can also arise for groups of cells encoding a linear correlation structure, which may be a result, for example, of spatial relationships or a linear differentiation process. Consider a simple model for a linear lineage structure, where a random expression profile of $p$ genes, $X \in \{-1,1\}^p$, is chosen as the root of the lineage, where the state of each gene is either ON or OFF. Then, at every subsequent time step, a new expression profile is formed by flipping the state of a single gene, which is chosen at random (from ON to OFF or vice versa). The eigenvalue distribution of the covariance matrix of such linear sequence of expression profiles forms a power-law tail of large eigenvalues, similarly to the bifurcating lineage structure (Fig. 4*A*). The same power-law tail emerges when simulating a more complex, realistic model for a linear trajectory of single cells suited to the sampling and noise statistics of scRNA-seq (*SI Appendix*, Fig. S7). Therefore, we expect cells that are sequenced along a linear lineage trajectory to generate corresponding power-law structures. To test this, we examined a set of mouse bone marrow DCs progenitors, differentiating from macrophage DC progenitors to common DC progenitors and then to pre-DCs and sequenced as single cells (41). We find that, indeed, the DC spectra follows the predicted power law of large eigenvalues (Fig. 4*B*).

Our linear model of gene expression trajectory was in fact not specific to lineage relationships. Therefore, we expect the existence of power-law tails of large eigenvalues to generalize to other biological scenarios that generate correlations between single cells along a linear structure. Such linear relationships can be induced, for example, by spatial structures exhibiting one-dimensional symmetry. This is the case in the intestinal epithelium, which is composed of repeating crypt-to-villus units, where cells originate at the bottom of the crypt and gradually migrate along the villus axis until they shed off at the tip. It was shown that enterocytes gradually change their expression profiles along the one-dimensional villus axis (42, 43). Indeed, when examining a single-cell dataset of enterocytes sequenced along the crypt-to-villus axis (44), we find a power-law signature in the large eigenvalues of the covariance matrix of the single-cell data (Fig. 4*C*). As above, the eigenvalue structure of single cells inferred (42) to originate from the same spatial zone along the crypt-to-villus axis is distinct from the eigenvalue structure of groups of randomly selected cells (*SI Appendix*, Fig. S8).

**Dynamics of Covariance Structure.** There are several predictions by the model that can be made for a dynamically proliferating and differentiating population of cells (*SI Appendix*). Following in time a population of cells starting from a single-cell type, which progressively differentiates into various cell types, would reflect an increase in the number of effective branching events in the cellular tree. As we progress in developmental time, we would expect 1) the eigenvalue distribution of the covariance matrix to progressively resemble a power law, as the power-law tail induced by the tree structure in the data relies on the number of branching events being sufficiently large (*SI Appendix*), and 2) the largest eigenvalue to increase, as it is predicted to grow exponentially with the number of branching events (*SI Appendix*). It is also of interest to examine the change in the slope of eigenvalues versus rank, which is a function of $m$, the effective (mean) length of the branches (*SI Appendix*) but can potentially also reflect other structural features of the differentiation process. To study this, we analyzed the results of an experiment in which fibroblasts were reprogrammed to iPSCs, in which single-cell sequencing data were collected for ∼251,000 cells, each characterized by an expression profile composed of ∼19,000 genes, at 39 time points over a period of 18 d (36). The single cells on day 0 included only mouse embryonic fibroblasts, while on day 18 they included a wide distribution over multiple differentiated cell types (*SI Appendix*, Fig. S9). As we expected, we find that the spectral structure of the single-cell data gradually changes over the course of the experiment (Fig. 5). The largest eigenvalue and absolute slope value gradually increase (Fig. 5 *A*–*C*). In addition, the eigenvalue distribution deviates from the null MP distribution more substantially as time progresses (Fig. 5 *D*–*F*, see *SI Appendix* for statistical analysis). This deviation is again lost when cell–cell correlations are diminished by permuting the data over cells (Fig. 5 *D* and *E*, *Insets*).

## Discussion

In this paper, we presented a random matrix theory-based approach to identify lineage effects from scRNA-seq data, without relying on gene-specific prior knowledge. We showcased these predictions on single cells sequenced from the mouse epidermis, mouse lung, whole *Hydra*, *Drosophila* embryo, and cells undergoing iPSC reprogramming. Furthermore, we showed that we could use this approach to identify groups of genes enriched for lineage or developmental processes. The de novo identification of genes dominating such hierarchical processes could also enhance the inference of more accurate differentiation or developmental trees based on single-cell data, which is an ongoing challenge in the field (10–17). Power-law tails of large eigenvalues are not only a characteristic of hierarchical structures, such as lineage but can also arise along linear trajectories. This notion led to the identification of power-law signatures in

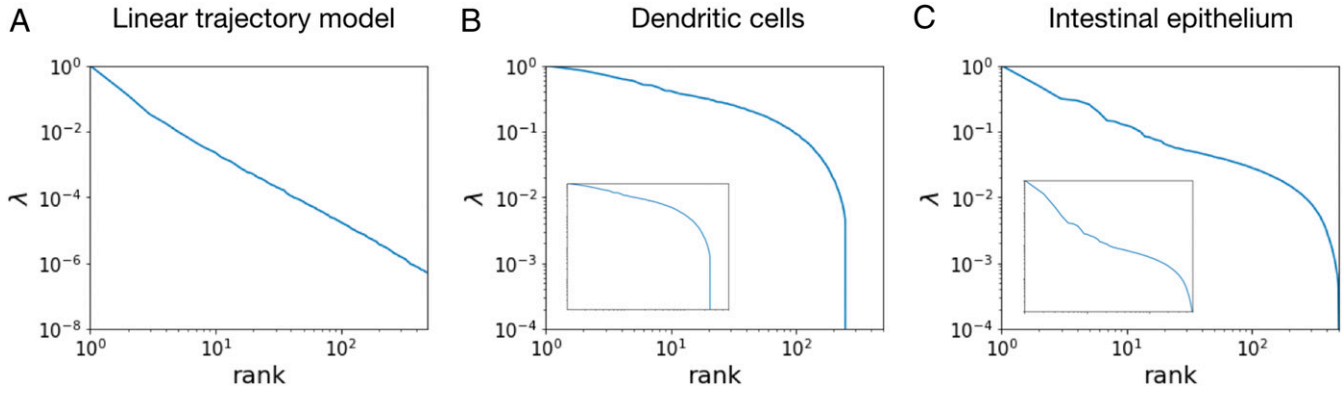


**Fig. 4.** The large eigenvalues of the covariance matrices of single cells along a linear trajectory follow a power-law tail. Eigenvalues versus their rank for (*A*) a linear trajectory model of single cells gradually changing their state along a one-dimensional axis, (*B*) linear differentiation trajectory of DCs from macrophage DC progenitors to common DC progenitors to pre-DCs, and (*C*) cells along the crypt-to-villus axis in the intestinal epithelium. In *B* and *C*, we show the ranked eigenvalues starting from the third eigenvalue and in the insets the whole set of ranked eigenvalues. Each set of eigenvalues is normalized by the largest respective eigenvalue.

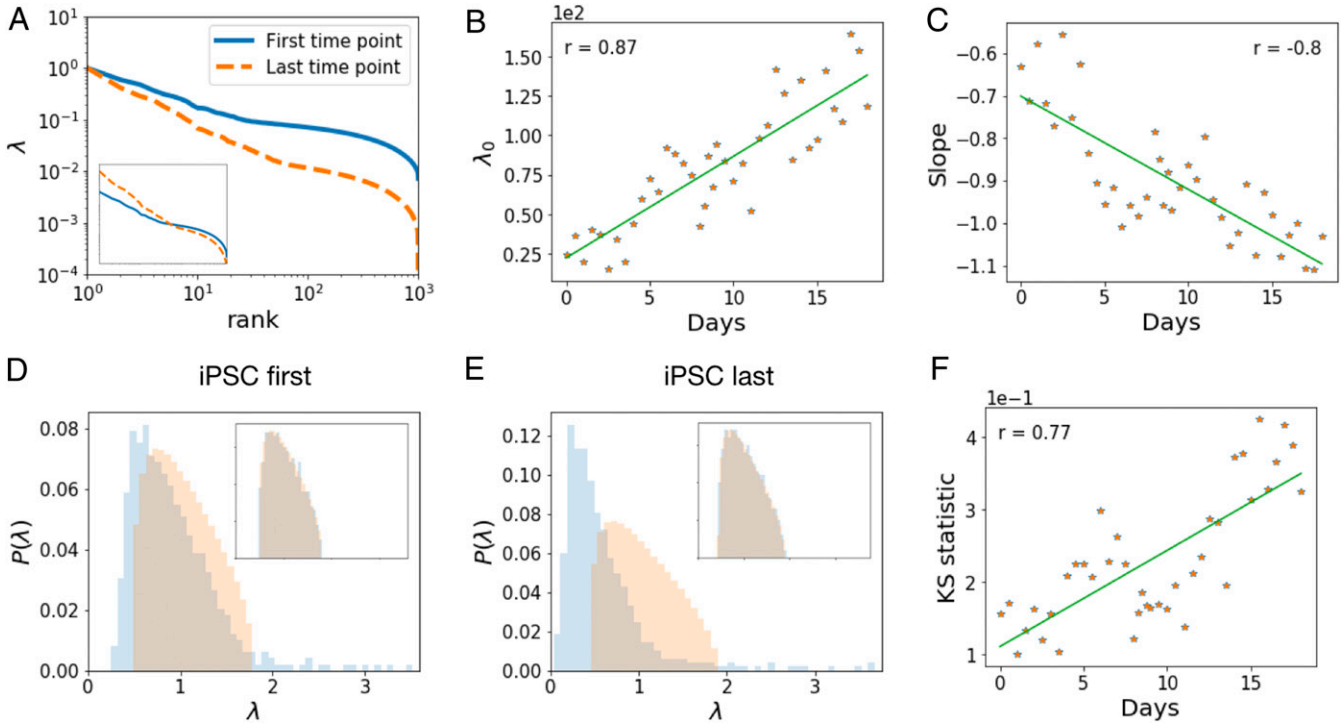


**Fig. 5.** Correlation statistics change along differentiation for the iPSC reprogramming system. (*A*) Eigenvalues versus their rank for the early time point (day 0, blue curve) and the late time point (day 18, orange dotted curve). Each set of eigenvalues is normalized by the largest respective eigenvalue. (*Inset*) Unnormalized eigenvalues versus rank. As cells gradually transition from mouse embryonic fibroblasts on day 0 to a population of various differentiated cell types on day 18 (*SI Appendix*), for the respective eigenvalue distribution of their covariance matrix, the largest eigenvalue increases (linear fit in green, r: 0.87, *P* value: 8.3e-13) (*B*) and the slope decreases (linear fit in green, r: −0.8, *P* value: 1.1e-9) (*C*). The eigenvalue distributions (blue histograms) and their respective MP distributions (orange histograms) for the first (day 0) time point (*D*) and the last (day 18) time point (*E*). (*Insets*) Eigenvalue distributions for the respective permuted single-cell data (blue) and associated MP distributions (orange). (*F*) The value of the KS statistic comparing the eigenvalue distribution and respective MP distribution increases over time (linear fit in green, r: 0.77, *P* value: 1.3e-8).

linearly differentiating DCs and, furthermore, in the crypt-to-villus axis in the intestinal epithelium, where the cells exhibit correlations due to their spatial structure, which is effectively one-dimensional. This framework allows us to identify when signals such as lineage or spatial relationships dominate the data and when they are lost or are too weak to begin with, such as exhibited by yeast colonies.

Given the observation of a power-law signature in the covariance eigenvalue distribution of a single-cell dataset, can we distinguish between different underlying biological signals that could have generated it (e.g., lineage structure versus spatial structure)? And can we rule out it being a technical or experimental artifact? Here, we showed that several common characteristics (both technical and biological) of scRNA-seq data, including noise statistics, batch effects, and clustered structure, generally do not generate such power-law signatures. Additionally, we showed that it is possible to identify "topologically informative genes" directly from single-cell data based on spectral signatures, without prior knowledge. This set of genes (or the output ranking over the genes) can then be queried for enriched biological processes, which can aid in the identification of the dominant biological signals underlying the data. While the focus of this work was the distinct spectral signature of lineage, we envision that this approach could be generalized for diverse biological processes, lying on low dimensional structures in high dimensional gene expression space. For example, we expect the cell cycle to form a cyclic structure in the high dimensional gene expression space (45, 46), physical spatial structures to form one-dimensional to three-dimensional grid-like structures (4, 47), and differentiation pathways to form tree-like structures (13). We are currently working on extending our framework beyond lineage and regulatory interactions to different geometric universality classes. This would lead to a better understanding of how diverse biological processes, including spatial relationships, cell-to-cell communication, and response to environmental signals, are reflected in single-cell spectra and how such processes can be disentangled from one another.

**Data Availability.** The single-cell RNA-seq datasets used for the current study were all previously published and can be accessed from the Gene Expression Omnibus (GEO) database with the following GEO accession numbers: GSE67602 for the epidermis (34), GSE52583 for the lung (35), GSE95025 for the *Drosophila* embryo (6), GSE121617 for the *Hydra* (37), GSE125162 for the yeast colonies (38), GSE122662 for the iPSC reprogramming dataset (36), GSE60783 for the DCs (41), and GSM2644349 and GSM2644350 for the intestinal epithelium (44) and their corresponding zone-reconstruction (42). Source code is available at GitHub, https://github.com/mornitzan/spectral_sc.

**ACKNOWLEDGMENTS.** We thank Lucy Colwell for introducing us to the ideas of spectral signatures in protein sequences as well as to Andrew Murray, Sergey Ovchinnikov, Yohai Bar Sinai, Edvin Memet, Nir Friedman, Aviv Regev, Raul Rabadan, and Michael Nicholson for important conversations. M.N. acknowledges support from the James S. McDonnell Foundation, Schmidt Futures, Israel Council for Higher Education, and the John Harvard Distinguished Science Fellows Program within the Faculty of Arts and Sciences Division of Science of Harvard University. M.P.B. acknowledges support from NSF Grants DMS-1715477 and ONR N00014-17-1-3029 as well as the Simons Foundation.

1. E. Papalexi, R. Satija, Single-cell RNA sequencing to explore immune cell heterogeneity. *Nat. Rev. Immunol.* **18**, 35–45 (2018).
2. A. Wagner, A. Regev, N. Yosef, Revealing the vectors of cellular identity with single-cell genomics. *Nat. Biotechnol.* **34**, 1145–1160 (2016).
3. V. Svensson, R. Vento-Tormo, S. A. Teichmann, Exponential scaling of single-cell RNA-seq in the past decade. *Nat. Protoc.* **13**, 599–604 (2018).
4. M. Nitzan, N. Karaiskos, N. Friedman, N. Rajewsky, Gene expression cartography. *Nature* **576**, 132–137 (2019).
5. R. Satija, J. A. Farrell, D. Gennert, A. F. Schier, A. Regev, Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).
6. N. Karaiskos *et al*., The *Drosophila* embryo at single-cell transcriptome resolution. *Science* **358**, 194–199 (2017).
7. Z. Liu *et al*., Reconstructing cell cycle pseudo time-series via single-cell transcriptome data. *Nat. Commun.* **8**, 22 (2017).
8. S. Tritschler *et al*., Concepts and limitations for learning developmental trajectories from single cell genomics. *Development* **146**, dev170506 (2019).
9. W. Saelens, R. Cannoodt, H. Todorov, Y. Saeys, A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* **37**, 547–554 (2019).
10. E. Marco *et al*., Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. *Proc. Natl. Acad. Sci. U.S.A.* **111**, E5643–E5650 (2014).
11. V. Moignard *et al*., Decoding the regulatory network of early blood development from single-cell gene expression measurements. *Nat. Biotechnol.* **33**, 269–276 (2015).
12. L. Haghverdi, M. Büttner, F. A. Wolf, F. Buettner, F. J. Theis, Diffusion pseudotime robustly reconstructs lineage branching. *Nat. Methods* **13**, 845–848 (2016).
13. L. Haghverdi, F. Buettner, F. J. Theis, Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics* **31**, 2989–2998 (2015).
14. J. Chen, A. Schlitzer, S. Chakarov, F. Ginhoux, M. Poidinger, Mpath maps multi-branching single-cell trajectories revealing progenitor cell progression during development. *Nat. Commun.* **7**, 11988 (2016).
15. M. Setty *et al*., Characterization of cell fate probabilities in single-cell data with Palantir. *Nat. Biotechnol.* **37**, 451–460 (2019).
16. Z. Ji, H. Ji, TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic Acids Res.* **44**, e117 (2016).
17. K. Street *et al*., Slingshot: Cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* **19**, 477 (2018).
18. A. Dixit *et al*., Perturb-Seq: Dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell* **167**, 1853–1866.e17 (2016).
19. A. Pratapa, A. P. Jalihal, J. N. Law, A. Bharadwaj, T. M. Murali, Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nat. Methods* **17**, 147–154 (2020).
20. F. Buettner *et al*., Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.* **33**, 155–160 (2015).
21. D. Kotliar *et al*., Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq. *eLife* **8**, e43803 (2019).
22. L. Aparicio, M. Bordyuh, A. J. Blumberg, R. Rabadan, A random matrix theory approach to denoise single-cell data. *Patterns (N Y)* **1**, 100035 (2020).
23. G. Eraslan, L. M. Simon, M. Mircea, N. S. Mueller, F. J. Theis, Single-cell RNA-seq denoising using a deep count autoencoder. *Nat. Commun.* **10**, 390 (2019).
24. J. Wang *et al*., Data denoising with transfer learning in single-cell transcriptomics. *Nat. Methods* **16**, 875–878 (2019).
25. D. van Dijk *et al*., Recovering gene interactions from single-cell data using data diffusion. *Cell* **174**, 716–729.e27 (2018).
26. C. Qin, L. J. Colwell, Power law tails in phylogenetic systems. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 690–695 (2018).
27. N. Papadopoulos, P. R. Gonzalo, J. Söding, PROSSTT: Probabilistic simulation of single-cell RNA-seq data for complex differentiation processes. *Bioinformatics* **35**, 3517–3519 (2019).
28. V. A. Marčenko, L. A. Pastur, Distribution of eigenvalues for some sets of random matrices. *Math. USSR-Sb.* **1**, 457–483 (1967).
29. N. Santhanam, J. Dingel, O. Milenkovic, "On modeling gene regulatory networks using Markov random fields" in *2009 IEEE Information Theory Workshop on Networking and Information Theory* (IEEE, 2009), pp. 156–160.
30. H. C. Nguyen, J. Berg, Mean-field theory for the inverse Ising problem at low temperatures. *Phys. Rev. Lett.* **109**, 050602 (2012).
31. T. R. Lezon, J. R. Banavar, M. Cieplak, A. Maritan, N. V. Fedoroff, Using the principle of entropy maximization to infer genetic interaction networks from gene expression patterns. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 19033–19038 (2006).
32. A. De Martino, D. De Martino, An introduction to the maximum entropy approach and its application to inference problems in biology. *Heliyon* **4**, e00596 (2018).
33. L. Zappia, B. Phipson, A. Oshlack, Splatter: Simulation of single-cell RNA sequencing data. *Genome Biol.* **18**, 174 (2017).
34. S. Joost *et al*., Single-cell transcriptomics reveals that differentiation and spatial signatures shape epidermal and hair follicle heterogeneity. *Cell Syst.* **3**, 221–237.e9 (2016).
35. B. Treutlein *et al*., Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* **509**, 371–375 (2014).
36. G. Schiebinger *et al*., Optimal-tansport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell* **176**, 928–943 (2019).
37. S. Siebert *et al*., Stem cell differentiation trajectories in *Hydra* resolved at single-cell resolution. *Science* **365**, eaav9314 (2019).
38. C. A. Jackson, D. M. Castro, G.-A. Saldi, R. Bonneau, D. Gresham, Gene regulatory network reconstruction using single-cell RNA sequencing of barcoded genotypes in diverse environments. *eLife* **9**, e51254 (2020).
39. E. Eden, R. Navon, I. Steinfeld, D. Lipson, Z. Yakhini, GOrilla: A tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* **10**, 48 (2009).
40. C. L. Smith, J. A. Blake, J. A. Kadin, J. E. Richardson, C. J. Bult; Mouse Genome Database Group, Mouse genome database (MGD)-2018: Knowledgebase for the laboratory mouse. *Nucleic Acids Res.* **46**, D836–D842 (2018).
41. A. Schlitzer *et al*., Identification of cDC1- and cDC2-committed DC progenitors reveals early lineage priming at the common DC progenitor stage in the bone marrow. *Nat. Immunol.* **16**, 718–728 (2015).
42. A. E. Moor *et al*., Spatial reconstruction of single enterocytes uncovers broad zonation along the intestinal villus axis. *Cell* **175**, 1156–1167.e15 (2018).
43. M. Nitzan, N. Karaiskos, N. Friedman, N. Rajewsky, Gene expression cartography. *Nature* **576**, 132–137 (2019).
44. K. S. Yan *et al*., Intestinal enteroendocrine lineage cells possess homeostatic and injury-inducible stem cell activity. *Cell Stem Cell* **21**, 78–90.e6 (2017).
45. G. La Manno *et al*., RNA velocity of single cells. *Nature* **560**, 494–498 (2018).
46. M. S. Kowalczyk *et al*., Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells. *Genome Res.* **25**, 1860–1872 (2015).
47. M. Adler, Y. Korem Kohanim, A. Tendler, A. Mayo, U. Alon, Continuum of gene-expression profiles provides spatial division of labor within a differentiated cell type. *Cell Syst.* **8**, 43–52.e5 (2019).