nature methods

REVIEW ARTICLE

https://doi.org/10.1038/s41592-021-01171-x

The triumphs and limitations of computational methods for scRNA-seq

Peter V. Kharchenko

The rapid progress of protocols for sequencing single-cell transcriptomes over the past decade has been accompanied by equally impressive advances in the computational methods for analysis of such data. As capacity and accuracy of the experimental techniques grew, the emerging algorithm developments revealed increasingly complex facets of the underlying biology, from cell type composition to gene regulation to developmental dynamics. At the same time, rapid growth has forced continuous reevaluation of the underlying statistical models, experimental aims, and sheer volumes of data processing that are handled by these computational tools. Here, I review key computational steps of single-cell RNA sequencing (scRNA-seq) analysis, examine assumptions made by different approaches, and highlight successes, remaining ambiguities, and limitations that are important to keep in mind as scRNA-seq becomes a mainstream technique for studying biology.

Transcriptional states provide a high-resolution, integrative view of genome activity, placing them at the core of functional genomics studies. Yet the precise state of every cell differs, reflecting its functional role, history, and stochastic fluctuations. scRNA-seq measurements in combination with computational analysis can reveal the transcriptional basis for heterogeneous cell states (Figs. 1 and 2 and Box 1). As scRNA-seq protocols and applications have progressed rapidly over the past decade (Fig. 3a), numerous analysis methods have also been developed. Though computational approaches vary, most formulate (1) a statistical model of the measurement, (2) a representation of the data in reduced dimensions, and (3) an approximation of the expression manifold (Box 2), with a set of discrete transcriptional subpopulations being the simplest and the most common approximation. The problems motivating these steps, and the specific solutions and their assumptions, are the subject of this review.

Statistical view of a cell

The single-cell estimates of the transcriptional state have much higher uncertainty than do their bulk counterparts, given that low amounts of starting material in each cell make limitations and distortions of the measurement more apparent. scRNA-seq protocols capture only a fraction of the molecules physically present in the cell, ranging from 5% to 20% for high-throughput protocols[1] and 30% to 40% for some well-based assays[2]. The number of captured molecules also varies between cells, commonly by more than an order of magnitude, making expression profiles of some cells much less certain than are those of others. In principle, the uncertainty in the state of the individual cells can be compensated for by measuring more cells (Fig. 3b): as long as the data provides sufficient signal to recognize groups of similar cells, computational methods can borrow information across cells to infer a well-resolved view of the transcriptional state. At the limit of low sensitivity, however, this trade-off is impractical owing to the cost and labor involved[1] (Extended Data Fig. 1a). In practice, the balance between sensitivity and the number of cells is dictated by the choice of the protocol, and considerable effort is being invested into improving molecular capture rates for high-throughput protocols.

The sparse and uncertain nature of scRNA-seq measurements makes it more practical to take a probabilistic perspective of the underlying expression state, considering, for instance, the likelihood that a transcript is expressed at a particular level, given the observed data (Fig. 3c). Through the Bayes theorem (Box 2), this is tied to the characteristics of the scRNA-seq assay itself—the probability of observing a specific number of reads for a transcript expressed at a given level in a cell. A variety of statistical models have been proposed to capture this relationship[3–5]. Most are extensions of the negative binomial models commonly used for bulk RNA-seq analysis, incorporating a more flexible mean–variance relationship or adding mixture components to account for an excess frequency of zeros—potentially representing a failure to detect a transcript in a given cell rather than true lack of expression. Improvements in scRNA-seq protocols, most notably incorporation of the unique molecular identifiers (UMIs), have largely attenuated such drop-out failures, allowing simpler negative binomial models without zero inflation[6] to be used. This implies that, despite the high abundance of zeros in the data, their overall frequency is not greater than what would be expected from a relatively simple, sparse sampling of messenger RNA molecules.

Common to nearly all statistical models is the assumption that the scRNA-seq measurement represents a random sampling of the mRNA molecules present in a cell. The real sampling rates, however, show gene-specific bias, which can be seen from the systematic difference of transcript-detection rates under different scRNA-seq protocols. For instance, Extended Data Fig. 1b–d illustrates that expression discrepancies between the 10x Chromium and Drop-seq platforms are persistent across cell lines. These biases, however, typically have little impact on the downstream analysis if the measurements being compared have been performed with the same scRNA-seq protocol[7].

Comparing transcriptional states

A statistical model of the scRNA-seq measurement dictates a recipe for testing whether a transcript's expression level shows a statistically significant difference between any two cells or two sets of cells. For instance, if the posterior probability distributions of the underlying expression level of a given gene have been estimated for each cell, the differential expression test can be performed by simply calculating their overlap—the probability that the expression levels in the two cells are the same[4,8] (Fig. 3c). Many specialized differential

Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. ✉e-mail: peter_kharchenko@hms.harvard.edu

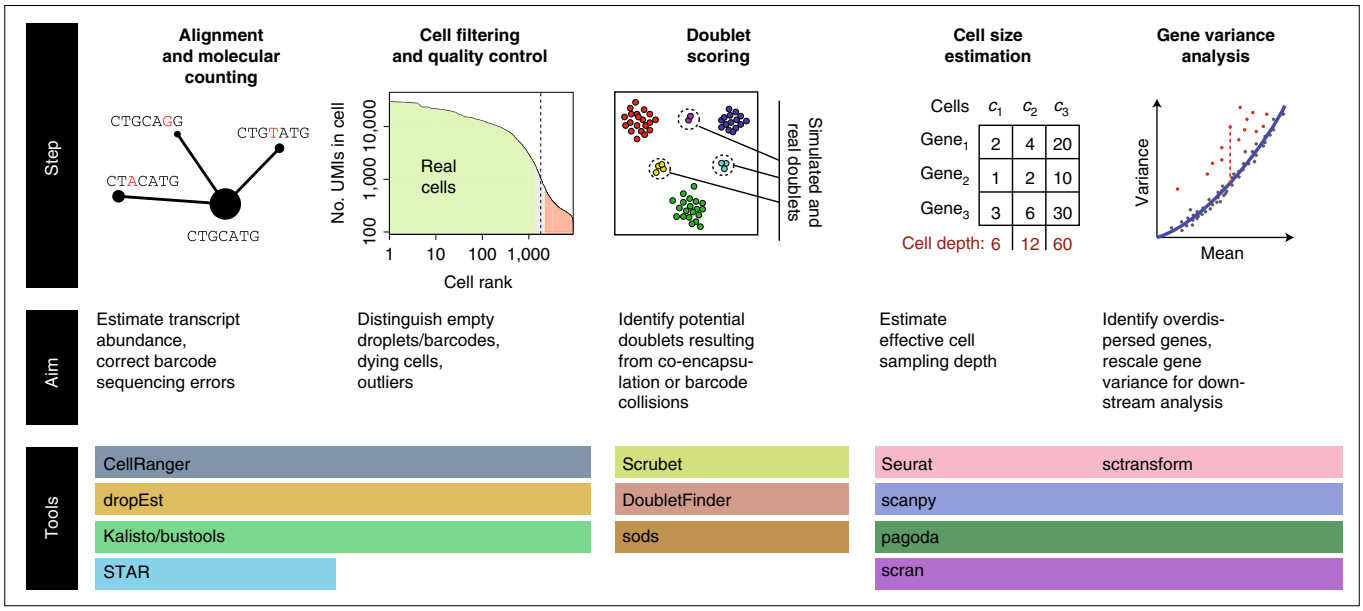


Fig. 1 | Key preprocessing steps in single-cell RNA-seq analysis. The diagram shows major steps involved in analysis of individual scRNA-seq datasets, along with the relevant software tools (see Box 1 and Luecken et al.[86] for other practical considerations).

expression methods have been developed on the basis of different parametric models and nonparametric descriptors[9]. When comparing larger cell populations (for example, >100 cells), the advantages of sophisticated parametric models fade[10], and standard non-parametric tests, such as the Wilcoxon test, can provide sufficient statistical power while making fewer assumptions. A corollary to the increase in statistical power is that comparisons of the mean expression between sufficiently large populations will report most genes as being differentially expressed (Fig. 3d). Hence other considerations, such as the magnitude of the difference in expression or the definition of the subpopulations themselves, become more important. Beyond mean expression, single-cell data enable comparison of other characteristics of the expression distribution, such as expression variability[11] or distribution shape[12,13]. For example, Martinez-Jimenez et al. demonstrate that CD4$^+$ T lymphocytes from older mice exhibit increased variability of expression, even though the mean expression levels of some of these genes remain similar[14]. Finally, differential tests are particularly sensitive to batch effects, and most existing methods can control for simple batch structure. Analysis of complex experimental designs, such as those involving multiple sample categories or matched samples, necessitates more elaborate control for covariates. While this remains an active research area, popular approaches discard single-cell variation within cell types and take advantage of bulk RNA-seq tests[15].

Complementary to differential expression analysis is the quantification of expression similarity or distance between cells. Such measures underlie most downstream analyses, including grouping of cells into clusters, fitting trajectories, and visualizing data. Two categories of distances have been used: traditional distances (for example, Euclidean, L1, and Canberra) that generally aim to quantify the amount of transcriptional difference from one cell to another, and measures of statistical deviation of two cells from equality (for example, Poisson-based similarity[16] and Jensen–Shannon divergence[17]). Neither category is grounded in biology, and they can show counterintuitive behaviors. The latter will vary depending on the depth of coverage of the cells, and the former will, in some way, equate changes along different dimensions. For example, an L1 measure will give the same weight to a 100-molecule shift in a level of a structural protein as it would to a transcription factor. A more fundamental problem, however, arises from the high dimensionality of the expression state. As the number of dimensions increases, traditional distance metrics, such as Euclidean distance, gradually lose their ability to distinguish the closest and most distant points[18]. In the case of ~10,000 dimensional noisy scRNA-seq measurements, this 'curse of dimensionality' is quite apparent and degrades the effectiveness of most alternative measures, such as cosine similarity (Extended Data Fig. 1e). Fortunately, transcriptional variation can be effectively described by a lower-dimensional space in which cell distances can then be calculated.

The quest for reduced dimensions

The effective dimensionality of scRNA-seq data is notably lower than that allowed by all possible combinatorial patterns of gene expression. This is in part because of the limited scope of scRNA-seq datasets: for instance, if the dataset profiles a tissue with only three major cell types, then most of the transcriptional variation will be explained by the two axes capturing the difference between these cell types. More fundamentally, lower effective dimensionality reflects the coordination of genes arising from the regulatory logic of a cell. For example, if a set of ten transcripts are, in most cases, activated by the same transcription factor, then their expression can be more parsimoniously described by a single variable instead of ten. The most common approach for dimensionality reduction, principal component analysis (PCA), aims to do exactly that: find a limited number of linear combinations of transcripts that captures as much variance in the dataset as possible (Extended Data Fig. 2a). These top principal components (PCs) describe a low-dimensional hyperplane on which main transcriptional variation between cells takes place.

Two properties of the scRNA-seq count data, however, complicate the application of PCA. First, the variance of individual transcript depends heavily on its expression magnitude (Fig. 3e), and the top PCs will focus on the detailed fluctuations of highly expressed transcripts at the expense of broader transcriptional patterns. To overcome this, the analysis methods compare the observed variance of each transcript to that expected from its expression magnitude. While some of the initial methods have estimated the expected variance on the basis of spike-in controls, such as used Evaluation of

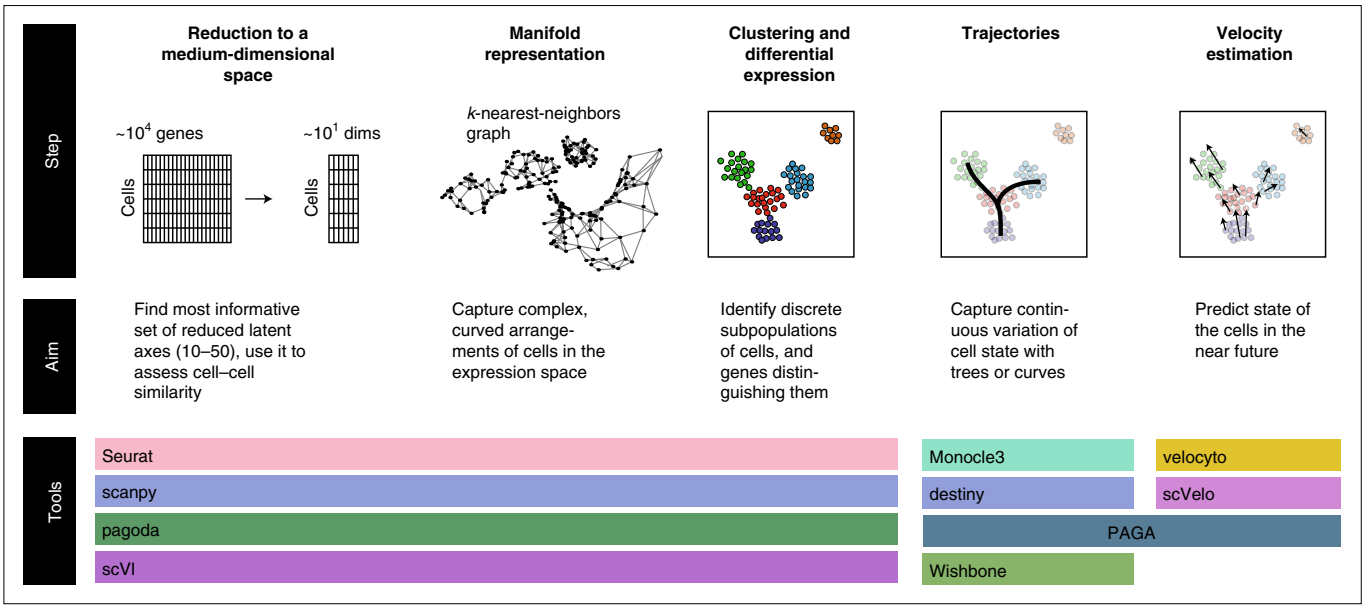


Fig. 2 | Key analysis steps in single-cell RNA-seq analysis. The diagram shows major steps involved in analysis of individual scRNA-seq datasets, along with the relevant software tools (see Box 1 and Luecken et al.[86] for other practical considerations).

the External RNA Controls Consortium (ERCC) sequences[19], this approach gradually fell out of favor, as designing and delivering equal amounts of spike-in molecules that would behave in the same way as endogenous mRNA presented its own challenges. In lieu of extrinsic controls, most analysis methods estimate the expected variance on the basis of the behavior of all observed transcripts (Fig. 3g). If a transcript shows excess variance over that expectation, its pattern of expression likely distinguishes major cell subpopulations (Fig. 3f–i). The PCA can then be restricted to a small set of highly variable transcripts[3]. The statistical model can also be used to normalize the variance of each transcript to match the residual variance observed in excess of what is expected from the measurement model[20–22]. As the variance is estimated considering all measured cells, the most variable transcripts will be focused on the most striking subpopulation differences (for example, epithelial versus immune cells), whereas capturing more subtle transcriptional shifts (for example, $CD8^+$ versus $CD4^+$ T cells) may require re-analysis of focused subpopulations (Extended Data Fig. 2b).

The second complication for PCA arises due to sparsity in the scRNA-seq count data. As PCA optimizes the decomposition of the covariance matrix, it works optimally on symmetrically-distributed data. Transformations such as log, asinh and other variance-stabilizing transformations[22,23] are typically applied to the normalized count values to make the distribution of expression magnitudes appear closer to normal distribution. A high frequency of zero counts, however, shows up as a singular spike, yielding a profoundly skewed distribution even after such transformations (Fig. 3j). As some cells have many more zeros than do others, PCA is likely to identify this technical difference between cells of high and low coverage as one of the top PCs. This effect can be mitigated by weighting down zero entries[20] or explicitly regressing out the dimension separating cells according to their total[24]. The latter correction is further complicated by the fact that some biologically-distinct subpopulations show systematic differences in depth due to a difference in the total amounts of mRNA. Some methods, therefore, use am iterative approach by first identifying major subpopulations and then regressing out the depth within them[25].

The limitations of PCA detailed above fundamentally stem from the fact that the optimization of the low-dimensional basis (that is, PCs) is disconnected from the underlying statistical structure of the scRNA-seq data. Thus, dimensionality reduction by PCA or other traditional methods[26,27] is likely to be detracted by noise or technical variation in the data at the expense of capturing true biological variation. A more general approach of factor analysis (FA) can be used to couple the factorization with an appropriate statistical model of the data. For instance, zero-inflated factor analysis (ZIFA) accommodates frequent occurrence of zeros in scRNA-seq data by modeling expression magnitudes as a mixture of normal distribution and a separate spike of zeros[28]. Beyond the normal approximation, a more accurate statistical description of the scRNA-seq data can be provided by count models, such as negative binomial distribution. While factorization of a matrix into multiple count-process components remains a challenging topic in statistics[29], several methods use a related hierarchical approach, modeling each entry of a matrix using a single count process whose parameters are driven by a linear combination of components. Such models are easier to compute and have been previously used in the context of bulk RNA-seq[30]. For instance, one such implementation, ZINB-WaVE, uses a generalized linear model to drive both mean expression and dropout parameters (Box 2) of the zero-inflated negative binomial[31].

The mapping between the transcript counts and reduced dimensions does not have to be limited to linear relationships. Similarly, the statistical parameters used to describe the properties of scRNA-seq measurements, such as amount of overdispersion, can depend on the properties of different transcripts (for example, GC content and subcellular localization) or cell types (for example, lysis susceptibility and enzymatic content). Accounting for such an extensive gamut of potential relationships, however, is challenging. To map more complex nonlinear relationships, several groups have proposed using autoencoder neural networks[8,32,33] (Box 2). Neural networks provide a convenient computational approach for learning complex nonlinear multidimensional functions[34]. For example, Extended Data Fig. 2e,f uses a simple neural network to learn an analytical function approximating an existing *t*-distributed stochastic neighbor embedding (*t*-SNE) embedding. As *t*-SNE embeddings are based on empirical optimization of the relative positions of neighboring cells, there is no obvious analytical function connecting

Box 1 | Select software tools

Tools for alignment, barcode correction, count matrix estimation, and quality control include:

- CellRanger (https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/installation): supports 10x Chromium datasets (commercial product)
- dropEst (https://github.com/hms-dbmi/dropEst): supports multiple droplet-based protocols
- STAR (https://github.com/alexdobin/STAR): aligner (used internally by CellRanger and dropEst), also has built-in options for count matrix estimation
- Optimus (https://data.humancellatlas.org/pipelines/optimus-workflow): supports 10x Chromium v2 and v3 datasets, designed for Human Cell Atlas
- Kallisto/bustools (https://www.kallistobus.tools): fast processing using pseudoalignment

Cell filter and doublet identification tools include:

- EmptyDrops (https://rdrr.io/github/MarioniLab/DropletUtils/man/emptyDrops.html): uses a classifier to distinguish 'empty' cells that look like the low-end tail of the cell size distribution
- Scrublet (https://github.com/AllonKleinLab/scrublet): python-based, doublet simulation and doublet scoring
- doubletFinder (https://github.com/chris-mcginnis-ucsf/DoubletFinder): R-based, doublet simulation and doublet scoring
- scds (https://github.com/kostkalab/scds): fast doublet scoring implementation

Tools for normalization, dimensionality reduction, and clustering and differential expression include:

- Seurat (https://satijalab.org/seurat/): the most popular analysis toolkit, R-based
- scanpy (https://github.com/theislab/scanpy): the most popular python-based toolkit
- scVI (https://github.com/YosefLab/scVI): latent space identification using variational neural net
- pagoda2 (https://github.com/hms-dbmi/pagoda2): fast, R-based processing
- SAUCIE (https://www.krishnaswamylab.org/projects/saucie): a neural-net-based dimensionality reduction, using maximal mean discrepancy penalty

Tools for trajectory fitting include:

- Monocle3 (https://cole-trapnell-lab.github.io/monocle3/): third iteration of the Monocle package, including updated tree utilities
- Slingshot (https://github.com/kstreet13/slingshot): tree fitting with improved pseudotime estimation
- PAGA (https://github.com/theislab/paga): tree/graph fitting approach combined with cell aggregation, also supports cluster-based velocity estimates
- Wishbone (https://dpeerlab.github.io/dpeerlab-website/wishbone.html): a bifurcation analysis method
- Destiny, DPT (https://github.com/theislab/destiny/): dimensionality reduction and trajectory fitting using diffusion maps[82]

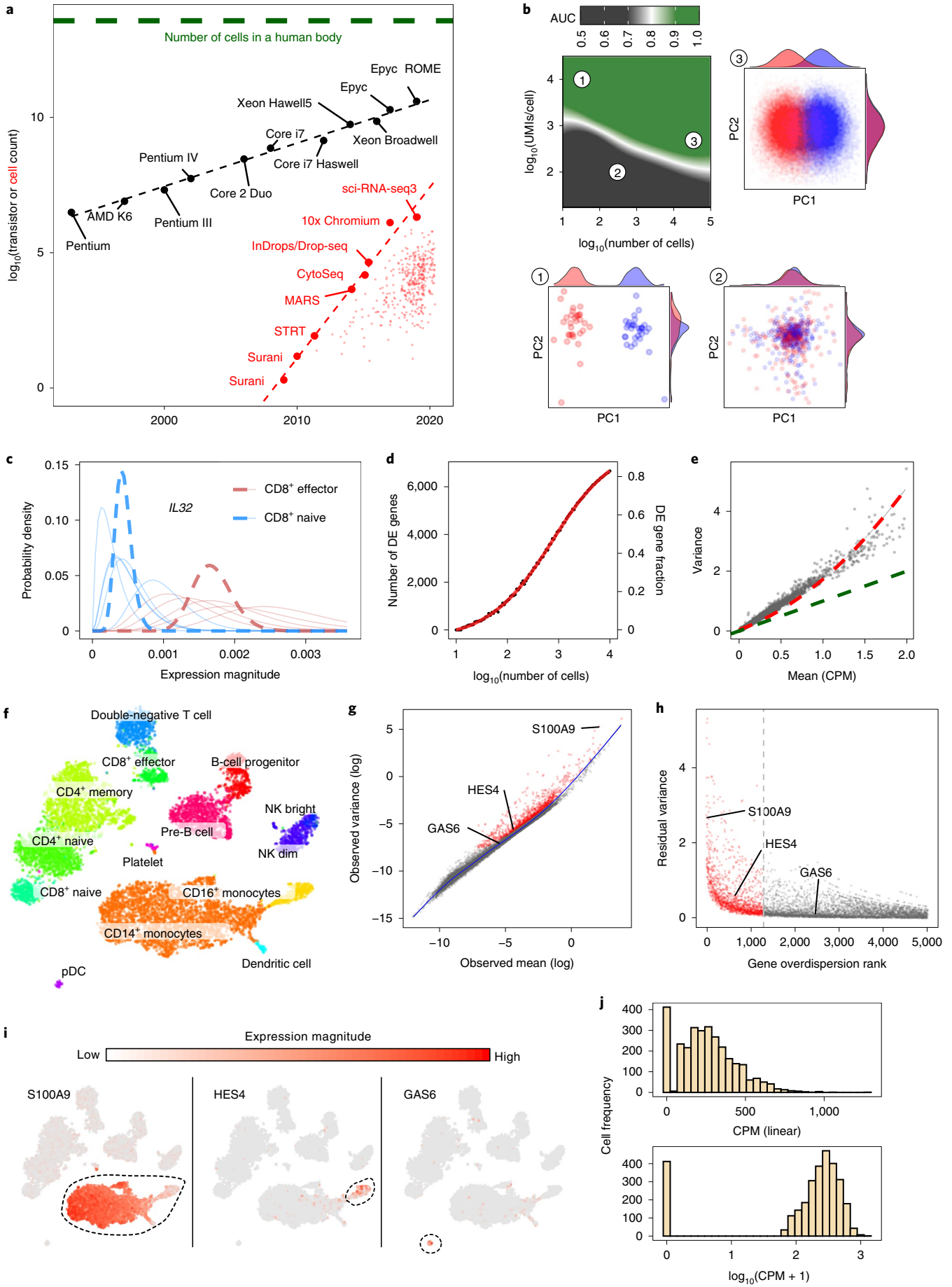
Tools for velocity estimation include:

- velocyto (http://velocyto.org/): reference python/R implementation
- scVelo (https://scvelo.readthedocs.io/): new implementation using curve-based phase portrait fit

the expression state with the resulting *t*-SNE coordinates, yet a sequence of transformations encoded by the neural network is sufficiently powerful to provide a good approximation. An autoencoder is a neural network designed to learn an effective low-dimensional representation by finding functions that map data to and from low-dimensional space in a way that yields optimal reconstruction of the original data (Extended Data Fig. 2d). Because these functions can be nonlinear, the resulting reduced dimensions can more effectively capture the underlying structure of the populations than can linear approaches (Extended Data Fig. 2c). Nonlinear mapping, however, makes it more difficult to interpret to the latent states. One possible compromise is using an asymmetric autoencoder, combining nonlinear encoding with a more interpretable linear decoding[35].

Lopez et al.[8] (scVI) demonstrated how neural networks can be combined with probabilistic models to propagate uncertainty of expression measurements into the latent space (Box 2). Taking the zero-inflated negative binomial model of the scRNA-seq measurement, scVI builds a variational autoencoder—an approach that formulates a probabilistic distribution on latent parameters—to fit nonlinear mapping between the high-dimensional transcript space and the latent space of reduced dimensions. Companion networks keep

Fig. 3 | scRNA-seq basics. a, Beating Moore's law. The number of cells measured by landmark scRNA-seq datasets over years (red), compared with the increase in the CPU transistor counts (black). The set of all published scRNA-seq studies[83] is shown with small red dots. The estimated number of cells in a human body is shown by a green dashed line. **b**, Shallow coverage of each cell can be compensated for by measuring more cells. The ability to distinguish two cell populations, assessed by the area under the receiver operating characteristic curve (ROC AUC) measure, is shown as a function of the number of measured cells (*x* axis) and the mean cell depth (*y* axis). Examples of three different simulations (1–3) within different parts of this design parameter space are shown on PCA projections. **c**, Probabilistic view of scRNA-seq estimates. Posterior probability of *IL32* gene expression magnitude is shown for five cells from two different CD8$^+$ T cell populations (red and blue, thin lines). Joint posteriors assessing the mean expression magnitude within each subpopulation are shown by thick dashed lines. **d**, Comparing CD4$^+$ T cells and CD14$^+$ monocytes, the plot shows the number (*y* axis, left) and the fraction (*y* axis, right) of the genes passing a 1% statistical significance threshold for differential expression (DE) as a function of the number of cells compared from each population (*x* axis). **e**, The scatter plot shows for each gene (dots) the mean (*x* axis) and variance (*y* axis) of the normalized UMI counts (CPM, counts per million) in CD4$^+$ T cells. The Poisson expected value is shown in green, with a quadratic-based negative binomial fit shown in red. **f–i**, Variance normalization and most variable genes. **f**, A *t*-SNE embedding of a primary peripheral blood mononuclear cell (PBMC) dataset with cell annotations. NK, natural killer, separated into CD56 bright and dim subsets. pDC, plasmacytoid dendritic cell. **g**, Mean–variance relationship of different genes (dots) in the PBMC dataset is shown for log-transformed expression estimates. The genome-wide relationship, as captured by smoothed regression, is shown by the blue line. Genes whose variance is significantly higher than the genome-wide trend are shown as red dots. **h**, Residual variance is shown for the top 5,000 overdispersed genes, ordered by the statistical significance (*x* axis). **i**, Expression pattern of several example genes, with circles highlighting the subpopulations distinguished by the genes. **j**, Distribution of normalized expression magnitudes (CPM) for the *CTSH* gene across all CD14$^+$ monocytes is shown on the linear scale (top) and after log transformation (bottom) with a pseudocount.


a
Number of cells in a human body
log10(transistor or cell count)
Pentium
AMD K6
Pentium III
Pentium IV
Core 2 Duo
Core i7
Core i7 Haswell
Xeon Hawell5
Xeon Broadwell
Epyc
Epyc ROME
sci-RNA-seq3
10x Chromium
InDrops/Drop-seq
CytoSeq
MARS
STRT
Surani
Surani
2000
2010
2020
b
AUC
log10(UMIs/cell)
log10(number of cells)
PC1
PC2
c
Probability density
IL32
CD8+ effector
CD8+ naive
Expression magnitude
d
Number of DE genes
DE gene fraction
log10(number of cells)
e
Variance
Mean (CPM)
f
Double-negative T cell
CD8+ effector
B-cell progenitor
CD4+ memory
NK bright
Pre-B cell
CD4+ naive
Platelet
NK dim
CD8+ naive
CD16+ monocytes
CD14+ monocytes
Dendritic cell
pDC
g
Observed variance (log)
Observed mean (log)
S100A9
HES4
GAS6
h
Residual variance
Gene overdispersion rank
S100A9
HES4
GAS6
i
Expression magnitude
Low
High
S100A9
HES4
GAS6
j
Cell frequency
CPM (linear)
log10(CPM + 1)

Box 2 | Abbreviations and terms

CPM – Counts per million.

UMI – Unique molecular identifier. A randomized nucleotide sequence incorporated into the complementary DNA in the initial steps of RNA-seq protocol. Because this sequence is then carried in the subsequent amplification steps, it can be used to recognize multiple sequencing reads originating from the same physical mRNA transcript.

Cell state – Current molecular configuration of a cell. In most cases, the term is used to refer to a particular facet of the overall molecular configuration. In this manuscript, for instance, cell state refers to the repertoire of transcripts currently present in the cell. Since the full combinatorial set of possible states is exceedingly large, the term usually refers to a greatly simplified description of a state, such as its expected position in reduced dimensions.

Expression manifold – A smooth, low-dimensional surface on and around which the physiologically viable cellular states tend to be positioned.

Dropout – In the context of scRNA-seq, this refers to a stochastic failure to detect a transcript that is expressed in a cell.

Bayes theorem – This captures the reciprocal relationship between the conditional probabilities of two events A and B: $p(A \mid B) = p(B \mid A) \times p(A) / p(B)$. The theorem is useful in probabilistic handling of experimental evidence—for instance, evaluating the probability $p(A \mid B)$ that a given transcript is really expressed in a cell (A) given the molecule counts detected by the assay (B).

Autoencoder – A type of network designed to transform a signal into some restricted form (for example low-dimensional representation) and then reconstruct it back into its original form. The parameters of such a network can be thus optimized in an unsupervised manner by minimizing the difference between the original and reconstructed signals.

Latent space – The space of model variables that are not directly observed, for instance if a space can be defined by a latent variable capturing the cell cycle phase of a cell. Although it is not measured directly, it can be inferred from gene expression or other measured variables.

Ergodic process – A dynamic process whose statistical properties can be equivalently captured either by (1) sampling different time points or (2) by sampling different realizations. The first scenario would mean, for instance, that if one were to observe a single cell over time, it would eventually visit all possible cell states. The second scenario would mean that if one were to take an scRNA-seq snapshot of many cells, their distribution over the state space would be the same as the distribution observed over time in the scenario 1.

track of the uncertainty of the mapping to and from the reduced dimensions. This is notable because it enables downstream analysis, such as clustering, trajectory mapping, and differential expression, to account for the limitations and uncertainty of the underlying scRNA-seq measurement.

Approximating cell expression manifolds with neighbor graphs

The protagonist of the H.G. Wells' 1897 novel *The Invisible Man* complained that his shape would be exposed by the falling snow if it settled on him. In the same way, the shape of the underlying expression manifold—a smooth low-dimensional surface on which the observed states of the cells lie—can be approximated by observing the distribution of the measured cells. An effective way of doing this is to connect sets of nearby cells, forming a nearest-neighbor graph. If the sampling of cells is sufficiently dense, one can then follow the shape of the manifold by traversing this graph like a wire mesh (Fig. 4a).

The neighbor graph representation[36], which brings significant computational and analytical advantages, has been utilized in most scRNA-seq analyses[37–39]. The graphs are typically constructed by connecting each cell to its k nearest neighbors (kNN graph), using highly optimized 'approximate nearest neighbor' search algorithms, optionally applying additional refinements such as estimation of a shared-neighbor graph[40]. More elaborate graph-construction approaches have been developed to integrate multiple datasets, in order to overcome or correct for different types of batch effects in such designs[41]. For instance, Haghverdi et al.[42] used reciprocal-best-hit matching to establish additional edges connecting nodes corresponding to cells from different datasets. Many techniques for analyzing the structure of these graphs have also been developed in computer science and adopted for the analysis of scRNA-seq data. For instance, major features of the underlying expression manifold can be identified through spectral analysis of the graph. Specifically, the eigenvectors of the graph Laplacian matrix with smallest eigenvalues will correspond to connected components and main structural axes of the graph[43]. Adopting this Laplacian eigenmaps approach to scRNA-seq analysis, Haghverdi et al.[39,44] have shown that it is effective at deriving refined low-dimensional representation of the expression manifold, facilitating both visualization and analysis of cellular heterogeneity.

The neighbor graph representation empowers many, if not most, downstream analysis methods, including identification of subpopulations, dynamic processes, or mapping between datasets. Its ability to capture the structure of the data can be readily appreciated when looking at t-SNE or UMAP embeddings: both methods essentially visualize the neighbor graph, positioning directly connected cells nearby on the page[45,46]. In datasets with many subpopulations, linear dimensionality reduction methods, such as PCA, need many more than two dimensions to capture the relative placement of different subpopulations, as expression differences between subpopulations lie in different planes and can vary significantly in magnitude. In contrast, t-SNE and UMAP largely discount global distances, distance magnitudes, and directions, instead preserving the local neighborhood relationships. These relationships are effective at capturing subpopulations, continuous trajectories, and other structures in the data, even when constrained to the two dimensions of the page[47,48].

Clustering cells

Many methods for clustering transcriptionally similar cells have been developed or adopted from computer science (see Kiselev et al.[49] for a recent review). The most popular approach uses neighbor graph to identify 'communities' of cells that tend to connect to each other more often than to the cells outside of the cluster. Fast algorithms are available for detecting such communities. For example, Louvain clustering[50] or its more recent extension, Leiden clustering[51], can be effectively ran on graphs containing millions of cells, and is utilized by many analysis packages[38,52–54].

Cell clusters serve as units of interpretation, enabling effective navigation of a dataset and guiding downstream analysis, such as differential expression. Their nature, however, is approximate and does not carry an inherent biological meaning beyond transcriptional similarity. The clusters may separate distinct cell types, capture subtle transcriptional shifts within the same cell type, or over-partition a uniform cell subpopulation based on weak stochastic variation.

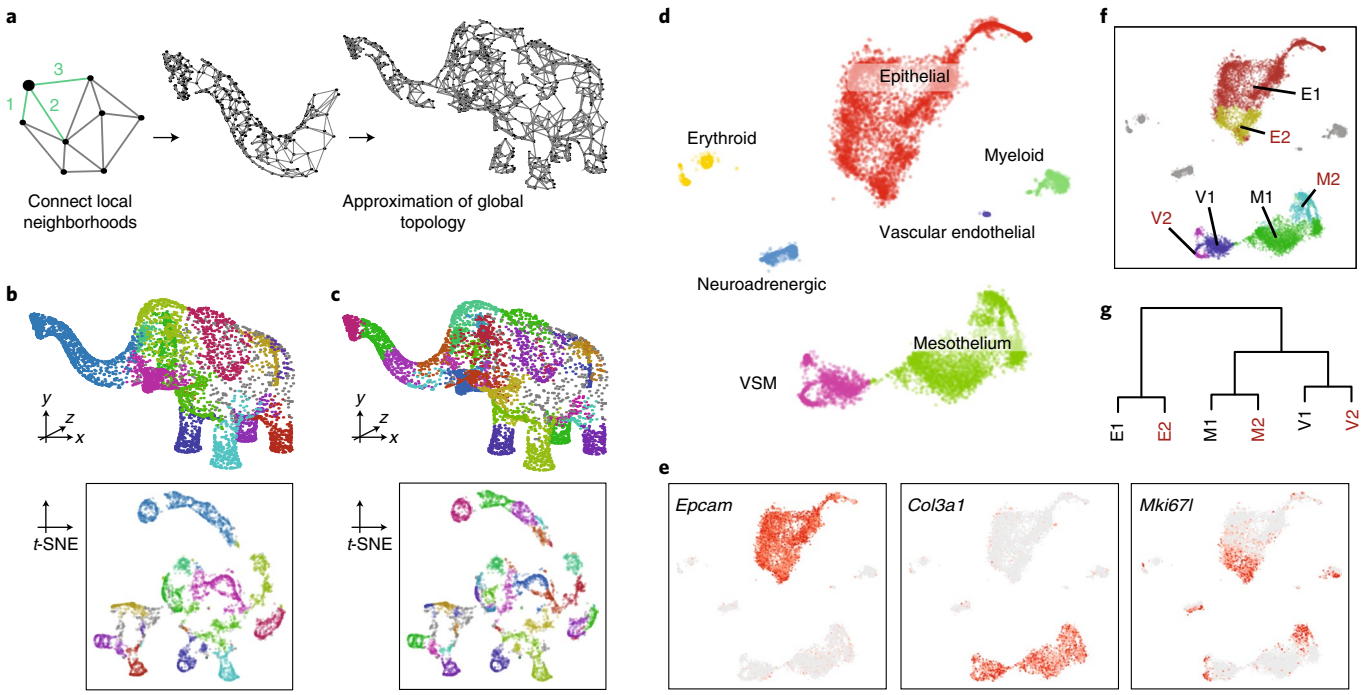


Fig. 4 | Approximating and partitioning complex manifolds. a, Complex, curved surfaces can be well approximated by neighborhood graphs. A simple graph connects each point with its *k* closest neighbors (kNN graph). As more points and regions are measured, the complex structure of the object can be revealed. **b**, The elephant graph (in **a**) is clustered using the Leiden clustering algorithm[51] (resolution $r = 0.5$). The resulting clusters are shown as colors on the 3D model (top) and *t*-SNE embedding (bottom) of the data. **c**, Clustering resolution is arbitrary. Similar to **b**, the plots show clustering with increased resolution ($r = 3$). The clusters are smaller but capture equally valid anatomical elements. **d**, Overview of the E12.5 fetal pancreas dataset[84] shown on a *t*-SNE embedding. **e**, Expression of genes illustrating epithelial (*Epcam*), mesenchymal (*Col3a1*) and cell cycle (*Mki67*) signatures. **f**, Coarse clustering of the mesenchymal and epithelial populations separating cycling cells within each group. **g**, Hierarchical clustering of the coarse clusters (**f**) on the basis of the cosine similarity of the whole-transcriptome average profiles is driven by the cell-type differences. Neither *t*-SNE embedding nor hierarchical clustering by themselves reveal that the cycling clusters (E2, V2, M2) are distinguished from their corresponding populations by a very similar cell cycle signature.

The resolution of the clusters depends on the specifics of the method employed, and for many methods can be explicitly modulated by the user (Fig. 4b,c). Indeed, there are cases where the narrative and the level of interpretation of the data would make it appropriate to group, for example, all T lymphocytes into a single cluster. And other cases where separation of detailed T cell subtypes would be of primary interest. The hierarchical clustering methods, including community detection algorithms like walktrap[55], capture this type of correspondence between different levels of cluster resolution, which is useful for describing relationships between major and minor cell types. Simple hierarchical representations, however, can easily break down when other sources of transcriptional variation, such as cell cycle or cell activity, are present in the data (Fig. 4d–g).

Analysis of dynamic processes

In many biological contexts, such as organism development or response to stimuli, the dynamics of transcriptional states is of primary interest. Current scRNA-seq protocols are designed to capture a static snapshot of cellular states at a particular point in time. Some features of the underlying dynamical process, however, can often be deduced. Consider, for example, a photograph of an ant trail. One can trace the paths being taken by the ants by following the filaments of ant density (Fig. 5a). The search for an optimal path is typically formulated assuming a specific class of topologies, such as trees (for example, principal graph problem[56], Fig. 5c,d) or curves (principal curve problem[57]). This general approach of tracing cell density in a reduced-dimensional space was initially utilized by Monocle[58] and has been further elaborated by many methods (see Saelens et al.[59] for a recent comparison). It has been successfully applied to capture branching cell differentiation trajectories and other dynamic processes in numerous biological contexts and is by far the most common approach for inferring transcriptional dynamics. Some caution needs to be taken when applying such tools, as they will typically aim to span all of the provided cells with trajectories, regardless of whether they actually participate in a dynamic process. For example, Extended Data Fig. 1f shows computationally optimal spanning tree for the PBMC populations, yet interpreting it as a dynamic process would be incorrect. It is, therefore, important to determine the identity of the subpopulations first and restrict the trajectory modeling to the appropriate part of the dataset. In interpreting the results, it is also important to keep in mind that some details of the reported trajectory may be uncertain[60]. The uncertainty can stem from a limited number of cells supporting a particular feature (for example, a tree branch with very few cells). Most methods fit trajectories in low-dimensional representations of the data, such PCs, diffusion components[44], or even on the two-dimensional embeddings[58]. Other methods simplify the problem by connecting cell clusters instead of individual cells[61–64]. It is important to keep in mind that the specific behavior of the clustering or dimensionality reduction algorithm can also have a significant impact on the resulting trajectory.

At its core, the 'ant trail' tracing approach depends on the ability to observe cells at all points along the trajectories. In other words, it assumes that the underlying dynamic process is stationary and ergodic (Box 2), so that a single sufficiently large scRNA-seq measurement would cover all possible intermediate cellular states. This would not be the case, for example, in processes synchronized by developmental time or external stimuli, such as neuronal activation

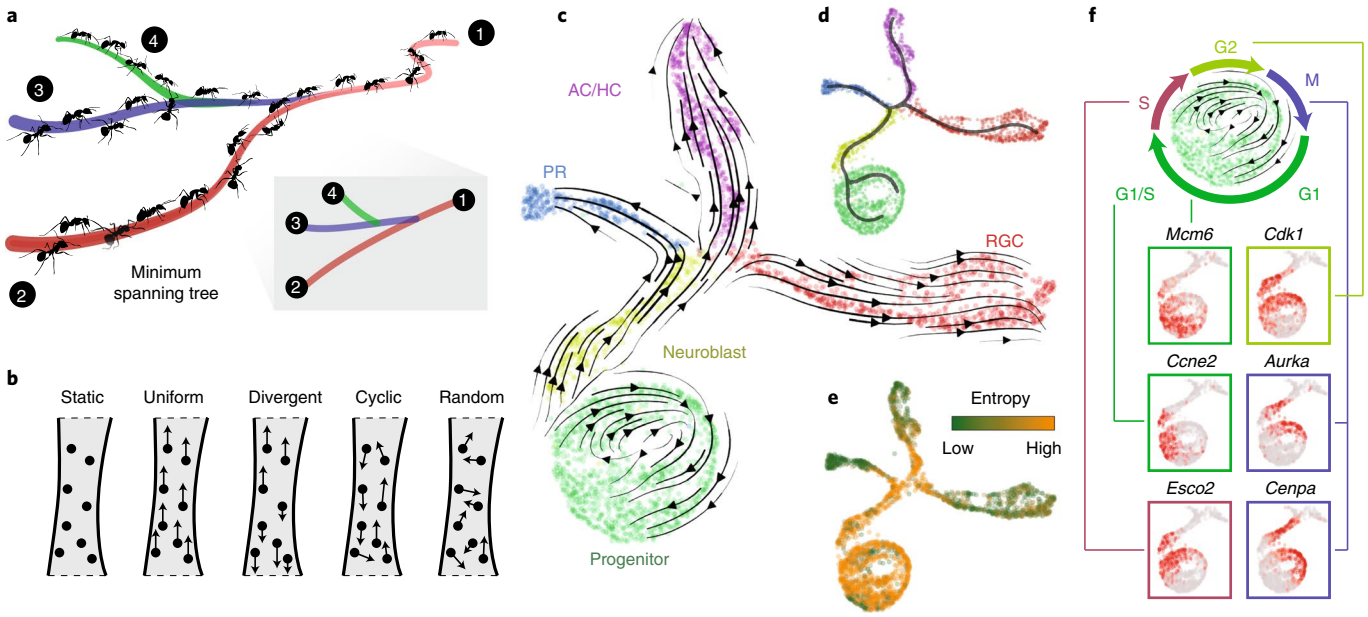


Fig. 5 | Approximating dynamical processes. a, A cartoon of an ant trail, with the inset showing the resulting principal tree capturing the structure of the ant density fibers. **b**, Several incompatible scenarios of cell dynamics are shown for the same stretch of cell density. **c**, Transcriptional dynamics of developing retina. An UMAP embedding of the E15.5 mouse retinal dataset from Lo Giudice et al.[85] is shown, together with the RNA velocity directions estimated by scVelo[71]. PR, photoreceptors; RGC, retinal ganglion cells; AC/HC, amacrine/horizontal cells. **d**, Retina trajectory approximated by a principal tree. **e**, Expression entropy tends to be high in progenitor and low in differentiated populations. **f**, The circular topology and velocity estimates of the progenitor populations is driven by the cell cycle process. Expression patterns of relevant genes are shown on the relevant part of the embedding.

by a pulse of light[65]. In such scenarios, the cells would propagate as a wavefront, and a single measurement would show a discrete cluster of cells instead of contiguous trajectories. Asynchrony can be reintroduced by measuring mixtures of different samples or multiple time points. It should be noted, however, that except for a handful of methods designed specifically for the multi-time-point designs[66,67], trajectory inference methods cannot take sample time into account.

The density-based methods described above aim to approximate central trajectories along which the cells proceed. That leaves considerable uncertainty about the details, as very different scenarios can give rise to the same density patterns[68] (Fig. 5b). The most obvious ambiguity is the direction of the flow. Most analyses rely on external knowledge to assign directions, for instance using previously identified markers to recognize the start or end point of the process. One systematic criterion has been proposed for cell differentiation trajectories by Grun and colleagues[69], based on a general trend that stem/progenitor populations tend to express a wider range of transcripts than do more differentiated cell types. Quantifying Shannon entropy of the transcripts expressed in each cell, the approach assigns direction so that cell entropy decreases with differentiation (Fig. 5e). Such an entropy heuristic has been shown to work on a number of developmental examples, but does not generalize to other types of dynamic processes.

Independent of the 'ant trail' patterns of cell density, additional information about dynamical processes can be gleaned by inferring kinetics of mRNA lifecycle from static scRNA-seq snapshots. Examining the balance of mRNA molecules at different stages of their lifecycle (that is, nascent transcription, processing, and degradation), one can infer the dynamics of a cell around the time the snapshot was taken. For example, if one observes a cell producing more nascent pre-mRNA molecules than is needed to compensate for degradation of the existing mRNAs, then we can predict that the abundance of that transcript would increase at the next point in time. In a collaboration between my and Sten Linnarsson's groups, we have shown that rare intronic reads captured by the common scRNA-seq protocols can be used to quantify the relative abundance unspliced pre-mRNA and mature (spliced) mRNA for many transcripts[70]. These quantities can be combined in a simple model of mRNA kinetics to estimate the first time derivative of the transcriptional state—termed RNA velocity—which can then be used to approximate what the state of the cell would have been shortly before or after the time of the measurement. RNA velocity can reveal the patterns of complex flows, including branches, cycles, or counterflows (Fig. 5c,f). While this approach does not require the assumptions of ergodicity associated with the density-based methods, it has other limitations. For one, the underlying dynamical process has to occur on a timescale comparable with that of mRNA splicing kinetics. The estimates are currently limited to transcripts with sufficiently long intronic regions (those incorporating internal priming sites). The estimates can carry considerable uncertainty, and commonly average across cell neighborhoods to improve robustness. The RNA velocity approach, nevertheless, demonstrates that temporal dynamics can be inferred from a more detailed consideration of the molecular state in a static molecular snapshot of a cell. More elaborate parameter fitting techniques have been developed to improve the accuracy of such inference[71]. The predictive ability will likely improve with protocols that capture different aspects of mRNA lifecycle, epigenetic, or protein states. For instance, two groups have recently demonstrated how 4-thiouridine pulse labeling can be used to obtain more accurate RNA dynamics estimates in the context of cell culture and in vitro perturbations[72,73]

Going forward: integrating samples, molecular modalities, and physical space

To date, most analytical efforts in single-cell genomics have been aimed at overcoming the challenges posed by inherently sparse and noisy, yet informative, nature of the data. This is also the focus of the current review. As robustness and accessibility of the scRNA-seq assays have increased over the past five years, scRNA-seq has become an important tool for biological investigations. Tackling real-world

problems necessitates more elaborate experimental designs. A study of a disease, for example, would typically involve comparison across many samples from different groups of individuals. Other studies would involve analysis of longitudinal sample collections and multiple tissues or sample types. The progress in scRNA-seq has spurred single-cell measurements of other molecular modalities, such as DNA methylation[74], chromatin accessibility[75], or protein abundance[76]. Spatially resolved transcriptomics is being developed to provide valuable information about the context of each cell[77–80]. All these advances require development of improved analysis methods that consider additional layers of information. Most of the lessons learned from the analysis of individual scRNA-seq datasets, however, remain applicable.

For example, a number of approaches have been developed to 'align' single-cell datasets, that is to identify corresponding cell populations across collections of datasets[42,52] (see Luecken et al.[41] for a recent comparison). Most such methods rely on the same normalization and dimensionality-reduction steps, extending neighborhood graphs representation to include the mapping of cells between datasets for downstream integrative analysis[24,42,54]. These methods represent a generalization of the earlier batch-correction problem. For instance, analyses of datasets involving other molecular modalities, such as chromatin accessibility, have been carried out using the same normalization and integration techniques[24,54,75,81]. Such approaches are likely to remain important, even as new computational methods are developed to keep up with the increasing volumes and complexity of the datasets, as well as new single-cell assays.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at [https://doi.org/10.1038/s41592-021-01171-x].

Received: 8 November 2018; Accepted: 29 April 2021;
Published online: 21 June 2021

References

1. Ding, J. et al. Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. *Nat. Biotechnol.* **38**, 737–746 (2020).
2. Hagemann-Jensen, M. et al. Single-cell RNA counting at allele and isoform resolution using Smart-seq3. *Nat. Biotechnol.* **38**, 708–714 (2020).
3. Brennecke, P. et al. Accounting for technical noise in single-cell RNA-seq experiments. *Nat. Methods* **10**, 1093–1095 (2013).
4. Kharchenko, P. V., Silberstein, L. & Scadden, D. T. Bayesian approach to single-cell differential expression analysis. *Nat. Methods* **11**, 740–742 (2014).
5. Vu, T. N. et al. Beta-Poisson model for single-cell RNA-seq data analyses. *Bioinformatics* **32**, 2128–2135 (2016).
6. Svensson, V. Droplet scRNA-seq is not zero-inflated. *Nat. Biotechnol.* **38**, 147–150 (2020).
7. Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S. & Vert, J. P. A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat. Commun.* **9**, 284 (2018).
8. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **15**, 1053–1058 (2018).
9. Wang, T., Li, B., Nelson, C. E. & Nabavi, S. Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. *BMC Bioinf.* **20**, 40 (2019).
10. Soneson, C. & Robinson, M. D. Bias, robustness and scalability in single-cell differential expression analysis. *Nat. Methods* **15**, 255–261 (2018).
11. Vallejos, C. A., Richardson, S. & Marioni, J. C. Beyond comparisons of means: understanding changes in gene expression at the single-cell level. *Genome Biol* **17**, 70 (2016).
12. Nabavi, S., Schmolze, D., Maitituoheti, M., Malladi, S. & Beck, A. H. EMDomics: a robust and powerful method for the identification of genes differentially expressed between heterogeneous classes. *Bioinformatics* **32**, 533–541 (2016).
13. Korthauer, K. D. et al. A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. *Genome Biol.* **17**, 222 (2016).
14. Martinez-Jimenez, C. P. et al. Aging increases cell-to-cell transcriptional variability upon immune stimulation. *Science* **355**, 1433–1436 (2017).
15. Crowell, H. L. et al. muscat detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data. *Nat. Commun.* **11**, 6077 (2020).
16. Jaitin, D. A. et al. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* **343**, 776–779 (2014).
17. Ntranos, V., Kamath, G. M., Zhang, J. M., Pachter, L. & Tse, D. N. Fast and accurate single-cell RNA-seq analysis by clustering of transcript-compatibility counts. *Genome Biol.* **17**, 112 (2016).
18. Aggarwal, C. C., Hinneburg, A. & Keim, D. A. in *Database Theory — ICDT 2001*. (eds Van den Bussche, J. & Vianu, V.) 420–434 (Springer Berlin Heidelberg, 2001).
19. Jiang, L. et al. Synthetic spike-in standards for RNA-seq experiments. *Genome Res.* **21**, 1543–1551 (2011).
20. Fan, J. et al. Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. *Nat. Methods* **13**, 241–244 (2016).
21. Eling, N., Richard, A. C., Richardson, S., Marioni, J. C. & Vallejos, C. A. Correcting the mean-variance dependency for differential variability testing using single-cell RNA sequencing data. *Cell Syst.* **7**, 284–294 (2018).
22. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.* **20**, 296 (2019).
23. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014).
24. Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902 e1821 (2019).
25. Azizi, E. et al. Single-cell map of diverse immune phenotypes in the breast tumor microenvironment. *Cell* **174**, 1293–1308 (2018).
26. Shao, C. & Hofer, T. Robust classification of single-cell transcriptome data by nonnegative matrix factorization. *Bioinformatics* **33**, 235–242 (2017).
27. Zhu, X., Ching, T., Pan, X., Weissman, S. M. & Garmire, L. Detecting heterogeneity in single-cell RNA-seq data by non-negative matrix factorization. *PeerJ* **5**, e2888 (2017).
28. Pierson, E. & Yau, C. ZIFA: dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biol* **16**, 241 (2015).
29. Zhou, M. Nonparametric Bayesian negative binomial factor analysis. *Bayesian Analysis* **13**, 1065–1093 (2018).
30. Zhang, L. & Mallick, B. K. Inferring gene networks from discrete expression data. *Biostatistics* **14**, 708–722 (2013).
31. Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S. & Vert, J. P. Publisher Correction: A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat. Commun.* **10**, 646 (2019).
32. Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S. & Theis, F. J. Single-cell RNA-seq denoising using a deep count autoencoder. *Nat. Commun.* **10**, 390 (2019).
33. Amodio, M. et al. Exploring single-cell data with deep multitasking neural networks. *Nat. Methods* **16**, 1139–1145 (2019).
34. Aggarwal, C. C. *Neural Networks and Deep Learning: A Textbook*. (Springer International Publishing, 2018).
35. Svensson, V., Gayoso, A., Yosef, N. & Pachter, L. Interpretable factor models of single-cell RNA-seq via variational autoencoders. *Bioinformatics* **36**, 3418–3421 (2020).
36. Islam, S. et al. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res.* **21**, 1160–1167 (2011).
37. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).
38. Levine, J. H. et al. Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell* **162**, 184–197 (2015).
39. Haghverdi, L., Buettner, F. & Theis, F. J. Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics* **31**, 2989–2998 (2015).
40. Jarvis, R. A. & Patrick, E. A. Clustering using a similarity measure based on shared near neighbors. *IEEE Trans. Comput.* **C-22**, 1025–1034 (1973).
41. Luecken, M. D. et al. Benchmarking atlas-level data integration in single-cell genomics. Preprint at *bioRxiv* https://doi.org/10.1101/2020.05.22.111161 (2020).
42. Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* **36**, 421–427 (2018).
43. Van Mieghem, P. *Graph Spectra for Complex Networks*. (Cambridge University Press, 2010).
44. Haghverdi, L., Buttner, M., Wolf, F. A., Buettner, F. & Theis, F. J. Diffusion pseudotime robustly reconstructs lineage branching. *Nat. Methods* **13**, 845–848 (2016).
45. Maaten, L. V. D. & Hinton, G. Visualizing data using *t*-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
46. McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. Preprint at https://arxiv.org/abs/1802.03426 (2018).
47. Amir el, A. D. et al. viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nat. Biotechnol.* **31**, 545–552 (2013).

48. Becht, E. et al. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* **37**, 38–44 (2018).
49. Kiselev, V. Y., Andrews, T. S. & Hemberg, M. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat. Rev. Genet.* **20**, 273–282 (2019).
50. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech.* **2008**, P10008 (2008).
51. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* **9**, 5233 (2019).
52. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).
53. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
54. Barkas, N. et al. Joint analysis of heterogeneous single-cell RNA-seq dataset collections. *Nat. Methods* **16**, 695–698 (2019).
55. Pons, P. & Latapy, M. Computing communities in large networks using random walks. *J. Graph Algorithms Appl.* **10**, 191–218 (2006).
56. Gorban, A. N. & Zinovyev, A. Y. in *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques* 28–59 (IGI Global, 2010).
57. Hastie, T. & Stuetzle, W. Principal curves. *J. Am. Stat. Assoc.* **84**, 502–516 (1989).
58. Trapnell, C. et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).
59. Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* **37**, 547–554 (2019).
60. Soldatov, R. et al. Spatiotemporal structure of cell fate decisions in murine neural crest. *Science* **364**, eaas9536 (2019).
61. Ji, Z. & Ji, H. TSCAN: pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic Acids Res.* **44**, e11 (2016).
62. Shin, J. et al. Single-cell RNA-seq with waterfall reveals molecular cascades underlying adult neurogenesis. *Cell Stem Cell* **17**, 360–372 (2015).
63. Street, K. et al. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* **19**, 477 (2018).
64. Wolf, F. A. et al. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol.* **20**, 1–9 (2019).
65. Hrvatin, S. et al. Single-cell analysis of experience-dependent transcriptomic states in the mouse visual cortex. *Nat. Neurosci.* **21**, 120–129 (2018).
66. Schiebinger, G. et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell* **176**, 928–943(2019).
67. Tran, T. N. & Bader, G. D. Tempora: cell trajectory inference using time-series single-cell RNA sequencing data. *PLoS Comput. Biol.* **16**, e1008205 (2020).
68. Weinreb, C., Wolock, S., Tusi, B. K., Socolovsky, M. & Klein, A. M. Fundamental limits on dynamic inference from single-cell snapshots. *Proc. Natl Acad. Sci. USA* **115**, E2467(2018).
69. Grun, D. et al. De novo prediction of stem cell identity using single-cell transcriptome data. *Cell Stem Cell* **19**, 266–277 (2016).
70. La Manno, G. et al. RNA velocity of single cells. *Nature* **560**, 494–498 (2018).
71. Bergen, V., Lange, M., Peidli, S., Wolf, F. A. & Theis, F. J. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat. Biotechnol.* **38**, 1408–1414 (2020).
72. Cao, J., Zhou, W., Steemers, F., Trapnell, C. & Shendure, J. Sci-fate characterizes the dynamics of gene expression in single cells. *Nat. Biotechnol.* **38**, 980–988 (2020).
73. Erhard, F. et al. scSLAM-seq reveals core features of transcription dynamics in single cells. *Nature* **571**, 419–423 (2019).
74. Welch, J. D. et al. Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell* **177**, 1873–1887(2019).
75. Cusanovich, D. A. et al. A single-cell atlas of in vivo mammalian chromatin accessibility. *Cell* **174**, 1309–1324(2018).
76. Stoeckius, M. et al. Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* **14**, 865–868 (2017).
77. Lein, E., Borm, L. E. & Linnarsson, S. The promise of spatial transcriptomics for neuroscience in the era of molecular cell typing. *Science* **358**, 64–69 (2017).
78. Moffitt, J. R. et al. Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science* **362**, eaau5324 (2018).
79. Eng, C.-H. L. et al. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature* **568**, 235–239 (2019).
80. Rodriques, S. G. et al. Slide-seq: a scalable technology for measuring genome-wide expression at high spatial resolution. *Science* **363**, 1463–1467 (2019).
81. Lake, B. B. et al. Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nat. Biotechnol.* **36**, 70–80 (2018).
82. Angerer, P. et al. destiny: diffusion maps for large-scale single-cell data in R. *Bioinformatics* **32**, 1241–1243 (2016).
83. Svensson, V., da Veiga Beltrame, E. & Pachter, L. A curated database reveals trends in single-cell transcriptomics. *Database (Oxford)* **2020**, baaa073 (2020).
84. Byrnes, L. E. et al. Lineage dynamics of murine pancreatic development at single-cell resolution. *Nat. Commun.* **9**, 3922 (2018).
85. Lo Giudice, Q., Leleu, M., La Manno, G. & Fabre, P. J. Single-cell transcriptional logic of cell-fate specification and axon guidance in early-born retinal neurons. *Development* **146**, dev178103 (2019).
86. Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* **15**, e8746 (2019).

Methods

Balance between the number of cells and depth. Two cell populations were simulated using splatter[87], using the following parameters: group.prob = c(0.5, 0.5), nGenes = 5,000, batchCells = 5×10^5, lib.loc = 12. A total of ten such matrices were generated with different random seeds. For each round of simulations, the simulated dataset was randomly subsampled to include the desired number of genes and molecules (2× the number of cells were simulated to perform test/validation, as described below). A total of ten randomized subsampling rounds were performed for each of the ten simulated matrixes. The ability to distinguish the simulated subpopulations was evaluated as follows: the count matrix was variance normalized using pagoda2 (n.odgenes = 1,000) and first two principal components were estimated. Linear discriminant analysis model (MASS::lda) was built using 50% of the cells, and the ROC AUC was evaluated on the other 50% of the cells (using ROCR R package).

Estimating expression magnitude posteriors. A Poisson model of measurement was assumed. 5 cells were randomly selected from the $CD8^+$ naive and $CD8^+$ effector populations. The *IL32* gene was used as an example. For each cell, the posterior probability of the true *IL32* expression level being at λ was evaluated as $p(c \mid \lambda) = \text{dpois}(c, s \times \lambda)$, where dpois() is the Poisson distribution probability density function, c is the number of UMIs of the *IL32* gene observed in a given cell, and s is the total number of UMIs observed in that cell. Joint posterior for each population was calculated as a product of individual cell posteriors.

Number of differentially expressed genes as a function of the number of cells. The 66k PBMC dataset was downloaded from 10x Genomics. Annotation was determined based on alignment with the 10k PBMC dataset (using conos) and manual examination of markers. Differential expression was tested between $CD4^+$ T cell and $CD14^+$ monocyte clusters. To evaluate how the number of differentially expressed genes changes as the size of the populations being compared increases, random sets of $CD4^+$ T cells and $CD14^+$ monocytes were drawn from these populations without replacement. The size ranged from 10 to 10,000 cells per population. The Wilcoxon rank-sum test was used to assess differential expression, with genes considered differentially expressed if the test P value was below 1% (no multiple hypothesis corrections were applied).

Mean–variance relationships. The 10k PBMC dataset from 10x Genomics was used. To illustrate the mean–variance relationship on the linear scale (Fig. 3e), cells from the $CD4^+$ naive population were used. The UMI counts were normalized as $c' = c \times 5{,}000 / s$, where s is the total number of UMIs observed in a given cell. The quadratic fit for the negative binomial was calculated using the formula $y \sim \text{offset}(x) + I(x^2)$, where y is the variance and x is the mean expression.

For Fig. 3f–i, the dataset variance normalization was performed using pagoda2.

Trajectories, entropy, and RNA velocity. The trajectories were fit in the two dimensions of the embedding, using SimplePPT[88], as implemented in Soldatov et al.[60]. Expression entropy was calculated using the entropy package in R, randomly subsampling each cell to 1,000 UMIs. RNA velocity estimation was performed using the scVelo package[71].

Systematic detection bias of different scRNA-seq platforms. Measurements of different cell lines on different scRNA-seq platforms were taken from Tian et al.[89]. 10x Chromium and Drop-seq platforms were compared. Differentially detected genes, determined using the H1975 cell line data, were those showing a tenfold difference in expression magnitude (Extended Data Fig. 1b). This set of genes was then examined visually in the scatter plots for the other two cell lines.

Sensitivity of the distance metrics. To demonstrate the relative sensitivity of different distance measures in spaces of increasing dimensions, a set of 100 points was randomly drawn from a k-dimensional uniform distribution. A total of 1,000 such draws were performed for each k. Extended Data Fig. 1e shows median distance range ratio across the 1,000 draws.

Clustering. The E12.5 timepoint from the fetal pancreas dataset[84] (GSM3140915) was used. Normalization and embedding were performed using pagoda2. To show subpopulations associated with cell cycle, Leiden clustering was performed with resolution $r = 5$, and clusters with high Mki67 signal were aggregated within each major population (Fig. 4f). The hierarchical clustering of the resulting clusters (Fig. 4g) was calculated using correlation distance on CPM values estimated from aggregated transcript counts for each cluster, using 'ward.D2' linkage in R 'hclust()' procedure (Figs. 1 and 2).

Data availability

The following scRNA-seq datasets were used in creating example figures:

- 10x Genomics PBMC 10k (https://support.10xgenomics.com/single-cell-gene-expression/datasets/3.0.0/pbmc_10k_v3).
- 10x Genomics PBMC 66k (https://support.10xgenomics.com/single-cell-gene-expression/datasets/3.1.0/5k_pbmc_NGSC3_aggr).
- Fetal pancreas: E12.5 timepoint data from Byrnes et al.[84] were downloaded from GEO (GSM3140915).
- Mouse developing retina: 10x Chromium replicate from Lo Giudice et al.[85] was downloaded from GEO (GSM3466902).
- Cell lines: Benchmarking data measuring different cell lines on different platforms, taken from Tian et al.[89], were downloaded from GEO (GSE118767).
- Metadata on the single-cell RNA-seq experiments were taken from http://www.nxn.se/single-cell-studies/.

Code availability

The notebooks and scripts for the figures presented in the paper can be found on the author's website: http://pklab.med.harvard.edu/peterk/review2020/.

References

87. Zappia, L., Phipson, B. & Oshlack, A. Splatter: simulation of single-cell RNA sequencing data. *Genome Biol.* **18**, 174 (2017).
88. Mao, Q., Yang, L., Wang, L., Goodison, S. & Sun, Y. in *Proceedings of the 2015 SIAM International Conference on Data Mining* 792–800 (SIAM, 2015).
89. Tian, L. et al. Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. *Nat. Methods* **16**, 479–487 (2019).

Acknowledgements

P.V.K was supported by the NHLBI R01HL131768 award from NIH and CAREER (NSF-14-532) award from NSF.

Competing interests

P.V.K. serves on the scientific advisory boards of Celsius Therapeutics and Biomage Inc.

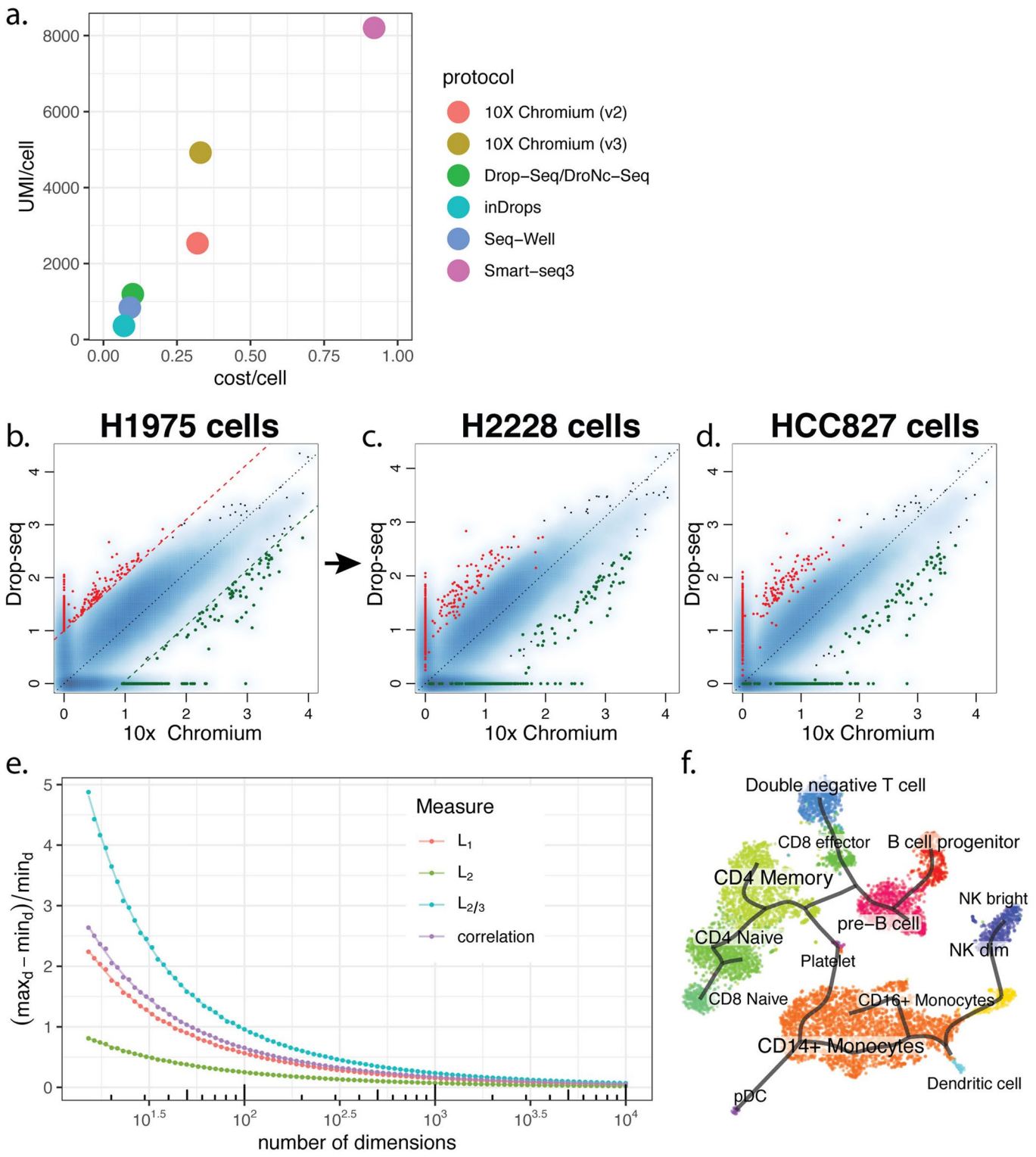
Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41592-021-01171-x.

Correspondence should be addressed to P.V.K.

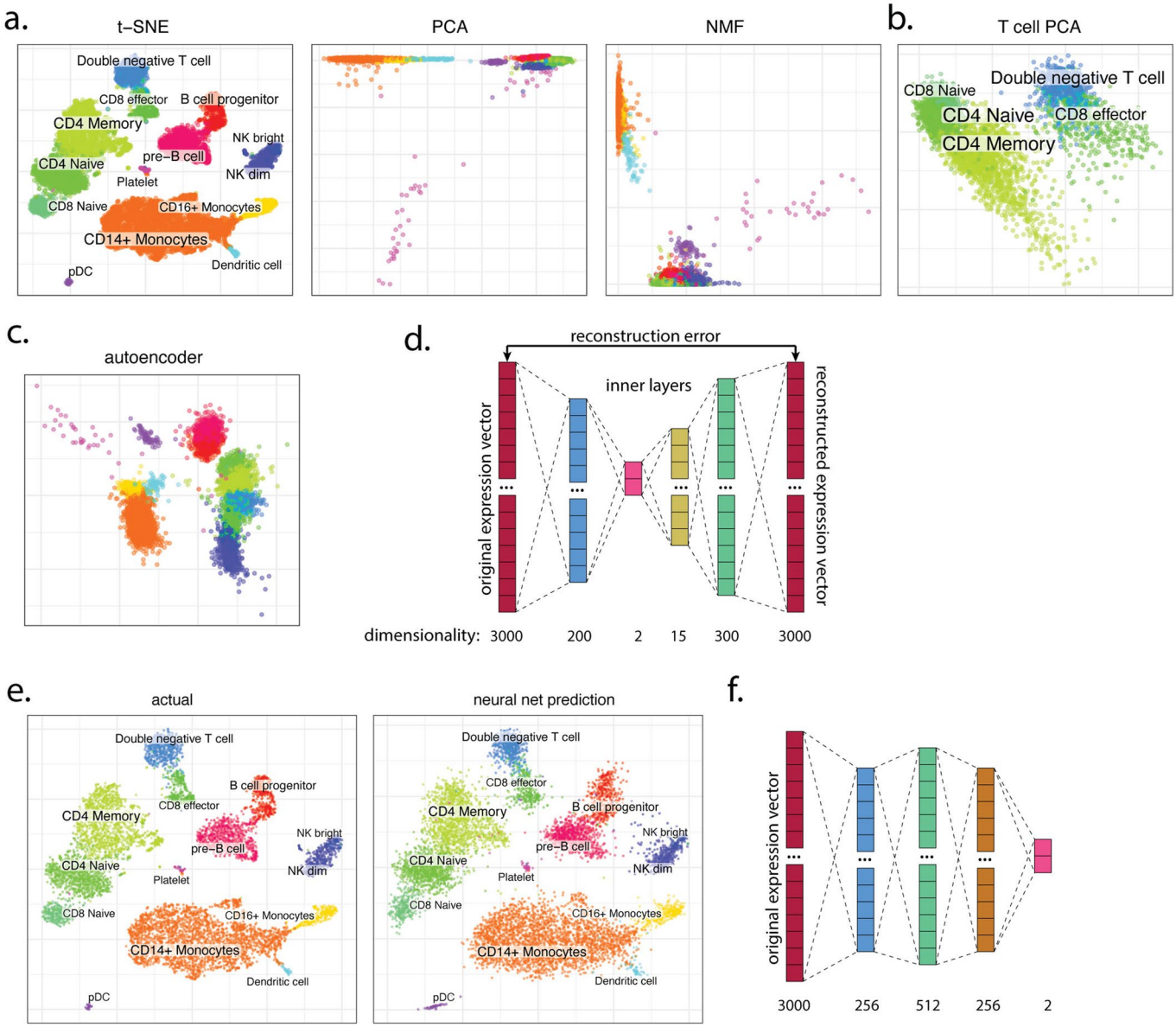
Peer review information *Nature Methods* thanks Martin Hemberg, Michael Morgan and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Lei Tang was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Reprints and permissions information is available at www.nature.com/reprints.



Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Properties of scRNA-seq measurements. a. Dependency between cost per cell (x axis) and the expected depth (UMIs per cell, y axis) is shown for a number of popular methods, largely based on the assessment by Ding. et al.[1]. **b–d**. Systematic transcript-specific bias of different scRNA-seq protocols. **b**. The scatter plot shows average log10(CPM+1) values for different genes (each dot represents a gene), as assessed using 10x Chromium (x axis) or dropseq (y axis) platforms. Genes showing higher (red) or lower (green) expression (above 10-fold threshold) are highlighted. **c, d**. Similar scatter plots shown for other two cell lines: H2228 (b) and HCC827 (c) cells. The set of differential genes determined from analysis of the H1975 cell line (a) is shown. Most of the genes that showed large discrepancy in the detection rate in H1975 results also show same discrepancy in the other two cell lines, illustrating stable detection bias between the two platforms. **e**. The ability to distinguish nearest neighbors decreases as the dimensionality of the space increases. The difference between closest (min_d) and furthest (max_d) points from the origin, normalized by min_d (y axis) is shown for different distance measures as a function of increasing number of dimensions (x axis). For each dimensionality n, a set of 100 random points are drawn from the n-dimensional uniform distribution, and a median of 1000 draws is shown. The distinction between closest and furthest points approaches 0 at high dimensions. In other words, relative to the origin, in high-dimensional space the points appear to be distributed on the surface of a high-dimensional sphere. **f**. Principal tree fit to the PBMC10k dataset. The tree shows computationally optimal spanning of the PBMC populations, yet the interpreting it as a dynamic process is incorrect.



Extended Data Fig. 2 | Dimensionality reduction and neural networks. a. A t-SNE embedding of the PBMC10k dataset (left); projection of cells onto the first two principal components (middle); projection of cells onto first two basis of the non-negative matrix factorization (right); **b**. Projection of cells onto the first two principal components, based on re-analysis of a subset of the PBMC10k dataset that contains only T lymphocytes. Given this restricted cellular context, the first two components are much better at capturing separation between different subsets of T cells, compared to the PCA on the full dataset shown in the previous panel. **c**. Visualization of the PBMC10k dataset in the 2D latent space determined by an autoencoder structure shown in (d). **d**. The architecture of an autoencoder used to reduce dimensions of the PBMC10k dataset in the previous panel. The autoencoder starts with a vector of top 3000 most variable genes, and then for each cell transforms this expression profile through a series of non-linear transformations, first into increasingly narrow dimensions, culminating in a two-dimensional middle layer, and then back into a full 3000-dimensional vector. The values of the two-dimensional middle layer are shown in (d). The parameters of the transformations connecting each layer are optimized so that they minimize the discrepancy between the original expression vector (leftmost layer) and the reconstructed vector (rightmost layer). **e, f**. Using neural networks to learn non-linear mapping from high-dimensional expression state to the coordinates of a t-SNE embedding. As t-SNE embeddings are based on empirical optimization of the relative positions of neighboring cells, there is no obvious analytical function connecting the expression state with the resulting t-SNE coordinates. Neural networks, however, can be used to approximate highly nonlinear and noisy functions. Here, a neural network with an architecture shown in (f) was used to approximate such a function. The parameters of the transformations connecting the layers were optimized based on a training set of 3000 cells, and then an additional set of 3000 test cells was used to illustrate the resulting fit. The left panel in (e) shows the actual positions of the 3000 test cells, and the right plot shows the positions predicted by the trained network.