Title

Revealing the vectors of cellular identity with single-cell genomics.

Journal

Nature Biotechnology, 34(11)

Authors

Wagner, Allon

Regev, Aviv

Yosef, Nir

Publication Date

2016-11-08

DOI

10.1038/nbt.3711

Peer reviewed

HHS Public Access

Author manuscript

Nat Biotechnol. Author manuscript; available in PMC 2017 November 08.

Published in final edited form as:

Nat Biotechnol. 2016 November 08; 34(11): 1145–1160. doi:10.1038/nbt.3711.

Revealing the vectors of cellular identity with single-cell genomics

Allon Wagner[1], **Aviv Regev**[2,3,5], and **Nir Yosef**[1,4,5]

[1]Department of Electrical Engineering and Computer Science and the Center for Computational Biology, University of California, Berkeley, USA

[2]Howard Hughes Medical Institute, Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

[3]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA

[4]Ragon Institute of Massachusetts General Hospital, Massachusetts Institute of Technology, and Harvard University, Boston, Massachusetts, USA

Abstract

Single-cell genomics has now made it possible to create a comprehensive atlas of human cells. At the same time, it has reopened definitions of a cell's identity and type and of the ways in which they are regulated by the cell's molecular circuitry. Emerging computational analysis methods, especially in single-cell RNA sequencing (scRNA-seq), have already begun to reveal, in a data-driven way, the diverse simultaneous facets of a cell's identity, from a taxonomy of discrete cell types to continuous dynamic transitions and spatial locations. These developments will eventually allow a cell to be represented as a superposition of 'basis vectors', each determining a different (but possibly dependent) aspect of cellular organization and function. However, computational methods must also overcome considerable challenges—from handling technical noise and data scale to forming new abstractions of biology. As the scale of single-cell experiments continues to increase, new computational approaches will be essential for constructing and characterizing a reference map of cell identities.

To understand cells—the basic unit of life—we must not only catalog them and their molecular profiles but also determine the factors that shape them. A cell's identity, which is reflected in its molecular profile, is formed by the instantaneous intersection of multiple factors. These include its position in a taxonomy of cell types, the progress of multiple time-dependent processes that take place simultaneously, its response to signals from its local environment, and the precise location and neighborhood in which it resides (Fig. 1a). The factors that together span the space of possible cell states can be likened to the basis vectors

Correspondence should be addressed to A.R. (aregev@broadinstitute.org).
[5]These authors contributed equally to this work.

COMPETING FINANCIAL INTERESTS

A.R. is a member of the Scientific Advisory Board for Thermo Fisher Scientific and Syros Pharmaceuticals and a consultant for Driver Group. A.W. and N.Y. declare no competing financial interests.

that span a linear space, yet, unlike basis vectors, they may be intricately dependent on one another (Fig. 1b and Box 1).

Until recently, most genomic profiling studies have analyzed cell populations, although even cells of the same 'type' can exhibit substantial heterogeneity, reflecting finer sub-types, regulated functional variation, or inherent stochasticity[1–6]. However, over the past several years, rapid technological advances have enabled genome-wide profiling of RNA[7–10], DNA[11–17], protein[18–21], epigenetic modifications[22–26], chromatin accessibility[27,28], and other molecular events[29] in single cells. The scale and precision of such studies has continued to increase, reaching tens of thousands of cells for massively parallel scRNA-seq[30,31] and millions of cells for mass cytometry signature protein measurements[32–34]. We use the term single-cell genomics to describe all genome-scale measurements, from DNA to RNA to proteome.

Large-scale single-cell data allow us to address biological questions that were previously out of reach. First, we can now explore the identity of an individual cell and the factors underlying it through the comprehensive lens of the cell's unique molecular profile (Fig. 1b). By decomposing this profile to its separate components, it should be possible to determine, in a data-driven way, the specific physiological and molecular features of each of these factors, without relying on prior definitions, hypotheses, or markers. Second, a project to construct a comprehensive atlas of all human cell types and sub-types—including their activity states, dynamic transitions, physical locations, and lineage relationships through development—has become a tangible goal. Even preliminary progress toward such an atlas would help elucidate the organization and function of tissues in health and disease. In addition, single-cell data allow us to study the regulatory circuitry that governs cells at a resolution that had been impossible with data collected from bulk cell populations. Finally, single cells are the basic component of complex tissues. Through deconvolution of complex samples[30], such as tumor biopsies[35], one may infer their cellular milieu, and characterize the rare[5], functionally important[1], and unknown[36] cell types they contain.

Although the new single-cell data types harbor a wealth of information, they also pose specific analytical and technical challenges. The analytical challenges include (1) designing experiments and performing power analysis (e.g., how many cells do we need to profile for a given task? At what depth?); (2) preprocessing to distinguish biological from technical variation, especially false-negative gene detections (dropouts); (3) inferring the key aspects of a cell atlas, from discrete sub-types to continuous spatiotemporal ordering of cells; and (4) deriving molecular mechanisms from cell-to-cell variation. In each of these areas, we must grapple with common technical challenges, such as noise, sparsity, and false negatives; ever-increasing scale, which defies many traditional implementations of basic tasks in genomics; partial dependencies between the multiple facets of a cell's identity (its type, state, position, etc.), such that variation in one biological dimension may be a confounder for another; and the need for accessible and interpretable visualizations. Novel computational methods are needed to overcome these challenges and exploit the biological signals in single-cell data (Fig. 2, bottom).

Here we review key questions, progress, and open challenges in the development of computational methods in single-cell functional genomics, focusing primarily on scRNA-seq (we do not discuss single-cell genome analysis, as it was recently reviewed elsewhere[37]). We first distinguish key sources of variation in single cells, and experimental and computational strategies to tease them apart and to mitigate the effects of technical (unwanted) variation in order to explore the biological variation in the data. We highlight key current methods that can characterize the diverse factors involved in determining cellular identity, including cell type taxonomies, continuous phenotypes, temporal progression (on linear, bifurcating, or cyclic trajectories), and spatial position in the tissue. We close with areas of substantial opportunity and challenges for future research, including emerging methods that harness single-cell data to dissect the molecular circuitry, unique challenges associated with studying the single-cell epigenome, and open problems associated with the increasing scale of single-cell experiments, the integration of diverse single-cell assays, and the use of these data to illuminate the organization of complex tissues.

Addressing technical variation in single-cell RNA-seq

We distinguish three sources of variation in scRNA-seq (Fig. 2, top). The first is technical variation, which is due to factors such as differences in cell integrity and lysis, RNA capture and cDNA conversion, and detection[38,39]. The second is allele-intrinsic variation, namely stochastic factors intrinsic to the molecular mechanisms that control gene expression[40–42]. For example, the bursting statistics of transcriptional initiation coupled to variable rates of mRNA degradation can lead to fluctuations in transcript levels over time in one cell, and to differences between otherwise ‘identical’ cells measured at a single time point. This inherent stochasticity does not correlate between two alleles of the same gene. The third is allele-extrinsic variation, due to factors[42,43] extrinsic to the process of transcription, such as the presence of certain regulators or differences in stable chromatin state. These factors contribute to establishing differences between cell types or states, either stably[30] or transiently, but are correlated between two alleles of the same gene[41].

Although most studies aim to understand allele-extrinsic variation and its function, technical and allele-intrinsic variations are major confounders. Some technical variation (Fig. 3) is common to both scRNA-seq and bulk (population) RNA-seq, whereas several other factors — including zero inflation due to false negatives, overamplification, and cell doublets—are specific to the technical variation between single-cell profiles. In some cases, the extent of technical variation is affected by biological differences, undermining definitions of quality and limiting our ability to remove technical variation. For example, because smaller cells typically harbor less RNA, they appear to be lower in quality. Similarly, some cell types may be harder to capture or lyse. Finally, some cells are characterized by transcriptional profiles functionally dominated by very high expression of a few transcripts, whereas others have far more complex transcriptomes. Indeed, a recent study[44] reported that technical quality features were highly correlated with biological cell type.

Handling known and unknown sources of technical variation

Batch effects and technical variation due to unknown (unmodeled) confounders are common in both scRNA-seq and bulk RNA-seq. Batch effects have long been recognized as a source of non-biological variation in gene expression profiles between sets of samples that were prepared or processed together[45]. Whenever possible, careful experimental design, such as a uniform distribution of replicates across batches or within a plate, can help mitigate these effects (Fig. 3a). However, such designs are not always practical in scRNA-seq, where initial sample processing steps (e.g., dissociation and sorting) have a major impact and where samples must often be processed fresh. As an alternative, statistical methods can model batches as known covariates (Fig. 3d), and either remove batch terms[46–48] or directly incorporate them into downstream analyses (e.g., for detecting differentially expressed genes[49,50]).

As in bulk RNA-seq[51], normalization of technical library-to-library differences is a prerequisite for further analysis. However, scRNA-seq (where a library equals a cell) faces complexities that do not occur in bulk experiments because of technical factors that depend on biological differences between cells. In bulk RNA-seq it is generally (yet not always[52]) safe to assume that the starting quantities of RNA are uniform across libraries and to scale read counts accordingly with size factors[53] or relative expression units, such as RPKM[54] (reads per kilobase of exon per million mapped reads), FPKM[55] (fragments per kilobase of exon per million mapped fragments), or TPM[56,57] (transcripts per million). In contrast, the starting quantities of RNA in single-cell libraries depend both on technical factors and on biological factors, such as cell size and type. The distribution of transcript levels can also vary between cells. For example, the transcriptome of some ‘professional’ secretory cells is mostly devoted to a few exceptionally highly expressed transcripts. Such biological factors can be considered a nuisance factor in some contexts but part of the relevant biological variability in others[58]. A further complication arises from the 3′-end coverage bias characteristic of some of the popular scRNA-seq protocols[8,9]. When a transcript’s expression is normalized by its length, as in the RPKM, FPKM, and TPM units, the bias leads to an underestimation of the expression of longer transcripts[59].

Several studies have addressed the most prominent cell-to-cell bias[60], namely the combined effect of the cell-specific library complexity and the amount of cellular RNA, which together result in large variations in the number of genes detected in each cell (see below on dropouts). One approach is to scale each cell with two alternative[59,61] or cumulative[62] size factors (instead of one as in bulk[53]) to account for the technical and biological components of the bias. However, the high frequency of undetected genes (dropouts) may undermine[63] the utility of standard size factors. Instead, one may pool the counts of multiple cells together to decrease the prevalence of undetected genes, obtain robust size factors for the cell pools, and finally deconvolve the pools’ size factors into cell size factors[63]. Some approaches directly incorporate the scaling terms in a statistical model for the data rather than use them in a separate preprocessing step. For example, BASiCS[61,224] incorporates scale factors into a Bayesian model of read counts[62], MAST[58] incorporates the number of detected genes, a proxy of this bias, into generalized linear models, and BISCUIT[220] models and infers cell-specific scaling parameters simultaneously with cell clusters.

Other approaches correct the biases without explicitly modeling cell-specific size factors. The main axes of variation (i.e., first few principal components) in scRNA-seq data are often dominated by additional technical factors[1,60], such as library complexity[64], proportion of reads that were successfully aligned to the genome, and proportion of detected transcripts (Fig. 3b). While several studies have opted to remove these axes of variation entirely[27,65], this may result in eliminating important aspects of the biological signal. An alternative (N.Y., A.R. and colleagues[1]) is to define an overall quality score for every cell (taken as the first (few) principal component of the matrix of the quality metrics per cell) and explicitly remove its effect using global scaling normalization[66]. A less-supervised strategy (Fig. 3d) handles technical confounders as latent (i.e., unobserved) *unknown factors*, which can be inferred from the expression data, using negative control genes that are assumed a priori to be uniformly expressed by all the cells in the sample. The main axes of variation over these genes are designated as a "technical variability" component that is then either subtracted from the data or taken into account in the downstream analysis[49,50,67]. Finally, when *a priori control genes* are unavailable, but samples from different conditions are available, one may regress the data on the biological covariates of interest, and then designate the main axes of variation of the residual (i.e., the component of the data that is orthogonal to the biological covariates) as the technical variability component[67,68].

Alleviating technical biases using experimental cues

Experimental measures, especially spike-in controls, can help capture and account for some, but not all, of the technical confounders in scRNA-seq. Spike-ins are exogenous RNA sequences (e.g., 92–96 sequences from the External RNA Control Consortium (ERCC)[69]) that are added in known quantities during library preparation and are assumed to be unaffected by the biological covariates. They thus constitute a well-defined set of negative controls for the purpose of adjusting for the difference in total RNA content between cells[61,70], as well as for quality diagnostics of libraries[71] and experiments[72]. A series of studies[38,61,73] decomposed biological from technical variability in scRNA-seq data by using spike-ins to learn a model for technical noise. For example, since the technical noise affecting a gene depends on its expression (weakly expressed genes are more technically noisy[8]), spike-ins can be used to parametrize a model of this dependence, and in turn to construct a statistical test for detecting biologically variable genes[61]. However, although spike-in controls provide important information[59], they do not control for any technical variation preceding their introduction, especially tissue acquisition, dissociation, sorting and lysis[30]. The utility of the ERCC spike-ins in modeling technical noise is further curtailed by differences between them and endogenous RNA[59,74], and it has been reported[67] that their counts may be affected by biological (rather than technical) factors in an experiment and that technical covariates may affect spike-ins differently than genes.

Pool/split experimental designs can further assist in evaluating the magnitude of observed variation that should be attributed to technical factors[70,73]. In a typical pool/split design, source material (e.g., RNA) from several single cells is pooled together and then is split into independently processed libraries, each containing an amount of source material equivalent to that normally obtained from a single cell. Whereas the variation between single-cell libraries is a combination of biological and technical factors, the variation between the split

libraries derived from the same pool of source material is purely technical (assuming that the amount of pooled starting material is high enough to make uneven stochastic splitting of transcripts between the libraries unlikely). In some cases (such as with relatively big cells), splitting can be applied to a single-cell lysate. Notably, studies of allelic gene expression[75,221] split the lysate of *individual* cells into two independently processed libraries and compared the library pairs to estimate the abundance of biallelic expressed genes that are falsely called as monoallelic expressed due to dropouts.

Another vexing challenge in scRNA-seq is posed by 'doublets'—pairs of cells that are analyzed together due to failure of cell sorters[76], microfluidics[30], or droplets[30,31]. These can manifest as high-quality cells by many measures (more transcripts, more complex libraries) and lead to erroneous biological conclusions, such as false detection of 'hybrids'—cells that seem to be an intermediate type between two distinct cell populations but are in fact a mixture of RNA from two or more cells (A.R. and colleagues[30]). Some experimental platforms may provide mechanisms for visual inspection[2]; a general-purpose internal control is to mix cells from two species together[30,31] to estimate the proportion of characteristics of doublets. This, however, can be challenging to implement in some settings due to either cost or sample type.

False negatives and overamplification

Imperfections in the capture, conversion, and amplification[77,78] of the minute quantities of RNA in an individual cell into a cDNA library lead to challenges in regard to both false negatives (expressed but undetected transcripts) and false positives (transcripts with inflated expression levels) for individual transcripts, which may confound the understanding of biological variation. For example, false negatives can be erroneously interpreted as allele-specific expression[38,75] or as evidence for a wild-type homozygote rather than a genetic variant (or mutation)[77]. Inflated expression can lead to a false sense of bimodality in expression or misestimation of the amount of actual transcript per cell. While adjustments in the experimental protocol for obtaining cDNA from single cells can result in considerably improved yields[9,79] and mitigation of this technical variation, such variation remains a major challenge in single-cell genomics. A combination of experimental and computational methods has emerged to address this key challenge.

Massive whole-transcriptome amplification (WTA) inevitably amplifies any source of noise already present in the data and can introduce additional sources of noise, such as false-positive detections due to DNA polymerase errors occurring at an early PCR amplification cycle (Fig. 3c). An elegant way to address these is through the use of random molecular tags (RMTs), also called unique molecular identifiers (UMIs). RMTs are short random barcodes (4–20 base pairs) attached to the 3′ or 5′ end of cDNA molecules before amplification[3,80–83]. After amplification, the number of unique barcodes associated with reads aligned to a genomic position, rather than the number of aligned reads *per se*, serves as the count for the presence of each cDNA molecule in the original sample. Thus, RMTs and UMIs should provide an absolute molecular count for the number of converted cDNA molecules, in contrast to the standard units employed in the absence of UMIs (CPM, FPKM, and TPM). Such counting should not be done naively, however: base incorporation errors in

PCR can create spurious RMTs, and therefore barcodes with low copy numbers should be handled carefully[80]. In addition, if an RMT-tagged transcript molecule is overamplified, we expect it to generate, as a result of sequencing errors, RMT sequences that are close to each other (short edit distance) and should be collapsed[30] to avoid over-counting of transcripts. Finally, in “barcode collision” (i.e., when two copies of a transcript receive the same barcode, which makes them appear as a single copy), a highly expressed RNA species with copy number n can be associated with only $k < n$ unique barcodes; however, this effect can be corrected by explicitly modeling the probability for collision events[84]. Finally, while RMTs are a compelling method, current protocols are limited to ‘end-counting’ and cannot provide full-length transcript information. When the main goal is to estimate transcript levels, this is acceptable, but when information on transcript sequence (e.g., in tumor samples) or splice isoforms is desired, it may be less appropriate. Such cases would benefit from the development of new methods for molecular barcoding of full-length transcripts.

At the other end of the spectrum, the limited efficiency of RNA capture and conversion into cDNA leads to prevalent *dropout* events (false negatives): transcripts that are expressed in the cell but are entirely undetected in its mRNA profile[8] (Fig. 3c). Across a set of cells, this manifests as bimodality of gene expression (with one mode at zero). While in some cases such bimodality is due to stochastic (‘bursty’) gene expression[40] or more stable variation between cell types and states[3], in many others it is dominated by false negatives[70,85]. Indeed, the number of undetected transcripts is correlated to batch effects[60] and other metrics of library quality, and transcripts that are weakly expressed when detected are also more frequently undetected[1].

Several strategies have been proposed to address false negatives, including zero-inflated models, false-negative curves, and data imputation. In zero-inflated models (Fig. 3e) gene expression is modeled as a mixture of two distributions: one in which the transcript is successfully amplified and detected at a level that correlates with its true abundance, and another in which the transcript is undetected because of technical effects. These were first introduced[86] for modeling single-cell quantitative PCR (qPCR) as a mixture of log-normal and a point mass at zero with a mixing proportion that is the frequency of expression of the gene across all cells. This approach was used for scRNA-seq with a detection threshold (in log-space TPM)[2] or as a mixture model for read counts in which the first component is a negative binomial and the second is a low-magnitude Poisson, rather than a fixed zero, to account for background noise[85]. Notably, the mixing proportion for a gene in the latter model depends on its expected expression magnitude across all cells, conforming to the empirical observation that highly expressed genes are less prone to dropouts. Other studies[58,86,87] suggested a hurdle approach[88], in which the probability of detecting a gene and its measured expression level, conditioned on its detection, are modeled independently as generalized linear models (GLMs). The GLMs can readily consider complex experimental designs and technical covariates. In this approach, one can independently test the association of either detection probability or expression level with a variable of interest, or combine the evidence from both tests. Once a zero-inflated model is fitted, it can be used for differential expression analysis that accounts for the dropout errors[58,85–87], to inform other downstream analyses, such as gene set enrichment[58,89], and to alleviate the distortions

resulting from dropouts when dimensionality reduction is performed on scRNA-seq[90] or single-cell qPCR[91] data (where dropouts manifest as censored values[86]).

Another strategy, false-negative curves[1,2,92] (Fig. 3f), accounts for the fact that false negatives are determined both by the gene's actual expression level and by the individual cell's 'quality' in scRNA-seq (which may be affected by lysis, size, WTA success, and other unknown factors). This approach uses a set of transcripts that are expected to be expressed in all cells (e.g., housekeeping genes) and for each cell c independently plots the binary detection outcome (0/1) of those transcripts as a function of their expected expression μ (set for a transcript either as its average expression across all cells in which it was detected or as its expression in a matching bulk RNA library). The resulting curve is fit well by a logistic regression $F_c(\mu)$, whose odds quantify the cell's technical detection efficiency. Given a transcript whose expected expression, conditioned on it being detected, is μ^*, and that is undetected in cell c, the value $1 - F_c(\mu^*)$ is the probability that the transcript represents a technical dropout. This probability can be used to mitigate dropout effects by weighing down the contribution of zero values, for instance when computing covariance between genes, when computing principal components of the expression matrix, or in gene set enrichment analysis (GSEA) (N.Y., A.R. and colleagues[1]).

A third approach to address the false-negative problem is through data imputation. Here, coexpression patterns across cells are used to learn a model of the expression of a gene of interest, and subsequently the predicted estimates from the model are used instead of the measured levels, including instead of the zero expression of potential dropouts. Focusing on a small subset of 'landmark' genes of interest, one study (by A.R. and colleagues[93]) modeled the expression of each gene as a linear combination of the expression of other, highly variable genes. The model was trained across all cells in the data set using the least absolute shrinkage and selection operator (LASSO) method to guarantee sparseness. Zero expression values were then replaced by the predictions of the trained model and used for the downstream analysis. While this approach was originally applied for spatial mapping (below), it can be generalized in principle to other applications. Bayesian models can implicitly perform data imputation when they distinguish between the observed read counts of a gene and its latent 'true' expression, even when they are not explicitly zero-inflated. For example, the BASiCS model[62], which generalizes previous work[73], allows a gene with nonzero expression to have an observed read count of zero due to the combined effects of Poissonian sampling noise, capture inefficiency, and low cell-specific mRNA contents.

Revealing the vectors of cellular identity

One of the most exciting applications of single-cell genomics is as a means to identify and understand the factors that jointly define a cell's identity (Fig. 1 and Box 1). These factors may not only be the discrete categories that are often assumed when classifying cells into major types[4,30,94] but may also represent a continuous spectrum[1], or a combination of discrete and continuous categories[36,95–97,227]. First, in contexts such as development and physiology, some facets of cellular identity are transient in time and space. Temporal processes may, for example, progress along one or more trajectories (e.g., differentiation); oscillate continuously between cellular states (e.g., the cell cycle[98,99] or circadian

rhythm[100]); or be influenced by the physical position and neighborhood of the cell[71,93]. Moreover, even within a type, cells may span a continuous range of functional phenotypes (e.g., T cells of a single type, but with a range of inflammatory versus regulatory phenotypes, see N.Y., A.R. and colleagues[1]). While each such facet of a cell's identity is often considered separately, they are at least partly interdependent. Cataloging sources of biological variation, and understanding how they combine to determine a cell's identity, is an integral task in the compilation of a human cell atlas. We expect that, eventually, the measured genomic profile of a cell will be used not only to assign it to predefined categories, but also to quantify its identity with respect to sources of biological variation. By analogy, the sources of biological variation that determine a cell's identity are akin to basis vectors that span a linear space—namely, their combinations produce all possible points in the space. However, unlike basis vectors in algebra, they can in fact be dependent, which further complicates their identification and interpretation, and poses a problem of statistical identifiability. For instance, cells in a given position may be more quiescent compared to other positions, which complicates the inference of the biological variation that should be attributed to cell cycle versus spatial position[101].

In the following sections we review current computational methods to infer prominent facets of cellular identity, which include discrete cell types, continuous phenotypes, dynamic processes, and spatial location.

Distinguishing discrete cell types and subtypes

Whereas cell types were traditionally defined based on criteria such as morphology, physiology, and marker protein expression, single-cell analysis provides a means of systematically detecting cellular subtypes that cannot be defined by a handful of markers, or for which markers are not yet known[102]. Once a cellular subpopulation has been detected, statistical analysis can help identify defining markers, which can subsequently be validated by orthogonal experimental approaches, such as profiling of morphology and histology and in functional assays. Classification of cells into discrete types from single-cell profiles is a problem of unsupervised clustering in high dimensions. The key, inter-related challenges include (1) adapting methods to the exponentially increasing scale of single-cell data; (2) ensuring that the resulting classification is reproducible across experiments and platforms; (3) finding the proper granularity and detecting hierarchies of types and subtypes where they exist, especially when cell type frequency varies by multiple orders of magnitude from the most abundant to the rarest sub-type; (4) distilling molecular markers and signatures to characterize each cell type and/or cluster; (5) matching the resulting classes to legacy knowledge, and using semi-supervised methods where such knowledge exists; and (6) visualizing, sharing, and comparing classifications. Solutions to many of these challenges are only beginning to emerge.

Clustering in high-dimensional space is obstructed by the instability of distance metrics in high dimension[103] (which is a facet of the 'curse of dimensionality'). As a result, dimensionality reduction with linear or nonlinear approaches has been used extensively as an initial step. Among linear approaches, principal-component analysis (PCA) produces a deterministic and interpretable projection of the genomic profiles into the lower dimension,

and is highly scalable. It has been used repeatedly in single-cell analysis, including iteratively[104], having its axes inferred from bulk data and then applied to single-cells[218], and combined with an expectation-maximization (EM) algorithm to fit a finite mixture of Gaussians to PCA-reduced expression profiles[98]. A powerful, nonlinear alternative[5,105–107] is *t*-distributed stochastic neighbor embedding[108] (t-SNE; introduced for single-cell analysis as viSNE[105]), especially when computed with an efficient approximation[109] compatible with the scale of large genomic data sets[110]. Unlike PCA, t-SNE does not learn an explicit mapping between the high- and low- dimension spaces; points that are close in the high-dimensional space will be close (with high probability) in the low-dimensionality embedding, but more global relations are not directly interpretable, and held-out or additional data cannot be simply embedded in the same space in which previous data had already been embedded. A t-SNE variant that learns a mapping based on a deep neural network has been suggested[111]; while it is applicable only to massive data sets because of its large parameter space, as single-cell data grow it may increase in utility. A related alternative, ACCENSE[106], down-samples the data in a density-dependent manner (as in SPADE[94], described below), uses t-SNE to embed only the remaining cells, and finds clusters, by applying a kernel density transformation to the t-SNE map and seeking the local maxima that represent cluster centers. A *post hoc* procedure to ACCENSE can embed the remaining cells in the t-SNE space[112] to verify that they do not change the density map substantially.

Finally, PCA and t-SNE can be combined[108]. For example, a recent study (by A.R. and colleagues[30]) first performed PCA (using only highly variable genes), retained only principal components that explained significantly more variance than two null models (and were probably less susceptible than raw expression to experimental noise), and then projected cells into a two-dimensional t-SNE map on the basis of their principal component scores, followed by a density-based approach (DBSCAN[113]) to discern cell clusters. Another study (A.R. and colleagues[114]) proposed an iterative t-SNE variant. Once t-SNE computes a two-dimensional map of the cells, only features whose highest expression is obtained in cells that are close to one another in the map, as opposed to cells scattered across it, are retained, and then t-SNE is recomputed. A very recent approach, SIMLR[222], learns a cell-to-cell similarity matrix S that is based on a combination of multiple kernels (rather than a predefined distance metric) by assuming that there exist C discernible cell clusters and exploiting the nearly block-diagonal structure they induce on S. The learned S is then provided to t-SNE (which usually computes similarities in the high-dimension with a Gaussian kernel) for dimensionality reduction, followed by clustering and visualization. The retained features are the ones that determine the cell relationships captured by the first t-SNE map, and consequently the second map has more discernible clusters whose biological significance can be further explored. Importantly, the reduced dimensionality data are less noisy than the high-dimensional data but lose some of the biological variance. Therefore, clustering cells based on their coordinates in two- or three-dimensional PCA or t-SNE maps may be insufficient. It is sometimes useful to cluster the cells in a reduced space that still has several dozen features (depending on the complexity of the data) and retains more of the biological information and then present them in a two- or three-dimensional map for visual inspection (A.R. and colleagues[115]).

Other studies refrain from dimensionality reduction altogether. BackSPIN[4] (based on the SPIN) algorithm[116]) circumvents the lack of useful information in most genes with hierarchical biclustering: cells are partitioned into clusters by discerning groups of genes that are highly correlated in subsets of the cells. BackSPIN is divisive, such that genes that are assigned to a particular cluster are excluded from consideration when determining other clusters (nonexclusive biclustering methods have been described in other contexts[117]). After classifying the cells into types, a regression model infers posterior probability distributions of gene expression in each class, modeling linear contributions from known covariates (sex, age, and cell diameter), the cell type, and negative binomial noise. The PhenoGraph algorithm[34] takes a different approach, building on the Louvain clustering algorithm from network theory[118]: it finds the *k*-nearest neighbors of every cell in the high-dimension space by Euclidean distance, and constructs a graph in which nodes are cells and the edges between them are weighted by the Jaccard similarity of their *k*-neighborhoods. It then partitions the cells by detecting *communities*[119] in the graph—groups of nodes that are more densely connected than expected by chance in the same graph; the partition's modularity[120] is maximized by an efficient greedy heuristic[118].

Interestingly, scRNA-seq reads can be used to cluster cells based on features other than explicitly quantified gene expression. A recent study proposed a radical redesign of the standard RNA-seq workflow[121]: by replacing read alignment with pseudoalignment, i.e., by only identifying the transcripts from which a read could have originated without determining the exact sequence alignments between them, the time required to perform quantification (assigning expression values to genes) is decreased by two orders of magnitude without compromising accuracy. Beyond facilitating the quantification of massive single-cell data sets, this redesign paves the way to another idea—doing away with quantification altogether. Pseudoalignment divides (under reasonable assumptions[121]) reads into equivalence classes, each consisting of the set of transcripts the read could have originated from. By counting the number of reads a cell has in each of the classes (transcript-compatibility counts, TCC), one obtains a high-dimensional representation of the cells with features other than explicitly quantified gene expression[122]. While this feature space is not as biologically interpretable as gene expression space, it can produce[122] a similar cell clustering while sidestepping the time-consuming quantification task and avoiding the need to define a statistical model for read generation. Therefore, this approach reverses the order of quantification and clustering: one first clusters cells in the TCC space, and then quantifies gene expression only from representative cells in each cluster, or pooled data from the entire cluster, to assign a biological interpretation to the clusters.

A challenge in clustering is posed by the combination of two contradictory factors: (1) the massive number of cells profiled, which can lead to spurious small clusters, and (2) the orders of magnitude of difference in cell proportions between a common and rare cell type in the same tissue (e.g., in the bone marrow, hematopoietic stem cells) are present at <1:100,000; neutrophils are >50%). Absent prior knowledge, it can be very challenging to distinguish a new rare cell type from a spurious signal. Several strategies have been proposed to tackle this challenge. RaceID2[223] (an improvement of RaceID[5]) looks for saturation in the decrease of within-cluster dispersion as the number of clusters increases to determine the optimal number of clusters, conducts *k*-medoids clustering, and then

systematically searches each cluster for outlier cells—ones in which at least a user-specified number of genes are distinctly expressed with respect to the rest of the cluster. These outliers are candidate representatives of rare cell types. SPADE[94], most often applied to mass cytometry data[19,21,124], down-samples cells in the dense regions of the high-dimensional space (that correspond to abundant cell types), such that rare types become as prevalent as abundant types. This increases the probability that rare cell types will form their own clusters rather than becoming outliers to clusters of more abundant types, albeit at the cost of a decreased biological signal-to-noise ratio. SPADE subsequently clusters the remaining cells with an agglomerative hierarchical algorithm, structures the clusters as a minimal spanning tree, and visualizes the tree with the Fruchterman-Reingold algorithm[125]. It associates pruned cells with the cluster to which their nearest unpruned neighbor belongs.

Analyzing continuous phenotypic spectra within types

Categorization of cells into discrete types is a powerful abstraction and has been the focus of much of the single-cell genomic research thus far[4,30,34,36,114,115]. However, for some facets of a cell's identity it is more apt to speak of a continuous phenotypic spectrum within a type than of discrete cell types (Fig. 1b). Continuous facets can be characterized by combining dimensionality reduction with enrichment for functional annotations. For example, scRNA-seq profiles obtained from T helper 17 (T_H17) lymphocytes differentiated *in vitro* did not reveal distinct subtypes, but their first principal component was highly correlated with a cell's ability to trigger an autoimmune response (N.Y., A.R. and colleagues[1]). In leukemic bone marrow[105], t-SNE[108] embedding of mass cytometry profiles showed that while some markers were expressed in distinct clusters, others formed a continuous gradient. Finally, even when a cell population partitions into discrete types, hybrid cells—single cells that are a mix of two or more types—may be observed[95,97] (we refer here to true biological hybrids, in contrast to false hybrids resulting from doublets, as discussed above). In normal physiological settings, it has been proposed[96], following a series of studies[126–129], that hybrids may be 'generalist' cells that balance multiple cellular objectives: in this model genomic profiles are confined within a low-dimensional polytope, whose vertices correspond to key cellular tasks; cells lying near one of the vertices specialize in its corresponding task, whereas hybrids positioned toward the center of the polytope perform multiple tasks suboptimally. In pathological settings, especially cancer[97,130,131], such hybrid states may reflect either transitions (below) or the intrinsic abnormal mixture of different functional modules (N.Y., A.R. and colleagues[130]).

Such continuous states and vertices can be characterized based on prior knowledge. For example, given a gene set annotated with some function, PAGODA[89] scores each cell with respect to that function with its principal component 1 (PC1) score in a PCA limited to that gene set. If the variance explained by that principal component is significantly higher than expected, then the gene set represents an aspect of heterogeneity in the data. PAGODA subsequently combines gene sets that represent similar aspects of heterogeneity (i.e., similar PC1 loadings or PC1 cell scores) to form a more succinct representation. Another approach's[1,92] input is a *gene signature* consisting of a set of 'plus' and 'minus' genes that are highly and weakly expressed in a condition of interest. It scores each cell with the difference in the average expression of 'plus' versus 'minus' genes in that cell to reveal the

heterogeneity of cells with respect to the gene signature. As in other analyses, this procedure can be weighted to control for dropouts.

Prior knowledge on cell states can be used to interpret the biological relevance of the data's main axes of variation. One study of T cell differentiation compiled an extensive set of gene signatures associated with relevant cell states and kept only those that significantly varied across cells in its data (using a one-versus-all GSEA test). It then computed the correlation of the single-cell signature scores with the projection of cells to each principal component, allowing annotation of the principal components with respect to functionality. The scored signatures provided biological interpretation for the main axes of variation, such as reflecting a smooth transition from a naive-like (unexposed) T cell state to an effector (exposed) phenotype (N.Y., A.R. and colleagues[1]). One can similarly interpret a principal component by ranking genes according to their loadings and testing which known gene sets are enriched at the top or the bottom of the ranked list[132]. The high co-linearity in gene expression data implies that almost all genes will have nonzero loadings in every principal component. This motivates a rigorous approach for determining which genes are significantly associated with a given principal component[133], which has been applied to single-cell data (N.Y., A.R. and colleagues[2]). Similar methods can be applied to interpret the biological relevance of the low-dimensional map of the data produced by any projection algorithm, such as t-SNE. FastProject (N.Y. and colleague[92]) scales this process by visualizing and statistically testing the behavior of multiple gene signatures across multiple projection algorithms, discerning signatures that can illuminate particular projections.

Mapping dynamic processes

Cells undergo dynamic transitions, including short-term responses to environmental signals, cell differentiation, and ongoing oscillations. Each dynamic process is typically reflected in the cell's molecular profile, such that single-cell analysis of RNA or protein can position a cell in a temporal trajectory. Importantly, a cell participates simultaneously in multiple trajectories. For example, in a given moment in time a cell may be responding to an environmental signal while being in a certain differentiation state and going through a certain phase of the light–dark cycle. It is often assumed (implicitly or explicitly) that some of these processes are largely independent, but this may not be the case (e.g., cell cycle and differentiation). Studying a continuous dynamic process through bulk genomic assays requires to artificially break and measure the process at discrete time points, and also to synchronize the population of cells. While this can sometimes be achieved with an external stimulus, other cases require sophisticated experimental means[134]. In contrast, single-cell genomics provides a snapshot of the entire dynamic process. Since cells are unsynchronized, the set of single cells captured at any time point will stochastically contain cells positioned in different instantaneous time points along the temporal trajectory[59,135,136] (Fig. 1b). Computational analysis can then use these data to infer a near-continuous view of the temporal progression.

Pioneering computational methods recovered the temporal ordering by creating a graph that connects cells by their profiles' similarity and finding an optimal path on this graph starting from a user-specified source. This path introduces the notion of 'pseudo-time'—a scalar

measure of a cell's progress along the temporal trajectory (different from real time because all cells are sampled at the same time point). The Wanderlust algorithm[135] reconstructs a linear temporal trajectory in an unsupervised manner. It first builds a graph in which vertices are cells, and cells are connected to the *k* most similar cells (by cosine distance) with edges that are weighted by the dissimilarity. The temporal trajectory is based on shortest-path distances in this graph from a user-specified origin cell. Since noise accumulates with each step, making longer paths less reliable, Wanderlust randomly chooses a set of waypoint vertices uniformly to be used in distance computation (below). In addition, to prevent 'short circuits' due to the occasional proximity in the graph of two temporally distant cells, Wanderlust randomizes an ensemble of graphs by randomly choosing only *l* of each vertex's *k* neighbors, operating independently on each graph in the ensemble, and then averaging the temporal position of each cell across the ensemble; short circuits are expected to appear only in a small fraction of the graphs, and their effect is thus averaged out. For each graph independently, the trajectory score of cells is initialized to their respective distances to the origin cell. Then, Wanderlust updates the trajectory score for every cell based on the average of the waypoints' trajectory scores, weighted by their distance to the cell. This step updates the trajectory scores of the waypoints as well, and the algorithm iterates until convergence.

Another algorithm, Monocle[137] (extending previous work in microarrays[138]), reconstructs a tree describing the biological process and assigns each cell a pseudo-time. Monocle builds a complete graph in which each cell is a vertex and the edge between every pair of cells is weighted by their distance in a low-dimensional space (computed with independent component analysis, (ICA)). It then constructs a minimal spanning tree (MST) of the graph and assumes that the longest path through the MST corresponds to the main temporal trajectory. More generally, the *k* longest backbones (with *k* specified by the user) correspond to the biological branches, whereas the other branches of the tree are considered technical noise. Monocle's implementation relies on a PQ tree[139]—a concise representation of the temporal ordering of cells that allows for uncertainty. StemID[223] infers likely edges in a differentiation lineage tree by clustering cells in low-dimension and drawing an edge between the medoids of every two clusters. For every cell *c* whose cluster's medoid is *m*, the vector connecting *m* to *c* is projected onto all edges going out from *m*. The cell is then assigned to the edge on which it had the longest projection relative to the edge's length, and the projection's length is used as a measure of the cell's progression on the edge. Only edges that are significantly more uniformly covered than expected by an empirical null model are retained in the final cell fate tree. Several studies[140–142] used diffusion maps[143] to study differentiation trajectories. By representing cells as isotropic Gaussians around their measured expression, the interference of these Gaussians created high probability density paths between cells, which are interpreted as transition probabilities between cellular states along a trajectory. Replacing the Gaussian kernel with a more general form[140] allows dropouts to be addressed in the resulting statistical model. All these methods assume that the temporal transition between cellular states is smooth and that all intermediate cellular states are represented in the available cells.

The SCUBA algorithm[144] takes a slightly different departure point (taken also by Wave-Crest[218]) and assumes that single-cell profiles are available along a time course, such that cells are a priori sampled from distinct time points, but still allows for an asynchronous

process. SCUBA aims to detect in an unsupervised manner bifurcation events in which differentiating cells split between alternative fates. First, SCUBA uses *k*-means to cluster the cells of the first time point, while determining the number of clusters with the gap statistic[123], and assigns every cell in the second time point to the parental cluster most similar to it. To determine whether a split has occurred, SCUBA partitions the progeny of every parental cluster in two and uses the gap statistic to decide whether the split explains the data better than the original single cluster. The process is iterated until SCUBA reconstructs a binary tree of cellular development across all time points, which it refines by likelihood maximization. To study the gene expression dynamics associated with a bifurcation event, SCUBA reduces the high-dimensional gene expression into the bifurcation direction, defined as the one-dimensional line connecting the centers of two clusters differentiated from a common parental cluster, and fits a potential function that describes the gene expression dynamics along that line. Importantly, this allows SCUBA to predict genetic perturbations that bias the proportions of the bifurcation toward one of the branches.

Another approach to capture bifurcations, Wishbone[145], extends Wanderlust and allows the trajectory to bifurcate once into two alternative fates. Instead of sampling an ensemble, Wishbone mitigates the effect of short circuits by projecting the data to low dimension with a diffusion map[143] and considering only the top diffusion components to define the Euclidean distance between nodes. A graph is then constructed based on these distances, and waypoints are selected as in Wanderlust, only with a medoid-based refinement to eliminate outlier cells from becoming waypoints. Similarly to Wanderlust, Wishbone computes a temporal trajectory starting in an a priori known progenitor cell and refines it iteratively using the waypoints, but prevents waypoints in one branch from influencing the other. It computes the *perspective* of each waypoint, which is an approximate distance of every cell from the progenitor, when the waypoint is taken as anchor. When two waypoints each lie in a different branch, their perspectives will diverge. Wishbone capitalizes on this observation by constructing a dissimilarity matrix between the waypoints and using a spectral clustering approach to identify the branches and their bifurcation point.

The cell cycle is the primary example of an oscillatory process that is readily detected from single-cell profiles. In pioneering work based on a few markers, ergodic rate analysis (ERA) was used to accurately infer trajectories along the cell cycle from single-cell measurements of fixed steady-state populations[136]. Several studies inferred the cell cycle phase of individual cells from single-cell RNA-seq data[30,35,97,99], relying on transcriptional signatures of discrete phases from earlier bulk profiling experiments[146]. This is a prime example of the power of single-cell profiling: while bulk profiles of cells synchronized for the cell cycle have proven challenging to perform and compare across cells and species[134], single-cell profiles provide a near-continuous sampling and show high conservation across cell types and species[30,97,99]. More recently, the Oscope[147] method was developed to detect genes that oscillate with time, as in the cell cycle, and to order cells by their phase in the cycle. Oscope allows for multiple orthogonal oscillatory processes to take place concurrently, each involving different genes, and the relative order of the cells may be different with respect to each of the cycles. Oscope assumes that an unsynchronized set of cells measured in a scRNA-seq experiment contains a dense sample of time points along any cycle. It fits gene pairs with two phase-shifted sinusoidal functions across all cells, retains

the best fits as candidate oscillatory genes, and clusters the cells, using the fit errors of gene pairs as their dissimilarity and picking the optimal number of clusters by maximizing silhouette width. Each gene cluster defines an independent oscillatory process, and cells are ordered with respect to their phase in it cycle with the nearest-insertion heuristic of the traveling salesperson problem[148].

The cell cycle also epitomizes the challenge of dissecting multiple co-occurring processes in the same single cell, especially when these processes are not fully independent of each other. The cell cycle often has an enormous impact on cellular gene expression[146,149,150] (although some[87] have challenged this view), but this impact can extend to other processes, and in turn be affected by them. Thus, the large variation induced by the cell cycle in some scRNA-seq experiments may conceal other important sources of biological variation[101], especially cellular differentiation processes[137,141,151,152]. On the other hand, removing its impact altogether (e.g., regressing it out) may eliminate important facets of biological variation, for example, in differentiation processes[98]. This challenge is likely to affect other processes (in addition to cell cycle) as we investigate them further.

Nevertheless, several studies have attempted to handle cell cycle covariates in scRNA-seq analysis. In particular, scLVM[101], a statistical framework developed to account for unwanted and a priori known covariates from scRNA-seq data, has been effectively used to handle the cell cycle as a known covariate. scLVM is based on Gaussian process latent variable models (GPLVMs)[153] and is inspired by the two-step removal of unwanted variation (RUV-2) discussed above[49]. However, instead of using a set of negative control genes, as in RUV-2, scLVM uses a set of genes annotated a priori as associated with a confounding factor, such as the cell cycle (simple removal of the annotated genes is insufficient since many unannotated ones are affected as well). A GPLVM approach maximizes the likelihood of a latent cell cycle variable (a one-dimensional variable per cell), derives its cell-to-cell covariance structure, and either takes it into account in downstream analyses (e.g., gene–gene correlations) or eliminates the latent cell cycle factor from the data (before visualization or clustering). Applied to T helper 2 (T_H2) cell differentiation, scLVM helped highlight a set of correlated genes enriched in T_H2 cell differentiation and distinguish two subpopulations—neither of which is detectable when the cell cycle covariates are present[101].

Inference of spatial location

Cells operate within complex tissues where their spatial context – from their physical position to the identity of neighboring cells – is critical to their function (N.Y. and A.R.[225]). Unfortunately, most high-throughput single-cell approaches require the dissociation of tissues into single-cell suspensions and therefore lose this spatial information. While pioneering approaches for *in situ* sequencing[154,155], transcriptome *in vivo* analysis[156], and multiplex FISH[157–159] of single-cell RNA are emerging (see also a recent review[160]), their throughput and accessibility to the research community have not yet matched those of more established protocols. Conversely, spatial information can be readily obtained for a limited number of ‘landmark’ genes by traditional experimental methods such as *in situ* hybridization and histochemistry; indeed, for many model systems there are extensive catalogs of such information.

This has inspired the development of two methods, Seurat (A.R. and colleagues)[93] and a method by Achim *et al.*[71] that combine limited spatial cues for landmark genes with single-cell RNA-seq of cells from the same type of dissociated tissue to infer the spatial location of the dissociated cells through the patterns reflected by the landmark genes. For example, Seurat[93] takes as input scRNA-seq and *in situ* hybridization for a small set of spatial landmark genes, converted into a spatial reference map of binary assignments ('on/off') in discrete spatial bins specified by the user. On the basis of the landmark genes, Seurat computes a posterior probability for each cell to have originated from each bin. Because of the limited number of landmarks expressed at any spatial position, dropouts in scRNA-seq could severely compromise the mapping. Seurat addresses this by first learning a (L1-constrained) linear model for each landmark transcript from the scRNA-seq profiles themselves and then imputing the landmark gene value in each cell before mapping. It then transforms the imputed continuous values for the landmark genes into binary assignments by fitting the distribution of each landmark gene with a mixture of two Gaussians. Finally, it constructs a multivariate normal model for the expression of the landmark genes in each bin, which it uses to derive the aforementioned posterior probabilities. A similar method was developed independently and in parallel[71]. Starting with a similar, but higher-resolution, spatial reference map, and a set of sequenced cells, the method computes a specificity score for the expression of each landmark gene m in every cell c, and then a correspondence score of each cell–voxel pair: if a landmark gene m is detected in cell c, then the score increases or decreases depending on m's association with this voxel in the reference map, weighted by the specificity of the pair (c, m), namely the extent in which m is particularly expressed in c relative to the other cells. Given the expected dropouts, no penalty is given when m is undetected in scRNA-seq. The statistical significance of the correspondence scores is assessed through a permutation test.

Such mapping approaches critically depend on the nature of the reference map. An algorithm typically cannot map cells at a resolution better than that of its reference map (with some exceptions related to gradients), and it cannot map cells if they do not evince a discernible pattern that the map associates with a particular spatial locus. The number of required landmarks is related to this spatial complexity, introducing the notion of power to detect spatial patterns[93]. Successful application of reference-based mapping approaches has so far been limited to tissues with canonical structure, such as a developing embryo, which is faithfully reproduced in replicate samples. This is required in order to obtain both landmark gene expression for a reference map (obtained from one or more samples) and scRNA-seq profiles (obtained from a separate physical sample). While this is a likely situation in most normal tissues (e.g., brain structures[71]) and physiological processes (e.g., embryogenesis[93]), it is not the case in pathological tissues, such as tumors, where each tissue is idiosyncratic. A different hurdle to successful application of such approaches occurs in normal tissues, such as the retina[161,162], in which functionally similar cells are in fact mosaically organized. Higher-order computational approaches will be required to address these challenges.

When a detailed reference map is not available, some spatial information can be gleaned from gene markers of spatial axes[163]. For example, studies of the otocyst[164,165], a spherical organ, placed gene expression profiles from dissociated cells on a three-dimensional sphere by taking each cell's coordinates in the first three principal components and projecting it to

the unit sphere (this approach creates a hollow sphere and does not accommodate multilayered spheres). To relate PCA space to real-world geometry, the dorsal–ventral axis was defined by the direction of the vector connecting the origin to the centroid of cells expressing a dorsal gene marker, and further heuristics determined anterior–posterior and medial–lateral axes.

Analyzing cellular circuitry from cell–cell variation

Cellular identity and function are governed by elaborate regulatory circuits, many of which act through control of RNA expression[166,167]. Deciphering circuitry typically relies on combining observations across diverse samples with manipulation of one or more regulators to observe their effect. Single-cell genomics addresses this challenge in two important ways. First, it provides a sheer scale of individual samples (up to tens of thousands single cells in one experiment, whereas bulk genomics typically measures several dozens of samples) increasing statistical power[97]. Second, because of both stochastic and regulated differences between the cells, each cell forms its own ‘perturbation system’, in which multiple regulators are subtly, or more strongly, perturbed, possibly reducing the need for additional experimental perturbations[3,41].

Leveraging this insight, several studies have analyzed the covariation structure between transcripts or proteins across single cells to infer regulatory relations[1–3,41,135,137,168–172]. In analogy to simple observational approaches for inferring gene regulation, correlating the expression profiles of genes or proteins opens the way to elucidating the regulatory mechanisms that control them[2,3,41,169,170]. For example, an analysis (N.Y., A.R. and colleagues[3]) of the covariation in the expression of ~600 transcripts across 18 dendritic cells detected a module of antiviral gene expression, which included two key transcription factors, STAT2 and IRF7. These factors were predicted to control the module because variation in their protein expression levels was expected to propagate downstream and create variation in their targets’ mRNA expression, as well in as their own mRNA expression through autoregulation. Notably, analysis of covariation across cells effectively circumvents many challenges associated with dropouts because the probability that a gene will drop out, once conditioned on its expression, is independent from biological gene-to-gene covariation. Similar approaches have been used in other biological settings[1], and enhanced resolution, for example from pseudotemporal ordering, has helped recover such dependencies[135,137]. A related strategy to identify co-regulated genes is by comparing their expression between two conditions, each of which is represented by multiple individual cells from multiple samples.

As in circuit analysis in bulk genomic profiles, regulatory predictions from observational expression profiles can be combined in turn with other mechanistic (e.g., chromatin immunoprecipitation followed by sequencing (ChIP-seq) on bulk profiles) or genetic (e.g., perturbations followed by RNA-seq) profiles, to increase power to detect regulatory relations or to help resolve the molecular basis of interaction or causality. A straightforward approach is to first identify subsets of genes of interest (e.g., based on principal components that correlate with important biological properties[1] or on an inferred ‘pseudo-time’ ordering[137]) and then use other data sets, such as those for *cis*-regulatory motifs, ChIP-seq, or

perturbation followed by RNA-seq, to identify transcription factors whose targets are enriched in those sets.

More recent studies have begun to determine regulatory circuit topology beyond pairwise interaction. One approach[141] uses Boolean models (also used in the past for bulk profile analysis[173]) in which a cell's state is represented by a binary vector indicating for each gene whether it is 'on' (1) or 'off' (0), and states form nodes in a network in which they are connected by an edge if their respective binary vectors differ by exactly one bit. Paths through this graph are perceived as gradual transitions through a sequence of cellular states. Finally, the network's paths (or binary state transitions) are used to identify an 'executable' Boolean model that expresses each gene as a Boolean function (using Boolean operators such as AND, OR, NOT) of other genes, using tools from computer program synthesis[174,175].

Notably, when the number of single cells analyzed becomes very large, the extent of noise in the measurement of any given transcript can introduce substantial challenges for simple calculations of pairwise covariation, from scatter plots to linear correlation. Two complementary methods, DREVI and DREMI[32], tackled these challenges through visualization and quantification, respectively, in the context of single-cell mass cytometry (CyTOF) data for dozens of phosphoproteins across hundreds of thousands of cells. Considering a given pair of proteins (say X and Y), it has been observed that a simple analysis of their joint density $f(X, Y)$ is dominated by the most common cell types, but misses parts of the dynamic range that transpire only in rare subsets of cells but are key to uncovering the biologically meaningful signal of the relationship between proteins. Therefore, the two methods use a conditional density $f(Y|X)$, which is equivalent to up-weighting ranges of X values that are represented only by rare subtypes of cells, allowing them to have the same effect as the abundant subtypes. To achieve this, DREVI computes a two-dimensional kernel density estimation for $f(X, Y)$ via a heat diffusion formulation[176], derives $f(Y|X)$ from it, and visualizes $f(Y|X)$ as a heat map. DREMI then provides a quantitative score for the strength of this (directional) dependence of Y on X by estimating their mutual information using the conditional, rather than the joint, distribution. Finally, the conditional density distribution is fitted with a curve to obtain a parametric description of how Y changes as a function of X, revealing important aspects of their dependence, such as activation thresholds or saturation.

There are several crucial impediments to the inference of regulatory networks from single-cell transcript abundances. First, transcript abundance is affected not only by allele-extrinsic factors, but also by *allele-intrinsic* stochasticity, resulting in its divergence across cells exposed to the same extrinsic triggers[42,43] (see above and Fig. 2). While the statistically strongest correlations are often indicative of underlying regulatory circuits (N.Y., A.R. and colleagues[1]), we expect that novel computational approaches, possibly relying on experimental cues[41], will emerge to accommodate this challenge. Second, protocols that lyse entire cells measure total mRNA, and consequently they mix the fluctuating nuclear RNA and the more stable cytoplasmic RNA[177,178] and combine transcriptional and post-transcriptional effects in a single measurement. This can be addressed in the future by integrative models that rely on additional experiments, e.g., a combination of RNA-FISH

(fluorescence *in situ* hybridization) and single-cell ATAC-seq[27,28] (assay for transposase-accessible chromatin with high-throughput sequencing) to measure transcription *per se*; or new methods for measuring single-cell nascent RNA levels) and/or by bulk data on RNA transcription rates or stability (A.R. and colleagues[179]). Third, the number of mRNA molecules may sometimes not correlate well with the numbers of their protein products at the same time point[180–184], and consequently regulatory networks ought to be considered at both the transcript and the protein levels[171]. Advances in single-cell proteomics will no doubt contribute to studies of regulatory mechanisms. On the other hand, scRNA-seq does make it possible to study the mechanisms underlying allele-intrinsic variation by examining differences in other moments of the distribution (e.g., variance), which can shed light on the kinetics of the transcription process[168,171,172].

Challenges in single-cell epigenome analysis

Although most high-throughput single-cell profiling studies to date have focused on RNA and protein, recent progress in epigenome analysis[185] has included assays for single-cell ChIP-seq (Drop-ChIP[26]), single-cell ATAC-seq[27,28], single-cell Hi-C[29] (a genome-wide derivative of the chromosome conformation capture (3C) technique), and single-cell DNA methylation[22–25]. These methods hold promise for elucidating the epigenetic controls of cellular identity[186,187]. In particular, once haplotypes are resolved[188–190], single-cell epigenomic profiles should allow correlation of events such as the methylation patterns of a gene's promoters and enhancers, or of enhancers targeted by the same transcription factor[191]. This would not be possible using bulk epigenomic data, which provide averages of events recorded in multiple cells. Similarly, single-cell Hi-C can derive a coherent three-dimensional chromosomal topology[29] unattainable from bulk 3C data, which averages the chromosomal conformations in millions of cells.

However, single-cell epigenomic data are very sparse: a single cell contains only one copy of each allele molecule, and owing to the size of the genome and the limited efficiency of current single-cell epigenetics protocols, only a small subset of loci is measured in any given cell. This leads to unique computational challenges in detecting signal from very sparse data. For example, because the probability of detecting a signal (finding a read) at a particular locus is very low, clustering of single-cell ChIP-seq profiles using locus-based correlations between them is highly sensitive to algorithm parameters and global technical attributes, such as mean single-cell coverage[26]. Low capture ratios coupled to massive amplification would also lead to epihaplotypes, similarly to falsely allelic scRNA-seq, where a gene could be erroneously classified as monoallelically expressed because of dropouts[38,75], and to allelic dropouts[77] in single-cell DNA-seq due to which heterozygous variants could be falsely classified as homozygous.

To overcome this intrinsic sparseness, several studies[26,28] have reasoned that functionally related genomic elements, such as all the sites bound by the same transcription factor, would co-vary within the same cell (analogous to the observation above on transcript co-regulation to infer circuits), providing a higher-order aggregate signal that allows one to group cells by their shared inferred state, as well as to determine the extent of variation in the activity of the associated factors between cells[26,28]. As single-cell epigenome studies become more

common, more attention will be devoted to developing methods to account for their technical noise[191], similar to the body of work that had already emerged for scRNA-seq.

Outlook

Single-cell genomics has opened a new frontier for understanding biological systems at multiple levels, including cell types and states, underlying molecular circuits, and tissue organization. Innovative, efficient, robust, and scalable computational analysis methods are essential to delivering on this promise. Pioneering work over the past few years has provided an initial toolbox of algorithms. Some of these methods adapt well-established tools previously used in bulk genomics to the particular challenges of noise and scale in single-cell data[1,44,89,90,105], whereas others are designed specifically for single-cell data[71,93,135,137]. Despite this progress, the field is still in its infancy. Experimental and computational biologists are still learning the characteristics of single-cell data and the proper ways to accommodate them, while at the same time these very characteristics change rapidly as experimental technologies advance. An important task for the future is to define a set of best practices through comparisons of different statistical methods and experimental platforms. We anticipate substantial development of computational methods in the next few years to tackle challenges such as the growing scale of the data, the need for effective visualizations, the emergence of new single-cell assays, and integration of either diverse data types[219] or multiple levels of biological organization, from single cells to cell–cell circuits to tissues.

The challenge posed by increased scale for both processing and visualization should not be underestimated. Using the latest protocols for massively parallel scRNA-seq[30,31] and mass cytometry[32–34], a single laboratory can now readily collect tens of thousands to hundreds of thousands of single-cell RNA-seq profiles and tens of millions of single-cell protein profiles. Data at this scale present difficulties in basic processing[192] (e.g., calculating a covariation matrix; clustering) and in assessing the statistical significance and robustness of results. It is also a challenge to effectively visualize such magnitudes of data, and improved visualization methods will be crucial to fully exploit the potential of single-cell data to lead to biological discovery. For example, it is standard to visualize the first principal components of the data, yet in single-cell data, unlike many other data types, those capture only a minute fraction of the variation in the data (note that t-SNE[108] too often considers only top principal components in high-dimensional data). On the other hand, the increase in scale presents computational opportunities. It allows benefitting from ‘big data’ approaches, most notably deep learning[193], developed in fields such as image and text analysis, in which data sets tend to have many more samples than were available in genomics thus far.

Many of the methods described here are applicable across different single-cell data types[145]. However, certain data types require distinct computational strategies. For instance, while data sparsity currently characterizes all single-cell assays, it is particularly challenging in epigenomic profiles, as discussed above. Another prominent domain that requires specific computational methods is genetic variation between cells, whether in normal or pathological contexts. While our focus has been on single-cell transcriptomics, important strides have also been made in single-cell DNA sequencing[77,194] as well as in analysis of copy-number

variations (CNVs)[11] and point mutations[15,195] from single-cell DNA data. These methods have been used to decipher pathological evolution in tumors[11,15,195–197] and somatic mosaicism[198] in healthy tissues and in non-cancerous disease[199–203]. Some of the challenges highlighted here, including experimental noise[16,204], can strongly affect estimates of heterogeneity due to false positives, false negatives (including allelic dropouts[77]), and data sparsity; these will require new methods and adaptation of existing ones (e.g., in phylogenetic reconstruction[205]). Notably, the transcript sequences in scRNA-seq data (the 'expressed exome') can also be used to derive valuable genetic information. This includes identification of large-scale CNVs, inferred by considering skews in gene expression levels along large chromosomal windows[97]; identification of fusion transcripts[206,207]; inference of T- or B-cell receptors (TcRs or BcRs, respectively) sequences from the respective transcripts[208,217]; and even mutation calling[97,226], albeit with similar challenges related to dropouts. Genetic inference from scRNA-seq also provides a means of connecting genetic and functional states in the same cell.

These possibilities highlight the exciting potential of integrating different molecular profiles at the single-cell level, either by direct experimental measurement or by computational inference. Emerging experimental frameworks now allow simultaneous measurement of multiple omic data types from the same single cell. Early successes include parallel measurement of the transcriptome and either the genome (DR-seq[209] and G&T-seq[210]), the methylome (scM&T-seq[211], an adaptation of G&T-seq), or the proteome[212–214]. Multiple-omic protocols will not only allow results to be corroborated by distinct data types (e.g., CNV calling from DNA-seq[11] and scRNA-seq data[97]) but also elucidate how information flows in biological circuits from DNA to RNA to protein. For example, the framework of expression quantitative trait loci (eQTLs[215]) can be re-cast in single-cell analysis[168], where every cell is now an individual, and dependencies between single-cell genome variants and transcriptome variation allow inferring causality from the former to the latter. This can boost the power of eQTL detection[168] and analysis of factors that affect expression variance and temporal dynamics. It can also provide insight into tumors, as analysis of multiple single cells from one tumor sample can be used to identify the genetic basis of transcriptional state variation between malignant cells[197,226]. Similarly, simultaneous analysis of protein and RNA or of epigenomic profiles and RNA can provide mechanistic explanations that relate variations in the state of regulator proteins[3] or *cis*-regulatory loci[27,28] to variation in target transcripts.

Even without new experimental methods, computational techniques can help relate distinct data types across single cells, as was effectively shown for relating spatially resolved *in situ* hybridization data to scRNA-seq to derive a spatial mapping of cells to positions[71,93]. This suggests a more general framework whereby some types of data (e.g., *in situ* hybridization, protein expression), often available for only a subset of the variables, can be used to link individual cells with new metadata (e.g., cell type, spatial position). Extending these approaches could help connect, for example, CyTOF profiles and RNA-seq profiles to define cell states.

Finally, single-cell genomics will help elucidate how cells are organized into multicellular systems, addressing questions such as the spatial organization of tissues and the direct and

short-range molecular interactions between cells. Two key experimental and computational advances are needed. First, we need to determine the molecular profiles of single cells without dissociating them from their tissue context. Single-cell profiles will allow us to determine which cells likely interact with each other and through which molecules, and additional spatial resolution at the sub-cellular level can help determine direct interactions. While experimental methods are emerging for this purpose[155,159], they will require sophisticated accompanying computational analysis methods. Second, we need to generalize methods for studying the intracellular functional circuitry to ones that can infer the functional circuitry of *interacting* cells. Just as for intracellular molecular circuits, a physical description is illuminating but insufficient to understand the operation of a circuit and how it creates a biological phenotype. By combining single-cell tissue genomics with computational models for molecular circuits that relate cells to each other in space and time, from their molecular connections and up to their functional impacts, we can build an integrated understanding of the way cells fulfill their function in health and disease.

Acknowledgments

We thank Eric Lander, Alex K. Shalek, Russell B. Fletcher, Oren Ram, and David Stafford for helpful discussions, and Leslie Gaffney and Anna Hupalowska for artwork. A.W. and N.Y. were supported in part by the BRAIN Initiative grant U01 MH105979 from the US National Institute of Mental Health. A.R. is an Investigator of the Howard Hughes Medical Institute and was supported by the Klarman Cell Observatory at the Broad Institute, NIH grant P50 HG006193, Koch Institute Support (core) grant P30-CA14051 from the National Cancer Institute, NIH BRAIN grant 1U01MH105960-01, NCI grant 1U24CA180922, and NIAID grant 1U24AI118672-01.

References

1. Gaublomme JT, et al. Single-cell genomics unveils critical regulators of Th17 Cell pathogenicity. Cell. 2015; 163:1400–1412. [PubMed: 26607794]
2. Shalek AK, et al. Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. Nature. 2014; 510:363–369. [PubMed: 24919153]
3. Shalek AK, et al. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. Nature. 2013; 498:236–240. [PubMed: 23685454]
4. Zeisel A, et al. Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science. 2015; 347:1138–1142. [PubMed: 25700174]
5. Grün D, et al. Single-cell messenger RNA sequencing reveals rare intestinal cell types. Nature. 2015; 525:251–255. [PubMed: 26287467]
6. Altschuler SJ, Wu LF. Cellular heterogeneity: do differences make a difference? Cell. 2010; 141:559–563. [PubMed: 20478246]
7. Tang F, et al. mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods. 2009; 6:377–382. [PubMed: 19349980]
8. Ramsköld D, et al. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. Nat Biotechnol. 2012; 30:777–782. [PubMed: 22820318]
9. Picelli S, et al. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. Nat Methods. 2013; 10:1096–1098. [PubMed: 24056875]
10. Hashimshony T, Wagner F, Sher N, Yanai I. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. Cell Reports. 2012; 2:666–673. [PubMed: 22939981]
11. Navin N, et al. Tumour evolution inferred by single-cell sequencing. Nature. 2011; 472:90–94. [PubMed: 21399628]
12. Zong C, Lu S, Chapman AR, Xie XS. Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. Science. 2012; 338:1622–1626. [PubMed: 23258894]

13. Xu X, et al. Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. Cell. 2012; 148:886–895. [PubMed: 22385958]
14. Hou Y, et al. Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm. Cell. 2012; 148:873–885. [PubMed: 22385957]
15. Wang Y, et al. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. Nature. 2014; 512:155–160. [PubMed: 25079324]
16. Leung ML, Wang Y, Waters J, Navin NE. SNES: single nucleus exome sequencing. Genome Biol. 2015; 16:55. [PubMed: 25853327]
17. Lohr JG, et al. Whole-exome sequencing of circulating tumor cells provides a window into metastatic prostate cancer. Nat Biotechnol. 2014; 32:479–484. [PubMed: 24752078]
18. Bandura DR, et al. Mass cytometry: technique for real time single cell multitarget immunoassay based on inductively coupled plasma time-of-flight mass spectrometry. Anal Chem. 2009; 81:6813–6822. [PubMed: 19601617]
19. Bendall SC, et al. Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. Science. 2011; 332:687–696. [PubMed: 21551058]
20. Chattopadhyay PK, et al. Quantum dot semiconductor nanocrystals for immunophenotyping by polychromatic flow cytometry. Nat Med. 2006; 12:972–977. [PubMed: 16862156]
21. Bodenmiller B, et al. Multiplexed mass cytometry profiling of cellular states perturbed by small-molecule regulators. Nat Biotechnol. 2012; 30:858–867. [PubMed: 22902532]
22. Guo H, et al. Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. Genome Res. 2013; 23:2126–2135. [PubMed: 24179143]
23. Smallwood SA, et al. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. Nat Methods. 2014; 11:817–820. [PubMed: 25042786]
24. Farlik M, et al. Single-cell DNA methylome sequencing and bioinformatic inference of epigenomic cell-state dynamics. Cell Reports. 2015; 10:1386–1397. [PubMed: 25732828]
25. Guo H, et al. The DNA methylation landscape of human early embryos. Nature. 2014; 511:606–610. [PubMed: 25079557]
26. Rotem A, et al. Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. Nat Biotechnol. 2015; 33:1165–1172. [PubMed: 26458175]
27. Cusanovich DA, et al. Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. Science. 2015; 348:910–914. [PubMed: 25953818]
28. Buenrostro JD, et al. Single-cell chromatin accessibility reveals principles of regulatory variation. Nature. 2015; 523:486–490. [PubMed: 26083756]
29. Nagano T, et al. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature. 2013; 502:59–64. [PubMed: 24067610]
30. Macosko EZ, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell. 2015; 161:1202–1214. [PubMed: 26000488]
31. Klein AM, et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell. 2015; 161:1187–1201. [PubMed: 26000487]
32. Krishnaswamy S, et al. Systems biology. Conditional density-based analysis of T cell signaling in single-cell data. Science. 2014; 346:1250689. [PubMed: 25342659]
33. Sen N, et al. Single-cell mass cytometry analysis of human tonsil T cell remodeling by varicella zoster virus. Cell Reports. 2014; 8:633–645. [PubMed: 25043183]
34. Levine JH, et al. Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. Cell. 2015; 162:184–197. [PubMed: 26095251]
35. Tirosh I, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science. 2016; 352:189–196. [PubMed: 27124452]
36. Tasic B, et al. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. Nat Neurosci. 2016; 19:335–346. [PubMed: 26727548]
37. Gawad C, Koh W, Quake SR. Single-cell genome sequencing: current state of the science. Nat Rev Genet. 2016; 17:175–188. [PubMed: 26806412]

38. Kim JK, Kolodziejczyk AA, Ilicic T, Teichmann SA, Marioni JC. Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression. Nat Commun. 2015; 6:8687. [PubMed: 26489834]
39. Kolodziejczyk AA, Kim JK, Svensson V, Marioni JC, Teichmann SA. The technology and biology of single-cell RNA sequencing. Mol Cell. 2015; 58:610–620. [PubMed: 26000846]
40. Raj A, Peskin CS, Tranchina D, Vargas DY, Tyagi S. Stochastic mRNA synthesis in mammalian cells. PLoS Biol. 2006; 4:e309. [PubMed: 17048983]
41. Stewart-Ornstein J, Weissman JS, El-Samad H. Cellular noise regulons underlie fluctuations in *Saccharomyces cerevisiae*. Mol Cell. 2012; 45:483–493. [PubMed: 22365828]
42. Raj A, van Oudenaarden A. Nature, nurture, or chance: stochastic gene expression and its consequences. Cell. 2008; 135:216–226. [PubMed: 18957198]
43. Swain PS, Elowitz MB, Siggia ED. Intrinsic and extrinsic contributions to stochasticity in gene expression. Proc Natl Acad Sci USA. 2002; 99:12795–12800. [PubMed: 12237400]
44. Ilicic T, et al. Classification of low quality cells from single-cell RNA-seq data. Genome Biol. 2016; 17:29. [PubMed: 26887813]
45. Leek JT, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet. 2010; 11:733–739. [PubMed: 20838408]
46. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007; 8:118–127. [PubMed: 16632515]
47. Benito M, et al. Adjustment of systematic microarray data biases. Bioinformatics. 2004; 20:105–114. [PubMed: 14693816]
48. Cole, M., et al. Single Cell Genomics. Cambridge UK: Sep. 2016 SCONE: correcting and evaluating the influence of unwanted variation on single-cell RNA-seq data [poster]. https://github.com/YosefLab/scone
49. Gagnon-Bartsch JA, Speed TP. Using control genes to correct for unwanted variation in microarray data. Biostatistics. 2012; 13:539–552. [PubMed: 22101192]
50. Leek JT. svaseq: removing batch effects and other unwanted noise from sequencing data. Nucleic Acids Res. 2014; 42:e161–e161.
51. Bullard JH, Purdom E, Hansen KD, Dudoit S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics. 2010; 11:94. [PubMed: 20167110]
52. Lovén J, et al. Revisiting global gene expression analysis. Cell. 2012; 151:476–482. [PubMed: 23101621]
53. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010; 11:R106. [PubMed: 20979621]
54. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008; 5:621–628. [PubMed: 18516045]
55. Trapnell C, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010; 28:511–515. [PubMed: 20436464]
56. Wagner GP, Kin K, Lynch VJ. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory Biosci. 2012; 131:281–285. [PubMed: 22872506]
57. Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN. RNA-Seq gene expression estimation with read mapping uncertainty. Bioinformatics. 2010; 26:493–500. [PubMed: 20022975]
58. Finak G, et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biol. 2015; 16:278. [PubMed: 26653891]
59. Stegle O, Teichmann SA, Marioni JC. Computational and analytical challenges in single-cell transcriptomics. Nat Rev Genet. 2015; 16:133–145. [PubMed: 25628217]
60. Hicks SC, Teng M, Irizarry RA. On the widespread and critical impact of systematic bias and batch effects in single-cell RNA-Seq data. bioRxiv. 2015
61. Brennecke P, et al. Accounting for technical noise in single-cell RNA-seq experiments. Nat Methods. 2013; 10:1093–1095. [PubMed: 24056876]

62. Vallejos CA, Marioni JC, Richardson S. BASiCS: Bayesian analysis of single-cell sequencing data. PLOS Comput Biol. 2015; 11:e1004333. [PubMed: 26107944]
63. Lun ATL, Bach K, Marioni JC. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. Genome Biol. 2016; 17:75. [PubMed: 27122128]
64. Levin JZ, et al. Comprehensive comparative analysis of strand-specific RNA sequencing methods. Nat Methods. 2010; 7:709–715. [PubMed: 20711195]
65. Alter O, Brown PO, Botstein D. Singular value decomposition for genome-wide expression data processing and modeling. Proc Natl Acad Sci USA. 2000; 97:10101–10106. [PubMed: 10963673]
66. Risso D, Schwartz K, Sherlock G, Dudoit S. GC-content normalization for RNA-Seq data. BMC Bioinformatics. 2011; 12:480. [PubMed: 22177264]
67. Risso D, Ngai J, Speed TP, Dudoit S. Normalization of RNA-seq data using factor analysis of control genes or samples. Nat Biotechnol. 2014; 32:896–902. [PubMed: 25150836]
68. Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet. 2007; 3:1724–1735. [PubMed: 17907809]
69. Jiang L, et al. Synthetic spike-in standards for RNA-seq experiments. Genome Res. 2011; 21:1543–1551. [PubMed: 21816910]
70. Marinov GK, et al. From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. Genome Res. 2014; 24:496–510. [PubMed: 24299736]
71. Achim K, et al. High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. Nat Biotechnol. 2015; 33:503–509. [PubMed: 25867922]
72. Munro SA, et al. Assessing technical performance in differential gene expression experiments with external spike-in RNA control ratio mixtures. Nat Commun. 2014; 5:5125. [PubMed: 25254650]
73. Grün D, Kester L, van Oudenaarden A. Validation of noise models for single-cell transcriptomics. Nat Methods. 2014; 11:637–640. [PubMed: 24747814]
74. Grün D, van Oudenaarden A. Design and analysis of single-cell sequencing experiments. Cell. 2015; 163:799–810. [PubMed: 26544934]
75. Deng Q, Ramsköld D, Reinius B, Sandberg R. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. Science. 2014; 343:193–196. [PubMed: 24408435]
76. Jaitin DA, et al. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. Science. 2014; 343:776–779. [PubMed: 24531970]
77. Wang Y, Navin NE. Advances and applications of single-cell sequencing technologies. Mol Cell. 2015; 58:598–609. [PubMed: 26000845]
78. Saliba AE, Westermann AJ, Gorski SA, Vogel J. Single-cell RNA-seq: advances and future challenges. Nucleic Acids Res. 2014; 42:8845–8860. [PubMed: 25053837]
79. Hashimshony T, et al. CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. Genome Biol. 2016; 17:77. [PubMed: 27121950]
80. Islam S, et al. Quantitative single-cell RNA-seq with unique molecular identifiers. Nat Methods. 2014; 11:163–166. [PubMed: 24363023]
81. Kivioja T, et al. Counting absolute numbers of molecules using unique molecular identifiers. Nat Methods. 2011; 9:72–74. [PubMed: 22101854]
82. Fu GK, et al. Molecular indexing enables quantitative targeted RNA sequencing and reveals poor efficiencies in standard library preparations. Proc Natl Acad Sci USA. 2014; 111:1891–1896. [PubMed: 24449890]
83. Shiroguchi K, Jia TZ, Sims PA, Xie XS. Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. Proc Natl Acad Sci USA. 2012; 109:1347–1352. [PubMed: 22232676]
84. Fu GK, Hu J, Wang PH, Fodor SPA. Counting individual DNA molecules by the stochastic attachment of diverse labels. Proc Natl Acad Sci USA. 2011; 108:9026–9031. [PubMed: 21562209]
85. Kharchenko PV, Silberstein L, Scadden DT. Bayesian approach to single-cell differential expression analysis. Nat Methods. 2014; 11:740–742. [PubMed: 24836921]

86. McDavid A, et al. Data exploration, quality control and testing in single-cell qPCR-based gene expression experiments. Bioinforma. 2013; 29:461–467.

87. McDavid A, et al. Modeling bi-modality improves characterization of cell cycle on gene expression in single cells. PLOS Comput Biol. 2014; 10:e1003696. [PubMed: 25032992]

88. Dalrymple ML, Hudson IL, Ford RPK. Finite Mixture, Zero-inflated Poisson and Hurdle models with application to SIDS. Comput Stat Data Anal. 2003; 41:491–504.

89. Fan J, et al. Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. Nat Methods. 2016; 13:241–244. [PubMed: 26780092]

90. Pierson E, Yau C. ZIFA: dimensionality reduction for zero-inflated single-cell gene expression analysis. Genome Biol. 2015; 16:241. [PubMed: 26527291]

91. Buettner F, Moignard V, Göttgens B, Theis FJ. Probabilistic PCA of censored data: accounting for uncertainties in the visualization of high-throughput single-cell qPCR data. Bioinformatics. 2014; 30:1867–1875. [PubMed: 24618470]

92. DeTomaso D, Yosef N. FastProject: a tool for low-dimensional analysis of single-cell RNA-Seq data. BMC Bioinformatics. 2016; 17:315. [PubMed: 27553427]

93. Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. Nat Biotechnol. 2015; 33:495–502. [PubMed: 25867923]

94. Qiu P, et al. Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. Nat Biotechnol. 2011; 29:886–891. [PubMed: 21964415]

95. Antebi YE, et al. Mapping differentiation under mixed culture conditions reveals a tunable continuum of T cell fates. PLoS Biol. 2013; 11:e1001616. [PubMed: 23935451]

96. Korem Y, et al. Geometry of the gene expression space of individual cells. PLOS Comput Biol. 2015; 11:e1004224. [PubMed: 26161936]

97. Patel AP, et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science. 2014; 344:1396–1401. [PubMed: 24925914]

98. Pollen AA, et al. Molecular identity of human outer radial glia during cortical development. Cell. 2015; 163:55–67. [PubMed: 26406371]

99. Kowalczyk MS, et al. Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells. Genome Res. 2015; 25:1860–1872. [PubMed: 26430063]

100. Lande-Diner L, Stewart-Ornstein J, Weitz CJ, Lahav G. Single-cell analysis of circadian dynamics in tissue explants. Mol Biol Cell. 2015; 26:3940–3945. [PubMed: 26269583]

101. Buettner F, et al. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. Nat Biotechnol. 2015; 33:155–160. [PubMed: 25599176]

102. Shapiro E, Biezuner T, Linnarsson S. Single-cell sequencing-based technologies will revolutionize whole-organism science. Nat Rev Genet. 2013; 14:618–630. [PubMed: 23897237]

103. Beyer, KS., Goldstein, J., Ramakrishnan, R., Shaft, U. Proceedings of the 7th International Conference on Database Theory; Springer; 1999. p. 217-235.

104. Usoskin D, et al. Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. Nat Neurosci. 2015; 18:145–153. [PubMed: 25420068]

105. Amir AD, et al. viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. Nat Biotechnol. 2013; 31:545–552. [PubMed: 23685480]

106. Shekhar K, Brodin P, Davis MM, Chakraborty AK. Automatic classification of cellular expression by nonlinear stochastic embedding (ACCENSE). Proc Natl Acad Sci USA. 2014; 111:202–207. [PubMed: 24344260]

107. Wilson NK, et al. Combined single-cell functional and gene expression analysis resolves heterogeneity within stem cell populations. Cell Stem Cell. 2015; 16:712–724. [PubMed: 26004780]

108. van der Maaten L, Hinton GE. Visualizing high-dimensional data using t-SNE. J Mach Learn Res. 2008; 9:2579–2605.

109. van der Maaten L. Accelerating t-SNE using tree-based algorithms. J Mach Learn Res. 2014; 15:3221–3245.

110. Mahfouz A, et al. Visualizing the spatial gene expression organization in the brain through non-linear similarity embeddings. Methods. 2015; 73:79–89. [PubMed: 25449901]
111. Maaten, L. Learning a parametric embedding by preserving local structure. In: Dyk, DV., Welling, M., editors. Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics (AISTATS-09); 2009. p. 384-391.http://machinelearning.wustl.edu/mlpapers/papers/AISTATS09_Maaten
112. Berman GJ, Choi DM, Bialek W, Shaevitz JW. Mapping the stereotyped behaviour of freely moving fruit flies. J R Soc Interface. 2014; 11:20140672. [PubMed: 25142523]
113. Ester, M., Kriegel, HP., Sander, J., Xu, X. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96); Portland, Oregon, USA. AAAI Press; 1996. p. 226-231.
114. Habib N, et al. Div-Seq: single-nucleus RNA-Seq reveals dynamics of rare adult newborn neurons. Science. 2016; 353:925–928. [PubMed: 27471252]
115. Shekhar K, et al. Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics. Cell. 2016; 166:1308–1323. [PubMed: 27565351]
116. Tsafrir D, et al. Sorting points into neighborhoods (SPIN): data analysis and visualization by ordering distance matrices. Bioinformatics. 2005; 21:2301–2308. [PubMed: 15722375]
117. Madeira SC, Oliveira AL. Biclustering algorithms for biological data analysis: a survey. IEEE/ACM Trans Comput Biol Bioinform. 2004; 1:24–45. [PubMed: 17048406]
118. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J Stat Mech Theory Exp. 2008; 2008:P10008.
119. Newman MEJ. Communities, modules and large-scale structure in networks. Nat Phys. 2012; 8:25–31.
120. Newman MEJ, Girvan M. Finding and evaluating community structure in networks. Phys Rev E. 2004; 69:026113.
121. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016; 34:525–527. [PubMed: 27043002]
122. Ntranos V, Kamath GM, Zhang JM, Pachter L, Tse DN. Fast and accurate single-cell RNA-seq analysis by clustering of transcript-compatibility counts. Genome Biol. 2016; 17:112. [PubMed: 27230763]
123. Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic. J R Stat Soc Ser B Stat Methodol. 2001; 63:411–423.
124. Horowitz A, et al. Genetic and environmental determinants of human NK cell diversity revealed by mass cytometry. Sci Transl Med. 2013; 5:208ra145.
125. Fruchterman TMJ, Reingold EM. Graph drawing by force-directed placement. Softw Pract Exper. 1991; 21:1129–1164.
126. Shoval O, et al. Evolutionary trade-offs, Pareto optimality, and the geometry of phenotype space. Science. 2012; 336:1157–1160. [PubMed: 22539553]
127. Hart Y, et al. Inferring biological tasks using Pareto analysis of high-dimensional data. Nat Methods. 2015; 12:233–235. 3, 235. [PubMed: 25622107]
128. Tendler A, Mayo A, Alon U. Evolutionary tradeoffs, Pareto optimality and the morphology of ammonite shells. BMC Syst Biol. 2015; 9:12. [PubMed: 25884468]
129. Sheftel H, Shoval O, Mayo A, Alon U. The geometry of the Pareto front in biological phenotype space. Ecol Evol. 2013; 3:1471–1483. [PubMed: 23789060]
130. Novershtern N, et al. Densely interconnected transcriptional circuits control cell states in human hematopoiesis. Cell. 2011; 144:296–309. [PubMed: 21241896]
131. Gupta PB, et al. Stochastic state transitions give rise to phenotypic equilibrium in populations of cancer cells. Cell. 2011; 146:633–644. [PubMed: 21854987]
132. Wagner F. GO-PCA: an unsupervised method to explore gene expression data using prior knowledge. PLoS One. 2015; 10:e0143196. [PubMed: 26575370]
133. Chung NC, Storey JD. Statistical significance of variables driving systematic variation in high-dimensional data. Bioinformatics. 2015; 31:545–554. [PubMed: 25336500]

134. Bar-Joseph Z, Gitter A, Simon I. Studying and modelling dynamic biological processes using time-series gene expression data. Nat Rev Genet. 2012; 13:552–564. [PubMed: 22805708]
135. Bendall SC, et al. Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. Cell. 2014; 157:714–725. [PubMed: 24766814]
136. Kafri R, et al. Dynamics extracted from fixed cells reveal feedback linking cell growth to cell cycle. Nature. 2013; 494:480–483. [PubMed: 23446419]
137. Trapnell C, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol. 2014; 32:381–386. [PubMed: 24658644]
138. Magwene PM, Lizardi P, Kim J. Reconstructing the temporal ordering of biological samples using microarray data. Bioinformatics. 2003; 19:842–850. [PubMed: 12724294]
139. Booth KS, Lueker GS. Testing for the consecutive ones property, interval graphs, and graph planarity using PQ-tree algorithms. J Comput Syst Sci. 1976; 13:335–379.
140. Haghverdi L, Buettner F, Theis FJ. Diffusion maps for high-dimensional single-cell analysis of differentiation data. Bioinformatics. 2015; 31:2989–2998. [PubMed: 26002886]
141. Moignard V, et al. Decoding the regulatory network of early blood development from single-cell gene expression measurements. Nat Biotechnol. 2015; 33:269–276. [PubMed: 25664528]
142. Angerer P, et al. destiny: diffusion maps for large-scale single-cell data in R. Bioinformatics. 2016; 32:1241–1243. [PubMed: 26668002]
143. Coifman RR, et al. Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps. Proc Natl Acad Sci USA. 2005; 102:7426–7431. [PubMed: 15899970]
144. Marco E, et al. Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. Proc Natl Acad Sci USA. 2014; 111:E5643–E5650. [PubMed: 25512504]
145. Setty M, et al. Wishbone identifies bifurcating developmental trajectories from single-cell data. Nat Biotechnol. 2016; 34:637–645. [PubMed: 27136076]
146. Whitfield ML, et al. Identification of genes periodically expressed in the human cell cycle and their expression in tumors. Mol Biol Cell. 2002; 13:1977–2000. [PubMed: 12058064]
147. Leng N, et al. Oscope identifies oscillatory genes in unsynchronized single-cell RNA-seq experiments. Nat Methods. 2015; 12:947–950. [PubMed: 26301841]
148. Rosenkrantz D, Stearns R, Lewis PII. An analysis of several heuristics for the traveling salesman problem. SIAM J Comput. 1977; 6:563–581.
149. Cho RJ, et al. Transcriptional regulation and function during the human cell cycle. Nat Genet. 2001; 27:48–54. [PubMed: 11137997]
150. Zopf CJ, Quinn K, Zeidman J, Maheshri N. Cell-cycle dependence of transcription dominates noise in gene expression. PLOS Comput Biol. 2013; 9:e1003161. [PubMed: 23935476]
151. Shin J, et al. Single-cell RNA-seq with waterfall reveals molecular cascades underlying adult neurogenesis. Cell Stem Cell. 2015; 17:360–372. [PubMed: 26299571]
152. Llorens-Bobadilla E, et al. Single-cell transcriptomics reveals a population of dormant neural stem cells that become activated upon brain injury. Cell Stem Cell. 2015; 17:329–340. [PubMed: 26235341]
153. Lawrence N. Probabilistic non-linear principal component analysis with Gaussian process latent variable models. J Mach Learn Res. 2005; 6:1783–1816.
154. Ke R, et al. In situ sequencing for RNA analysis in preserved tissue and cells. Nat Methods. 2013; 10:857–860. [PubMed: 23852452]
155. Lee JH, et al. Highly multiplexed subcellular RNA sequencing in situ. Science. 2014; 343:1360–1363. [PubMed: 24578530]
156. Lovatt D, et al. Transcriptome in vivo analysis (TIVA) of spatially defined single cells in live tissue. Nat Methods. 2014; 11:190–196. [PubMed: 24412976]
157. Lubeck E, Cai L. Single-cell systems biology by super-resolution imaging and combinatorial labeling. Nat Methods. 2012; 9:743–748. [PubMed: 22660740]
158. Lubeck E, Coskun AF, Zhiyentayev T, Ahmad M, Cai L. Single-cell in situ RNA profiling by sequential hybridization. Nat Methods. 2014; 11:360–361. [PubMed: 24681720]

159. Chen KH, Boettiger AN, Moffitt JR, Wang S, Zhuang X. RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. Science. 2015; 348:aaa6090. [PubMed: 25858977]
160. Crosetto N, Bienko M, van Oudenaarden A. Spatially resolved transcriptomics and beyond. Nat Rev Genet. 2015; 16:57–66. [PubMed: 25446315]
161. Rockhill RL, Euler T, Masland RH. Spatial order within but not between types of retinal neurons. Proc Natl Acad Sci USA. 2000; 97:2303–2307. [PubMed: 10688875]
162. Masland RH. The neuronal organization of the retina. Neuron. 2012; 76:266–280. [PubMed: 23083731]
163. Scialdone A, et al. Resolving early mesoderm diversification through single-cell expression profiling. Nature. 2016; 535:289–293. [PubMed: 27383781]
164. Durruthy-Durruthy R, et al. Reconstruction of the mouse otocyst and early neuroblast lineage at single-cell resolution. Cell. 2014; 157:964–978. [PubMed: 24768691]
165. Durruthy-Durruthy R, Gottlieb A, Heller S. 3D computational reconstruction of tissues with hollow spherical morphologies using single-cell gene expression data. Nat Protoc. 2015; 10:459–474. [PubMed: 25675210]
166. Kim HD, Shay T, O'Shea EK, Regev A. Transcriptional regulatory circuits: predicting numbers from alphabets. Science. 2009; 325:429–432. [PubMed: 19628860]
167. Yosef N, Regev A. Impulse control: temporal dynamics in gene transcription. Cell. 2011; 144:886–896. [PubMed: 21414481]
168. Wills QF, et al. Single-cell gene expression analysis reveals genetic associations masked in whole-tissue experiments. Nat Biotechnol. 2013; 31:748–752. [PubMed: 23873083]
169. Tay S, et al. Single-cell NF-kappaB dynamics reveal digital activation and analogue information processing. Nature. 2010; 466:267–271. [PubMed: 20581820]
170. Xue Z, et al. Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. Nature. 2013; 500:593–597. [PubMed: 23892778]
171. Munsky B, Neuert G, van Oudenaarden A. Using gene expression noise to understand gene regulation. Science. 2012; 336:183–187. [PubMed: 22499939]
172. Kim JK, Marioni JC. Inferring the kinetics of stochastic gene expression from single-cell RNA-sequencing data. Genome Biol. 2013; 14:R7. [PubMed: 23360624]
173. Karlebach G, Shamir R. Modelling and analysis of gene regulatory networks. Nat Rev Mol Cell Biol. 2008; 9:770–780. [PubMed: 18797474]
174. Fisher, J., Köksal, AS., Piterman, N., Woodhouse, S. Synthesising executable gene regulatory networks from single-cell gene expression data. In: Kroening, D., Păsăreanu, CS., editors. Computer Aided Verification—27th International Conference, CAV 2015; San Francisco, California, USA. July 18–24, 2015; Springer; 2015. p. 544-560.Proceedings, Part I
175. Köksal, AS., et al. Synthesis of Biological Models from Mutation Experiments. Proceedings of the 40th Annual ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages; ACM; 2013. p. 469-482.
176. Botev ZI, Grotowski JF, Kroese DP. Kernel density estimation via diffusion. Ann Stat. 2010; 38:2916–2957.
177. Battich N, Stoeger T, Pelkmans L. Control of transcript variability in single mammalian cells. Cell. 2015; 163:1596–1610. [PubMed: 26687353]
178. Bahar Halpern K, et al. Nuclear retention of mRNA in mammalian tissues. Cell Reports. 2015; 13:2653–2662. [PubMed: 26711333]
179. Rabani M, et al. High-resolution sequencing and modeling identifies distinct dynamic RNA regulatory strategies. Cell. 2014; 159:1698–1710. [PubMed: 25497548]
180. Taniguchi Y, et al. Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. Science. 2010; 329:533–538. [PubMed: 20671182]
181. Liu Y, Beyer A, Aebersold R. On the dependency of cellular protein levels on mRNA abundance. Cell. 2016; 165:535–550. [PubMed: 27104977]
182. Schwanhäusser B, et al. Global quantification of mammalian gene expression control. Nature. 2011; 473:337–342. [PubMed: 21593866]

183. Li JJ, Bickel PJ, Biggin MD. System wide analyses have underestimated protein abundances and the importance of transcription in mammals. Peer J. 2014; 2:e270. [PubMed: 24688849]

184. Jovanovic M, et al. Immunogenetics. Dynamic profiling of the protein life cycle in response to pathogens. Science. 2015; 347:1259038. [PubMed: 25745177]

185. Clark SJ, Lee HJ, Smallwood SA, Kelsey G, Reik W. Single-cell epigenomics: powerful new methods for understanding gene regulation and cell identity. Genome Biol. 2016; 17:72. [PubMed: 27091476]

186. Cedar H, Bergman Y. Linking DNA methylation and histone modification: patterns and paradigms. Nat Rev Genet. 2009; 10:295–304. [PubMed: 19308066]

187. Zhou VW, Goren A, Bernstein BE. Charting histone modifications and the functional organization of mammalian genomes. Nat Rev Genet. 2011; 12:7–18. [PubMed: 21116306]

188. Dixon JR, et al. Chromatin architecture reorganization during stem cell differentiation. Nature. 2015; 518:331–336. [PubMed: 25693564]

189. Landan G, et al. Epigenetic polymorphism and the stochastic formation of differentially methylated regions in normal and cancerous tissues. Nat Genet. 2012; 44:1207–1214. [PubMed: 23064413]

190. Shipony Z, et al. Dynamic and static maintenance of epigenetic memory in pluripotent and somatic cells. Nature. 2014; 513:115–119. [PubMed: 25043040]

191. Schwartzman O, Tanay A. Single-cell epigenomics: techniques and emerging applications. Nat Rev Genet. 2015; 16:716–726. [PubMed: 26460349]

192. Xin, RS., et al. Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data; ACM; 2013. p. 13-24.(ed.)

193. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015; 521:436–444. [PubMed: 26017442]

194. Navin NE. The first five years of single-cell cancer genomics and beyond. Genome Res. 2015; 25:1499–1507. [PubMed: 26430160]

195. Gawad C, Koh W, Quake SR. Dissecting the clonal origins of childhood acute lymphoblastic leukemia by single-cell genomics. Proc Natl Acad Sci USA. 2014; 111:17947–17952. [PubMed: 25425670]

196. Potter NE, et al. Single-cell mutational profiling and clonal phylogeny in cancer. Genome Res. 2013; 23:2115–2125. [PubMed: 24056532]

197. Meyer M, et al. Single cell-derived clonal analysis of human glioblastoma links functional and genomic heterogeneity. Proc Natl Acad Sci USA. 2015; 112:851–856. [PubMed: 25561528]

198. Biesecker LG, Spinner NB. A genomic view of mosaicism and human disease. Nat Rev Genet. 2013; 14:307–320. [PubMed: 23594909]

199. Cai X, et al. Single-cell, genome-wide sequencing identifies clonal somatic copy-number variation in the human brain. Cell Reports. 2014; 8:1280–1289. [PubMed: 25159146]

200. Evrony GD, et al. Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. Cell. 2012; 151:483–496. [PubMed: 23101622]

201. McConnell MJ, et al. Mosaic copy number variation in human neurons. Science. 2013; 342:632–637. [PubMed: 24179226]

202. Gole J, et al. Massively parallel polymerase cloning and genome sequencing of single cells using nanoliter microwells. Nat Biotechnol. 2013; 31:1126–1132. [PubMed: 24213699]

203. Knouse KA, Wu J, Whittaker CA, Amon A. Single cell sequencing reveals low levels of aneuploidy across mammalian tissues. Proc Natl Acad Sci USA. 2014; 111:13409–13414. [PubMed: 25197050]

204. Zhang CZ, et al. Calibrating genomic and allelic coverage bias in single-cell sequencing. Nat Commun. 2015; 6:6822. [PubMed: 25879913]

205. Kim KI, Simon R. Using single cell sequencing data to model the evolutionary history of a tumor. BMC Bioinformatics. 2014; 15:27. [PubMed: 24460695]

206. Suzuki A, et al. Single-cell analysis of lung adenocarcinoma cell lines reveals diverse expression patterns of individual cells invoked by a molecular target drug treatment. Genome Biol. 2015; 16:66. [PubMed: 25887790]

207. Weirather JL, et al. Characterization of fusion genes and the significantly expressed fusion isoforms in breast cancer by hybrid sequencing. Nucleic Acids Res. 2015; 43:e116–e116. [PubMed: 26040699]

208. Afik S, et al. Targeted reconstruction of T cell receptor sequence from single cell RNA-sequencing links CDR3 length to T cell differentiation state. bioRxiv. 2016

209. Dey SS, Kester L, Spanjaard B, Bienko M, van Oudenaarden A. Integrated genome and transcriptome sequencing of the same cell. Nat Biotechnol. 2015; 33:285–289. [PubMed: 25599178]

210. Macaulay IC, et al. G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. Nat Methods. 2015; 12:519–522. [PubMed: 25915121]

211. Angermueller C, et al. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. Nat Methods. 2016; 13:229–232. [PubMed: 26752769]

212. Frei AP, et al. Highly multiplexed simultaneous detection of RNAs and proteins in single cells. Nat Methods. 2016; 13:269–275. [PubMed: 26808670]

213. Albayrak C, et al. Digital quantification of proteins and mRNA in single mammalian cells. Mol Cell. 2016; 61:914–924. [PubMed: 26990994]

214. Darmanis S, et al. Simultaneous multiplexed measurement of RNA and proteins in single cells. Cell Reports. 2016; 14:380–389. [PubMed: 26748716]

215. Albert FW, Kruglyak L. The role of regulatory variation in complex traits and disease. Nat Rev Genet. 2015; 16:197–212. [PubMed: 25707927]

216. Risso D, et al. Power gain: how normalization affects reproducibility and biological insight of RNA-seq studies in neuroscience [v1; not peer reviewed]. F1000Research ISCB Comm J. 2015; 4:411.

217. Stubbington MJT, et al. T cell fate and clonality inference from single-cell transcriptomes. Nat Meth. 2016; 13:329–332.

218. Chu LF, et al. Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. Genome Biol. 2016; 17:1–20. [PubMed: 26753840]

219. Cadwell CR, et al. Electrophysiological, transcriptomic and morphologic profiling of single neurons using Patch-seq. Nat Biotech. 2016; 34:199–203.

220. Prabhakaran, S., Azizi, E., Carr, A., Pe'er, D. Dirichlet Process Mixture Model for Correcting Technical Variation in Single-Cell Gene Expression Data. Proceedings of the 33nd International Conference on Machine Learning, ICML 2016; New York City, NY, USA. June 19–24, 2016; 2016. p. 1070-1079.

221. Reinius B, et al. Analysis of allelic expression patterns in clonal somatic cells by single-cell RNA-seq. Nat Genet. 2016 advance online publication.

222. Wang B, Zhu J, Pierson E, Ramazzotti D, Batzoglou S. Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. bioRxiv. 2016

223. Grün D, et al. De Novo Prediction of Stem Cell Identity using Single-Cell Transcriptome Data. Cell Stem Cell. 2016; 19:266–277. [PubMed: 27345837]

224. Vallejos CA, Richardson S, Marioni JC. Beyond comparisons of means: understanding changes in gene expression at the single-cell level. Genome Biol. 2016; 17:1–14. [PubMed: 26753840]

225. Yosef N, Regev A. Writ large: Genomic Dissection of the Effect of Cellular Environment on Immune Response. Science. 2016

226. Tirosh I, et al. Single-cell RNA-seq supports a developmental hierarchy in IDH-mutant oligodendroglioma. Nature. 2016

227. Gokce O, et al. Cellular Taxonomy of the Mouse Striatum as Revealed by Single-Cell RNA-Seq. Cell Rep. 2016; 16:1126–1137. [PubMed: 27425622]

Box 1

The many facets of a cell's identity

We define a cell's *identity* as the outcome of the instantaneous intersection of all factors that affect it. We refer to the more permanent aspects in a cell's identity as its *type* (e.g., a hepatocyte typically cannot turn into a neuron) and to the more transient elements as its *state*. Cell types are often organized in a hierarchical taxonomy, as types may be further divided into finer subtypes; such taxonomies are often related to a cell fate map, reflecting key steps in differentiation. Cell *states* arise transiently during time-dependent processes, either in a *temporal progression* that is unidirectional (e.g., during differentiation, or following an environmental stimulus) or in a *state vacillation* that is not necessarily unidirectional and in which the cell may return to the origin state. Vacillating processes can be *oscillatory* (e.g., cell-cycle or circadian rhythm) or can transition between states with no predefined order (e.g., due to stochastic, or environmentally controlled, molecular events). These time-dependent processes may occur transiently within a stable cell type (as in a transient environmental response), or may lead to a new, distinct type (as in differentiation). A cell's identity is also affected by its *spatial context* that includes the cell's absolute *location*, defined as its position in the tissue (for example, the location of a cell along the dorsal ventral axis determines its exposure to a morphogen gradient), and the cell's *neighborhood*, which is the identity of neighboring cells.

The cell's identity is manifested in its molecular contents. Genomic experiments measure these in *molecular profiles*, and computational methods infer information on the cell's identity from the measured molecular profiles (inevitably, the molecular profile also reflects allele-intrinsic and technical variation that must be handled properly by computational methods before any analysis is done). We refer to this as inferring *facets* of the cell's identity (or the *factors* that created it) to stress that none describes it fully, but each is an important, distinguishable aspect.

By analogy, we relate the facets to the *vectors* that span the space of cell identities. In many cases, computational analysis methods find such basis vectors directly (as discussed in main text) and these indeed relate well to biological facets of identity. However, this idealized definition, and the present computational tools, are likely to be insufficient to capture the true nature of this space. In particular, basis vectors in algebra are defined to be independent of each other, but facets of a cell's identity that we would like to distinguish and identify separately— such as its type, location, and state—may be largely dependent on one another. For example, the spatial position of a cell in a solid organ is a fixed element of its identity that is usually distinguished from its 'type' but is nevertheless not independent of cell type. In another example, whereas a cell cycle phase may have invariant characteristics across systems[30,35,99], the ability of a cell to enter the cell cycle and the duration of the phase can depend on cell type and can influence other temporal processes like differentiation. As the field of single-cell genomics develops, it may be possible to define abstractions, possibly employing data-driven categories rather ones imposed by prior conceptions, that both are mathematically precise and reflect the key biological components.

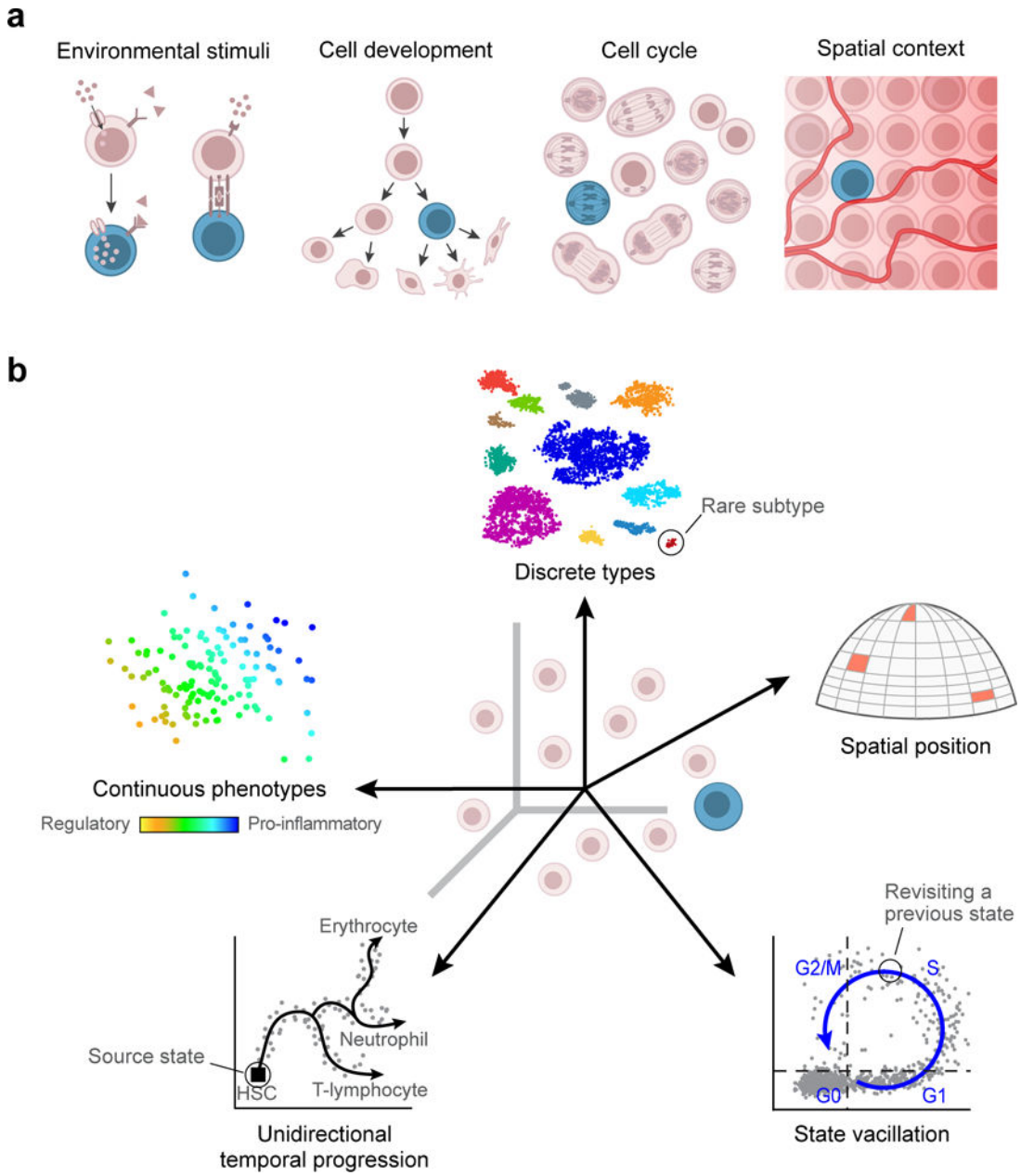

Figure 1.

(a) A cell participates simultaneously in multiple biological contexts. The illustration depicts a particular cell (highlighted in blue) as it experiences multiple concurrent contexts that shape its identity simultaneously (from left to right): environmental stimuli, such as nutrient availability or the binding of a signaling molecule to a receptor; a specific state on a developmental trajectory; the cell cycle; and a spatial context, which determines its physical environment (e.g., oxygen availability), cellular neighbors, and developmental cues (e.g., morphogen gradients).

(b) The biological factors affecting the cell combine to create its unique, instantaneous identity, which is captured in the cell's molecular profile. Computational methods dissect the molecular profile and tease apart facets of the cell's identity, which are akin to 'basis

vectors' that span a space of possible cellular identities. Key examples include (counterclockwise from top): (1) division into discrete types (e.g., cell populations in the retina (A.R. and colleagues[30])); cell type frequency can vary by multiple orders of magnitude from the most abundant to the rarest subtype; (2) continuous phenotypes (e.g., the pro-inflammatory potential of each individual T cell, quantified through a gene expression signature derived from bulk pathogenic T cell profiles (N.Y., A.R. and colleagues[1])); (3) temporal progression (e.g., normal differentiation, such as hematopoiesis); (4) temporal vacillation between cellular states (e.g., oscillation through cell cycle; data taken from A.R. and colleagues[99]); (5) physical locations: a schematic representation of an embryo at 50% epiboly (only half is shown), divided into discrete spatial bins; independent *in situ* hybridization data of landmark genes allows inferring spatial bins (highlighted) from which single cells had likely originated (figure adapted from A.R. and colleagues[93]). The scatterplots represent single cells (dots) projected onto two dimensions (e.g., first two principal components or using t-SNE).

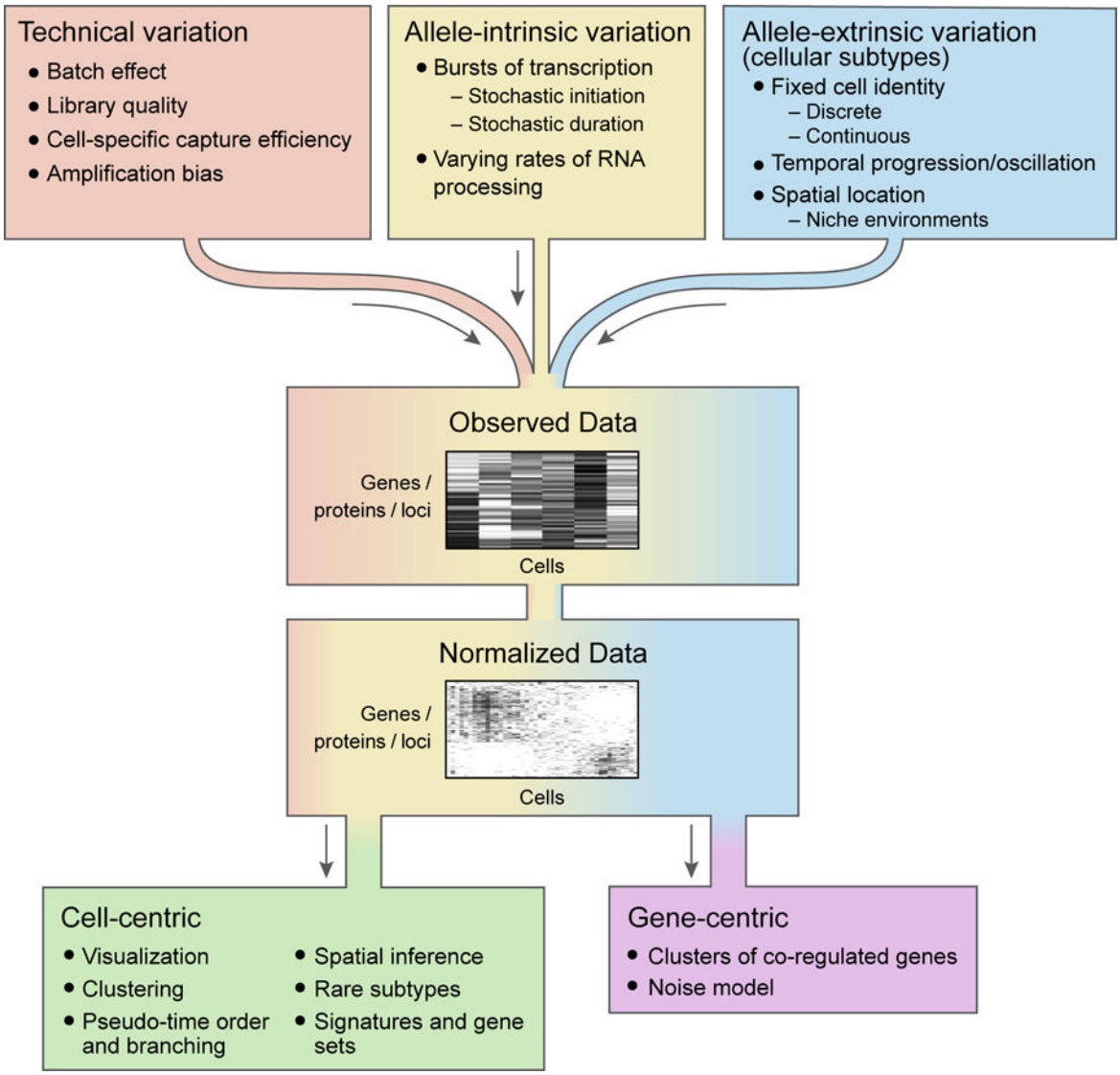

Figure 2.

Biological and technical factors combine to determine the measured genomic profiles of single cells; computational methods remove technical effects and tease apart facets of the biological variation. The sources of variation that affect single-cell genomics data are (1) technical factors that reflect variance due to the experimental process (e.g., batch effects); (2) factors that are intrinsic to the process under study (e.g., transcription) and reflect stochastic fluctuations (e.g., transcriptional or translational bursts in mRNA or proteins) that do not correlate between two alleles of the same gene; and (3) factors that are extrinsic to the process under study, reflecting the presence of different cell types and states (e.g., concentrations of key transcription, translation, or metabolic factors). Computational methods are needed to remove the nuisance technical variation (although they typically cannot completely eliminate it) before the biological variation can be confidently explored. Most single-cell studies explore allele-extrinsic factors and can be classified as either cell-centric or gene-centric. Cell-centric analyses aim to catalog the cells into phenotypic groups,

whether discrete (e.g., clustering) or continuous (e.g., temporal ordering). Gene-centric analyses aim to understand the dynamics and regulation of the generating mechanism (e.g., transcriptional circuits).

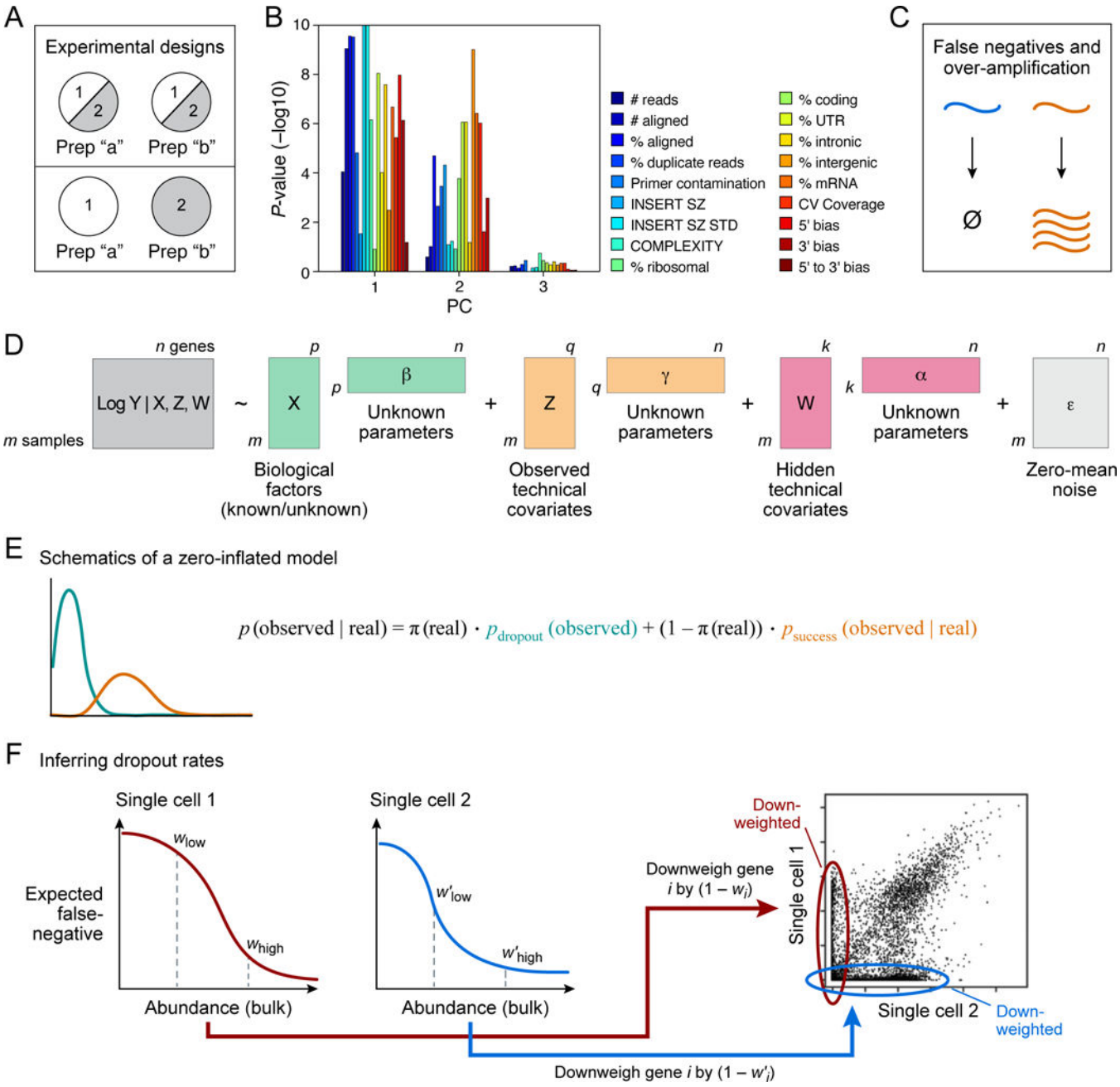

Figure 3.

Technical confounders of single-cell RNA-seq and computational methods to handle them. (**a**) Batch effects. This source of technical variability can be mitigated by careful experimental design. The upper panel shows a design in which two biological conditions ("1" and "2"; for example, wild type and knockout) are distributed evenly between two technical batches ("Prep a" and "Prep b"). This allows statistical methods to account for the batch effect. In contrast, in the lower panel, the biological variation cannot be separated from the batch effect. (**b**) Library quality. The primary principal components of single-cell gene expression correlate strongly with quality metrics such as number of aligned reads and

library complexity (A.R. and colleagues[64]). A typical example is provided. The *y* axis shows the $-\log_{10} P$ value of the Spearman correlation between each of 18 quality metrics (color coded) and one of the primary principal components of the unnormalized expression data (FPKM units; data is taken from N.Y., A.R. and colleagues[1], in which the quality metrics are described; SZ = size, STD = standard deviation). (**c**) Dropouts and amplification bias. Because of the minute quantities of starting material in single cells, expressed transcripts may not be detected because they stochastically failed to amplify; on the other hand, the massive amplification exaggerates any source of technical noise. (**d**) Latent technical confounders. These can be identified using matrix factorization (notation follows ref. 49; visualization adapted from ref. 216). The observed expression (Y, m samples by n genes) is often assumed to be a linear combination of biological and technical factors for statistical tractability[49,50,67,68]. It can be generally modeled as a sum of (**a**) X: p biological factors (either known a priori—for example, genetic background—or latent, in which case p is unknown); (**b**) Z: n known technical covariates (e.g., experimental batches); (**c**) W: k latent factors of technical noise (k is unknown); (**d**) random noise ε with zero mean. α, β, γ determine the influence of each factor on every gene, with β representing the biology of interest and α, γ, being nuisance factors that need to be properly handled before β can be inferred. (**e**) The prevalence of dropouts is modeled through a zero-inflated model: gene expression is modeled as a mixture of two distributions: the ‘real’ one, observed when a transcript is successfully amplified, that reflects the true mRNA abundance ($p$success, in orange)) and a ‘dropout’ that occurs when a transcript fails to amplify ($p$dropout, in teal). The mixing ratio π depends on the transcript’s real expression since it has been empirically observed that weakly expressed transcripts are more prone to dropouts[1]. (**f**) Modeling dropout probabilities based on empirical data. Left and middle, false-negative rate curves (computed for each cell) describe the probability for a dropout event (y axis) as a logistic function of transcript abundance in the corresponding bulk population. Right, the inferred rates weigh down the effect of possible dropout events. Each dot represents the expression of one gene in two arbitrary single cells (x and y axes). Undetected genes (circled) are weighed down when computing the correlation between the expression profiles of the two cells (data obtained from N.Y., A.R. and colleagues[3]).