Perspective

Defining cell types and states with single-cell genomics

Cole Trapnell

Department of Genome Sciences, University of Washington, Seattle, Washington 98105, USA

A revolution in cellular measurement technology is under way: For the first time, we have the ability to monitor global gene regulation in thousands of individual cells in a single experiment. Such experiments will allow us to discover new cell types and states and trace their developmental origins. They overcome fundamental limitations inherent in measurements of bulk cell population that have frustrated efforts to resolve cellular states. Single-cell genomics and proteomics enable not only precise characterization of cell state, but also provide a stunningly high-resolution view of transitions between states. These measurements may finally make explicit the metaphor that C.H. Waddington posed nearly 60 years ago to explain cellular plasticity: Cells are residents of a vast "landscape" of possible states, over which they travel during development and in disease. Single-cell technology helps not only locate cells on this landscape, but illuminates the molecular mechanisms that shape the landscape itself. However, single-cell genomics is a field in its infancy, with many experimental and computational advances needed to fully realize its full potential.

Since Robert Hooke first observed the multicellular structure of plants and animals under his microscope in 1665, biologists have aimed to catalog and classify cells by form and function. How many different types of cells are there in our bodies? What does each type do? How does this diversity arise? How do the different types of cells collaborate in a tissue, and ultimately, an organism? Although much has been learned over the past three and a half centuries, these fundamental questions still captivate us today.

Cataloging the cells of the human body is a maddeningly difficult problem. Human bodies are frequently said to have 210 different types of cells. However, a single type of cell from this taxonomy is still bewilderingly diverse. For example, muscle cells can be divided by functional differences such as contraction speed and subcategorized by unique gene expression programs. Should these subcategories be declared distinct cell types? What differences, be they functional, regulatory, or morphological, are sufficient to define an organism's cellular taxonomy?

Distinguishing cells presumes the ability to measure the genes and functions that set them apart. However, many cell types or subtypes have few (if any) reliable markers that can be used to experimentally purify them for further study. Even cells that can be purified on the basis of well-established markers will contain hidden diversity. Perhaps, for example, "CD14+ monocytes" actually consist of multiple subpopulations that share CD14 expression in common. Surely, any group of cells will vary in the pathways that are active, the genes that are expressed, and the functions that are being performed at any given instant in time. How much variation is to be expected within a given type? How could such variation even be detected unless markers for these subpopulations were already known?

The challenges we face in classifying and cataloging the various cells in the human body are even more daunting when we consider how they arise during development. Every cell in an adult arises from a single zygote through a sequence of cell divisions and "fate decisions," in which a cell makes a transition from one type or state to another. For the most part, the states a cell can pass through and the genes that govern its choices remain unknown. A developing embryo is a highly organized community of rapidly proliferating cells undergoing continuous morphological and functional changes. These changes are driven by intricate gene expression programming, which itself responds swiftly to an ever-changing milieu of morphogen gradients and cell-to-cell signals. Even if we could rigorously define cell types and stable cellular states, how can we make sense of such a dynamic biological situation?

The advent of single-cell genomics represents a turning point in cell biology. For the first time, we can assay the expression level of every gene in the genome across thousands of individual cells in a single experiment. Such experiments can be performed on mixed populations of cells without the need to experimentally purify or separate the cells by type, eliminating the need for markers that uniquely distinguish them. Doing so may enable not only rigorous and unbiased classification of cell types and states, but also the construction of comprehensive systems biology models that predict the behavior of cells during development. Single-cell genomics will also likely lead to the discovery of the genes and pathways that govern cell fate decisions and transitions.

In this Perspective, I review the current state of single-cell genomics, highlight some areas of ongoing technical development, and describe what are, in my opinion, the major analytic obstacles to realizing the potential of these assays.

Defining cell types and states requires single-cell assays

Single-cell measurements help overcome several key obstacles that have frustrated efforts to precisely define cellular states and catalog them in development and disease. Except in rare instances where cells can be precisely synchronized, bulk measurements destroy crucial information by averaging signals from individual cells together. One example of this is described by Simpson's Paradox (Simpson 1951), which is well known in statistics but rarely discussed in cell biology. Consider an experiment aiming to assess whether two genes show correlated expression levels across a population of cells. Finding such pairs of genes is often a major goal of

Corresponding author: coletrap@uw.edu
Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.190595.115. Freely available online through the *Genome Research* Open Access option.

single-cell genomics experiments, as some cell fate decisions are governed by a single pair of mutually exclusive regulators. For example, the myeloid/erythroid fate decision in common myeloid progenitor cells is made via the mutually antagonistic transcription factors PU.1 (encoded by *SPI1*) and GATA1 (Rekhtman et al. 1999). In the hypothetical experiment, one might observe the pattern on the left side of Figure 1A and conclude that the genes in question are expressed mutually exclusively. However, if the population consisted of a mixture of two separate groups of cells, this conclusion might be incorrect. Grouping the cells properly by type, and *then* performing the analysis could reveal the genes are in fact positively correlated, not negatively correlated. That is, failing to properly compartmentalize the data by cell type leads to a *qualitatively incorrect* interpretation. The misleading effects of Simpson's Paradox are likely to be pervasive in modern experiments using bulk assays.

Another crucial reason that single-cell measurements are necessary to define cell states is that bulk measurements confound changes due to gene regulation with those due to shifts in cell type composition. Consider an experiment aimed at studying the effect of a drug on a tissue composed of two cell types. (Fig. 1B) Suppose a certain gene's expression is measured via bulk-cell analysis before and after treatment with the drug. Upon measuring a major increase in the gene's expression, one might surmise that the drug causes an up-regulation of the gene in the relevant cells. However, it might instead be that the drug stimulates cell division in one of the two subpopulations. Bulk assays cannot discriminate between these two cases.

Time series studies of gene expression, which are foundational for examining dynamic processes in development and disease progression, also suffer from averaging in bulk studies. For example, differentiating cells might transition through a sequence of intermediate states on the way to becoming fully mature. However, they typically do not do so in a synchronized manner, so sampling a population of them at any particular moment in time will yield a mixture of cells from different stages, as illustrated in Figure 1C. Tracking the time-average expression of a gene of interest might yield a qualitatively misleading picture of that gene's developmental regulation. If we could reorder the cells on the horizontal axis, we would see a faithful representation of the gene's expression dynamics during development.

Bulk measurements are thus fundamentally constrained by averaging. Accurately defining the cell types and states in our bodies and explaining how they arise in development and disease demands single-cell measurements.

Technological advances in cellular state measurement

Next-generation sequencing has proven to be a remarkably sensitive means of monitoring gene expression, epigenetic configuration, nuclear structure, and other aspects of cellular state. However, most assays require a minimum level of input material that exceeds that of a single cell. Over the past several years, a number of sequencing-based assays have been optimized to work at the level of individual cells. These improvements have primarily stemmed from breakthroughs in amplification techniques, modifications to reverse transcriptase that improve processivity and enable controlled template switching, and the development of instruments that physically capture and isolate individual cells.

Most single-cell genomics assays have been adapted from similar techniques developed for analyzing bulk-cell populations (Table 1). The most widely used of these is RNA-seq, which measures global gene expression by reverse transcribing RNA into cDNA and sequencing it (Cloonan et al. 2008; Mortazavi et al. 2008; Nagalakshmi et al. 2008). Counting the reads that originate from each gene yields a measure of its expression. Single-cell versions of the RNA-seq protocol isolate individual cells in microfluidic capillaries (Wu et al. 2013), by serial dilution, or via flow

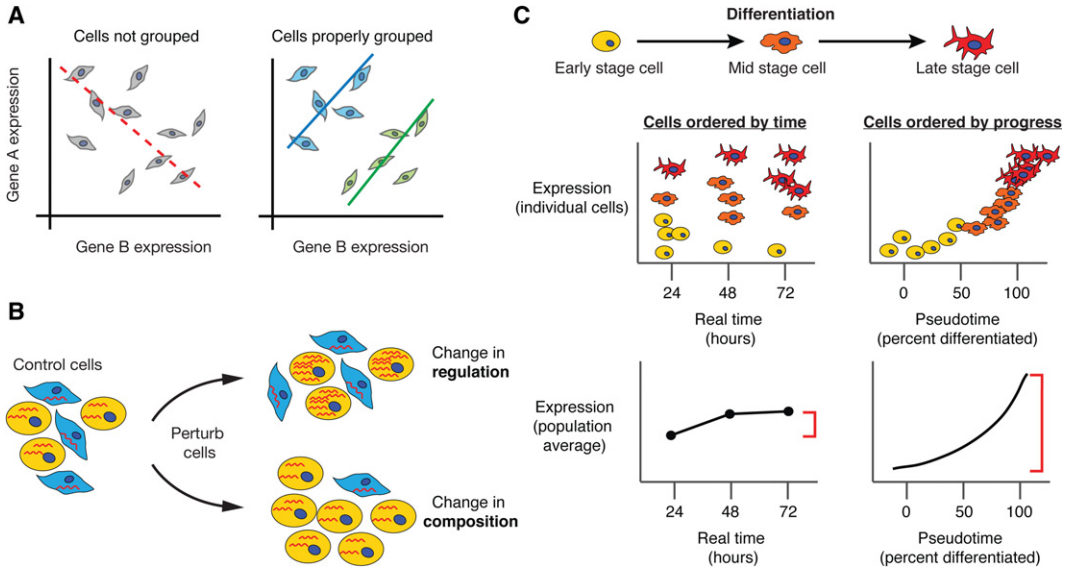

Figure 1. Single-cell measurements preserve crucial information that is lost by bulk genomics assays. (*A*) Simpson's Paradox describes the misleading effects that arise when averaging signals from multiple individuals. (*B*) Bulk measurements cannot distinguish changes due to gene regulation from those that arise due to shifts in the ratio of different cell types in a mixed sample. (*C*) Time series experiments are affected by averaging when cells proceed through a biological process in an unsynchronized manner. A single time point may contain cells from different stages in the process, obscuring the dynamics of relevant genes. Reordering the cells in "pseudotime" according to biological progress eliminates averaging and recovers the true signal in expression (Trapnell et al. 2014).

Table 1. Single-cell genomics assays and papers describing selected protocol

Assay	Measures	References
RNA-seq	Gene expression	Tang et al. (2009); Islam et al. (2011); Hashimshony et al. (2012); Jaitin et al. (2014); Fan et al. (2015); Klein et al. (2015); Macosko et al. (2015); Nakamura et al. (2015)
DNA-seq	Genotype	Navin et al. (2011)
DNA-seq +RNA-seq	Genotype and gene expression	Dey et al. (2015); Macaulay et al. (2015)
In situ RNA-seq	Gene expression, subcellular mRNA localization	Larsson et al. (2010); Lee et al. (2014)
Bisulfite-sequencing	DNA methylation	Guo et al. (2013); Smallwood et al. (2014)
Hi-C	Chromatin conformation	Nagano et al. (2013)
ATAC-seq	DNA accessibility	Buenrostro et al. (2015); Cusanovich et al. (2015)

sorting into multiwell plates. Each cell can thereafter be lysed to recover its RNA. Generating cDNA libraries from picograms of RNA has required major technical innovations and remains an area of considerable development effort (Tang et al. 2009; Islam et al. 2011; Hashimshony et al. 2012; Ramsköld et al. 2012; Picelli et al. 2014). A recent review provides a thorough discussion of the molecular biology underlying recent protocols (Saliba et al. 2014).

Analyzing single-cell RNA-seq data is substantially more difficult than bulk experiments. The size and scale of the data puts considerable strain on most algorithms: Where bulk experiments typically have a few dozen samples at most, typical single-cell studies capture hundreds (Shalek et al. 2013; Trapnell et al. 2014; Treutlein et al. 2014) or even thousands (Shalek et al. 2014; Fan et al. 2015; Klein et al. 2015; Macosko et al. 2015) of cells. Furthermore, single-cell expression measurements are often highly variable, and separating technical from biological variability is essential. New normalization techniques based on exogenous "spike-in" standards or molecular barcoding (Fu et al. 2014) may be needed to attribute changes in expression across cell populations to genuine biology (Brennecke et al. 2013; Grün et al. 2014). For a detailed discussion of these issues, see Stegle et al. (2015); Kolodziejczyk et al. (2015). Finally, single-cell RNA-seq enables a biologist to ask and answer questions that simply cannot be approached with bulk data and for which no tool currently exists. New algorithms and software tools may be required, both to obtain accurate results and also to exploit the unique possibilities of these measurements.

Integrating imaging with single-cell genomics will help connect cell form, function, and communication with gene regulation. Remarkably, individual cells maintain similar concentration of individual mRNA species despite wide variation in cell volume, suggesting that global transcriptional regulation is tightly connected to the physical properties of the cell (Padovan-Merhar et al. 2015). Mats Nilsson and colleagues used rolling circle amplification of padlock probes to resolve subcellular localization of multiple highly similar RNA species in individual cells (Larsson et al. 2010). More recently, Lee et al. (2014) devised FISSEQ, which obtains mRNA abundances and subcellular localization for thousands of genes by performing cDNA sequencing-by-synthesis reactions within single cells rather than on a sequencer's glass flowcell. SeqFISH and MERFISH are combinatorial fluorescence in situ hybridization strategies that use successive rounds of labeling, imaging, and photobleaching to recover mRNA abundances and subcellular localization simultaneously (Lubeck and Cai 2012; Chen et al. 2015). These techniques could in principle illuminate gene networks that mediate cell–cell communication in solid tissues. Several groups have reconstructed spatial position of individual cells directly from their transcriptomes via statistical inference in developing mouse and zebrafish embryos (Durruthy-Durruthy et al. 2014; Achim et al. 2015; Satija et al. 2015). The intuition behind such algorithms is that cells from the same region of a tissue are receiving similar paracrine signals and reside at similar positions within morphogen gradients, which should induce a common signature in these cells. An algorithm that can identify those signatures should be able to orient all cells with respect to one another. Single-cell analysis is already helping to map novel cell–cell interactions under controlled, in vitro settings (Shalek et al. 2014). Single-cell genomic analysis of such interactions in vivo will provide a radical advance in our understanding of how cells collaborate in solid tissues.

RNA-seq is only one of several aspects of cell state to be measured with single-cell genomics. Peter Fraser and colleagues adapted Hi-C, a technique for measuring chromatin conformation (Lieberman-Aiden et al. 2009) to work in single cells (Nagano et al. 2013). Hi-C measures physical proximity between each pair of sites in the genome in three dimensions, producing a "contact map" that can be used to identify looping interactions between regulatory elements and gene loci. Single-cell Hi-C might be used to investigate a key question in development, i.e., does the highly structured packing of DNA into the nucleus, which is disrupted as part of every cell division, contribute to cell fate decisions? Do stable cellular states have corresponding "signature" chromatin conformations? Single-cell Hi-C has so far only been applied to a small number of cells, but with enough observations, it might be possible to answers such questions.

If RNA-seq measures the "output" of each gene locus, epigenetic sequencing assays such as bisulfite sequencing and DNase I hypersensitivity sequencing track a locus's "input" signals. Whole-genome bisulfite sequencing originally introduced to assay DNA methylation (Lister et al. 2008) has been successfully adapted to work in single cells (Guo et al. 2013) through "reduced representation" mapping (Meissner et al. 2008), which assays DNA methylation at CpG islands. This snapshot of the methylome constitutes an epigenetic signature that distinguishes different gene regulatory states. Histone modifications, such as H3K4 trimethylation, which marks active gene promoters, can be assayed genome-wide by performing chromatin immunoprecipitation and then sequencing the recovered DNA (ChIP-seq). ChIP-seq for chromatin marks or transcription factors can be used to create highly informative descriptions of the gene regulatory state of a cell population (Johnson et al. 2007; Mikkelsen et al. 2007). Although ChIP-seq has not yet been adapted to work in single cells, it is being optimized for smaller and smaller input (Lara-Astiaso et al. 2014).

Just as each cell type displays characteristic gene expression programming, cell types are identified by distinct sets of genomic regulatory sequences that are accessible to DNA binding profiles (Stergachis et al. 2013). There are now several genome-wide assays for accessibility (Crawford et al. 2006; Giresi et al. 2007; Hesselberth et al. 2009; Buenrostro et al. 2013). We recently adapted ATAC-seq, which uses an engineered transposase to construct a

sequencing library from accessible DNA elements (Buenrostro et al. 2013), for use in single cells (Buenrostro et al. 2015; Cusanovich et al. 2015). Tn5 not only cleaves DNA at nucleosome-free regions, it also inserts DNA sequencing library adapters in situ. Thus, treating chromatin with Tn5 produces a library of fragments that mostly come from nucleosome free regions, albeit with lower genomic resolution than DNase I hypersensitivity sequencing (DNase-seq). Remarkably, the Tn5 reaction can take place within intact nuclei, allowing them to be barcoded for single-cell sequencing without ever physically isolating individual nuclei. Physical isolation is currently a requirement of most single-cell genomics protocols and generally results in cost and effort that scales linearly with the number of cells. In contrast, cost and effort for the "cellular indexing" strategy introduced by Cusanovich et al. (2015) scale sublinearly. This scalability enables the profiling of hundreds, thousands, or in principle, tens of thousands of cells in a single experiment.

Assaying gene expression, epigenetic state, nuclear structure, and other aspects of individual cells has enormous advantages over bulk assays, but it also comes with risks and challenges that are poorly understood. One basic difficulty stems from disaggregating a solid tissue into a suspension of single cells. For example, tissue samples are typically treated with enzymes such as collagenase to degrade proteins that hold the cells together. Experiments that use single-cell genomics to survey cell types and frequencies require that the disaggregation be unbiased. A procedure that leads to lysis of one cell type in the tissue but not another would dramatically skew the results of a single-cell genomics analysis. Furthermore, the very act of dissociation could substantially perturb the state of the cells. For example, cells that rely on integrin signaling or communication via elements of the extracellular matrix might be greatly disturbed between disaggregation and measurement. Careful comparisons with bulk assays will be needed to determine the extent of bias introduce by these and other problems that arise from the need to "singularize" tissues.

Ultimately, single-cell measurements of transcriptome and epigenome state will help construct a global view of gene regulation in development and disease. What if we could profile both the transcriptome and the epigenome in an individual cell? Such measurements might unlock the "histone code," revealing how histone modification regulates transcription (Strahl and Allis 2000). However, integrating several types of single-cell measurement poses an enormous experimental and analytic challenge. The amount of DNA and RNA in a single cell is minute: How much might one lose by splitting it among multiple assays? Nevertheless, two recent studies demonstrated how single cells can be genotyped and profiled for global gene expression simultaneously (Dey et al. 2015; Macaulay et al. 2015). Cocapture of different layers of gene regulatory information in individual cells will power statistical models of how upstream regulatory activity corresponds to transcriptional output.

A dynamical systems view of the cell

In 1957, C.H. Waddington proposed a powerful metaphor for how differentiated cell types emerge from a single, totipotent cell (Waddington 1957). The "epigenetic landscape" envisions a ball rolling down a hill, encountering ridges and furrows in the hill as it goes. These features restrict its progress, ultimately forcing it to wind up at or near one of several locations at the base of the hill. Waddington used this metaphor to provide an intuition for how the network of biochemical and gene regulatory interactions in our cells endows them with both robustness and plasticity. Waddington was inspired by *dynamical systems*, which are mathematical frameworks for predicting the qualitative behavior of complex phenomena, such as the orbits of planets, over time. The metaphor has been refined and formalized over time to explicitly describe a cell's state at a given time as a point in "gene regulation space." As the cell regulates gene expression during development, it moves through the space along a *trajectory*. The shape of the trajectory is a function of (1) the cell's starting position in the space, and (2) the differential equations that dictate changes in the cell's gene expression levels that will occur at the next moment in time, given where it currently is in the space. The central question posed by Waddington is: Where will the cell end up?

The landscape analogy is appealing, but we are a long way from using dynamical systems to describe cell fate transitions in most developmental contexts. Nevertheless, some important steps have been taken on the road to constructing useful models of cell state transitions. Stuart Kauffmann theorized that not only can cell fate transitions be described by a dynamical system, but that what we think of as distinct cell types correspond to *attractors* of that system (Kauffman 1993). An attractor is a location in the state space of a dynamical system that will move toward and to which it will return if perturbed. That is, trajectories flow toward attractors. Later work by Sui Huang showed that a simple regulatory network with just two genes that inhibit one another but promote their own expression is sufficient to generate a hypothetical landscape with three stable cell fates (Huang et al. 2007). Two of these states are highly stable, but the third is only weakly stable, and a small perturbation would induce cells in this state to move to one of the other two (Fig. 2A). This situation is reminiscent of a stem cell, which is poised to choose one of several differentiated fates, but will do so only upon some sort of stimulus. Furthermore, modifying the network by weakening the tendency of the genes to promote their own expression makes the "stem-like" state less and less stable, until finally it disappears. Mathematical models describing lateral inhibition mediated by Notch-Delta interactions and the resulting changes in cell state have been experimentally validated and improved to include more members of the pathway (Collier et al. 1996; Sprinzak et al. 2010; Boareto et al. 2015). James Ferrell used nonlinear dynamics to model mitotic oscillators and examined bistability in *Xenopus* oocyte maturation (Ferrell and Machleder 1998). Thus, although dynamical systems models have been restricted to just a few genes, they have captured remarkably rich behaviors commonly observed in development.

Single-cell clustering and trajectory analysis

Single-cell genomics offers a means of precisely quantifying the state of individual cells and thus may enable the construction of explicit, genome-scale dynamical cellular models. Early single-cell transcriptomic studies lend support to the idea that cells are occupants of a vast, complex landscape of possible states and raise doubts that cell types are precisely defined, discrete entities (Buganim et al. 2012; Moignard et al. 2013, 2015; Xue et al. 2013; Durruthy-Durruthy et al. 2014; Kumar et al. 2014; Pollen et al. 2014; Shalek et al. 2014; Trapnell et al. 2014; Treutlein et al. 2014; Buettner et al. 2015). Even within a single cell type, there is considerable variation in global gene expression, consistent with the idea that most cells reside within wells of attraction. Time series experiments of differentiation have observed cells transitioning between a starting state and one or more end states, with many cells distributed along a trajectory between them (Bendall

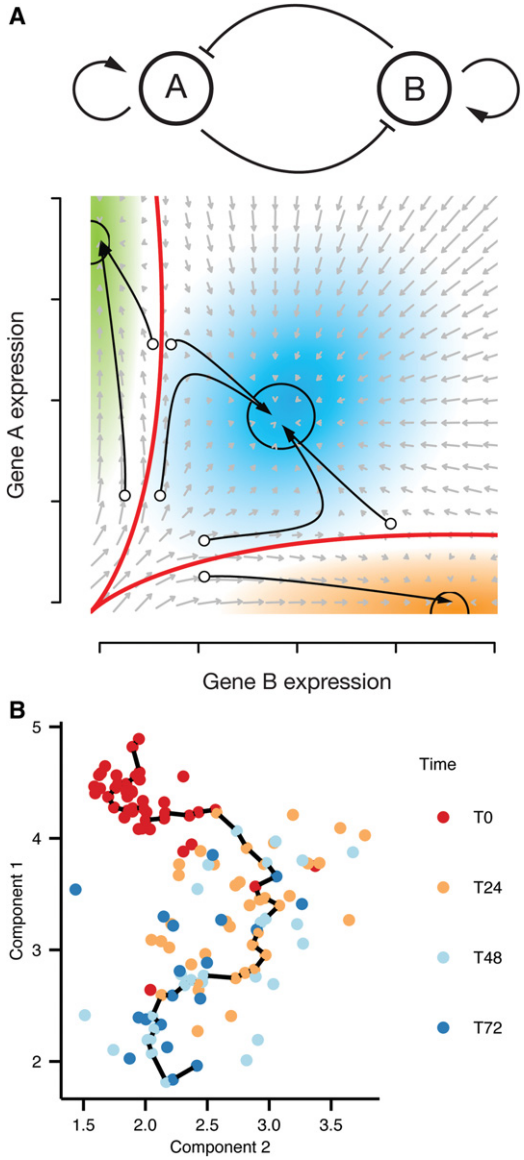

Figure 2. Single-cell "trajectories" shed light on gene regulation. (*A*) An idealized regulatory network consisting of two genes can have three distinct stable states. If the ratio of A to B is sufficiently high, the system will fall into a state in which only A is expressed (green region). Likewise, cells expressing predominantly B will eventually express only B. However, cells with roughly equal expression of A and B will remain in a "poised" state (blue) region. The shaded areas are referred to as "basins of attraction," which determine where cells at different initial positions (white circles) will ultimately rest at equilibrium. (*B*) Gene expression profiles for individual differentiating cells can be informatically organized into trajectories, potentially revealing regulatory network structure and cell fate dynamics.

et al. 2014; Trapnell et al. 2014; Buettner et al. 2015; Moignard et al. 2015).

Unsupervised clustering might be an effective way to identify attractors and thus define cell types and stable states using single-cell genomics data. For example, consider an RNA-seq experiment that has captured a mixture of cells from a solid tumor. The tumor might contain a mixture of three types of cells: invasive tumor cells, noninvasive cells, and stromal cells. We would like to characterize the expression profile of each and assign the cells to the proper type. Each cell captured in the experiment can be represented as a point in a high-dimensional expression space; that is, if there are 30,000 genes in the genome, each cell is a point in a 30,000-dimensional space. Clustering the cells amounts to measuring the distances between all pairs of points, then grouping them into neighborhoods based on mutual proximity. Although there are many algorithms for solving this task, they all face a sobering mathematical reality: the "curse of dimensionality." Distances between points become more and more similar as the dimension of the space they reside within increases. Specifically, the distance between a point and its most distant neighbor approaches the distance between the point and its nearest neighbor (Beyer et al. 1999). If all the cells are equidistant from one another, how are we to cluster them?

Fortunately, in most biological systems, we do not have to shoulder the full burden posed by 30,000-dimensional measurements. Most genes can be grouped together into "modules" with expression levels that are highly correlated across cells in the experiment. Although a module might contain 1000 genes, if they are tightly correlated, we only need one representative gene to describe them all. The cells in a single-cell RNA-seq experiment can often be represented in a space with far fewer than 30,000 dimensions: We only need one dimension for each module of genes. As for unsupervised clustering, there are many algorithms that aim to reduce the dimensionality of data. Principal components analysis (PCA) is by far the most widely used, and it has already proven effective in a number of single-cell genomics studies. By "projecting" the cells down to the first two principal components, we are describing each cell as a point in two dimensions instead of 30,000. For example, several groups have studied developing embryos (Hashimshony et al. 2012; Grindberg et al. 2013; Xue et al. 2013) with single-cell transcriptomics by projecting them onto the first two principal components. In PCA space, cells from two-cell embryos are close to one another and closer to those from four-cell embryos than eight-cell embryos, with later stages still further away. Are two dimensions sufficient to capture all of the subtleties and nuances contained in a real experiment? Probably not, but reducing dimensionality might enable grouping the cells by type, which might be required for other downstream analyses.

If clustering single cells reveals the attractors of the cell state space, can it also be used to find transition paths between stable states? Can we reconstruct cellular trajectories that correspond, for example, to cell differentiation, progression through the cell cycle, or other dynamic processes? Recently, several groups have developed algorithms for not only clustering cells by type, but also organizing them by temporal or developmental stage. These algorithms have aims similar to prior strategies (Magwene et al. 2003; Qiu et al. 2011) for staging embryos, tumors, and other settings, where progression is not a simple function of time. The Monocle algorithm introduced the notion of *pseudotime*, a quantitative measure of biological progression through a process such as cell differentiation. Monocle first reduces the dimensionality of the expression data, then reconstructs a trajectory along which the cells are presumed to travel, and finally projects each cell onto this trajectory at the proper position. Each cell's pseudotime value is measured as the distance along the trajectory from its position back to the beginning (Fig. 2B). In order to describe complex differentiation processes in which cells make fate decisions, Monocle allows these trajectories to have a branched structure with multiple possible outcomes or "lineages" (Trapnell et al. 2014). Monocle is designed to work with single-cell RNA-seq data, but a similar

approach for proteomic data was recently described by Dana Pe'er and colleagues, whose algorithm, Wanderlust, organizes cells measured with high dimensional cytometry (CyTOF) using an entirely different algorithm than Monocle, and achieves similar ends: Bendall et al. (2014) captured millions of differentiating B cells and correctly ordered them by a proteomic measure of pseudotime. More recently, Moignard et al. (2015) used single-cell qPCR to capture the bifurcation between blood-forming and endothelial cells with yet a third algorithm based on diffusion maps. These algorithms all assume that cells are sampled at different points along a complex biological trajectory. Regulation of the cell cycle is intimately coupled to cell differentiation, which could confound trajectory analysis. Should the signal driven by the cell cycle be removed from the data to bring differentiation-dependent changes into focus? Should it be explicitly modeled, even if that means sacrificing power in downstream analyses? Buettner et al. (2015) introduced an algorithm that captures the signal due to the cell cycle (or in principle, other "confounding" phenomena) that is compatible with trajectory analysis. Although these techniques are still very new and a great deal of work remains to be done, it is clear that single-cell experiments can illuminate trajectories in a wide variety of biological settings.

Finding genes that govern cell state

Ultimately, most single-cell studies aim to identify and characterize the genes that determine where a cell is in the state space and drive a cell to transition from one state to another. Doing so requires first distinguishing the genes that are differentially expressed between states or during state transitions from genes that are constitutively expressed throughout the state space. Differentially regulated genes must then be further analyzed to identify those that are responsible for establishing cell state or drive transitions. Solving these two problems may sound straightforward. However, both are surprisingly difficult (the second far more than the first) and have been the subject of countless studies.

Differential expression analysis refers broadly to the task of identifying those genes with expression levels that depend on some variable, like cell type or state, time since perturbation, or genetic background. A biologist that wishes to identify such genes in an experiment collects RNA samples from two or more conditions and compares them with a program designed to find those that differ more than some critical threshold. Exactly what that threshold needs to be depends on how much measurement variability exists in the assay as well as how much uninteresting biological variability exists within each condition in the contrast. To control for uninteresting changes in expression, each condition can be replicated, either by measuring a sample multiple times (to assess technical variability) or repeating the experiment several times (to assess biological variability). The algorithm used to analyze the expression data must then learn a statistical model that accounts for this variability and use it to set the critical threshold for changes in gene expression. This problem has been extensively studied and is essentially solved for expression microarrays, conventional RNA-seq, and a number of other bulk assays.

Single-cell genomics demand new algorithms and software for identifying genes that are differentially expressed. Current single-cell protocols destroy a cell in order to generate a library from it, so a given cell can only be measured once. Although in principle, the lysate from a single cell could be split among several technical replicates, this capability has not yet been demonstrated. Thus, in single-cell genomics experiments, there are no "replicates" per se, which increases the risk that studies will be confounded by batch effects. As discussed above, controlling for library quality and accounting for technical variability is a major focus of technological developments. Analysis software will need to explicitly accommodate protocol improvements in order to distinguish biological variation from technical noise across cells. Single-cell experiments also typically have far more samples than bulk experiments. Due to cost and labor concerns, bulk expression experiments typically have no more than a few samples for each condition. Current single-cell studies profile hundreds or even thousands of cells in each condition and must therefore be analyzed with algorithms that scale to very large data sets.

A cell-state transition typically involves changes in hundreds or even thousands of genes. How do we identify the "master regulators" of the switch? Inferring regulatory networks from expression data captured with microarrays or RNA-seq is considered extremely hard, but single-cell genomics may make it vastly easier. In theory, variation across cells in the RNA level of a key regulatory gene such as a transcription factor should be reflected in variation in RNA levels of downstream targets. Unfortunately, computational methods for inferring regulatory networks have been hamstrung by two major issues. The first is another manifestation of the curse of dimensionality; because there are so many possible gene–gene interactions, even large-scale experiments lack enough data to reliably predict which genes interact. The second arises due to averaging, which destroys the crucial source of variation that any algorithm needs to accurately reconstruct gene regulatory networks from expression data. Moreover, typical experiments capture hundreds or even thousands of cells under highly controlled, consistent conditions. This constitutes orders of magnitude more data than conventional RNA-seq or microarray studies generate, and thus far more information than any existing network inference algorithm has ever been provided.

Several recent single-cell expression studies have constructed regulatory networks from the data using a variety of machine learning–based methods. For example, Buganim et al. (2012) constructed a Bayesian network to define the hierarchy of transcriptional regulators that drive reprogramming from mouse fibroblasts to the induced pluripotent state. After reconstructing a branched trajectory with diffusion maps, Moignard et al. (2015) synthesized a Boolean regulatory network model to pinpoint the genes driving the trajectory. Both of these studies have some important caveats. Notably, because they probe 48 or 96 different genes, the network might exclude some important regulators. Nevertheless, these studies are likely to be the first of many that aim to realize the long-sought goal of inferring regulatory relationships directly from expression data.

Conclusions and outlook

Single-cell genomics experiments will revolutionize our understanding of gene regulation during development and disease because they access a level of information that has simply never been available. Fundamental questions regarding how cell identity is defined and maintained will be, for the first time, answerable. Robust unsupervised clustering analysis will help catalog cell types and states. Trajectory analysis will track cells as they travel across the landscape of possible states. In time, we may even be able to construct complete regulatory networks and genome-scale dynamical systems of the cell from these measurements. However, much work remains to be done, both in the experimental and computational single-cell domains. Sensitivity of

single-cell RNA-seq and other sequencing assays needs further improvement. There are still aspects of genome regulation, such as transcription-factor binding activity, that are not yet measurable in single cells. Many basic statistical analysis questions, such as how to deal with batch effects and technical variability, have not been resolved. Algorithms to identify the critical regulators of cell fate decisions are only just beginning to appear. Despite rapid developments and considerable uncertainty surrounding protocols, instruments, and software, one thing is clear: Single-cell genomics will have a transformative impact on cell biology.

Acknowledgments

I am grateful to Jay Shendure and Vijay Ramani for a critical reading of the manuscript. I am supported by a Dale Frey Award for Breakthrough Scientists through the Damon Runyon Cancer Research Foundation as well as an Alfred P. Sloan Foundation Fellowship.

References

Achim K, Pettit JB, Saraiva LR, Gavriouchkina D, Larsson T, Arendt D, Marioni JC. 2015. High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. *Nat Biotechnol* **33:** 503–509.

Bendall SC, Davis KL, Amir el-AD, Tadmor MD, Simonds EF, Chen TJ, Shenfeld DK, Nolan GP, Pe'er D. 2014. Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell* **157:** 714–725.

Beyer K, Goldstein J, Ramakrishnan R, Shaft U. 1999. When is "nearest neighbor" meaningful? In *Database theory—ICDT'99* (ed. Beeri C, Buneman P), Vol. 1540 of *Lecture Notes in Computer Science*, pp. 217–235. Springer, Berlin, Heidelberg, Germany.

Boareto M, Jolly MK, Lu M, Onuchic JN, Clementi C, Ben-Jacob E. 2015. Jagged–Delta asymmetry in Notch signaling can give rise to a Sender/Receiver hybrid phenotype. *Proc Natl Acad Sci* **112:** E402–E409.

Brennecke P, Anders S, Kim JK, Kołodziejczyk AA, Zhang X, Proserpio V, Baying B, Benes V, Teichmann SA, Marioni JC, et al. 2013. Accounting for technical noise in single-cell RNA-seq experiments. *Nat Methods* **10:** 1093–1095.

Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* **10:** 1213–1218.

Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, Snyder MP, Chang HY, Greenleaf WJ. 2015. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523:** 486–490.

Buettner F, Natarajan KN, Casale FP, Proserpio V, Scialdone A, Theis FJ, Teichmann SA, Marioni JC, Stegle O. 2015. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat Biotechnol* **33:** 155–160.

Buganim Y, Faddah DA, Cheng AW, Itskovich E, Markoulaki S, Ganz K, Klemm SL, van Oudenaarden A, Jaenisch R. 2012. Single-cell expression analyses during cellular reprogramming reveal an early stochastic and a late hierarchic phase. *Cell* **150:** 1209–1222.

Chen KH, Boettiger AN, Moffitt JR, Wang S, Zhuang X. 2015. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* **348:** aaa6090.

Cloonan N, Forrest ARR, Kolle G, Gardiner BBA, Faulkner GJ, Brown MK, Taylor DF, Steptoe AL, Wani S, Bethel G, et al. 2008. Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat Methods* **5:** 613–619.

Collier JR, Monk NA, Maini PK, Lewis JH. 1996. Pattern formation by lateral inhibition with feedback: a mathematical model of delta-notch intercellular signalling. *J Theor Biol* **183:** 429–446.

Crawford GE, David S, Scacheri PC, Renaud G, Halawi MJ, Erdos MR, Green R, Meltzer PS, Wolfsberg TG, Collins FS. 2006. DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays. *Nat Methods* **3:** 503–509.

Cusanovich DA, Daza R, Adey A, Pliner HA. 2015. Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348:** 910–914.

Dey SS, Spanjaard B, Bienko M, van Oudenaarden A. 2015. Integrated genome and transcriptome sequencing of the same cell. *Nat Biotechnol* **33:** 285–289.

Durruthy-Durruthy R, Gottlieb A, Hartman BH, Waldhaus J, Laske RD, Altman R, Heller S. 2014. Reconstruction of the mouse otocyst and early neuroblast lineage at single-cell resolution. *Cell* **157:** 964–978.

Fan HC, Fu GK, Fodor SP. 2015. Expression profiling. Combinatorial labeling of single cells for gene expression cytometry. *Science* **347:** 1258367.

Ferrell JE Jr, Machleder EM. 1998. The biochemical basis of an all-or-none cell fate switch in *Xenopus* oocytes. *Science* **280:** 895–898.

Fu GK, Xu W, Wilhelmy J, Mindrinos MN, Davis RW, Xiao W, Fodor SP. 2014. Molecular indexing enables quantitative targeted RNA sequencing and reveals poor efficiencies in standard library preparations. *Proc Natl Acad Sci* **111:** 1891–1896.

Giresi PG, Kim J, McDaniell RM, Iyer VR, Lieb JD. 2007. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Res* **17:** 877–885.

Grindberg RV, Yee-Greenbaum JL, McConnell MJ, Novotny M, O'Shaughnessy AL, Lambert GM, Araúzo-Bravo MJ, Lee J, Fishman M, Robbins GE, et al. 2013. RNA-sequencing from single nuclei. *Proc Natl Acad Sci* **110:** 19802–19807.

Grün D, Kester L, van Oudenaarden A. 2014. Validation of noise models for single-cell transcriptomics. *Nat Methods* **11:** 637–640.

Guo H, Zhu P, Wu X, Li X, Wen L, Tang F. 2013. Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. *Genome Res* **23:** 2126–2135.

Hashimshony T, Wagner F, Sher N, Yanai I. 2012. CEL-Seq: single-cell RNA-seq by multiplexed linear amplification. *Cell Rep* **2:** 666–673.

Hesselberth JR, Chen X, Zhang Z, Sabo PJ, Sandstrom R, Reynolds AP, Thurman RE, Neph S, Kuehn MS, Noble WS, et al. 2009. Global mapping of protein-DNA interactions *in vivo* by digital genomic footprinting. *Nat Methods* **6:** 283–289.

Huang S, Guo YP, May G, Enver T. 2007. Bifurcation dynamics in lineage-commitment in bipotent progenitor cells. *Dev Biol* **305:** 695–713.

Islam S, Kjällquist U, Moliner A, Zajac P, Fan JB, Lönnerberg P, Linnarsson S. 2011. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res* **21:** 1160–1167.

Jaitin DA, Kenigsberg E, Keren-Shaul H, Elefant N, Paul F, Zaretsky I, Mildner A, Cohen N, Jung S, Tanay A, et al. 2014. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* **343:** 776–779.

Johnson DS, Mortazavi A, Myers RM, Wold B. 2007. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316:** 1497–1502.

Kauffman SA. 1993. *The origins of order: self-organization and selection in evolution*. Oxford University Press, Oxford, UK.

Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, Peshkin L, Weitz DA, Kirschner MW. 2015. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161:** 1187–1201.

Kolodziejczyk AA, Kim JK, Svensson V, Marioni JC, Teichmann SA. 2015. The technology and biology of single-cell RNA sequencing. *Mol Cell* **58:** 610–620.

Kumar RM, Cahan P, Shalek AK, Satija R, DaleyKeyser AJ, Li H, Zhang J, Pardee K, Gennert D, Trombetta JJ, et al. 2014. Deconstructing transcriptional heterogeneity in pluripotent stem cells. *Nature* **516:** 56–61.

Lara-Astiaso D, Weiner A, Lorenzo-Vivas E, Zaretsky I, Jaitin DA, David E, Keren-Shaul H, Mildner A, Winter D, Jung S, et al. 2014. Chromatin state dynamics during blood formation. *Science* **345:** 943–949.

Larsson C, Grundberg I, Söderberg O, Nilsson M. 2010. *In situ* detection and genotyping of individual mRNA molecules. *Nat Methods* **7:** 395–397.

Lee JH, Daugharthy ER, Scheiman J, Kalhor R, Yang JL, Ferrante TC, Terry R, Jeanty SS, Li C, Amamoto R, et al. 2014. Highly multiplexed subcellular RNA sequencing in situ. *Science* **343:** 1360–1363.

Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, et al. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326:** 289–293.

Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR. 2008. Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell* **133:** 523–536.

Lubeck E, Cai L. 2012. Single-cell systems biology by super-resolution imaging and combinatorial labeling. *Nat Methods* **9:** 743–748.

Macaulay IC, Haerty W, Kumar P, Li YI, Hu TX, Teng MJ, Goolam M, Saurat N, Coupland P, Shirley LM, et al. 2015. G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nat Methods* **12:** 519–522.

Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, et al. 2015. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161:** 1202–1214.

Magwene PM, Lizardi P, Kim J. 2003. Reconstructing the temporal ordering of biological samples using microarray data. *Bioinformatics* **19:** 842–850.

Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, Zhang X, Bernstein BE, Nusbaum C, Jaffe DB, et al. 2008. Genome-scale DNA

methylation maps of pluripotent and differentiated cells. *Nature* **454:** 766–770.
Mikkelsen TS, Manching K, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK, Koche RP, et al. 2007. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448:** 553–560.
Moignard V, Macaulay IC, Swiers G, Buettner F, Schütte J, Calero-Nieto FJ, Kinston S, Joshi A, Hannah R, Theis FJ, et al. 2013. Characterization of transcriptional networks in blood stem and progenitor cells using high-throughput single-cell gene expression analysis. *Nat Cell Biol* **15:** 363–372.
Moignard V, Woodhouse S, Haghverdi L, Lilly AJ, Tanaka Y, Wilkinson AC, Buettner F, Macaulay IC, Jawaid W, Diamanti E, et al. 2015. Decoding the regulatory network of early blood development from single-cell gene expression measurements. *Nat Biotechnol* **33:** 269–276.
Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5:** 621–628.
Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M. 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320:** 1344–1349.
Nagano T, Lubling Y, Stevens TJ, Schoenfelder S, Yaffe E, Dean W, Laue ED, Tanay A, Fraser P. 2013. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* **502:** 59–64.
Nakamura T, Yabuta Y, Okamoto I, Aramaki S, Yokobayashi S, Kurimoto K, Sekiguchi K, Nakagawa M, Yamamoto T, Saitou M. 2015. SC3-seq: a method for highly parallel and quantitative measurement of single-cell gene expression. *Nucleic Acids Res* **43:** e60.
Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, Cook K, Stepansky A, Levy D, Esposito D, et al. 2011. Tumour evolution inferred by single-cell sequencing. *Nature* **472:** 90–94.
Padovan-Merhar O, Nair GP, Biaesch AG, Mayer A, Scarfone S, Foley SW, Wu AR, Churchman LS, Singh A, Raj A. 2015. Single mammalian cells compensate for differences in cellular volume and DNA copy number through independent global transcriptional mechanisms. *Mol Cell* **58:** 1–15.
Picelli S, Faridani OR, Björklund AK, Winberg G, Sagasser S, Sandberg R. 2014. Full-length RNA-seq from single cells using Smart-seq2. *Nat Protoc* **9:** 171–181.
Pollen AA, Nowakowski TJ, Shuga J, Wang X, Leyrat AA, Lui JH, Li N, Szpankowski L, Fowler B, Chen P, et al. 2014. Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat Biotechnol* **32:** 1–9.
Qiu P, Gentles AJ, Plevritis SK. 2011. Discovering biological progression underlying microarray samples. *PLoS Comput Biol* **7:** e1001123.
Ramsköld D, Luo S, Wang YC, Li R, Deng Q, Faridani OR, Daniels GA, Khrebtukova I, Loring JF, Laurent LC, et al. 2012. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat Biotechnol* **30:** 777–782.
Rekhtman N, Radparvar F, Evans T, Skoultchi AI. 1999. Direct interaction of hematopoietic transcription factors PU.1 and GATA-1: functional antagonism in erythroid cells. *Genes Dev* **13:** 1398–1411.
Saliba AE, Westermann AJ, Gorski SA, Vogel J. 2014. Single-cell RNA-seq: advances and future challenges. *Nucleic Acids Res* **42:** 8845–8860.
Satija R, Farrell JA, Gennert D, Schier AF, Regev A. 2015. Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol* **33:** 495–502.
Shalek AK, Satija R, Adiconis X, Gertner RS, Gaublomme JT, Raychowdhury R, Schwartz S, Yosef N, Malboeuf C, Lu D, et al. 2013. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* **498:** 236–240.
Shalek AK, Satija R, Shuga J, Trombetta JJ, Gennert D, Lu D, Chen P, Gertner RS, Gaublomme JT, Yosef N, et al. 2014. Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature* **510:** 363–369.
Simpson EH. 1951. The interpretation of interaction in contingency tables. *J R Stat Soc Series B Stat Methodol* **13:** 238–241.
Smallwood SA, Lee HJ, Angermueller C, Krueger F, Saadeh H, Peat J, Andrews SR, Stegle O, Reik W, Kelsey G. 2014. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat Methods* **11:** 817–820.
Sprinzak D, Lakhanpal A, LeBon L, Santat LA, Fontes ME, Anderson GA, Garcia-Ojalvo J, Elowitz MB. 2010. *Cis*-interactions between Notch and Delta generate mutually exclusive signalling states. *Nature* **465:** 86–90.
Stegle O, Teichmann SA, Marioni JC. 2015. Computational and analytical challenges in single-cell transcriptomics. *Nat Rev Genet* **16:** 133–145.
Stergachis AB, Neph S, Reynolds A, Humbert R, Miller B, Paige SL, Vernot B, Cheng JB, Thurman RE, Sandstrom R, et al. 2013. Developmental fate and cellular maturity encoded in human regulatory DNA landscapes. *Cell* **154:** 888–903.
Strahl BD, Allis CD. 2000. The language of covalent histone modifications. *Nature* **403:** 41–45.
Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Bodeau J, Tuch BB, Siddiqui A, et al. 2009. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods* **6:** 377–382.
Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, Lennon NJ, Livak KJ, Mikkelsen TS, Rinn JL. 2014. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* **32:** 381–386.
Treutlein B, Brownfield DG, Wu AR, Neff NF, Mantalas GL, Espinoza FH, Desai TJ, Krasnow MA, Quake SR. 2014. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* **509:** 371–375.
Waddington CH. 1957. *The strategy of the genes: a discussion of some aspects of theoretical biology*. Allen & Unwin, London.
Wu AR, Neff NF, Kalisky T, Dalerba P, Treutlein B, Rothenberg ME, Mburu FM, Mantalas GL, Sim S, Clarke MF, et al. 2013. Quantitative assessment of single-cell RNA-sequencing methods. *Nat Methods* **11:** 41–46.
Xue Z, Huang K, Cai C, Cai L, Jiang CY, Feng Y, Liu Z, Zeng Q, Cheng L, Sun YE, et al. 2013. Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. *Nature* **500:** 593–597.

Defining cell types and states with single-cell genomics

Cole Trapnell

Genome Res. 2015 25: 1491-1498
Access the most recent version at doi:10.1101/gr.190595.115