Cells in a multicellular organism fulfill specific functions by enacting cell-type-specific programs of gene regulation. Single-cell RNA sequencing technologies have provided a transformative view of cell-type-specific gene expression, the output of cell-type-specific gene regulatory programs. This review discusses new single-cell genomic technologies that complement single-cell RNA sequencing by providing additional readouts of cellular state beyond the transcriptome. We highlight regression models as a simple yet powerful approach to relate gene expression to other aspects of cellular state, and in doing so, gain insights into the biochemical mechanisms that are necessary to produce a given gene expression output.

Successes and Limitations of sc-RNA-Seq and Pseudotemporal Analysis
Single-cell RNA sequencing (sc-RNA-seq) methods have allowed biologists to produce ‘molecular atlases’ of gene expression [1–19] that comprehensively catalog the repertoire of cell types present in a tissue, or even in a whole organism. These atlases have given us an unprecedented view into what each cell in an organism is doing at a given time. But the question of why a cell adopted one state and not another is difficult to answer with sc-RNA-seq alone. This review discusses new experimental methods that complement sc-RNA-seq by providing readouts of additional aspects of cellular state beyond the transcriptome and analytical methods that use this ‘multi-omic’ data to try to identify causal factors that regulate cell-state dynamics. Most of these methods are still in a proof-of-concept stage, needing additional technical development before being suitable for wider use. We therefore focus less on the biological settings the methods have been applied to and more on how the data from each method, in theory, might fit into a statistical model of gene regulation.

The idea of using single-cell data to gain insights into gene regulation precedes the development of multi-omic methods. In 2014, two software packages, Monocle [20] and Wanderlust [21], independently introduced the concept of ‘pseudotemporal analysis’, in which sc-RNA-seq data are collected for a population of cells undergoing a dynamic biological process and then computationally ordered into a trajectory that reflects the continuous changes in gene expression that occur from the beginning to the end of the process. Pseudotime trajectories allow one to identify genes that exhibit differential expression (DE; see Glossary) over the course of the biological process and cluster them based on their expression dynamics (i.e., genes with increasing, decreasing, or transient expression patterns). Identifying DE genes with a known regulatory function, such as transcription factors (TFs), can help prioritize follow-up experiments. For example, the original Monocle paper [20] identified candidate regulators of myogenesis based on pseudotime DE gene analysis and validated these candidates using RNAi.

Pseudotemporal analysis has been refined by methods including Monocle 2 [22], DPT [23], Wishbone [24], SLICER [25], and URD [18] that allow one to infer branches in pseudotime.

Highlights
Regression models offer a simple yet powerful framework for integrating single-cell transcriptomic, genetic, and epigenetic data to identify mechanisms of gene regulation.

New protocols for CRISPR loss-of-function screens read out gene expression and genetic perturbations in the same single cell. Regressing expression (phenotype) versus genotype can provide insights into gene function and epistasis.

Antibodies conjugated to barcoded oligonucleotides have been used to read out gene expression and protein epitope abundance in the same single cells. Regression modeling of such data may facilitate the reconstruction of cell signaling networks.

Emerging single-cell ATAC-seq technologies measure chromatin accessibility in single cells and can facilitate the identification of noncoding DNA elements, sequence features, and transcription factors that drive gene expression dynamics.

*Department of Genome Sciences, Room S333, Foege Building, Box 355065, Seattle, WA 98105, USA

1Correspondence: coletrap@uw.edu (C. Trapnell).
Branches in pseudotime correspond to ‘decision points’ in which a cell decides to progress toward one or two mutually exclusive fates. Branched pseudotime inference has been successfully applied to complex biological processes such as hematopoietic development [22] and zebrafish embryogenesis [18]. Methods such as Waddington-OT [26], RNA velocity analysis [27], topological data analysis [28], and Monocle 3 generalize pseudotime even further to support modeling trajectories in which cells may cycle through recurrent intermediate states before terminally differentiating.

The main limitation of pseudotemporal analysis of sc-RNA-seq data lies in the difficulty in identifying the causal factors that push a cell toward one lineage on a trajectory versus another. A fate decision may correlate with the expression of many lineage-specific TFs, making the relative importance of these factors unclear. Moreover, the expression of lineage-specific TFs is often not sufficient to establish a robust differentiation process. Experiments with direct reprogramming of fibroblasts to other lineages [29–32] have shown that to achieve efficient reprogramming, a suitable cell signaling context is necessary to potentiate the effects of lineage-specific TFs. When we apply sc-RNA-seq and pseudotime analysis to in vivo systems, we can observe the result of a cell’s gene regulatory network transducing signals from its environment: the cell appears to traverse a smooth gradient of gene expression that has been compared to the ‘epigenetic gradient’ of Waddington’s landscape [26,27]. But we do not directly observe the structure of the gene regulatory network, or the set of signals the cell has received.

The promise of single-cell multi-omic assays is that by modeling the statistical relationships between different aspects of a cell’s genetic and epigenetic states, we will be able to confirm specific causal factors that regulate the cell fate decisions that one can see in a pseudotime trajectory. We discuss four main families of assays. CRISPR knockout screens measure the impact of gene loss of function (LoF) on gene expression and enable the mapping of gene regulatory networks. Methods for paired quantitation of protein epitopes and RNAs allow one to correlate the state of cell signaling proteins with gene expression. Single-cell ATAC-seq measures chromatin accessibility, and when this information is integrated with transcriptomic data and sequence analysis, it can identify DNA elements important for cis-regulation. Lastly, single-molecule fluorescence in situ hybridization (FISH) and related methods can put gene expression in context of spatial position within a tissue section.

In our discussion of these four families of multi-omic assays, we emphasize statistical regression as a simple yet powerful means to integrate the diverse data types they produce into quantitative models of gene regulation. The approach of directly regressing multiple readouts of cellular state against each other is made possible by the enormous sample size of single-cell assays, in which every cell provides an independent observation of the state of a gene regulatory network. We anticipate that the engine of statistical regression, fueled by single-cell multi-omic data, will in the coming years enable the construction of comprehensive models of regulatory interactions between genes, proteins, noncoding DNA elements, and cell communities.

**Outlining the Architecture of Signaling Pathways with CRISPR Screens**

In development, signaling pathways such as BMP/TGF-β, Wnt, Notch, and Hedgehog are used pleiotropically across tissues and lineages to regulate cell fate decisions. These pleiotropic capabilities were highlighted by Loh et al. [33], who used sequential combinations of activation and inhibition of BMP, Wnt, and Notch signaling to differentiate human pluripotent stem cells into 12 mesodermal lineages. The competency of a cell to enact a lineage-specific response to

---

**Glossary**

- **Accessible chromatin**: DNA that is not wrapped around a nucleosome.
- **ATAC-seq**: assay for transposase accessible chromatin using sequencing.
- **ChIP-seq**: Chromatin immunoprecipitation followed by sequencing. Used to map transcription factor binding sites or histone modification domains.
- **CRISPR**: a method for epigenetically repressing a target gene using a nuclease-deficient Cas9 protein fused to a transcriptionally repressing KRAB domain.
- **CRISPR-induced indel**: a sequence insertion or deletion introduced at a target locus using a CRISPR/Cas9 system. Cas9 protein cuts DNA to make a double-strand break, which is repaired by non-homologous end joining, often introducing indels.
- **CUT&RUN**: cleavage under targets and releasing using nuclease. An alternative to ChIP-seq that requires substantially less input material and sequencing depth.
- **Differential expression**: a statistically significant different in the abundance of a given RNA between one or more datasets.
- **DNase-seq**: uses chromatin digestion with DNase I followed by sequencing to identify DNase hypersensitive sites, which correspond to regions of accessible chromatin.
- **DPT**: a pseudotime analysis method that uses diffusion maps for dimensionality reduction of sc-RNA-seq data.
- **Gene ontology enrichment analysis**: a statistical test to identify gene ontology terms that are associated with a larger proportion of a query set of genes than would be expected due to chance.
- **MNase digestion**: cutting DNA into small fragments using micrococcal nuclease (MNase), which has higher specificity for non-nucleosomal DNA than DNase I.
- **Monocle 2**: a software package for single-cell analysis in the R programming language. Includes an implementation of sc-RNA-seq pseudotime analysis that is based on the dimensionality reduction algorithm ‘DDRTree’.
- **Nested effects models**: regression models that assume that data points
a ubiquitously used signaling pathway is established by a variety of factors. A cell could, for example, express a specific subset of a family of related receptor proteins [34]; express lineage-specific TFs that physically interact with signal transducing transcriptional cofactors [35]; or restrict nuclear receptor binding to specific genomic locations that have a pre-accessible chromatin state [36,37].

The most straightforward way to show that a gene has a causal role in making a cell competent to respond to a signal, drug, or other perturbation is to show that loss of gene function (LoF) results in an abnormal response. This basic principle can be scaled to screen almost every gene in a genome using RNAi [38], a CRISPR-induced indel [39], or CRISPR-mediated epigenetic repression (CRISPRi) [40,41]. A caveat to these methods is that they do not produce complete LoF phenotypes. CRISPR will only make loss of function edits on both alleles in a minority of cells; and with RNAi and CRISPRi, gene expression is knocked down with variable efficiency. Even the phenotype of a bona fide loss of function mutant can be variable due to incomplete penetration. A high-throughput LoF screen is only interpretable if paired with a statistical model to assess the significance of a putative LoF phenotype.

Until recently, most large-scale LoF screens involved measuring a single quantitative metric, for example, the amount of fluorescence from a reporter gene or the fold change in cell count after a drug selection. A natural way to model such data is a regression model in which the quantitative phenotype is a linear function of the “genotype” (which genes are knocked down/out in an experiment or in a single cell). LoF in a gene can be considered to have a significant effect on the phenotype if the regression coefficient for that gene is significantly different from zero.

Recently, CRISPR LoF screens have been paired with sc-RNA-seq to give a multivariate, transcriptomic readout [41–45]. The data produced by these methods can also be interpreted using a phenotype ~ genotype regression model (Figure 1A). In this model, the response is a matrix: rows correspond to cells in the experiment, columns correspond to genes, and the entries in the matrix are log counts of number of mRNA molecules observed for a given gene in a given cell. Correspondingly, there is now a matrix of regression coefficients: each coefficient $\beta_i$ represents the effect of LoF in gene $j$ on the expression of gene $i$. A coefficient $\beta_i < 0$ indicates non-functional gene $j$ results in reduced expression of gene $i$ and therefore suggests that gene $j$ has a role in activating gene $i$. A coefficient $\beta_i > 0$ suggests that gene $j$ inhibits gene $i$.

Several methods for analyzing CRISPR LoF sc-RNA-seq data were recently developed by Dixit et al. [42], who applied them to investigate the role of TFs in bone marrow dendritic cell response to lipopolysaccharide stimulation. After setting up a phenotype ~ genotype regression model as described above, hierarchical clustering was applied to the regression coefficient matrix. Clustering identified ‘modules’ of TFs with similar LoF phenotypes, and modules of coregulated target genes, that they associated with biological pathways using Gene Ontology enrichment analysis [46]. This allowed the results of the experiment to be intuitively summarized as a graph of activation and inhibition relationships between TF modules and target gene modules.

CRISPR LoF screens in which two or more genes per cell are knocked out could potentially allow one to identify and quantify genetic interactions. A phenotype ~ genotype regression model could include interaction terms between genotype terms. Interaction terms could be used to identify genes that are part of a common pathway (interaction term $< 0$, indicating that are generated by a hierarchical process.

Pseudotime trajectory: a computational ordering of cells from a single-cell assay (i.e., sc-RNA-seq) that aims to reconstruct the continuous temporal dynamics of a gene regulatory process.

RNA FISH: RNA fluorescence in situ hybridization. Fluorophore-conjugated oligonucleotide probes are hybridized to mRNA molecules in a fixed sample, allowing them to be counted.

RNA: a biological pathway present in many eukaryotes in which short hairpin RNAs or short double-stranded RNAs trigger degradation and/or translational inhibition of complementary mRNA.

RNA velocity: an algorithm that analyzes the ratios of spliced mRNA to un-spliced pre-mRNAs in single-cell RNA-seq data to estimate the time derivative of cell gene expression profiles, that is, a cell’s ‘RNA velocity’.

Single-cell ATAC-seq: one of several protocols that adapt ATAC-seq to provide chromatin accessibility data for individual cells instead of a bulk cell population.

Single-cell bisulfite sequencing: one of several protocols that adapt bisulfite sequencing to profile CpG methylation in single cells. Bisulfite sequencing uses a chemical reaction to convert non-methylated cytosine nucleotides to uracil, which is read out as thymine when sequenced.

Single-cell THS-seq: an adaptation of the THS-seq protocol that provides chromatin accessibility data for individual cells instead of a bulk cell population.

SLICER: an algorithm for pseudotemporal analysis that examines shortest paths on a k-nearest-neighbor graph of cells.

THS-seq: transposome hypersensitive site sequencing. An alternative to ATAC-seq that uses in vitro transcription to amplify sequence from regions of accessible chromatin.

URD: an algorithm for pseudotemporal analysis. Similar to DPT, but supports reconstructing trajectories with multiple branches.

Waddington’s landscape: a metaphor to describe the process of
Several published combinatorial CRISPR screens have identified synthetic lethal genetic interactions by measuring changes of synthetic guide RNA (sgRNA) barcode abundances before and after a cell population is expanded in culture, often in the context of a drug selection [47–50]. The exponential nature of cell growth allows even small changes of cellular fitness to result in significant differences in sgRNA barcode abundances. In principle, single-cell CRISPR screens performed at large enough scale could be used to test all pairwise or even higher-order mutants from a library of candidate genes, potentially even without a phenotypic selection. Realizing this goal will require further improvements to the technique.

Accurately genotyping each cell is crucial to assessing the effect of mutations on molecular phenotype. In most published methods for CRISPR LoF screens with a sc-RNA-seq readout, sgRNAs are expressed on a plasmid from one promoter, and a barcode linked to the sgRNA during cloning is expressed from a different promoter as part of an mRNA. However, it was found that recombination between the sgRNA and the barcode can break the genotype-to-phenotype linkage, reducing statistical power [51]. One protocol, CROP-seq [45], avoids this problem by encoding cell barcodes by using unique CRISPR-GRNAs for each cell as a part of the CRISPR-Cas9 system. This allows for a one-to-one mapping of each cell to its sgRNA barcode, and the combined effect is less than the sum of the individual effects, or genes that have redundant functions (individual effects $\sim 0$ but interaction term $> 0$).

The combined effect is less than the sum of the individual effects, or genes that have redundant functions (individual effects $\sim 0$ but interaction term $> 0$).
problem by expressing the sgRNA from a promoter placed within the untranslated region of an mRNA (Figure 1B), allowing it to be read out directly in sc-RNA-seq data. However, CROP-seq’s design precludes the vector from coexpressing multiple sgRNAs, which may constrain screens of higher-order combinations of mutations.

Another problem is CRISPR cuts that do not always result in gene LoF. Suppose the probability of an edit causing LoF on one allele is \( p = \frac{5}{6} \) and a cell receives a sgRNA for two genes. The cell will be recorded as having a LoF genotype for both genes, but the chance that the cell will actually get a homozygous knockout for both genes is only \( p^2 \approx 20\% \). There is also a \( 2p^2(1 - p)^2 + 4p(1 - p)^3 + (1 - p)^4 = 21\% \) chance that the cell will have no LoF alleles for at least one of the two targeted genes. One study [42] used an expectation-maximization procedure to try to impute which cells got LoF edits versus non-LoF edits. While this procedure was successful in increasing their signal-to-noise ratio, it may lead to over-estimated regression coefficients in a scenario where a bona fide LoF has incomplete penetrance. Optimized sgRNA libraries for CRISPRi [52] may offer a solution to this problem by producing more reliable LoF than CRISPR cuts. However, to our knowledge, a head-to-head comparison of CRISPRi versus CRISPR cut has not yet been published.

**Paired Protein and Transcriptomic Readouts Using Antibody-Conjugated Oligos**

Gene regulatory networks are mediated by the proteins those genes encode. Proteomic methods that give a readout of protein abundance and state (i.e., the abundance of phosphorylated epitopes) can give key insights into the biochemical mechanisms that underlie the statistical properties of a gene regulatory network. For example, mass cytometry [53], a system that combines antibody labeling with mass spectrometry to enable protein epitope quantification in single cells, has been leveraged to perform de novo statistical inference of signaling pathway architecture [54,55]. Systems based on mass spectrometry however require a different experimental apparatus and technical skill-set from those based on DNA sequencing, which has impeded the integration of genomic and proteomic methods. To work around this technical gap, several groups have developed assays to quantify protein epitope abundances in single cells using DNA sequencing, and in tandem quantify RNA abundances for the same single cells.

In CITE-seq [56] and REAP-seq [57], antibodies are conjugated to single-stranded oligonucleotides (oligos) that contain a barcode for the antibody. These oligos are then reverse transcribed as part of standard sc-RNA-seq protocols, allowing antibody barcodes to be quantified as part of a transcriptomic readout. While these methods could potentially be used to quantify phospho-epitopes, the CITE-seq and REAP-seq papers did not directly demonstrate the capability. A similar antibody-conjugated-oligo system, ID-seq, was developed and used to profile the effects of ~300 kinase inhibitors on the abundance of 70 phospho-epitopes in the context of epidermal stem cell response to epidermal growth factor receptor signaling [58]. This study was at the level of cell populations, not single cells, but it is an impressive proof-of-concept nevertheless. Applying even simple regression models to such data could yield insights into the organization of signaling pathways. Coupled with a genetic LoF screen, more complex machine learning techniques, such as nested effects models [59], might be able to accurately and automatically reconstruct them from phospho-epitope quantification data (see Outstanding Questions).

**Interrogating Chromatin State at Single-Cell Resolution**

Chromatin state at noncoding DNA elements provides another key biochemical mechanism by which a cell establishes its gene regulatory network. A cell’s response to signaling inputs is in
large part determined by pre-existing chromatin state. So-called ‘pioneer’ TFs, named for their ability to bind to closed, nucleosome-bound chromatin and make it accessible to other TFs [60], are a minority among the repertoire of TFs expressed in a cell. Other TFs, including effectors of signaling pathways such as the glucocorticoid receptor, predominantly bind to accessible chromatin [36,61,62]. Studies on ‘dynamic assisted loading’ [37,63,64], a process in which TFs cooperate stochastically to displace nucleosomes, have shown that the ‘pioneer’/‘settler’ distinction is an simplification of reality; however, there remains substantial evidence that in processes such as myogenesis [65–67], adipogenesis [68], and hematopoiesis [69–72], binding of lineage-specific TFs is dependent on accessible chromatin states having been established in advance by more general TFs that are expressed in multiple lineages.

Sequencing-based assays for interrogating chromatin state have the potential to accelerate the process of developing mechanistic models for gene regulation, but several technological hurdles remain. Ideally, an experimenter should be able to (i) quantify the temporal dynamics of chromatin state in a biological process, (ii) relate chromatin state changes to TF binding, and (iii) relate chromatin state changes to gene expression.

**DNase-seq** [73], **ATAC-seq** [74,75], and **THS-seq** [76] assays enable chromatin accessibility to be profiled genome-wide, and **ChIP-seq** targeting histone modifications can provide additional information, for example, to distinguish active from poised enhancers [77]. In some cases, the temporal dynamics of chromatin state could be profiled with a simple time series of bulk assays. Bulk assays however will convolute the dynamics of different cell types if used to profile a heterogenous cell community. They can also be misleading for systems in which cells differentiate asynchronously, as population-level dynamics will not reflect the sequence of chromatin state changes that an individual cell goes through.

Just as sc-RNA-seq has resolved such problems for gene expression analysis, **sc-ATAC-seq** [78,79] and **single-cell THS-seq** [7] hold the potential to resolve them for chromatin accessibility analysis. These methods have already been applied to resolve cell types from heterogeneous tissues [7,80,81] and to reconstruct pseudotemporal trajectories of chromatin state change during differentiation [82,83]. First-generation single-cell assays for chromatin accessibility could be substantially improved in several ways. One method [79] relies on Fluidigm microfluidics that can only process on the order of hundreds of cells per experiment. Other methods [7,78] handle thousands of cells per experiment, but each cell’s chromatin accessibility profile is more sparsely sampled (sparse coverage can be mitigated by aggregating cells with similar accessibility profiles). Commercial kits have not yet been released for single-cell chromatin assays (10X Genomics has announced one in development at the 2018 Advances in Genome Biology and Technology conference), and analysis algorithms that exploit single-cell accessibility data are only starting to appear. Despite these challenges, the potential utility of these assays for quantitatively modeling gene regulation is tremendous.

A bulk or sc-ATAC-seq time series experiment can identify genomic sites that change in accessibility over time. Having identified such sites, a natural question to ask is which TFs are causing the chromatin state to change? One way to model this is as a simple logistic regression [82]: predict whether a site’s accessibility will increase, decrease, or remain unchanged on the basis of features associated with the site (Figure 2A). Ideally, these features would be direct measurements of TF binding. ChIP-seq TF profiling is expensive and requires large numbers of cells however. **CUT&RUN** [84], a new protocol that maps TF binding events using antibody-guided **MNase digestion**, requires fewer cells and lower sequencing depth than ChIP-seq; but it still requires a separate antibody and experiment for each TF one wants to examine.
An alternative to using TF binding profiles to predict chromatin accessibility dynamics is to use computational predictions of TF binding based on DNA sequence. In our experience, TF binding imputed from sequence motifs (position weight matrices) is a mediocre predictor of chromatin state changes and is limited by the fact that many TFs have the exact same motif. More advanced methods for predicting TF binding, such as gapped kernel support vector machines [85,86] or convolutional neural networks [87,88] may substantially improve our ability to explain chromatin state dynamics.

Quantitative models of chromatin state are a stepping stone toward models of gene expression, the ultimate ‘output’ of the various biochemical events that occur around regulatory DNA.
In vertebrates, the prevalence of distal regulatory sites is a major impediment to models of gene expression, as distal site to target gene relationships are difficult to predict. Promoter-to-distal-site contacts can be directly quantified using assays such as ChIA-PET [89–91], promoter capture HiC [92–94], and HiChIP [95,96]. HiChIP is a promising technology that dramatically reduces the input number of cells required compared to ChIA-PET and promoter capture HiC (~100 000 compared to ~10 million or more).

An alternative approach was recently developed in which an algorithm, Cicero [82], is used to computationally predict connections between promoters and distal sites based on patterns of co-accessibility (Figure 2B) in sc-ATAC-seq data. We believe this method may enable the reconstruction of cis-regulatory landscapes in heterogeneous tissue samples. As a proof-of-concept, Cicero [82] was used to map cis-regulatory landscapes in the context of myoblast differentiation. Using these maps, a regression model was trained in which changes in gene expression were predicted based on sequence motifs in the gene promoter and at linked distal sites (Figure 2C). Integrating distal sites into the model more than doubled the proportion of variance explained, and the motifs that were most predictive of expression changes were those of known myogenic TFs such as MYOD, MEF2C, and MEIS1.

We envision that future experiments will integrate sophisticated TF binding prediction, high-quality promoter-to-distal-site maps, and large-scale sc-ATAC-seq datasets to develop regression models that can predict gene expression dynamics and attribute them to TF activity at specific sequences of regulatory DNA. Early work on ‘co-assays’ that measure ‘inputs’ such as chromatin accessibility along with the ‘output’ of gene or protein expression in the same single cells could dramatically improve the power and accuracy of such models compared to methods that integrate data separate experiments. Currently, two such methods, scNMT-seq [97] and Pi-ATAC [98], exist, but they are limited by low throughput. A sufficiently advanced model of gene regulation would implicitly learn the ‘combinatorial logic’ that relates TF protein–protein interactions to expression outputs and allows widely expressed TFs to regulate tissue-specific genes. Such a model would also allow one to perform in silico mutagenesis experiments to predict the functional impacts of noncoding human genetic variation [88,99–101]. Follow-up validation of candidate enhancers with genetic deletions or CRISPRi [96,102–104] would be essential.

Single-cell assays also exist for other epigenetic features in addition to chromatin accessibility. Single-cell bisulfite sequencing [105–108] profiles DNA methylation across whole genomes and has recently adapted to support high throughputs using combinatorial indexing [109]. Since they profile the whole genome, these methods are expensive and require deep sequencing. In the latter study [109], the mean coverage of mappable CpG dinucleotides is 1.1% given a mean unique aligned read count per cell of >400 000. Other single-cell methods include single-cell ChIP-seq [110], which is limited by even greater sparsity than sc-ATAC (~1000 unique reads/cell versus ~10 000 for sc-ATAC), and single-cell Hi-C [111–115], which is well suited for answering questions about 3D genome structure at the megabase scale, but is less suitable (due to lack of resolution) for characterizing individual gene loci.

Relating Single-Cell States to Environmental Context

The single-cell assays discussed so far provide a wealth of tools for investigating how cells enact a response to a developmental or environmental signal. In many contexts however, we do not know what the most important signaling ligands are, or which cells are producing them. Organoid models [116,117] can provide a controlled environment for investigating cell signaling in development. Most organoid systems involve stimulating pluripotent stem cells (PSCs) with
signaling agonist/inhibitor molecules that mimic morphogen gradients in early development. What makes an organoid an organoid however is that once primed by this initial exogenous signaling, the cells, given suitable culture conditions, self-organize into organ-like structures that contain cells differentiated into multiple distinct lineages. sc-RNA-seq has been used to characterize the resulting heterogeneous cell populations [118–121].

Camp et al. [120] provide a model for how to follow up an observational sc-RNA-seq experiment to gain mechanistic insights into cell signaling. They made liver bud organoids by co-culturing human PSC (hPSC)-derived hepatic endoderm cells with mesenchymal and endothelial cells and showed with sc-RNA-seq that the development of hepatocytes in the organoid more closely resembled in vivo hepatogenesis than homotypic differentiation of hPSCs into hepatocytes. They then compared the signaling receptors and ligands expressed in each of the three cell types in the organoid to perform an in silico screen for potential cross-lineage signaling events. The predictions were validated with a multiplexed chemical screen in which miniaturized organoids were exposed to signaling inhibitors and the ratio of hepatic to endothelial cells measured using confocal imaging. This screen confirmed that inhibition of several pathways predicted in silico to be involved in cross-lineage communication affected hepatic differentiation.

Profiling single-cell transcriptomes with RNA FISH allows for analyses of signaling interaction between cell types to be performed in a native biological context. Unlike sequencing, RNA FISH can be applied directly to tissue sections without cell dissociation. New protocols such as MERFISH [122] and seqFISH [123] have scaled RNA FISH to profile hundreds of RNAs in the same experiment, enabling high-throughput quantification of signaling ligands, receptors, and other genes of interest in situ. Moreover, these imaging based readouts preserve the spatial and morphological information present in the sample.

Figure 3. Spatial Gene Expression Analysis with In Situ Hybridization and sc-RNA-Seq. A cartoon illustrating the analytical approach used by Halpern et al. [125], who profiled murine liver lobules with RNA fluorescence in situ hybridization (FISH) and single-cell ATAC-sequencing (sc-RNA-seq). Liver lobules are hexagonal structures with a central vein and portal veins at each vertex (shown as circles in the figure). The spatial axis of interest is the relative distance of a cell from the central versus portal vein. Halpern et al. profiled the spatial expression patterns of a handful of ‘landmark genes’ with FISH. The position of cells from sc-RNA-seq on the central-to-portal axis was imputed based on their expression of these landmark genes. Given the imputed cell positions, the spatial gene expression patterns of novel genes without FISH data could be estimated. Some genes, such as Hamp and Igfbp2, featured non-monotonic expression patterns, peaking in the middle between the pericentral and periportal regions. t-SNE is t-stochastic neighbor embedding.
Profiling the spatial expression patterns of a limited set of ‘landmark’ genes with in situ hybridizations can allow one to map the physical location of cells in a sc-RNA-seq assay and correspondingly impute the spatial expression patterns of novel genes. This approach was first demonstrated by Satija et al. [124], who integrated a database of in situ assays and sc-RNA-seq data to impute spatial expression patterns in early zebrafish embryos. Karaikos et al. [9] performed a similar analysis for early Drosophila embryos and were able to impute complex spatial expression patterns, such as stripes. Halpem et al. [125] applied this approach to the murine liver, imputing the relationship between gene expression and physical distance from the central vein versus portal nodes in liver lobules (Figure 3). In each system, variation in cell morphology with respect to spatial position is much less prominent than variation in gene expression, suggesting that cryptic spatial expression variation may exist in other biological contexts that have putatively homogenous cell populations. We anticipate that adding additional multi-omic assays, such as single-cell protein epitope quantification chromatin accessibility profiling, to spatial gene expression analysis will substantially advance our understanding of how cell signaling and morphogen gradients give rise to stereotyped patterns of gene expression in development.

Concluding Remarks and Future Directions
How does a complex animal endowed with consciousness arise from a single cell? This fundamental question has driven human inquiry as far back as Aristotle. Many general gene regulatory principles underlying the process of development, for example, cell signaling, lineage-specific TFs, and chromatin biology, were demonstrated well before the genomics era. Genomics gives us the opportunity to fill in all the details with high throughput to learn which proteins, which DNA sequences, and which cell types are necessary in any given sub-task within the grand program of development. With the advent of sc-RNA-seq, it is now feasible to make a comprehensive ‘parts list’ for an entire organism. The challenge that remains is how to scale traditional methods for interrogating the function of these parts, for example, genetic screens, phospho-proteomics, chromatin state profiling, or in situ analysis, to keep pace with the massive scale of observational data being generated (see Outstanding Questions).

Regression models provide a simple yet powerful tool that can leverage the scale and diverse readouts provided by single-cell multi-omic assays to construct quantitative models of gene regulation. As one profiles larger numbers of cells, one obtains more and more observations to fuel a regression model and therefore the statistical power to fit a model with more and more complex sets of features. In the coming years, single-cell datasets on the order of hundreds of thousands or millions of cells will become commonplace. We anticipate that single-cell analysis at this scale will allow us to model how interactions between genes, proteins, regulatory DNA, and cell communities establish the epigenetic landscape more comprehensively and elegantly than ever before.

References

Outstanding Questions
Could co-assaying single-cell phospho-epitope abundances and expression measurements in single cells be used to computationally reconstruct cell signaling network architecture?

Regression models relating gene expression dynamics to regulatory element sequence features can identify candidate transcription factors that establish the global dynamics of a biological process. But how can the key transcription factors and regulatory elements that explain the dynamics of individual, specific genes of interest be identified with high throughput?

Can the quantitative and statistical methods discussed here be adapted to emerging single-cell technologies that preserve spatial context, or will new computational frameworks be required?


Trends in Genetics

114. Stevens, T.J. et al. (2017) 3D structures of individual mammalian genomes studied by single-cell Hi-C. Nature 544, 59–64