Downloaded from genome.cshlp.org on May 2, 2026 . Published by Cold Spring Harbor Laboratory Press

Applications of single-cell genomics and computational strategies to study common disease and population-level variation

Benjamin J. Auerbach, Jian Hu, Muredach P. Reilly, et al.

Genome Res. 2021 31: 1728-1741
Access the most recent version at doi:10.1101/gr.275430.121

References This article cites 169 articles, 29 of which can be accessed free at:
http://genome.cshlp.org/content/31/10/1728.full.html#ref-list-1

Open Access Freely available online through the *Genome Research* Open Access option.

Creative Commons License This article, published in *Genome Research*, is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at http://creativecommons.org/licenses/by-nc/4.0/.

Email Alerting Service Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or **click here.**

To subscribe to *Genome Research* go to:
https://genome.cshlp.org/subscriptions

© 2021 Auerbach et al.; Published by Cold Spring Harbor Laboratory Press

Downloaded from genome.cshlp.org on May 2, 2026 . Published by Cold Spring Harbor Laboratory Press

Perspective

Applications of single-cell genomics and computational strategies to study common disease and population-level variation

Benjamin J. Auerbach,[1] Jian Hu,[2] Muredach P. Reilly,[3] and Mingyao Li[2]

[1]*Graduate Group in Genomics and Computational Biology, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania 19104, USA;* [2]*Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania 19104, USA;* [3]*Division of Cardiology, Department of Medicine, Columbia University Irving Medical Center, New York, New York 10032, USA*

The advent and rapid development of single-cell technologies have made it possible to study cellular heterogeneity at an unprecedented resolution and scale. Cellular heterogeneity underlies phenotypic differences among individuals, and studying cellular heterogeneity is an important step toward our understanding of the disease molecular mechanism. Single-cell technologies offer opportunities to characterize cellular heterogeneity from different angles, but how to link cellular heterogeneity with disease phenotypes requires careful computational analysis. In this article, we will review the current applications of single-cell methods in human disease studies and describe what we have learned so far from existing studies about human genetic variation. As single-cell technologies are becoming widely applicable in human disease studies, population-level studies have become a reality. We will describe how we should go about pursuing and designing these studies, particularly how to select study subjects, how to determine the number of cells to sequence per subject, and the needed sequencing depth per cell. We also discuss computational strategies for the analysis of single-cell data and describe how single-cell data can be integrated with bulk tissue data and data generated from genome-wide association studies. Finally, we point out open problems and future research directions.

Human physiology is shaped by trillions of cells. Although all cells contain nearly identical genomes, cells are programmed via the complex rules of genomic regulation, which requires the consideration of many variables, such as chromatin conformation, DNA methylation, histone modifications, etc., to take on unique cell states. These cell states, such as those associated with our common notions of cell types, enable cells to perform specific functions. Through the interaction of cells within local structures defined by tissues and across different local structures in organ systems, cells generate higher level functions of human physiology, for example, serum glucose regulation via cells of the pancreas, liver, and skeletal muscle.

Human diseases are often marked by abnormalities in high-level functions of human physiology that are caused by abnormalities in subpopulations of cells. One fundamental goal of human disease research is to identify the appropriate perturbations, for example, taking a drug or eating a certain diet, that will produce molecular changes in the subpopulation of cells to fix aberrant behavior; in doing so, such perturbations should produce downstream changes in higher-level physiology that will achieve improvement in health status. Moreover, precision medicine aims to achieve this goal by considering the influence of genetics (Ashley 2016). Achieving the ability to predict the effect of perturbations in humans to improve health will require an unmasking of the complex regulation of the cell and improved understanding of how cell interactions shape human physiology.

The advent of high-throughput single-cell genomics technologies has brought the scientific community one step closer toward meeting this fundamental goal (Linnarsson and Teichmann 2016). As single-cell RNA-sequencing (scRNA-seq) has been adopted earliest by the scientific community, its use has now become widespread and the technology has improved rapidly. At present, it is now common for laboratories to assay genome-wide transcriptomes of thousands of cells in a single scRNA-seq experiment (Aldridge and Teichmann 2020). Recent years have brought on continued development of single-cell technologies. The cost of single-cell experiments continues to cheapen. Technologies that enable the measurement of new information about single cells—for example, chromatin accessibility (Cusanovich et al. 2015; Lake et al. 2018; Preissl et al. 2018), protein quantification (Oikonomou et al. 2020; Brunner et al. 2021; Specht et al. 2021), spatial location (Moffitt et al. 2018; Eng et al. 2019; Takei et al. 2021), and RNA velocity (Qiu et al. 2020)—have been developed. Further, it has now become possible to profile multiple molecular modalities simultaneously within the same cell (Macaulay et al. 2017; Stoeckius et al. 2017; Cao et al. 2018; Chen et al. 2019; Zhu et al. 2019; Fiskin et al. 2020; Ma et al. 2020; Swanson et al. 2021; Xiong et al. 2021).

In this article, we first review the current state of single-cell studies in common disease and then discuss factors that need to be considered when designing a large-scale population-based single-cell study. We then describe computational strategies for the analysis of population-scale single-cell data. We conclude by summarizing lessons learned so far from existing single-cell studies of

Corresponding author: mingyao@pennmedicine.upenn.edu
Article and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.275430.121. Freely available online through the *Genome Research* Open Access option.

© 2021 Auerbach et al. This article, published in *Genome Research*, is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at http://creativecommons.org/licenses/by-nc/4.0/.

31:1728–1741 Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/21; www.genome.org

Downloaded from genome.cshlp.org on May 2, 2026 . Published by Cold Spring Harbor Laboratory Press

human disease and point out open questions and new opportunities for future research.

Applications of single-cell genomics to characterize cell state abnormalities in human disease

Although recent years have seen the development of single-cell technologies to survey new molecular modalities such as proteins and chromatin accessibility, scRNA-seq has been mostly widely used to study human disease because of its maturity. Since the first transcriptome-wide profiling of mRNA by high-throughput sequencing in a single cell was reported in 2009 (Tang et al. 2009), scRNA-seq has increasingly gained popularity owing to its ability to survey cell state diversity in an unbiased fashion. In the past few years, we have witnessed rapid development of scRNA-seq technology both in throughput and in detection sensitivity (Svensson et al. 2018). In particular, sample multiplexing and droplet-based approaches allow several thousands of cells to be assayed simultaneously. These technological advances and the increased adoption of scRNA-seq approaches have begun to shift the application of this method from descriptive analyses of cell heterogeneity closer toward the understanding of disease mechanisms.

scRNA-seq has been used in several contexts to characterize cell state differences between diseased and nondiseased individuals in cross-sectional designs. Type 2 diabetes (T2D) is among the disease fields that has adopted scRNA-seq since the earliest stages of scRNA-seq technology. In 2016 alone, there were six published studies that used scRNA-seq to profile the transcriptomes of pancreatic islets in healthy and T2D donors. Although the initial study only had less than 100 cells (Li et al. 2016), later studies have increased the number of cells as well as the number of donors (Baron et al. 2016; Segerstolpe et al. 2016; Wang et al. 2016; Xin et al. 2016; Lawlor et al. 2017). Most notable among these studies, Segerstolpe et al. (2016) profiled more than 2200 cells in six healthy and four T2D donors, the largest single-cell study in T2D at that time. Using the Smart-seq2 protocol, they generated transcriptional profiles of individual pancreatic endocrine and exocrine cells of healthy and T2D donors and simultaneously defined the transcriptional signatures of both abundant and rare cell types in the pancreas, including delta, gamma, epsilon, stellate, immune, and endothelial cells. Further analyses revealed cell type–specific gene expression and novel subpopulations, as well as gene correlations to body mass index and gene expression alterations in diabetes. After assigning cells to cell types, they observed that cell types grouped according to donor, yet they were able to identify subpopulations and cellular states after correcting for donor differences. Their computational analyses showed the power of cell type–resolved analyses and revealed cell type–specific gene expression programs, subpopulations, and transcriptional alterations in T2D. scRNA-seq has shown broad use and impact in other disease areas as well, such as Alzheimer's disease (AD). By analyzing single-nucleus RNA-seq (snRNA-seq) data from the prefrontal cortex of 48 individuals with varying degrees of AD pathology, Mathys et al. (2019) identified transcriptionally distinct subpopulations, as well as cell type–specific disease-associated gene expression changes from 80,660 cells. Notably, they found that female cells were overrepresented in disease-associated subpopulations and that transcriptional responses were substantially different between sexes in several cell types. The relatively large number of subjects enabled the investigation of sex effects on AD for the first time.

Single-cell genomics has also impacted many aspects of cancer research. Cancer cell populations are subject to high mutation rates and show high epigenetic plasticity, making tumor cell populations heterogeneous and especially sensitive to selective pressures. Understanding the landscape of genetic and epigenetic heterogeneity, as well as characterizing downstream effects on expression and cell state, will be crucial to better understand tumor initiation and progression. Single-cell genomics has shown utility for understanding both aspects of cancer. To characterize genetic heterogeneity, Navin et al. (2011) conducted the first single-cell DNA-seq (scDNA-seq) study in cancer. With the analysis of hundreds of single cells collected from two breast cancer patients, they identified a genetically diverse subpopulation of cells that do not travel to the metastatic site and revealed a punctuated model of clonal expansion. Since then, many other studies have used scDNA-seq to investigate cancer clonal evolution and have consistently corroborated the genetic plasticity of cancer (Wang et al. 2014; Garvin et al. 2015; Bakker et al. 2016; Kim et al. 2018; Laks et al. 2019; Andor et al. 2020; Velazquez-Villarreal et al. 2020). More recently, novel computational methods have enabled the study of copy number alterations in an allele- and haplotype-specific manner. For example, Zaccaria and Raphael (2021) developed CHISEL, a method for allele-specific copy number analysis relying on external phasing, and applied it to a detailed lineage reconstruction of a breast cancer; Wu et al. (2021a) developed Alleloscope, and through the analysis of multiple types of cancer, they found pervasive haplotype-specific copy number changes seeding minor subclones throughout the course of cancer evolution. Furthermore, Alleloscope allows the detection of haplotype-differentiated subclones in single-cell ATAC-seq (scATAC-seq) data to examine the interplay of genetic and epigenetic evolution. It is the first time that scATAC-seq has been used to study cancer clonal evolution in an allele- and haplotype-specific manner, which enables the dissection of the contributions of chromosomal instability and chromatin remodeling to tumor evolution.

Single-cell approaches have also been applied to better understand tumor initiation and progression through the lens of transcription. Multiple studies have used scRNA-seq to identify tumor progenitor cells and to study their transition toward malignant cell states. For example, Kim et al. (2020) collected 208,506 cells of cancerous and noncancerous lung tissue and used these data to map the trajectory of normal epithelial cells toward malignant cell states in lung adenocarcinoma. Couturier et al. (2020) collected 53,586 adult glioblastoma and 22,637 normal fetal brain cells to map the developmental lineages of glioblastoma cells, which identified glial progenitor–like cells within the tumor that are highly proliferative. Crucially, such analyses can enable the identification of candidate molecular signaling pathways and regulators underlying the transition toward malignant cell states, which may form the basis for therapeutic development. For example, Couturier et al. (2020) identified E2F4 pathway activation in glial progenitor–like cells and showed that inhibition of this pathway more effectively targets these cells than does traditional temozolomide chemotherapy used in glioblastoma.

Beyond characterizing intrinsic cell state changes toward malignant phenotypes in cancer, single-cell approaches have also been used to help better understand nonintrinsic immune factors associated with the malignant tumor microenvironment. In lung adenocarcinoma, Kim et al. (2020) identified exhausted $CD8^+$ T cells and identified macrophages and dendritic cells that express markers associated with immunosuppression, which may both play a crucial role in tumor progression. In kidney carcinoma,

Downloaded from genome.cshlp.org on May 2, 2026 . Published by Cold Spring Harbor Laboratory Press

Zhang et al. (2021) used scRNA-seq data from normal and tumor tissue to identify tumor epithelial expression associated with aberrant myeloid recruitment, and they further used cell–cell communication analyses to characterize mediators of myeloid recruitment.

Single-cell genomics can also be deployed as a powerful diagnostic or prognostic tool in human disease. Particularly, in cancer, tumor heterogeneity may underlie differential survival and response to therapy. This suggests measurements of the tumor cell state distribution from single-cell genomics assays may offer novel insights needed to better diagnose, prognose, and treat cancer. Indeed, Zhang et al. (2021) identified associations between the presence of macrophage subtypes and patient survival in renal cell carcinoma, and they further suggested the fraction of endothelial cells has prognostic value in therapy response. Through integrative analysis of scRNA-seq data at transcriptomic, genotypic, molecular, and phenotypic levels, Wang et al. (2021) identified two subtypes of peritoneal carcinomatosis that were prognostically independent of clinical variables, and they further constructed a 12-gene prognostic signature that was predictive of cancer survival and validated the signature in large-scale gastric adenocarcinoma cohorts.

Existing efforts to study the population-level germline genetic determinants of cell state abnormalities in human disease

As scRNA-seq has become cheaper and more widespread, more groups have shown interest in understanding the role that germline genetic variation plays as a determinant of gene expression. The pioneering work by Wills et al. (2013) illustrated how single-cell analyses can provide mechanistic insights of genetic variants on gene expression variation. Through innovative analysis of 92 genes in the Wnt signaling pathway in 1440 cells from 15 individuals, the investigators showed, for the first time, that many parameters of gene expression, such as expression mean, burst size, burst frequency, and coexpression between cells, are genetically heritable and are masked when examining whole-tissue expression across cells. Later studies by Jiang et al. (2017) and Larsson et al. (2019) further provided evidence of genetically determined bursting kinetics. In particular, through genome-wide analysis of allele-specific bursting kinetics in mouse blastocyst cells and human fibroblast cells, Jiang et al. (2017) showed that a noticeable fraction of genes shows *cis*-dependent burst frequency. Larsson et al. (2019) further showed that burst frequency is primarily encoded in enhancers, whereas burst size is encoded in core promoters. These studies show the power of allelic scRNA-seq for investigating the genetic impact on transcriptional kinetics. One of the main approaches to identify causal factors in human disease is GWAS, and eQTL analysis has been pivotal for the functional interpretation of disease-associated loci. However, as shown by Jiang et al. (2017) , traditional eQTL analysis with bulk RNA-seq misses many associations that are bursting related. Thus, scRNA-seq can be used to identify a more complete set of genetic variants influencing expression and, specifically, can identify GWAS variants with functional effects on bursting parameters.

scRNA-seq has also been used in contexts to study single-cell expression effects of known genetic risk variants. For example, GWAS has identified more than 30 AD genetic risk loci, many of which appear to be related to innate immunity and microglial function, including *APOE* and *TREM2* variants, which are associated with high genetic risks for sporadic AD (Guerreiro et al. 2013; Jonsson et al. 2013; Lambert et al. 2013; Efthymiou and Goate 2017; Neu et al. 2017; Kunkle et al. 2019; Bellenguez et al. 2021; Schwartzentruber et al. 2021). The *TREM2* R47H variant is associated with an approximately threefold increased risk for AD, whereas the *APOE* E4 variant is associated with an approximately three- to fourfold increased risk with one copy and an approximately 10- to 12-fold increased risk with two copies. How genetic risk factors, like *APOE* and *TREM2*, intersect with cellular responses to AD pathology in human tissues is not understood. Using snRNA-seq of 131,239 nuclei obtained from 15 postmortem human brains with varied *APOE* and *TREM2* genotypes and neuropathology, Nguyen et al. (2020a) identified distinct microglia subpopulations, including a subpopulation of CD163-positive amyloid-responsive microglia that are depleted in AD cases with *APOE* and *TREM2* risk variants. These results were validated in an expanded cohort of AD cases, showing that *APOE* and *TREM2* risk variants are associated with a significant reduction in CD163-postive amyloid-responsive microglia. This study showcased how genetic information, when integrated with single-cell transcriptomics, can advance our understanding of how genetic risk factors influence cellular responses to underlying pathologies.

Other studies have taken genome-wide approaches to identify the genetic determinants of disease-associated expression via single-cell eQTL studies. Sarkar et al. (2019) generated scRNA-seq data from induced pluripotent stem cells derived from 53 Yoruba individuals and investigated how genetic variants control gene expression variations both at the mean and the variance levels. Their analyses suggest that although the variance of gene expression is genetically controlled, the corresponding QTLs explain less phenotypic variance than eQTLs that control the mean expression. Although Wills et al. (2013) examined the relationship between coexpression and genetic variants, their study was limited by the small number of individuals and genes. Recently, van der Wijst et al. (2020) performed a similar study, but with 45 individuals and approximately 25,000 peripheral blood mononuclear cells. Through the construction of personalized coexpression networks, they identified genetic variants that significantly impact the coexpression of genes, implying that gene regulatory networks (GRNs) may vary across individuals. Because hundreds of genetic variants located in a few key regulatory pathways can contribute to complex diseases (Westra et al. 2013; Fagny et al. 2017), constructing personalized cell type–specific GRNs is a crucial step toward the understanding of genetic contributions to complex diseases. The recently formed Single-Cell eQTLGen Consortium will conduct GRN-based QTL analysis to examine genetic differences that change the architecture of the networks. Findings from such an analysis will enhance our basic understanding about the genetic contributions in gene expression and its regulation.

Efforts to detail the contribution of germline genetics to cell dynamics have also been made. Cuomo et al. (2020) studied the genetic determinants of iPS endoderm differentiation efficiency from 36,044 cells collected from 125 patient samples. More recently, in one of the largest scRNA-seq studies of humans to date, Jerber et al. (2021) studied the genetic determinants of iPS dopaminergic neuron differentiation from over 1 million cells collected from 215 human samples. Expanded efforts to study the genetic influence over other dynamic processes, such as differentiation, cell cycle, and circadian cycle, will greatly enhance our understanding of the context in which genetic variants exert their influence in disease.

Downloaded from genome.cshlp.org on May 2, 2026 . Published by Cold Spring Harbor Laboratory Press

Study design considerations for population-based single-cell studies

Although current studies have shown the power of single-cell technologies, these studies have been limited by the number of subjects. Nguyen et al. (2020a) were able to study the impact of *APOE* and *TREM2* with a limited number of subjects owing to the use of a genetic risk variant–enriched study design. Their success in identifying risk variant–dependent microglia subpopulations underscored the importance of study design. As the field is now moving into large-scale population-based single-cell studies, it becomes even more important to consider study design–related issues. Given a fixed budget, a key question to ask is how to allocate the limited budget while maximizing the information gain. Parameters that need to be considered include the number of subjects, the number of cells per subject, and the sequencing depth per cell. Determination of these parameters will depend on the goals of the study and in the selection of study subjects.

Sample selection

When the goal is to investigate the interaction between known genetic risk factors and cellular responses to disease, an appealing design is the genetic risk factor–enriched design as was performed by Nguyen et al. (2020a). When genetic information is available, selecting genetic risk factor–enriched individuals can substantially reduce the number of needed subjects (Fig. 1A). Although DNA genotyping needs to be performed for a large number of individuals when studying rare variants, the cost of DNA genotyping is much lower than that of scRNA-seq. When genetic risk factors are unknown, an alternative design is the extreme phenotype sampling design, which selects individuals that cover both extreme ends of a disease spectrum (Fig. 1B). There is a well-established inverse relationship between the allelic frequency of a given variant and its effect size on the phenotype (Lander and Botstein 1989; Peloso et al. 2016), and many studies have shown that extreme phenotypes tend to occur in extreme cases with an excess of rare variants. The extreme phenotype sampling design offers a cost-effective strategy for studying the interaction between rare variants and cellular responses.

Number of cells and sequencing depth per cell

After study subjects are determined, the next consideration is how many cells to sequence and the sequencing depth per cell. Shall we sequence a large number of cells with shallow sequencing depth per cell or deeply sequence a few cells for each subject? Although common cell types can be detected and their gene expression levels reliably measured with a relatively small number of cells (Heimberg et al. 2016; Zhang et al. 2020), to reliably detect rare cell types, a larger number of cells is needed. Thus, the number of cells per subject is largely determined by the frequency of the rarest cell type of interest. A number of software packages have been developed to estimate the number of cells that must be sampled in a single-cell sequencing experiment. For example, based on the user-specified frequency of the rarest cell population and the number of populations with approximately this frequency, SCOPIT (Davis et al. 2019) can estimate the number of cells for planning single-cell sequencing experiments. Schmid et al. (2020) developed scPower, a more general framework for single-cell power calculation, in which they showed that, for a fixed budget, the number of cells per individual is the major determinant of power of detecting rare cell types and differentially expressed genes, followed by the number of subjects and read depth. In general, shallow sequencing of high numbers of cells per individual leads to a higher overall power than does deep sequencing of fewer cells.

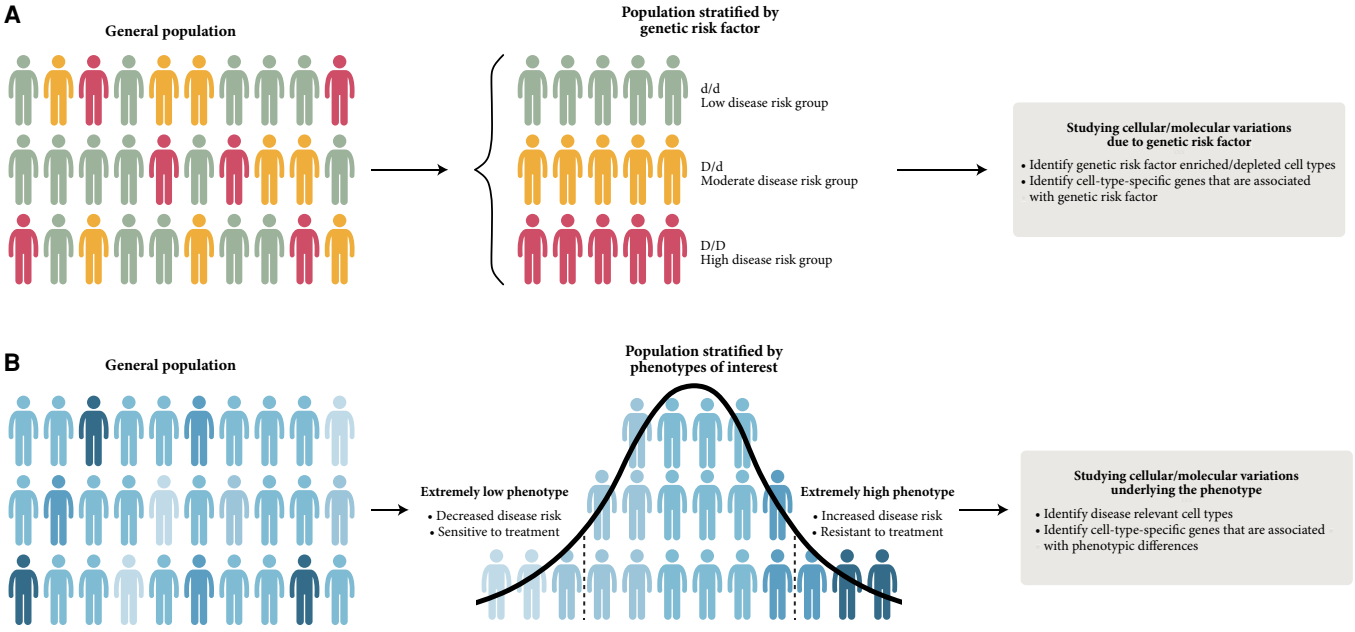

Figure 1. Sample selection strategy for population-based single-cell studies. (*A*) Genetic risk variant–enriched design in which individuals with the genetic risk variant are oversampled in order to achieve enough number of individuals that carry the genetic risk variant. (*B*) Extreme phenotype sampling design in which individuals with extremely low or extremely high phenotypes are selected. These extreme phenotype individuals are expected to carry more rare genetic risk variants than are individuals with intermediate phenotypes.

Downloaded from genome.cshlp.org on May 2, 2026 . Published by Cold Spring Harbor Laboratory Press

Sample collection design to mitigate batch effects

Like many high-throughput technologies (Leek et al. 2010), single-cell methods are susceptible to batch effects, which refer to systematic differences among samples processed in different batches (Hicks et al. 2018). Although batch effects can be minimized by a completely randomized experimental design (Bacher and Kendziorski 2016), such designs are often infeasible for studies that involve human tissues because practical considerations require tissue samples to be processed immediately to avoid tissue degradation. Furthermore, for studies that involve a large number of subjects, patients are recruited sequentially, and single-cell experiments may span several days, months, or years apart, introducing systematic nonbiological differences that can confound biological variations. Recently, Song et al. (2020) proposed two experimental designs, the reference panel and the chain-type designs, that can reduce the impact of batch effects from the study design stage. Under the reference panel design, one batch is required to include cells from all cell types to serve as the reference panel, whereas the other batches need to have at least two cell types. The requirement of a reference batch that includes all cell types makes it difficult to achieve in practice. An alternative and more practical design is the chain-type design, which requires two cell types to be shared between every two consecutive batches. A special form of this design is when two cell types are shared among all batches, a situation that is easy to meet in real studies. Song et al. (2020) mathematically proved that under these two experimental designs, true biological variability can be separated from batch effects.

Cost reduction by cell type deconvolution analysis in bulk RNA-seq

Although the cost of scRNA-seq has reduced in recent years, using scRNA-seq for all study subjects in a large-scale population-based study might still be cost prohibitive. Integrative analysis of scRNA-seq and bulk RNA-seq data offers an alternative approach that can substantially reduce the cost while returning cell type–specific gene expression information. Such integrative analysis relies on cell type deconvolution, which aims to infer cell type proportions from bulk transcriptomics data. Many methods have been developed that use scRNA-seq data to infer cell type proportions in bulk RNA-seq samples in the last few years (Newman et al. 2015, 2019; Du et al. 2019; Wang et al. 2019; Jew et al. 2020; Dong et al. 2021). The estimated cell type proportions can be treated as known, and further analyses that incorporate these proportions as covariates can infer cell type–specific gene expression in each subject, as is performed in CIBERSORTx (Newman et al. 2019); detect allelic expression imbalance, as is performed in BSCET (Fan et al. 2021); or detect cell type–interacting QTLs (Donovan et al. 2020; Kim-Hellmuth et al. 2020) or cell composition QTLs (Park et al. 2021). The estimated cell type proportions can also be used to compare cell type compositions between diseased cases and controls. Determining whether certain cell types are increased or decreased in proportion in a disease state is informative for understanding disease pathophysiology. For example, such analyses have detected the loss of beta cells in T2D (Wang et al. 2019; Dong et al. 2021), an increase of disease-associated microglia in AD (Buttner et al. 2020), and an increase of microglia in advanced age-related macular degeneration (Lyu et al. 2021).

Computational analysis and considerations for population-based single-cell studies

In this section, we describe analytical strategies of population-based single-cell studies. An overview of single-cell analysis workflow is shown in Figure 2. Analysis of single-cell data starts from data preprocessing and normalization. Imputation may also be performed when needed. As other papers have thoroughly reviewed these aspects (Bacher and Kendziorski 2016; Hie et al. 2020; Hou et al. 2020; Lytal et al. 2020; Wu and Zhang 2020; Zhang and Zhang 2020; Ahlmann-Eltze and Huber 2021; Melsted et al. 2021; Slovin et al. 2021), we will focus our discussion on the downstream statistical analyses.

Correction of batch effects

Large-scale single-cell data sets with many subjects contain batch-specific systematic variations that present a challenge to data analysis. Batch effects are inevitable in analyses of human tissue and are prevalent in many single-cell studies (Hicks et al. 2018; Lähnemann et al. 2020). Failure to remove batch effects can not only generate false-positive signals but also obscure true biological variations. As such, many methods have been developed to remove batch effects in single-cell data analysis. Batch effect correction can be performed either in the original high-dimensional gene expression space or the low-dimensional embedding space, for example, gene expression data projected down onto principal components from principal component analysis. Batch effect correction methods such as LIGER (Welch et al. 2019), Conos (Barkas et al. 2019), Harmony (Korsunsky et al. 2019), BBKNN (Polanski et al. 2020), and DESC (Li et al. 2020) remove batch effects only for the embedding space. Although useful for profiling the overall characteristics of cells such as clustering and trajectory reconstruction, these methods cannot be used for downstream gene-level analysis like differential expression (DE) and coexpression.

To be useful for gene-level analysis, batch effects need to be removed in the original high-dimensional gene expression space. However, this task is much more challenging than batch effect correction in the embedding space (Lucken et al. 2020). Popular methods such as Seurat 3.0 (Stuart et al. 2019) rely on the mutual nearest-neighbor (MNN) approach (Haghverdi et al. 2018) to remove batch effects for each gene, but MNN can only analyze two batches at a time. Its performance is affected by the order in which batches are corrected, and it quickly becomes computationally infeasible when the number of batches gets large. Scanorama (Hie et al. 2019) overcomes the computational issue of MNN by finding matching elements among all batches at once, which also makes it invariant to batch order. A more desirable approach, however, should remove batch effects in gene expression for all batches simultaneously. A few neural network–based methods have been developed for this purpose. For example, scVI (Lopez et al. 2018) removes batch effects by conditioning on batch information in a variational autoencoder, which learns a nonlinear embedding of cells; SAVERCAT (Huang et al. 2020) uses a conditional variational autoencoder to remove batch effects through explicit modeling of batch information as covariates; and CarDEC (Lakkis et al. 2021) uses a joint autoencoder together with iterative clustering to remove batch effects. Beyond the ability to model all batches simultaneously, an additional advantage of these neural network–based methods is their flexibility in achieving multiple tasks within the same framework. These approaches can not only remove batch effects in the original high-dimensional gene

Downloaded from genome.cshlp.org on May 2, 2026 . Published by Cold Spring Harbor Laboratory Press

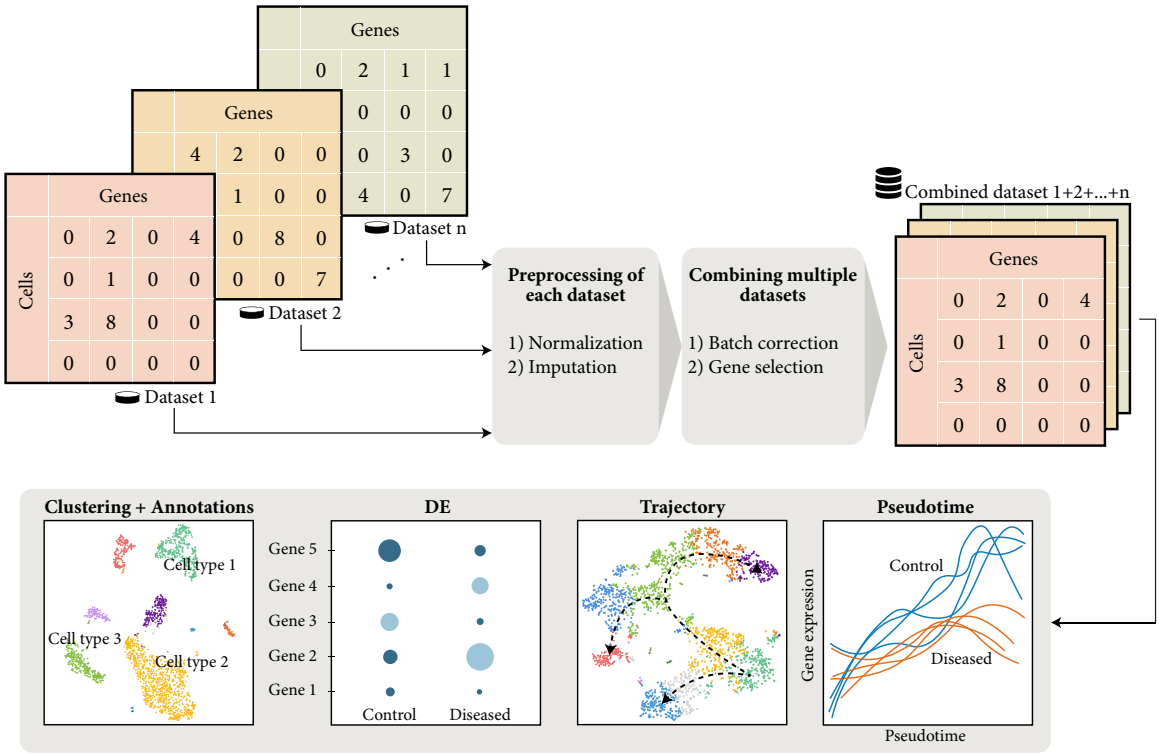

Figure 2. Overview of single-cell data analysis workflow. The typical workflow involves data preprocessing, combination of multiple single-cell data sets into a combined data set, clustering and cell type annotation, differential expression analysis, trajectory inference, and pseudotime analysis.

expression space but also remove batch effects in the low-dimensional embedding space to facilitate cell clustering. Moreover, these methods can impute gene expression, which may be desirable for downstream gene-level analyses.

Annotation of cell identities

Consistent annotation of cell identities is also a critical step in population-based single-cell studies. As such studies involve a large number of individuals, the data generation may span multiple years and across multiple laboratories. For such studies, it becomes infeasible to use unsupervised clustering algorithms as these algorithms require the reanalysis of all cells whenever new data become available. Moreover, unsupervised clustering algorithms may have difficulty resolving cell subtypes whose differences are biologically meaningful (Kiselev et al. 2019). One attractive approach to circumvent these issues is to rely on available, well-annotated single-cell data sets, such as those contained in Azimuth (Hao et al. 2021). Using these high-quality reference data, methods have been developed to identify and annotate cell types in new data. For example, scmap (Kiselev et al. 2018) projects cells in a query data set to a space determined by highly informative genes selected from a well-labeled labeled data set and then assigns cell identities for cells in the query data based on their correlation with average cell type–specific gene expression in the reference data. scANVI (Xu et al. 2021), a semisupervised variant of scVI (Lopez et al. 2018), annotates cell types in a query data set by leveraging any available cell annotations in a reference. Seurat 3.0 classifies cells in the query data by finding anchor cell pairs between a well-labeled reference and the unlabeled query data sets. Scmap learns cell type–specific gene expression information only in the reference but ignores useful information in the query data; thus, it is vulnerable to batch effects and platform differences between the reference and query data. Although Seurat 3.0 uses information both in the reference and the query data in the identification of anchor pairs, it does not specifically use cell type label information in the reference.

An ideal approach for cell identity annotation should be able to use cell type–specific gene expression information both in the reference and the unlabeled query data. Although reference data sets have become increasingly comprehensive, cell types/subtypes may exist in the unlabeled query data that were not previously detected in the reference, for example, owing to differences between the query and reference data in cell sample size or to differences in subject-specific covariates, etc. As such, approaches should carefully balance the contribution of each data type in cell type assignment annotation. As large single-cell references are continuously generating well-annotated reference data across various tissues, an ideal approach should also be able to combine multiple references together so that the users can learn from these comprehensive maps when annotating their own data. To address these issues, transfer learning–based approaches have been developed. For example, ItClust (Hu et al. 2020a) borrows ideas from supervised cell type classification algorithms but also leverages information in target data to ensure sensitivity in classifying cells that are only present in the target data through the use of an iterative transfer learning approach with neural networks. scArches (Lotfollahi et al. 2021) relaxed the requirement of having raw data from the reference. Through reusing neural network models by adding input nodes and weights and then fine-tuning those, it learns the joint latent representations of the reference and the query data, which allows the identification of rare cell states in the query

Downloaded from genome.cshlp.org on May 2, 2026 . Published by Cold Spring Harbor Laboratory Press

data that is otherwise difficult to detect. As the scale of single-cell studies continues to grow, we anticipate these transfer learning–based approaches will automate the labor-intensive clustering and annotation tasks and facilitate comparative analyses across tissues and disease conditions.

DE analysis

After cell identities are annotated, an important next step is to identify genes that are differentially expressed between conditions, for example, healthy versus diseased, within the same cell subpopulation. Although methods have been developed for DE analysis in scRNA-seq (Kharchenko et al. 2014; Finak et al. 2015; Korthauer et al. 2016; Jia et al. 2017), these methods ignore the effect of subject-specific covariates. Although subject-to-subject variation may have little impact on the identification of subpopulation-specific marker genes, their impact on DE analysis between different conditions within the same cell subpopulation is unknown. Through simulations, Crowell et al. (2020) investigated the performance of various methods in detecting DE genes in this situation. Interestingly, they found that the simple "pseudobulk" approach outperforms methods that are specifically designed for scRNA-seq. In such "pseudobulk" analysis, cell-level counts from a subpopulation are aggregated into a single observation per subject, which are then used as input for DE analysis using traditional bulk RNA-seq methods such as edgeR (Robinson et al. 2010), DESeq2 (Love et al. 2014), or limma-voom (Law et al. 2014). These aggregation-based DE methods not only are fast but also show a stable high performance across various scenarios, making them an appealing choice for large-scale scRNA-seq studies that involve many subjects. Notably, cell-level mixed models performed comparably to pseudobulk approaches in DE analysis, as the DE gene sets identified were similar. However, cell-level mixed models severely underestimated the expression differences of certain genes between different cell populations. For these genes, this is likely owing to the abundance of cells with zero counts, for which the gene's maximum likelihood estimate of the mean will be equal to zero for that cell. This becomes more likely for lowly expressed genes under sparse data settings, underscoring the need to model expression uncertainty. Perhaps future cell-level approaches can improve upon this issue by modeling gene expression uncertainty directly.

Differential splicing analysis

Previous studies have shown that genes showing changes in alternative splicing may reflect different biological processes from those with DE. For example, a recent scRNA-seq study in the adult mouse cortex found differences in splicing dynamics across cells were not explained by neuronal cell type definitions based on differences in isoform-agnostic transcript expression levels, suggesting that alternative splicing regulation might be orthogonal to transcriptional regulation in specifying neuronal identity and function (Feng et al. 2021). Therefore, differential alternative splicing may complement DE analysis in characterizing gene regulation. However, low sequencing depth, technical noise, and the lack of appropriate computational methods have precluded the investigation of splicing heterogeneity in most scRNA-seq studies. To date, only a few methods have been designed specifically for splicing analysis in scRNA-seq. Huang and Sanguinetti (2017) detected differential exon-usage by performing a pairwise comparison between every two cells. Song et al. (2017) quantified exon-inclusion levels based on junction-spanning reads. Qiu et al. (2017a) and Ntranos et al. (2019) detected differential transcript usage based on pre-estimated cell-specific isoform expressions or transcript compatibility counts. Hu et al. (2020b) detected differential alternative splicing by accounting for technical noise and low sequencing depth through grouping exons that originate from the same isoform(s). Although these methods have shown promising performance, they still have limited power for data without full-length transcript sequencing. Most single-cell studies use droplet-based technologies, for example, 10x Genomics (https://www.10xgenomics.com) or Drop-seq (Macosko et al. 2015), which have inherent limitations for splicing analysis owing to their sequencing of only the 3′ or 5′ end of the gene following fragmentation. Although Smart-seq2 (Picelli et al. 2014) can generate full-length transcripts, the lack of unique molecular identifiers (UMIs) makes it difficult to remove amplification bias. To fully characterize the splicing complexity of single cells, technologies with full-length transcriptome coverage and UMIs, such as ScISOr-Seq (Gupta et al. 2018; Joglekar et al. 2021) and SMART-seq3 (Hagemann-Jensen et al. 2020), are needed.

Trajectory analysis

A substantial portion of cell state variation can be explained by treating states as discrete; differences in our notion of "cell type" underlie large differences in cell morphology, function, and molecular composition across cells. Although treating cell states as discrete may be appropriate in many settings, cell state variation is best described as a continuum. Cells undergo gradual changes during cellular differentiation, as they transition from one cell type to another. Further, cell states can follow a continuum within a given cell type: Cell states are perturbed by both constant factors, such as the circadian clock, as well as asynchronous factors, such as the cell cycle. Characterizing continuous aspects of cell state and understanding the dynamics that give rise to them will be crucial to understand how cells function and how these functions can go awry in human disease.

Single-cell technologies provide a powerful tool to study continuous cell state variation. In particular, scRNA-seq has seen widespread use to characterize state differences owing to cell differentiation in both human developmental and adult-life contexts, for example, smooth muscle cell phenotypic switching during atherosclerosis (Wirka et al. 2019; Pan et al. 2020), subtype switching of macrophages during pathological cardiac hypertrophy (Ren et al. 2020), the transition of myeloid cells during the progression and regression of kidney disease (Conway et al. 2020), the transition from homeostatic microglia to amyloid-responsive microglia or motile microglia during AD progression (Nguyen et al. 2020a), and iPS-based models of cell type maturation (Cuomo et al. 2020; Jerber et al. 2021).

A key step in such analyses is the computational assignment of cells to continuous states, often referred to as trajectory or pseudotime assignment. For example, in the simple case of cells differentiating from one cell type into another, cells could be assigned a continuous value from zero to one, where zero indicates the starting cell state, one indicates the final cell state, and intermediate values indicate intermediate states. After the assignment of cells to continuous states, researchers can characterize molecular changes, such as mRNA expression associated with changes in cell state, and generate candidate mechanisms underlying these changes, for example, changes in transcription factor (TF) activity.

The growth of single-cell technologies has been accompanied by the development of several computational tools for trajectory inference (Saelens et al. 2019). The choice and use of such tools

Downloaded from genome.cshlp.org on May 2, 2026 . Published by Cold Spring Harbor Laboratory Press

designed for scRNA-seq require careful consideration. Before trajectory inference, the high-dimensional gene expression data may be transformed into a dimensionality-reduced representation. In the context of large-scale human disease studies, compressed representations can confer benefits as a tool for noise reduction when cells are shallowly sequenced and for improvement of the computational efficiency of trajectory assignment methods. Using this input, trajectory inference can be performed. A crucial consideration in selecting a trajectory method is the user's expectation of the underlying trajectory topology. For settings in which the user has no expectation of the trajectory topology, flexible tools have been developed that can detect a wide range of topologies, such as linear, circular, trees, and disconnected components (Ji and Ji 2016; Qiu et al. 2017b; Street et al. 2018; Wolf et al. 2019).

One consideration in the use of trajectory reconstruction methods is the relatively high degree of uncertainty of the trajectory shape and cell ordering (Saelens et al. 2019). This consideration becomes more crucial when the underlying trajectory has not been adequately sampled, that is, too few cells, which may produce unstable results owing to the similar likelihood of multiple topological hypotheses. Continued progress in trajectory inference methods to incorporate RNA velocity information (Lange et al. 2020) and the quantification of trajectory uncertainty (Lin et al. 2021) may aid in resolving such ambiguities and in interpreting results, respectively. Given the challenges associated with flexible trajectory models, for cases in which users have expectations of the trajectory topology, it is recommended they use methods with inductive biases that reflect this expectation. For example, variation across cells owing to cell cycle variation should be modeled by methods designed to detect circular topologies, such as reCAT (Liu et al. 2017).

Outlook and future research

Thus far, single-cell technologies have seen use in characterizing cell state differences among diseased and healthy individuals. Incorporating genetic information, groups have now begun to identify variants influencing cell states. Nonetheless, although single-cell technologies have rapidly advanced our ability to survey multiple molecular modalities describing cellular behavior, we remain far from the ability to predict how molecular and/or behavioral perturbations will influence high-level physiological features to improve human health. We believe the following four areas will see great strides toward this goal in the near future.

Modeling the effect of genotypic variation on transcriptional regulation

The development of precision medicine therapies will benefit from predictive models to interpret how genetic variants influence gene expression. At present, eQTL studies have largely modeled variants as having additive, linear effects on the expression of individual genes. In the presence of small sample sizes, this is a reasonable approach. However, as regulatory element interactions influence transcription, for example, enhancer–promoter interactions (Schoenfelder and Fraser 2019; Fitz et al. 2020) and enhancer cooperativity (Huang et al. 2018), models that consider regulatory element variants to contribute independently to changes in transcription kinetics are likely misspecified. Moreover, the lack of variance explained by eQTL models assuming additive linear effects (Price et al. 2011; Lloyd-Jones et al. 2017) suggests substantial model improvements are required not only to identify variants with effects on gene expression but also to faithfully capture how they affect gene expression.

Moving beyond purely additive linear models, convolutional neural networks (CNNs) appear to be a promising approach toward modeling the role of genomic variants in *cis* regulatory logic. In particular, CNNs have already shown great promise in modeling the contribution of promoter genetic variation on mean gene expression levels. Agarwal and Shendure (2020) first introduced their algorithm, Xpresso, a CNN designed to predict steady-state mean expression levels using sequence features of gene promoters and gene bodies. Motivated by the high correlation of gene expression across cell types, Xpresso first demonstrated an ability to detect sequence features describing expression variation in a cell type–agnostic fashion. This suggests that rules exist that generalize across cell type–specific contexts, and indeed, inspection of the model identified a number of genomic features associated with steady state expression including ORF exon density, 5′ UTR GC content, and promoter CpG content. Nonetheless, promoter-based models cannot explain all genetically determined expression variation. Notably when Xpresso was trained on cell type–specific expression with accompanying chromatin accessibility data, genes with the largest prediction residuals were those adjacent to stretch enhancers. This suggests that an ideal model of *cis* regulatory expression likely will require the consideration of multiple layers of regulatory control, such as the role of enhancer sequences, 3D genome configuration, and chromatin accessibility. In a recent preprint, Avsec et al. (2021) introduced a novel model architecture, dubbed Enformer, which jointly considers distal and proximal regulatory sequences in gene expression prediction. When applied to bulk human expression data from GTEx, Enformer shows a substantial improvement in our ability to predict expression from sequences. Moreover, the investigators point toward the use of in silico perturbations of the model to yield candidate *trans* regulators of distal regulatory activity. Although promising, future work remains to incorporate other layers of transcriptional control into genetically determined models of gene expression. Innovation in computational method development will be essential for the advancement in our understanding of the transcriptional regulation effects of variants in the context of human disease.

The development of models to interpret the transcriptional regulation effects of human variants will also benefit from continued development in experimental assays. The largest existing single-cell eQTL studies have assayed hundreds of individuals (Cuomo et al. 2020; Jerber et al. 2021). Although an achievement, this is a limited sample size relative to the space of regulatory variation observed in humans. The development of high-throughput base editor mutagenesis technologies holds great promise to probe the role of genetic variation. Hanna et al. (2021) recently introduced a cytosine base editor to study the effect of 52,034 ClinVar variants in 3584 genes. Future efforts to pair base editor mutagenesis with scRNA-seq will greatly advance our ability to explore the space of regulatory variation from human cells at scale.

Construction of GRNs

Evidence suggests *cis* regulatory variation only modestly explains gene expression variation (Liu et al. 2019). Although this may partially reflect misspecified models of how *cis* regulatory variants affect expression, it also points to the need to model the role of *trans* effects. Gene expression is regulated by the interaction of *cis* regulatory elements with TFs. The activity of TFs, themselves, depends

Downloaded from genome.cshlp.org on May 2, 2026 . Published by Cold Spring Harbor Laboratory Press

on their expression, which has its own regulatory logic. As such, faithfully modeling the role of *trans* effects on gene expression will require mapping cell type–specific GRNs that detail the gene targets of TFs. Mapping GRNs will enable researchers to better understand the underlying drivers of expression differences between cell states, such as differences in underlying TF levels. Further, they can inform predictions of how gene expression will change upon perturbations of TFs or upstream signaling pathways.

It has become increasingly common to estimate GRNs from steady-state scRNA-seq data, and several computational tools have been developed for this task. Although there are key nuances that distinguish each method, these tools generally construct GRNs by identifying gene pairs showing coexpression patterns within a given data set. GRNs are then represented as an undirected graph in which nodes represent genes and binary edges represent the presence or absence of relationships. To date, the application of GRN detection methods to scRNA-seq data has yielded results of mixed success. Using simulated scRNA-seq generated from ground truth GRNs, Nguyen et al. (2020b) recently showed that existing tools detect GRNs with success slightly better than random. This may reflect, in part, the inherent limitations of using scRNA-seq to detect GRNs. However, the general principles used by the best-performing GRN tools should form the basis for future computational developments. Notably, one of the earliest and most popular tools, SCENIC (Aibar et al. 2017), constructed networks most accurately across a variety of benchmarks. This is likely owing to SCENIC's inductive bias that predicted coregulated genes share motifs for an underlying TF, suggesting that methods incorporating domain knowledge may be better suited to construct GRNs using scRNA-seq. Nguyen et al. (2020b) also point out that existing tools assume GRN relationships are linear and that there are no interactions. This assumption may limit the power of GRN detection tools, as TF–TF interactions are known to significantly shape gene expression (Zeitlinger 2020). The interpretation of GRNs detected by existing tools is also challenging, as edges in GRN graphs are often undirected and may not represent functional relationships but, instead, correlations. Future approaches incorporating RNA velocity may help resolve the direction of GRN relationships from scRNA-seq.

Although GRN detection from steady-state scRNA-seq data has proved challenging, a promising alternative is the use of perturbation approaches paired with single-cell omics to map GRNs. Crucially, these approaches are high throughput in nature, allowing researchers to identify the regulatory targets of hundreds of TFs from a single tissue sample. Perturb-Seq (Dixit et al. 2016) first introduced the ability to generate a loss-of-function library of CRISPR guide RNAs to transfect a cell population and whose effects could be read out using single-cell transcriptomics. Using this technique, Dixit et al. (2016) were able to identify TF–gene regulatory relationships that were recapitulated using ChIP-seq. Depending on the loading concentration of guide RNAs, Perturb-Seq is also amendable to probing the transcriptional effects of higher order combinations of TF knockouts. A complementary approach to mapping GRNs is detailing how chromatin accessibility is perturbed by TF knockouts. Rubin et al. (2019) introduced Perturb-ATAC, which uses a loss-of-function library of CRISPR guide RNAs to assay their effects on single-cell chromatin accessibility. Using this approach, researchers may be able to preferentially identify TFs responsible for binding heterochromatin and promoting chromatin accessibility in particular cellular contexts. As such, Perturb-ATAC may be of particular relevance to help researchers identify pioneer factors that act as hubs in GRNs.

Integrative analysis of multiple molecular modalities and their correspondence with cell state

Although scRNA-seq has been predominantly used to characterize cell state differences between diseased and nondiseased individuals thus far, the emergence of single-cell multiomic technologies, wherein multiple molecular modalities are simultaneously profiled within the same cell, signifies an important next step in the study of human disease using single-cell approaches. Stoeckius et al. (2017) first introduced CITE-seq, an approach to jointly profile proteins and RNA in single cells. Since then, technological developments have made it possible to jointly profile the transcriptome in single cells with chromatin accessibility (Cao et al. 2018; Chen et al. 2019, Ma et al. 2020), DNA methylation (Gaiti et al. 2019; Luo et al. 2019), nucleosome occupancy (Pott 2017; Clark et al. 2018), chromatin occupancy (Xiong et al. 2021), or spatial location (Rodriques et al. 2019; Vickovic et al. 2019). Encouragingly, recent efforts show a trend toward increased detection sensitivity and cost reduction.

Single-cell multiomics will enable researchers to measure cell state on a more granular level, as different modalities may contain independent cell state information. Indeed, Hao et al. (2021) found protein information could segregate known T cell subtypes where mRNA could not, suggesting not only that multiomics can measure more granular aspects of cell state but also that these differences coincide with known aspects of biology that distinguish cell subtypes. Beyond independent information captured by individual modalities, multiomic data will also enable more meaningful measures of cell state via modeling of interactions between modalities that are known to modulate cell state, such as TF abundance and chromatin accessibility. Ultimately, the more granular cell state information provided by multiomic data will help researchers better distinguish between diseased and healthy cell states.

The development of tools to estimate cell state from single-cell multiomic data will be essential to maximize its utility. In brief, most tools estimate latent factors that maximize the joint probability of the observed data. Using these tools, researchers can study differences in cell states associated with disease and health. One of the earliest tools, LIGER (Welch et al. 2019), deploys an integrative nonnegative matrix factorization approach to estimate latent factors describing cell state. More recently, Argelaguet et al. (2020) introduced MOFA+, a Bayesian matrix factorization approach that uses priors to encourage the learning of sparse latent factors and loading matrices to improve their interpretability. Moving beyond linear approaches, Wu et al. (2021b) introduced BABEL, a nonlinear joint autoencoder approach. Although existing approaches to estimating cell state have made meaningful contributions to the analysis of multiomic data, they suffer from two main issues. First, models that aim to purely maximize the probability of the data are more likely to learn spurious statistical associations under sparse data settings and are less equipped to generalize to unseen data from new cell states. Indeed, Wu et al. (2021b) highlight their method's difficulty in generalizing to unseen cell states. Second, latent factors may be uninterpretable, making the identification of testable hypotheses for experimental follow-up challenging. In both respects, we believe future cell state estimation tools would benefit from using latent variable models based on underlying explanatory factors that reflect known biology. For example, the protein abundance of TFs is known to partially determine both a cell's chromatin accessibility and transcriptomic states; as such, multiomic chromatin accessibility and

Downloaded from genome.cshlp.org on May 2, 2026 . Published by Cold Spring Harbor Laboratory Press

transcriptomic data could be meaningfully described by a latent variable model wherein latent factors encode TF abundance. Evidence suggests such models are better equipped to deal with sparse data and generalize to unseen data and are more robust to learning spurious statistical associations (Bengio et al. 2013). Moreover, these more interpretable approaches can help identify testable hypotheses, such as the knockdown of a TF to perturb cells from a diseased to a healthy state.

Understanding how cells cooperate to give rise to tissue-level phenotypes

As we move closer toward understanding how processes are regulated within the cell and predict how molecular perturbations can direct changes in individual cell states, it is equally important to understand how changes in individual cells will contribute to changes in tissue-level and organismal-level physiology. For example, one goal in the treatment of atherosclerosis is the development of therapies to promote plaque stability. Atherosclerotic plaques are composed of a multitude of cell types, such as fibrochondrocytes, macrophages, smooth muscle cells, and lymphocytes (Wirka et al. 2019; Alencar et al. 2020; Pan et al. 2020). Although a great deal of work has detailed factors associated with plaque stability such as the role of inflammation, no working model exists of how cell types and their interactions relate to plaque stability. Such a working model would be valuable for identifying candidate molecular perturbations in specific cells to promote plaque stability. More broadly, in the future it may be fruitful to understand not just how individual cell states are perturbed in disease but also how these dysregulated cells jointly contribute to disrupted tissue-level physiology.

The advent of single-cell spatial transcriptomics (and other spatial omics methods) appears to be a promising experimental assay that will help researchers approach this task. In brief, sequencing-based spatial omic technologies deploy surfaces that are arrayed with barcodes corresponding to cellular position. After tissue permeabilization and sequencing, individual cellular locations can be ascertained based on the identity of the cell barcode. Using these data, researchers can build models predicting how cells interact to produce tissue-level physiological features. An ideal model for this task should take into account the spatial location of the cells, and interactions among cells should be a function of cell–cell proximity; namely, adjacent pairs of cells should be more likely to interact than distant pairs of cells (Hu et al. 2021). A natural choice to consider is the use of graph convolutional neural networks (GCNNs) in either regression-based or classification-based settings (Hu et al. 2020c). Using existing approaches to map cells' gene expression to cell states, individual cell states can be represented as nodes on a graph, where edges between nodes indicate that two cells are physically adjacent to one another. GCNNs can then use this graph as an input to predict either continuous or discrete aspects of the tissue of interest. At present, single-cell spatial transcriptomics may not yet be practical to do in large-scale human studies. Nonetheless, single-cell spatial transcriptomics combined with cost-effective histology may be a practical alternative to generate hypotheses of molecular perturbations to improve tissue-level measures. Using patient samples with matching spatial transcriptomics and histology, generative models such as those using graph convolutions can be trained to learn the joint relationship between the histology, spatial transcriptomics, and tissue-level information such as the stability of an atherosclerotic plaque. Given larger data of subjects with only collected histology and tissue-level measures, the generative model can first be used to predict the expression of individual cells for each sample. Using this predicted expression and the generative models, users can perform in silico perturbations of gene expression that produce improvements in tissue-level measures for each subject. Perturbations predicted to improve tissue-level measures that are shared across subjects, or subgroups of subjects, should then be prioritized for experimental follow-up.

Conclusion

Single-cell technologies have proven to be a valuable tool to understand human disease. Single-cell resolution enables researchers to characterize differences in cell states associated with disease status. This can be a powerful approach toward building an understanding of disease pathogenesis and its effects. At present, scRNA-seq has accounted for a substantial body of single-cell data collected. Using these data and the substantial body of supporting computational tools for their analysis, many groups have effectively detailed cell state differences underlying differences in human disease status. As this technology has grown more widespread, efforts to understand the genetic underpinnings of cell state differences have begun and continue to grow. The future of single-cell technologies in studying human disease appears promising, as new single-cell technologies to capture additional modalities such as chromatin accessibility, proteins, and spatial location have matured (Moffitt et al. 2018; Chen et al. 2019; Eng et al. 2019; Zhu et al. 2019; Ma et al. 2020; Specht et al. 2021; Takei et al. 2021; Thornton et al. 2021) and will enable researchers to detail factors underlying cell state differences not described by mRNA alone. Moreover, these technologies may help researchers further our understanding of the interaction between these factors in cell regulation. Maximizing the impact of single-cell technologies will require continued development in both experimental approaches to perturb cell states and in computational approaches to better understand their effects. Doing so will hopefully bring us closer to a better understanding of disease and how to treat it.

Competing interest statement

The authors declare no competing interests.

Acknowledgments

We thank Dr. Nancy Zhang for helpful discussions and comments on the manuscript. This work was supported by the following grants: National Human Genome Research Institute grant T32HG000046-21 (to B.J.A.); National Institute of General Medical Sciences grant R01GM125301 (to M.L.); National Eye Institute grants R01EY030192 (to M.L.), R01EY031209 (to M.L.), and R21EY031877 (to M.L.); and National Heart, Lung, and Blood Institute grants R21HL156234 (to M.L.), R01HL113147 (to M.L. and M.P.R.), and R01HL150359 (to M.L. and M.P.R.).

Author contributions: B.J.A. and M.L. wrote the manuscript with input from J.H. and M.P.R.

References

Agarwal V, Shendure J. 2020. Predicting mRNA abundance directly from genomic sequence using deep convolutional neural networks. *Cell Rep* **31:** 107663. doi:10.1016/j.celrep.2020.107663

Ahlmann-Eltze C, Huber W. 2021. Transformation and preprocessing of single-cell RNA-seq. bioRxiv doi:10.1101/2021.06.24.449781

Downloaded from genome.cshlp.org on May 2, 2026 . Published by Cold Spring Harbor Laboratory Press

Aibar S, González-Blas CB, Moerman T, Huynh-Thu VA, Imrichova H, Hulselmans G, Rambow F, Marine JC, Geurts P, Aerts J, et al. 2017. SCENIC: single-cell regulatory network inference and clustering. *Nat Methods* **14:** 1083–1086. doi:10.1038/nmeth.4463

Aldridge S, Teichmann SA. 2020. Single cell transcriptomics comes of age. *Nat Commun* **11:** 4307. doi:10.1038/s41467-020-18158-5

Alencar GF, Owsiany KM, Karnewar S, Sukhavasi K, Mocci G, Nguyen AT, Williams CM, Shamsuzzaman S, Mokry M, Henderson CA, et al. 2020. Stem cell pluripotency genes Klf4 and Oct4 regulate complex SMC phenotypic changes critical in late-stage atherosclerotic lesion pathogenesis. *Circulation* **142:** 2045–2059. doi:10.1161/CIRCULATIONAHA.120.046672

Andor N, Lau BT, Catalanotti C, Sathe A, Kubit M, Chen J, Blaj C, Cherry A, Bangs CD, Grimes SM, et al. 2020. Joint single cell DNA-seq and RNA-seq of gastric cancer cell lines reveals rules of in vitro evolution. *NAR Genom Bioinform* **2:** lqaa016. doi:10.1093/nargab/lqaa016

Argelaguet R, Arnol D, Bredikhin D, Deloro Y, Velten B, Marioni JC, Stegle O. 2020. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol* **21:** 111. doi:10.1186/s13059-020-02015-1

Ashley EA. 2016. Towards precision medicine. *Nat Rev Genet* **17:** 507–522. doi:10.1038/nrg.2016.86

Avsec Z, Agarwal V, Visentin D, Ledsam JR, Grabska-Barwinska A, Taylor KR, Assael Y, Jumper J, Kohli P, Kelley DR. 2021. Effective gene expression prediction from sequence by integrating long-range interactions. bioRxiv doi:10.1101/2021.04.07.438649

Bacher R, Kendziorski C. 2016. Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biol* **17:** 63. doi:10.1186/s13059-016-0927-y

Bakker B, Taudt A, Belderbos ME, Porubsky D, Spierings DC, de Jong TV, Halsema N, Kazemier HG, Hoekstra-Wakker K, Bradley A, et al. 2016. Single-cell sequencing reveals karyotype heterogeneity in murine and human malignancies. *Genome Biol* **17:** 115. doi:10.1186/s13059-016-0971-7

Barkas N, Petukhov V, Nikolaeva D, Lozinsky Y, Demharter S, Khodosevich K, Kharchenko PV. 2019. Joint analysis of heterogeneous single-cell RNA-seq dataset collections. *Nat Methods* **16:** 695–698. doi:10.1038/s41592-019-0466-z

Baron M, Veres A, Wolock SL, Faust AL, Gaujoux R, Vetere A, Ryu JH, Wagner BK, Shen-Orr SS, Klein AM, et al. 2016. A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure. *Cell Syst* **3:** 346–360.e4. doi:10.1016/j.cels.2016.08.011

Bellenguez C, Kucukali F, Jansen I, Andrade V, Morenau-Grau S, Amin N, Naj A, Grenier-Boley B, Martin RC, Holsman P, et al. 2021. New insights on the genetic etiology of Alzheimer's and related dementia. medRxiv doi:10.1101/2020.10.01.20200659

Bengio Y, Courville A, Vincent P. 2013. Representation learning: a review and new perspectives. In *International Conference on Learning Representations*, Scottsdale, AZ.

Brunner A-D, Thielert M, Vasilopoulou CG, Ammar C, Coscia F, Mund A, Hoerning OB, Bache N, Apalategui A, Lubeck M, et al. 2021. Ultra-high sensitivity mass spectrometry quantifies single-cell proteome changes upon perturbation. bioRxiv doi:10.1101/2020.12.22.423933

Buttner M, Ostner J, Muller CL, Theis FJ, Schubert B. 2020. scCODA: a Bayesian model for compositional single-cell data analysis. bioRxiv doi:10.1101/2020.12.14.422688

Cao J, Cusanovich DA, Ramani V, Aghamirzaie D, Pliner HA, Hill AJ, Daza RM, McFaline-Figueroa JL, Packer JS, Christiansen L, et al. 2018. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science* **361:** 1380–1385. doi:10.1126/science.aau0730

Chen S, Lake BB, Zhang K. 2019. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nat Biotechnol* **37:** 1452–1457. doi:10.1038/s41587-019-0290-0

Clark SJ, Argelaguet R, Kapourani C-A, Stubbs TM, Lee HJ, Alda-Catalinas C, Krueger F, Sanguinetti G, Kelsey G, Marioni JC, et al. 2018. scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nat Commun* **9:** 781. doi:10.1038/s41467-018-03149-4

Conway BR, O'Sullivan ED, Cairns C, O'Sullivan J, Simpson DJ, Salzano A, Connor K, Ding P, Humphries D, Stewart K, et al. 2020. Kidney single-cell atlas reveals myeloid heterogeneity in progression and regression of kidney disease. *J Am Soc Nephrol* **31:** 2833–2854. doi:10.1681/ASN.2020060806

Couturier CP, Ayyadhury S, Le PU, Nadaf J, Monlong J, Riva G, Allache R, Baig S, Yan X, Bourgey M, et al. 2020. Single-cell RNA-seq reveals that glioblastoma recapitulates a normal neurodevelopmental hierarchy. *Nat Commun* **11:** 3406. doi:10.1038/s41467-020-17186-5

Crowell HL, Soneson C, Germain PL, Calini D, Collin L, Raposo C, Malhotra D, Robinson MD. 2020. Muscat detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data. *Nat Commun* **11:** 6077. doi:10.1038/s41467-020-19894-4

Cuomo ASE, Seaton DD, McCarthy DJ, Martinez I, Bonder MJ, Garcia-Bernardo J, Amatya S, Madrigal P, Isaacson A, Buettner F, et al. 2020. Single-cell RNA-sequencing of differentiating iPS cells reveals dynamic genetic effects on gene expression. *Nat Commun* **11:** 810. doi:10.1038/s41467-020-14457-z

Cusanovich DA, Daza R, Adey A, Pliner HA, Christiansen L, Gunderson KL, Steemers FJ, Trapnell C, Shendure J. 2015. Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348:** 910–914. doi:10.1126/science.aab1601

Davis A, Gao R, Navin NE. 2019. SCOPIT: sample size calculations for single-cell sequencing experiments. *BMC Bioinformatics* **20:** 566. doi:10.1186/s12859-019-3167-9

Dixit A, Parnas O, Li B, Chen J, Fulco CP, Jerby-Arnon L, Marjanovic ND, Dionne D, Burks T, Raychowdhury R, et al. 2016. Perturb-Seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell* **167:** 1853–1866.e17. doi:10.1016/j.cell.2016.11.038

Dong M, Thennavan A, Urrutia E, Li Y, Perou CM, Zou F, Jiang Y. 2021. SCDC: bulk gene expression deconvolution by multiple single-cell RNA sequencing references. *Brief Bioinform* **22:** 416–427. doi:10.1093/bib/bbz166

Donovan MKR, D'Antonio-Chronowska A, D'Antonio M, Frazer KA. 2020. Cellular deconvolution of GTEx tissues powers discovery of disease and cell-type associated regulatory variants. *Nat Commun* **11:** 955. doi:10.1038/s41467-020-14561-0

Du R, Carey V, Weiss ST. 2019. deconvSeq: deconvolution of cell mixture distribution in sequencing data. *Bioinformatics* **35:** 5095–5102. doi:10.1093/bioinformatics/btz444

Efthymiou AG, Goate AM. 2017. Late onset Alzheimer's disease genetics implicates microglial pathways in disease risk. *Mol Neurodegener* **12:** 43. doi:10.1186/s13024-017-0184-x

Eng CL, Lawson M, Zhu Q, Dries R, Koulena N, Takei Y, Yun J, Cronin C, Karp C, Yuan GC, et al. 2019. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH. *Nature* **568:** 235–239. doi:10.1038/s41586-019-1049-y

Fagny M, Paulson JN, Kuijjer ML, Sonawane AR, Chen CY, Lopes-Ramos CM, Glass K, Quackenbush J, Platig J. 2017. Exploring regulation in tissues with eQTL networks. *Proc Natl Acad Sci* **114:** E7841–E7850. doi:10.1073/pnas.1707375114

Fan J, Wang X, Xiao R, Li M. 2021. Detecting cell-type-specific allelic expression imbalance by integrative analysis of bulk and single-cell RNA sequencing data. *PLoS Genet* **17:** e1009080. doi:10.1371/journal.pgen.1009080

Feng H, Moakley DF, Chen S, McKenzie MG, Menon V, Zhang C. 2021. Complexity and graded regulation of neuronal cell-type–specific alternative splicing revealed by single-cell RNA sequencing. *Proc Natl Acad Sci* **118:** e2013056118. doi:10.1073/pnas.2013056118

Finak G, McDavid A, Yajima M, Deng J, Gersuk V, Shalek AK, Slichter CK, Miller HW, McElrath MJ, Prlic M, et al. 2015. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol* **16:** 278. doi:10.1186/s13059-015-0844-5

Fiskin E, Lareau CA, Eraslan G, Ludwig LS, Regev A. 2020. Single-cell multimodal profiling of proteins and chromatin accessibility using PHAGE-ATAC. bioRxiv doi:10.1101/2020.10.01.322420

Fitz J, Neumann T, Steininger M, Wiedemann EM, Garcia AC, Athanasiadis A, Schoeberl UE, Pavri R. 2020. Spt5-mediated enhancer transcription directly couples enhancer activation with physical promoter interaction. *Nat Genet* **52:** 505–515. doi:10.1038/s41588-020-0605-6

Gaiti F, Chaligne R, Gu H, Brand RM, Kothen-Hill S, Schulman RC, Grigorev K, Risso D, Kim K-T, Pastore A, et al. 2019. Epigenetic evolution and lineage histories of chronic lymphocytic leukaemia. *Nature* **569:** 576–580. doi:10.1038/s41586-019-1198-z

Garvin T, Aboukhalil R, Kendall J, Baslan T, Atwal GS, Hicks J, Wigler M, Schatz MC. 2015. Interactive analysis and assessment of single-cell copy-number variations. *Nat Methods* **12:** 1058–1060. doi:10.1038/nmeth.3578

Guerreiro R, Wojtas A, Bras J, Carrasquillo M, Rogaeva E, Majounie E, Cruchaga C, Sassi C, Kauwe JS, Younkin S, et al. 2013. *TREM2* variants in Alzheimer's disease. *N Engl J Med* **368:** 117–127. doi:10.1056/NEJMoa1211851

Gupta I, Collier PG, Haase B, Mahfouz A, Joglekar A, Floyd T, Koopmans F, Barres B, Smit AB, Sloan SA, et al. 2018. Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells. *Nat Biotechnol* **36:** 1197–1202. doi:10.1038/nbt.4259

Hagemann-Jensen M, Ziegenhain C, Chen P, Ramsköld D, Hendriks GJ, Larsson AJM, Faridani OR, Sandberg R. 2020. Single-cell RNA counting at allele and isoform resolution using smart-seq3. *Nat Biotechnol* **38:** 708–714. doi:10.1038/s41587-020-0497-0

Downloaded from genome.cshlp.org on May 2, 2026 . Published by Cold Spring Harbor Laboratory Press

Haghverdi L, Lun ATL, Morgan MD, Marioni JC. 2018. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat Biotechnol* **36:** 421–427. doi:10.1038/nbt.4091

Hanna RE, Hegde M, Fagre CR, DeWeirdt PC, Sangree AK, Szegletes Z, Griffith A, Feeley MN, Sanson KR, Baidi Y, et al. 2021. Massively parallel assessment of human variants with base editor screens. *Cell* **184:** 1064–1080.e20. doi:10.1016/j.cell.2021.01.012

Hao Y, Hao S, Andersen-Nissen E, Mauck WM III, Zheng S, Butler A, Lee MJ, Wilk AJ, Darby C, Zager M, et al. 2021. Integrated analysis of multimodal single-cell data. *Cell* **184:** 3573–3587.e29. doi:10.1016/j.cell.2021.04.048

Heimberg G, Bhatnagar R, El-Samad H, Thomson M. 2016. Low dimensionality in gene expression data enables the accurate extraction of transcriptional programs from shallow sequencing. *Cell Syst* **2:** 239–250. doi:10.1016/j.cels.2016.04.001

Hicks SC, Townes FW, Teng M, Irizarry RA. 2018. Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics* **19:** 562–578. doi:10.1093/biostatistics/kxx053

Hie B, Bryson B, Berger B. 2019. Efficient integration of heterogeneous single-cell transcriptomes using scanorama. *Nat Biotechnol* **37:** 685–691. doi:10.1038/s41587-019-0113-3

Hie B, Bryson BD, Berger B. 2020. Leveraging uncertainty in machine learning accelerates biological discovery and design. *Cell Syst* **11:** 461–477.e9. doi:10.1016/j.cels.2020.09.007

Hou W, Ji Z, Ji H, Hicks SC. 2020. A systematic evaluation of single-cell RNA-sequencing imputation methods. *Genome Biol* **21:** 218. doi:10.1186/s13059-020-02132-x

Hu J, Li X, Hu G, Lyu Y, Susztak K, Li M. 2020a. Iterative transfer learning with neural network for clustering and cell type classification in single-cell RNA-seq analysis. *Nat Mach Intell* **2:** 607–618. doi:10.1038/s42256-020-00233-7

Hu Y, Wang K, Li M. 2020b. Detecting differential alternative splicing events in scRNA-seq with or without unique molecular identifiers. *PLoS Comput Biol* **16:** e1007925. doi:10.1371/journal.pcbi.1007925

Hu J, Li X, Coleman K, Schroeder A, Irwin DJ, Lee EB, Shinohara RT, Li M. 2020c. Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. bioRxiv doi:10.1101/2020.11.30.405118

Hu J, Schroeder A, Coleman K, Chen C, Auerbach BJ, Li M. 2021. Statistical and machine learning methods for spatially resolved transcriptomics with histology. *Comput Struct Biotechnol J* **19:** 3829–3841. doi:10.1016/j.csbj.2021.06.052

Huang Y, Sanguinetti G. 2017. BRIE: transcriptome-wide splicing quantification in single cells. *Genome Biol* **18:** 123. doi:10.1186/s13059-017-1248-5

Huang J, Li K, Cai W, Liu X, Zhang Y, Orkin SH, Xu J, Yuan GC. 2018. Dissecting super-enhancer hierarchy based on chromatin interactions. *Nat Commun* **9:** 943. doi:10.1038/s41467-018-03279-9

Huang M, Zhang Z, Zhang NR. 2020. Dimension reduction and denoising of single-cell RNA sequencing data in the presence of observed confounding variables. bioRxiv doi:10.1101/2020.08.03.234765

Jerber J, Seaton DD, Cuomo ASE, Kumasaka N, Haldane J, Steer J, Patel M, Pearce D, Andersson M, Bonder MJ, et al. 2021. Population-scale single-cell RNA-seq profiling across dopaminergic neuron differentiation. *Nat Genet* **53:** 304–312. doi:10.1038/s41588-021-00801-6

Jew B, Alvarez M, Rahmani E, Miao Z, Ko A, Garske KM, Sul JH, Pietiläinen KH, Pajukanta P, Halperin E. 2020. Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. *Nat Commun* **11:** 1971. doi:10.1038/s41467-020-15816-6

Ji Z, Ji H. 2016. TSCAN: pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic Acids Res* **44:** e117. doi:10.1093/nar/gkw430

Jia C, Hu Y, Kelly D, Kim J, Li M, Zhang NR. 2017. Accounting for technical noise in differential expression analysis of single-cell RNA sequencing data. *Nucleic Acids Res* **45:** 10978–10988. doi:10.1093/nar/gkx754

Jiang Y, Zhang NR, Li M. 2017. SCALE: modeling allele-specific gene expression by single-cell RNA sequencing. *Genome Biol* **18:** 74. doi:10.1186/s13059-017-1200-8

Joglekar A, Prjibelski A, Mahfouz A, Collier P, Lin S, Schlusche AK, Marrocco J, Williams SR, Haase B, Hayes A, et al. 2021. A spatially resolved brain region- and cell type-specific isoform atlas of the postnatal mouse brain. *Nat Commun* **12:** 463. doi:10.1038/s41467-020-20343-5

Jonsson T, Stefansson H, Steinberg S, Jonsdottir I, Jonsson PV, Snaedal J, Bjornsson S, Huttenlocher J, Levey AI, Lah JJ, et al. 2013. Variant of *TREM2* associated with the risk of Alzheimer's disease. *N Engl J Med* **368:** 107–116. doi:10.1056/NEJMoa1211103

Kharchenko PV, Silberstein L, Scadden DT. 2014. Bayesian approach to single-cell differential expression analysis. *Nat Methods* **11:** 740–742. doi:10.1038/nmeth.2967

Kim C, Gao R, Sei E, Brandt R, Hartman J, Hatschek T, Crosetto N, Foukakis T, Navin NE. 2018. Chemoresistance evolution in triple-negative breast cancer delineated by single-cell sequencing. *Cell* **173:** 879–893.e13. doi:10.1016/j.cell.2018.03.041

Kim N, Kim HK, Lee K, Hong Y, Cho JH, Choi JW, Lee JI, Suh YL, Ku BM, Eum HH, et al. 2020. Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. *Nat Commun* **11:** 2285. doi:10.1038/s41467-020-16164-1

Kim-Hellmuth S, Aguet F, Oliva M, Muñoz-Aguirre M, Kasela S, Wucher V, Castel SE, Hamel AR, Viñuela A, Roberts AL, et al. 2020. Cell type–specific genetic regulation of gene expression across human tissues. *Science* **369:** eaaz8528. doi:10.1126/science.aaz8528

Kiselev VY, Yiu A, Hemberg M. 2018. scmap: projection of single-cell RNA-seq data across data sets. *Nat Methods* **15:** 359–362. doi:10.1038/nmeth.4644

Kiselev VY, Andrews TS, Hemberg M. 2019. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat Rev Genet* **20:** 273–282. doi:10.1038/s41576-018-0088-9

Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, Baglaenko Y, Brenner M, Loh PR, Raychaudhuri S. 2019. Fast, sensitive and accurate integration of single-cell data with harmony. *Nat Methods* **16:** 1289–1296. doi:10.1038/s41592-019-0619-0

Korthauer KD, Chu LF, Newton MA, Li Y, Thomson J, Stewart R, Kendziorski C. 2016. A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. *Genome Biol* **17:** 222. doi:10.1186/s13059-016-1077-y

Kunkle BW, Grenier-Boley B, Sims R, Bis JC, Damotte V, Naj AC, Boland A, Vronskaya M, van der Lee SJ, Amlie-Wolf A, et al. 2019. Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, τ, immunity and lipid processing. *Nat Genet* **51:** 414–430. doi:10.1038/s41588-019-0358-2

Lähnemann D, Köster J, Szczurek E, McCarthy DJ, Hicks SC, Robinson MD, Vallejos CA, Campbell KR, Beerenwinkel N, Mahfouz A, et al. 2020. Eleven grand challenges in single-cell data science. *Genome Biol* **21:** 31. doi:10.1186/s13059-020-1926-6

Lake BB, Chen S, Sos BC, Fan J, Kaeser GE, Yung YC, Duong TE, Gao D, Chun J, Kharchenko PV, et al. 2018. Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nat Biotechnol* **36:** 70–80. doi:10.1038/nbt.4038

Lakkis J, Wang D, Zhang Y, Hu G, Wang K, Pan H, Ungar L, Reilly MP, Li X, Li M. 2021. A joint deep learning model enables simultaneous batch effect correction, denoising, and clustering in single-cell transcriptomics. *Genome Res* (this issue). doi:10.1101/gr.271874.120

Laks E, McPherson A, Zahn H, Lai D, Steif A, Brimhall J, Biele J, Wang B, Masud T, Ting J, et al. 2019. Clonal decomposition and DNA replication states defined by scaled single-cell genome sequencing. *Cell* **179:** 1207–1221.e22. doi:10.1016/j.cell.2019.10.026

Lambert JC, Ibrahim-Verbaas CA, Harold D, Naj AC, Sims R, Bellenguez C, DeStafano AL, Bis JC, Beecham GW, Grenier-Boley B, et al. 2013. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat Genet* **45:** 1452–1458. doi:10.1038/ng.2802

Lander ES, Botstein D. 1989. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121:** 185–199. doi:10.1093/genetics/121.1.185

Lange M, Bergen V, Klein M, Setty M, Reuter B, Bakhti M, Lickert H, Ansari M, Schniering J, Schiller B, et al. 2020. CellRank for directed single-cell fate mapping. bioRxiv doi:10.1101/2020.10.19.345983

Larsson AJM, Johnsson P, Hagemann-Jensen M, Hartmanis L, Faridani OR, Reinius B, Segerstolpe A, Rivera CM, Ren B, Sandberg R. 2019. Genomic encoding of transcriptional burst kinetics. *Nature* **565:** 251–254. doi:10.1038/s41586-018-0836-1

Law CW, Chen Y, Shi W, Smyth GK. 2014. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* **15:** R29. doi:10.1186/gb-2014-15-2-r29

Lawlor N, George J, Bolisetty M, Kursawe R, Sun L, Sivakamasundari V, Kycia I, Robson P, Stitzel ML. 2017. Single-cell transcriptomes identify human islet cell signatures and reveal cell-type–specific expression changes in type 2 diabetes. *Genome Res* **27:** 208–222. doi:10.1101/gr.212720.116

Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K, Irizarry RA. 2010. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet* **11:** 733–739. doi:10.1038/nrg2825

Li J, Klughammer J, Farlik M, Penz T, Spittler A, Barbieux C, Berishvili E, Bock C, Kubicek S. 2016. Single-cell transcriptomes reveal characteristic features of human pancreatic islet cell types. *EMBO Rep* **17:** 178–187. doi:10.15252/embr.201540946

Li X, Wang K, Lyu Y, Pan H, Zhang J, Stambolian D, Susztak K, Reilly MP, Hu G, Li M. 2020. Deep learning enables accurate clustering with batch effect removal in single-cell RNA-seq analysis. *Nat Commun* **11:** 2338. doi:10.1038/s41467-020-15851-3

Downloaded from genome.cshlp.org on May 2, 2026 . Published by Cold Spring Harbor Laboratory Press

Lin KZ, Lei J, Roeder K. 2021. Exponential-family embedding with application to cell developmental trajectories for single-cell RNA-seq data. *J Am Stat Assoc* **116:** 457–470. doi:10.1080/01621459.2021.1886106
Linnarsson S, Teichmann SA. 2016. Single-cell genomics: coming of age. *Genome Biol* **17:** 97. doi:10.1186/s13059-016-0960-x
Liu Z, Lou H, Xie K, Wang H, Chen N, Aparicio OM, Zhang MQ, Jiang R, Chen T. 2017. Reconstructing cell cycle pseudo time-series via single-cell transcriptome data. *Nat Commun* **8:** 22. doi:10.1038/s41467-017-00039-z
Liu X, Li YI, Pritchard JK. 2019. *Trans* effects on gene expression can drive omnigenic inheritance. *Cell* **177:** 1022–1034.e6. doi:10.1016/j.cell.2019.04.014
Lloyd-Jones LR, Holloway A, McRae A, Yang J, Small K, Zhao J, Zeng B, Bakshi A, Metspalu A, Dermitzakis M, et al. 2017. The genetic architecture of gene expression in peripheral blood. *Am J Hum Genet* **100:** 228–237. doi:10.1016/j.ajhg.2016.12.008
Lopez R, Regier J, Cole MB, Jordan MI, Yosef N. 2018. Deep generative modeling for single-cell transcriptomics. *Nat Methods* **15:** 1053–1058. doi:10.1038/s41592-018-0229-2
Lotfollahi M, Naghipourfar M, Luecken MD, Khajavi M, Büttner M, Wagenstetter M, Avsec Ž, Gayoso A, Yosef N, Interlandi M, et al. 2021. Mapping single-cell data to reference atlases by transfer learning. *Nat Biotechnol* doi:10.1038/s41587-021-01001-7
Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15:** 550. doi:10.1186/s13059-014-0550-8
Lucken MD, Buttner M, Chaichoompu K, Danese A, Interlandi M, Mueller MF, Strobl DC, Zappia L, Dugas M, Colome-Tatche M, et al. 2020. Benchmarking atlas-level data integration in single-cell genomics. bioRxiv doi:10.1101/2020.05.22.111161
Luo C, Liu H, Xie F, Armand EJ, Siletti K, Bakken TE, Fang R, Doyle WI, Hodge RD, Hu L, et al. 2019. Single nucleus multi-omics links human cortical cell regulatory genome diversity to disease risk variants. bioRxiv doi:10.1101/2019.12.11.873398
Lytal N, Ran D, An L. 2020. Normalization methods on single-cell RNA-seq data: an empirical survey. *Front Genet* **11:** 41. doi:10.3389/fgene.2020.00041
Lyu Y, Zauhar R, Dana N, Strang CE, Hu J, Wang K, Liu S, Pan N, Gamlin P, Kimble JA, et al. 2021. Implication of specific retinal cell-type involvement and gene expression changes in AMD progression using integrative analysis of single-cell and bulk RNA-seq profiling. *Sci Rep* **11:** 15612. doi:10.1038/s41598-021-95122-3
Ma S, Zhang B, LaFave LM, Earl AS, Chiang Z, Hu Y, Ding J, Brack A, Kartha VK, Tay T, et al. 2020. Chromatin potential identified by shared single-cell profiling of RNA and chromatin. *Cell* **183:** 1103–1116.e20. doi:10.1016/j.cell.2020.09.056
Macaulay IC, Ponting CP, Voet T. 2017. Single-cell multiomics: multiple measurements from single cells. *Trends Genet* **33:** 155–168. doi:10.1016/j.tig.2016.12.003
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. doi:10.1016/j.cell.2015.05.002
Mathys H, Davila-Velderrain J, Peng Z, Gao F, Mohammadi S, Young JZ, Menon M, He L, Abdurrob F, Jiang X, et al. 2019. Single-cell transcriptomic analysis of Alzheimer's disease. *Nature* **570:** 332–337. doi:10.1038/s41586-019-1195-2
Melsted P, Booeshaghi AS, Liu L, Gao F, Lu L, Min KHJ, da Veiga Beltrame E, Hjörleifsson KE, Gehring J, Pachter L. 2021. Modular, efficient and constant-memory single-cell RNA-seq preprocessing. *Nat Biotechnol* **39:** 813–818. doi:10.1038/s41587-021-00870-2
Moffitt JR, Bambah-Mukku D, Eichhorn SW, Vaughn E, Shekhar K, Perez JD, Rubinstein ND, Hao J, Regev A, Dulac C, et al. 2018. Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science* **362:** eaau5324. doi:10.1126/science.aau5324
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. doi:10.1038/nature09807
Neu SC, Pa J, Kukull W, Beekly D, Kuzma A, Gangadharan P, Wang LS, Romero K, Arneric SP, Redolfi A, et al. 2017. Apolipoprotein E genotype and sex risk factors for Alzheimer disease: a meta-analysis. *JAMA Neurol* **74:** 1178–1189. doi:10.1001/jamaneurol.2017.2188
Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, Hoang CD, Diehn M, Alizadeh AA. 2015. Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods* **12:** 453–457. doi:10.1038/nmeth.3337
Newman AM, Steen CB, Liu CL, Gentles AJ, Chaudhuri AA, Scherer F, Khodadoust MS, Esfahani MS, Luca BA, Steiner D, et al. 2019. Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat Biotechnol* **37:** 773–782. doi:10.1038/s41587-019-0114-2
Nguyen AT, Wang K, Hu G, Wang X, Miao Z, Azevedo JA, Suh E, Van Deerlin VM, Choi D, Roeder K, et al. 2020a. *APOE* and *TREM2* regulate amyloid-responsive microglia in Alzheimer's disease. *Acta Neuropathol* **140:** 477–493. doi:10.1007/s00401-020-02200-3
Nguyen H, Tran D, Tran B, Pehlivan B, Nguyen T. 2020b. A comprehensive survey of regulatory network inference methods using single cell RNA sequencing data. *Brief Bioinform* **22:** bbaa190. doi:10.1093/bib/bbaa190
Ntranos V, Yi L, Melsted P, Pachter L. 2019. A discriminative learning approach to differential expression analysis for single-cell RNA-seq. *Nat Methods* **16:** 163–166. doi:10.1038/s41592-018-0303-9
Oikonomou P, Salatino R, Tavazoie S. 2020. In vivo mRNA display enables large-scale proteomics by next generation sequencing. *Proc Natl Acad Sci* **117:** 26710–26718. doi:10.1073/pnas.2002650117
Pan H, Xue C, Auerbach BJ, Fan J, Bashore AC, Cui J, Yang DY, Trignano SB, Liu W, Shi J, et al. 2020. Single-cell genomics reveals a novel cell state during smooth muscle cell phenotypic switching and potential therapeutic targets for atherosclerosis in mouse and human. *Circulation* **142:** 2060–2075. doi:10.1161/CIRCULATIONAHA.120.048378
Park Y, He L, Davila-Velderrain J, Hou L, Mohammadi S, Mathys H, Peng Z, Bennett D, Tsai L-H, Kellis M. 2021. Single-cell deconvolution of 3,000 post-mortem brain samples for eQTL and GWAS dissection in mental disorders. bioRxiv doi:10.1101/2021.01.21.426000
Peloso GM, Rader DJ, Gabriel S, Kathiresan S, Daly MJ, Neale BM. 2016. Phenotypic extremes in rare variant study designs. *Eur J Hum Genet* **24:** 924–930. doi:10.1038/ejhg.2015.197
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. doi:10.1038/nprot.2014.006
Polanski K, Young MD, Miao Z, Meyer KB, Teichmann SA, Park JE. 2020. BBKNN: fast batch alignment of single cell transcriptomes. *Bioinformatics* **36:** 964–965. doi:10.1093/bioinformatics/btz625
Pott S. 2017. Simultaneous measurement of chromatin accessibility, DNA methylation, and nucleosome phasing in single cells. *eLife* **6:** e23203. doi:10.7554/eLife.23203
Preissl S, Fang R, Huang H, Zhao Y, Raviram R, Gorkin DU, Zhang Y, Sos BC, Afzal V, Dickel DE, et al. 2018. Single-nucleus analysis of accessible chromatin in developing mouse forebrain reveals cell-type-specific transcriptional regulation. *Nat Neurosci* **21:** 432–439. doi:10.1038/s41593-018-0079-3
Price AL, Helgason A, Thorleifsson G, McCarroll SA, Kong A, Stefansson K. 2011. Single-tissue and cross-tissue heritability of gene expression via identity-by-descent in related or unrelated individuals. *PLoS Genet* **7:** e1001317. doi:10.1371/journal.pgen.1001317
Qiu X, Hill A, Packer J, Lin D, Ma YA, Trapnell C. 2017a. Single-cell mRNA quantification and differential analysis with census. *Nat Methods* **14:** 309–315. doi:10.1038/nmeth.4150
Qiu X, Mao Q, Tang Y, Wang L, Chawla R, Pliner HA, Trapnell C. 2017b. Reversed graph embedding resolves complex single-cell trajectories. *Nat Methods* **14:** 979–982. doi:10.1038/nmeth.4402
Qiu Q, Hu P, Qiu X, Govek KW, Cámara PG, Wu H. 2020. Massively parallel and time-resolved RNA sequencing in single cells with scNT-seq. *Nat Methods* **17:** 991–1001. doi:10.1038/s41592-020-0935-4
Ren Z, Yu P, Li D, Li Z, Liao Y, Wang Y, Zhou B, Wang L. 2020. Single-cell reconstruction of progression trajectory reveals intervention principles in pathological cardiac hypertrophy. *Circulation* **141:** 1704–1719. doi:10.1161/CIRCULATIONAHA.119.043053
Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26:** 139–140. doi:10.1093/bioinformatics/btp616
Rodriques SG, Stickels RR, Goeva A, Martin CA, Murray E, Vanderburg CR, Welch J, Chen LM, Chen F, Macosko EZ. 2019. Slide-seq: a scalable technology for measuring genome-wide expression at high spatial resolution. *Science* **363:** 1463–1467. doi:10.1126/science.aaw1219
Rubin AJ, Parker KR, Satpathy AT, Qi Y, Wu B, Ong AJ, Mumbach MR, Ji AL, Kim DS, Cho SW, et al. 2019. Coupled single-cell CRISPR screening and epigenomic profiling reveals causal gene regulatory networks. *Cell* **176:** 361–376.e17. doi:10.1016/j.cell.2018.11.022
Saelens W, Cannoodt R, Todorov H, Saeys Y. 2019. A comparison of single-cell trajectory inference methods. *Nat Biotechnol* **37:** 547–554. doi:10.1038/s41587-019-0071-9
Sarkar AK, Tung PY, Blischak JD, Burnett JE, Li YI, Stephens M, Gilad Y. 2019. Discovery and characterization of variance QTLs in human induced pluripotent stem cells. *PLoS Genet* **15:** e1008045. doi:10.1371/journal.pgen.1008045
Schmid KT, Cruceanu C, Bottcher A, Lickert H, Binder EB, Thesis FJ, Heining M. 2020. Design and power analysis for multi-sample single cell genomics experiments. bioRxiv doi:10.1101/2020.04.01.019851
Schoenfelder S, Fraser P. 2019. Long-range enhancer–promoter contacts in gene expression control. *Nat Rev Genet* **20:** 437–455. doi:10.1038/s41576-019-0128-0

Downloaded from genome.cshlp.org on May 2, 2026 . Published by Cold Spring Harbor Laboratory Press

Schwartzentruber J, Cooper S, Liu JZ, Barrio-Hernandez I, Bello E, Kumasaka N, Young AMH, Franklin RJM, Johnson T, Estrada K, et al. 2021. Genome-wide meta-analysis, fine-mapping and integrative prioritization implicate new Alzheimer's disease risk genes. *Nat Genet* **53:** 392–402. doi:10.1038/s41588-020-00776-w

Segerstolpe A, Palasantza A, Eliasson P, Andersson EM, Andreasson AC, Sun X, Picelli S, Sabirsh A, Clausen M, Bjursell MK, et al. 2016. Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. *Cell Metab* **24:** 593–607. doi:10.1016/j.cmet.2016.08.020

Slovin S, Carissimo A, Panariello F, Grimaldi A, Bouché V, Gambardella G, Cacchiarelli D. 2021. Single-cell RNA sequencing analysis: a step-by-step overview. *Methods Mol Biol* **2284:** 343–365. doi:10.1007/978-1-0716-1307-8_19

Song Y, Botvinnik OB, Lovci MT, Kakaradov B, Liu P, Xu JL, Yeo GW. 2017. Single-cell alternative splicing analysis with expedition reveals splicing dynamics during neuron differentiation. *Mol Cell* **67:** 148–161.e5. doi:10.1016/j.molcel.2017.06.003

Song F, Chan GMA, Wei Y. 2020. Flexible experimental designs for valid single-cell RNA-sequencing experiments allowing batch effects correction. *Nat Commun* **11:** 3274. doi:10.1038/s41467-020-16905-2

Specht H, Emmott E, Petelski AA, Huffman RG, Perlman DH, Serra M, Kharchenko P, Koller A, Slavov N. 2021. Single-cell proteomic and transcriptomic analysis of macrophage heterogeneity using SCoPE2. *Genome Biol* **22:** 50. doi:10.1186/s13059-021-02267-5

Stoeckius M, Hafemeister C, Stephenson W, Houck-Loomis B, Chattopadhyay PK, Swerdlow H, Satija R, Smibert P. 2017. Simultaneous epitope and transcriptome measurement in single cells. *Nat Methods* **14:** 865–868. doi:10.1038/nmeth.4380

Street K, Risso D, Fletcher RB, Das D, Ngai J, Yosef N, Purdom E, Dudoit S. 2018. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* **19:** 477. doi:10.1186/s12864-018-4772-0

Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM III, Hao Y, Stoeckius M, Smibert P, Satija R. 2019. Comprehensive integration of single-cell data. *Cell* **177:** 1888–1902.e21. doi:10.1016/j.cell.2019.05.031

Svensson V, Teichmann SA, Stegle O. 2018. SpatialDE: identification of spatially variable genes. *Nat Methods* **15:** 343–346. doi:10.1038/nmeth.4636

Swanson E, Lord C, Reading J, Heubeck AT, Genge PC, Thomson Z, Weiss MD, Li XJ, Savage AK, Green RR, et al. 2021. Simultaneous trimodal single-cell measurement of transcripts, epitopes, and chromatin accessibility using TEA-seq. *eLife* **10:** e63632. doi:10.7554/eLife.63632

Takei Y, Yun J, Zheng S, Ollikainen N, Pierson N, White J, Shah S, Thomassie J, Suo S, Eng CL, et al. 2021. Integrated spatial genomics reveals global architecture of single nuclei. *Nature* **590:** 344–350. doi:10.1038/s41586-020-03126-2

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. doi:10.1038/nmeth.1315

Thornton CA, Mulqueen RM, Torkenczy KA, Nishida A, Lowenstein EG, Fields AJ, Steemers FJ, Zhang W, McConnell HL, Woltjer RL, et al. 2021. Spatially mapped single-cell chromatin accessibility. *Nat Commun* **12:** 1274. doi:10.1038/s41467-021-21515-7

van der Wijst M, de Vries DH, Groot HE, Trynka G, Hon CC, Bonder MJ, Stegle O, Nawijn MC, Idaghdour Y, van der Harst P, et al. 2020. The single-cell eQTLGen consortium. *eLife* **9:** e52155. doi:10.7554/eLife.52155

Velazquez-Villarreal EI, Maheshwari S, Sorenson J, Fiddes IT, Kumar V, Yin Y, Webb MG, Catalanotti C, Grigorova M, Edwards PA, et al. 2020. Single-cell sequencing of genomic DNA resolves sub-clonal heterogeneity in a melanoma cell line. *Commun Biol* **3:** 318. doi:10.1038/s42003-020-1044-8

Vickovic S, Eraslan G, Salmén F, Klughammer J, Stenbeck L, Schapiro D, Äijö T, Bonneau R, Bergenstråhle L, Navarro JF, et al. 2019. High-definition spatial transcriptomics for in situ tissue profiling. *Nat Methods* **16:** 987–990. doi:10.1038/s41592-019-0548-y

Wang Y, Waters J, Leung ML, Unruh A, Roh W, Shi X, Chen K, Scheet P, Vattathil S, Liang H, et al. 2014. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature* **512:** 155–160. doi:10.1038/nature13600

Wang YJ, Schug J, Won KJ, Liu C, Naji A, Avrahami D, Golson ML, Kaestner KH. 2016. Single-cell transcriptomics of the human endocrine pancreas. *Diabetes* **65:** 3028–3038. doi:10.2337/db16-0405

Wang X, Park J, Susztak K, Zhang NR, Li M. 2019. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nat Commun* **10:** 380. doi:10.1038/s41467-018-08023-x

Wang R, Dang M, Harada K, Han G, Wang F, Pool Pizzi M, Zhao M, Tatlonghari G, Zhang S, Hao D, et al. 2021. Single-cell dissection of intratumoral heterogeneity and lineage diversity in metastatic gastric adenocarcinoma. *Nat Med* **27:** 141–151. doi:10.1038/s41591-020-1125-8

Welch JD, Kozareva V, Ferreira A, Vanderburg C, Martin C, Macosko EZ. 2019. Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell* **177:** 1873–1887.e17. doi:10.1016/j.cell.2019.05.006

Westra HJ, Peters MJ, Esko T, Yaghootkar H, Schurmann C, Kettunen J, Christiansen MW, Fairfax BP, Schramm K, Powell JE, et al. 2013. Systematic identification of *trans* eQTLs as putative drivers of known disease associations. *Nat Genet* **45:** 1238–1243. doi:10.1038/ng.2756

Wills QF, Livak KJ, Tipping AJ, Enver T, Goldson AJ, Sexton DW, Holmes C. 2013. Single-cell gene expression analysis reveals genetic associations masked in whole-tissue experiments. *Nat Biotechnol* **31:** 748–752. doi:10.1038/nbt.2642

Wirka RC, Wagh D, Paik DT, Pjanic M, Nguyen T, Miller CL, Kundu R, Nagao M, Coller J, Koyano TK, et al. 2019. Atheroprotective roles of smooth muscle cell phenotypic modulation and the *TCF21* disease gene as revealed by single-cell analysis. *Nat Med* **25:** 1280–1289. doi:10.1038/s41591-019-0512-5

Wolf FA, Hamey FK, Plass M, Solana J, Dahlin JS, Göttgens B, Rajewsky N, Simon L, Theis FJ. 2019. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol* **20:** 59. doi:10.1186/s13059-019-1663-x

Wu Y, Zhang K. 2020. Tools for the analysis of high-dimensional single-cell RNA sequencing data. *Nat Rev Nephrol* **16:** 408–421. doi:10.1038/s41581-020-0262-0

Wu CY, Lau BT, Kim HS, Sathe A, Grimes SM, Ji HP, Zhang NR. 2021a. Integrative single-cell analysis of allele-specific copy number alterations and chromatin accessibility in cancer. *Nat Biotechnol* doi:10.1038/s41587-021-00911-w

Wu KE, Yost KE, Chang HY, Zou J. 2021b. BABEL enables cross-modality translation between multiomic profiles at single-cell resolution. *Proc Natl Acad Sci* **118:** e2023070118. doi:10.1073/pnas.2023070118

Xin Y, Kim J, Okamoto H, Ni M, Wei Y, Adler C, Murphy AJ, Yancopoulos GD, Lin C, Gromada J. 2016. RNA sequencing of single human islet cells reveals type 2 diabetes genes. *Cell Metab* **24:** 608–615. doi:10.1016/j.cmet.2016.08.018

Xiong H, Luo Y, Wang Q, Yu X, He A. 2021. Single-cell joint detection of chromatin occupancy and transcriptome enables higher-dimensional epigenomic reconstructions. *Nat Methods* **18:** 652–660. doi:10.1038/s41592-021-01129-z

Xu C, Lopez R, Mehlman E, Regier J, Jordan MI, Yosef N. 2021. Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Mol Syst Biol* **17:** e9620. doi:10.15252/msb.20209620

Zaccaria S, Raphael BJ. 2021. Characterizing allele- and haplotype-specific copy numbers in single cells with CHISEL. *Nat Biotechnol* **39:** 207–214. doi:10.1038/s41587-020-0661-6

Zeitlinger J. 2020. Seven myths of how transcription factors read the *cis*-regulatory code. *Curr Opin Syst Biol* **23:** 22–31. doi:10.1016/j.coisb.2020.08.002

Zhang L, Zhang S. 2020. Comparison of computational methods for imputing single-cell RNA-sequencing data. *IEEE/ACM Trans Comput Biol Bioinform* **17:** 376–389. doi:10.1109/TCBB.2018.2848633

Zhang MJ, Ntranos V, Tse D. 2020. Determining sequencing depth in a single-cell RNA-seq experiment. *Nat Commun* **11:** 774. doi:10.1038/s41467-020-14482-y

Zhang Y, Narayanan SP, Mannan R, Raskind G, Wang X, Vats P, Su F, Hosseini N, Cao X, Kumar-Sinha C, et al. 2021. Single-cell analyses of renal cell cancers reveal insights into tumor microenvironment, cell of origin, and therapy response. *Proc Natl Acad Sci* **118:** e2103240118. doi:10.1073/pnas.2103240118

Zhu C, Yu M, Huang H, Juric I, Abnousi A, Hu R, Lucero J, Behrens MM, Hu M, Ren B. 2019. An ultra high-throughput method for single-cell joint analysis of open chromatin and transcriptome. *Nat Struct Mol Biol* **26:** 1063–1070. doi:10.1038/s41594-019-0323-x