Trends in Genetics

CellPress
REVIEWS

Opinion

Co-expression in Single-Cell Analysis: Saving Grace or Original Sin?

Megan Crow[1] and Jesse Gillis[1,*]

As a fundamental unit of life, the cell has rightfully been the subject of intense investigation throughout the history of biology. Technical innovations now make it possible to assay cellular features at genomic scale, yielding breakthroughs in our understanding of the molecular organization of tissues, and even whole organisms. As these data accumulate we will soon be faced with a new challenge: making sense of the plethora of results. Early investigations into the replicability of cell type profiles inferred from single-cell RNA sequencing data have indicated that this is likely to be surprisingly straightforward due to consistent gene co-expression. In this opinion article we discuss the evidence for this claim and its implications for interpreting cell type-specific gene expression.

Highlights

Single-cell RNA sequencing approaches are vastly increasing in scale, with individual experiments routinely profiling thousands or even hundreds of thousands of cells.

Despite technical limitations associated with low-input sequencing, cell classification through unsupervised clustering is surprisingly replicable across studies. This can be attributed to the intrinsic low dimensionality of cell types dominating the variability seen in expression profiles.

Low dimensionality of expression profiles implies gene co-expression. An exploration of the history of co-expression highlights the perils of making gene-level inferences in light of collinearity, an issue that has previously arisen in cancer subtyping analysis.

Co-expression has been both the saving grace and the original sin of single-cell RNA-seq, enabling sample characterization at the cost of gene-level inference.

Single-Cell Rising

Single-cell RNA sequencing (scRNA-seq) technologies have exponentially increased in capacity over a few short years. Far from early studies of a few hand-picked cells, individual experiments now routinely run to thousands or even hundreds of thousands of cells [1]. This technical progress has fostered biological discovery at the single-cell level, including impressive approaches for whole-organism profiling [2–4] and cell lineage tracing [5]. Computational methods have proliferated in turn, and already more than 200 analysis tools have been catalogued as part of the scRNA-tools database [6].

Previous review articles have emphasized the novelty of the analytic challenges posed by single-cell data (e.g., [7,8]). By contrast, in this opinion article we aim to show the deep roots of scRNA-seq within the greater history of expression analysis, and particularly co-expression network analysis. We are motivated by recent evidence that single-cell studies show surprising replicability in spite of technical issues. Our thesis is that this can only be explained by robust gene co-expression. We discuss the link between low dimensionality in scRNA-seq and gene co-expression, and describe previous efforts to use co-expression for sample characterization in cancer. These examples clarify the major limitation of relying on co-expression for single-cell analysis: collinearity confounds gene-level inference (Box 1). For convenience, we focus on the simpler case of linear relationships, although more complex dependencies can be explored within the same framework. We conclude with a discussion of outstanding questions within this young field and highlight possible avenues for progress.

The Surprising Replicability of scRNA-Seq

Many single-cell studies are motivated by the hypothesis that characterizing the extent and causes of cellular heterogeneity will enable deeper understanding of biological systems [9]. One particular hope of single-cell approaches is that they will resolve the long-standing issue of whether differential expression in bulk tissue results from unequal cell type proportions or from changes to gene regulation within a cell type across samples. In recent years scRNA-seq has gained

[1]Cold Spring Harbor Laboratory, One Bungtown Road, Cold Spring Harbor, NY 11724, USA

*Correspondence: jgillis@cshl.edu (J. Gillis).

Box 1. Key Concepts

Collinearity and Co-expression

Collinearity (or multicollinearity) occurs when one explanatory variable can be linearly predicted using one or more of the other explanatory variables. In expression analysis, genes are the explanatory variables that characterize differences between sample groups, and they will be collinear if their expression profiles are correlated across samples. Correlation between genes is also known as co-expression, a well-established feature of high-throughput expression data. While perfectly benign as an observation, collinearity has important consequences for interpreting results. If 100 genes are correlated with cell differentiation, which gene is the driver? The data alone cannot tell us.

Low Dimensionality

If we consider expression data as a matrix with the form *N* columns x *P* rows, where the columns give different samples and the rows are genes, the dimension of each sample is *P*, the number of attributes listed for that sample. Most human scRNA-seq experiments use polyA-selection methods and map to protein-coding genes, making $P \sim 20\,000$, which is high relative to the number of samples (historically). Yet because genes are co-expressed, the number of dimensions necessary to characterize each sample is lower than it might at first appear: some of the dimensions (genes) are redundant. Principal component analysis (PCA) and other dimension reduction methods aim to extract the maximal amount of variability from a matrix using the fewest dimensions. Because these methods have been so successful at separating cell types, often requiring only 10–100 dimensions depending on the tissue and number of samples, there is a strong claim that cell types are low dimensional.

enormous popularity thanks to advances in microfluidics technology that enable high-throughput liquid handling and an economy of scale through barcoding strategies [1]. Some of the most prominent applications of scRNA-seq have been attempts to define all cell types within a tissue as a sort of molecular census [10–13] (Box 2). The early success of these strategies has prompted even greater interest in this approach, and the Human Cell Atlas project is a notable example of a large-scale effort to catalogue cell types with single-cell technology [14].

Alongside the many advances has come a greater appreciation of the potential pitfalls of low-input RNA sequencing, including technical variation caused by PCR amplification or signal drop-out [15–17] and prominent batch effects [18]. Further questions have been raised

Box 2. Inferring Cell Types from Expression Data

One of the primary tasks in single-cell transcriptomics has been to use expression data to characterize the heterogeneity of cells within a given cell type or tissue. A common workflow for this is to group the cells by their expression profiles and then compare expression values between groups. But what are these groups? Do they represent novel cell types or subtypes? Answering these questions requires us to formally define these terms, and field-wide standards in nomenclature have yet to be achieved. A standard working definition would discriminate between two key aspects of cell identity: cell type and cell state. Cell type refers to more permanent features of a cell's identity (a neuron does not become a red blood cell), whereas cell state is more variable and often reflects temporally limited processes (circadian rhythm, cell cycle). In this opinion article we are primarily concerned with studies that aim to define cell types from single cell data, which are often organized in a hierarchical taxonomy that can be further divided into subtypes.

The basic steps involved in scRNA-seq are as follows: cells are captured and lysed and mRNA is reverse transcribed to generate cDNA libraries, often including cell and molecular barcodes for multiplexing and reducing PCR-amplification bias, respectively; then, sequencing proceeds as usual. After quality control and normalization, sample clustering is a key step of almost any single-cell analysis pipeline. To obtain a robust representation of the underlying data, analysts rely on dimension reduction techniques such as PCA, often calculated based on a subset of highly variable genes. Distances are measured between cells based on their co-ordinates within this reduced space, and cells are counted as similar if they occupy similar positions (i.e., cluster).

This is the area of single-cell analysis that receives the most attention, and yet it is often the most opaque. Published protocols and bioinformatics packages suggest choosing the number of clusters that 'agrees with your intuition' (https://github.com/hb-gitified/cellrangerRkit/blob/master/vignettes/cellrangerrkit-PBMC-vignette-knitr.pdf) or that maximizes some measure of modularity [12,69]. Still others suggest taking the consensus across multiple parameter choices [70,71]. Without external data for validation, clustering is necessarily exploratory rather than confirmatory, and the risk of overfitting (finding idiosyncratic clusters) is high. These issues are partially resolved by resampling within the data but are best addressed through cross-dataset replicability analysis. Transcriptome-based classifications ultimately require biological validation via independent assays into cell identity and function.

regarding appropriate normalization and handling of biological confounders such as the cell cycle or transcriptional bursting [19–21]. Despite these challenges, it is becoming increasingly clear that cell profiles can be aligned across technical and biological sources of variation [22–29], and our own work has indicated that more than half of the computationally identified interneuron subtype profiles are highly replicable [30]. We and others [29] have also demonstrated high replicability among five studies of the human pancreas. How can these unexpected successes be explained?

Cell Types Are Low Dimensional

One plausible reason is that cell identity signals are highly robust. For example, we know a principal source of noise in single-cell data is incomplete sampling of the total mRNA pool, which means that a high proportion of genes are not detected within an individual sample. However, if many genes encode cell identity, then we will be able to read out this property regardless of individual gene drop-outs (Figure 1). The robustness of cell type transcriptional profiles was first suggested by early downsampling and multiplexing experiments that showed that cell type identification was possible with quite a small number of reads [31,32], and this message has been re-iterated by Drop-seq proponents [12,24].

In their 2016 paper, Heimberg and colleagues explored the conceptual basis of this phenomenon in detail [33]. Taking inspiration from signal processing where it is known that many natural

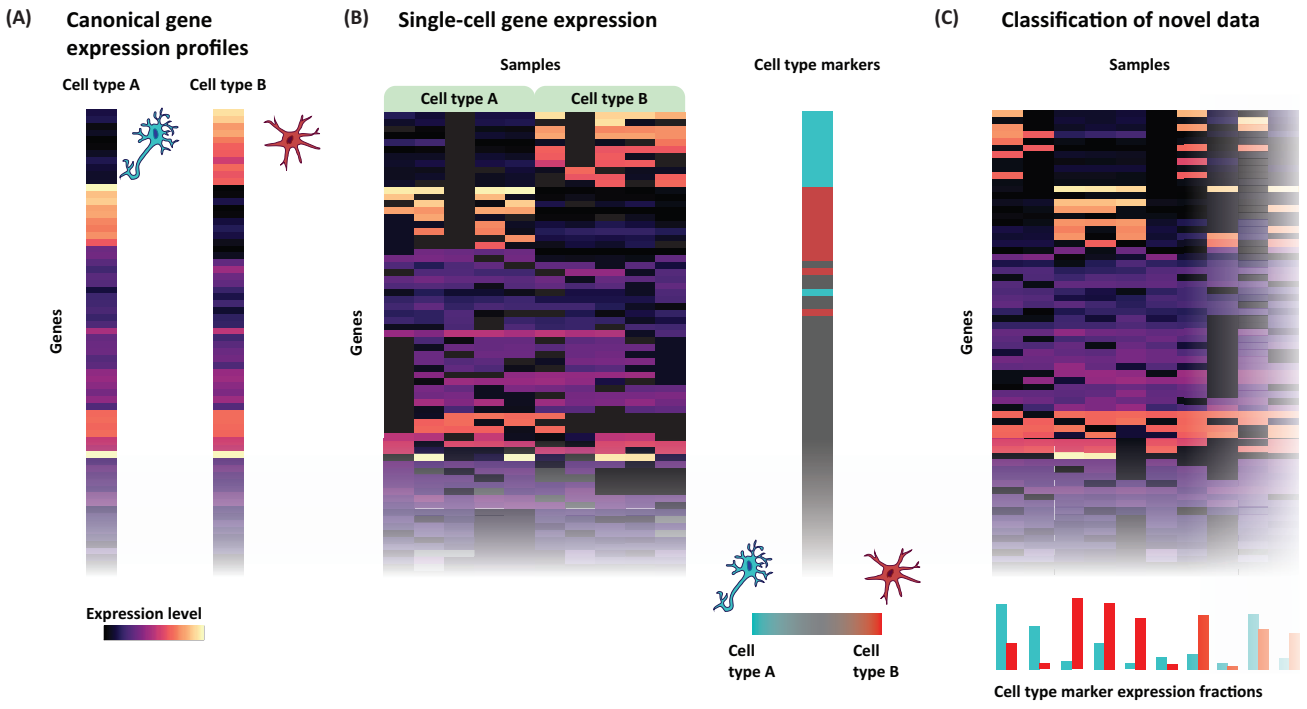

Figure 1. Cell Type Identity Is Encoded in Many Genes. (A) Schematic of ground-truth expression profiles for two cell types, A and B, where each row is a cell and the color indicates the expression level. Many genes are similarly expressed in both cell types, but a handful of markers are expressed exclusively in one cell type or the other. (B) A heatmap of single-cell RNA sequencing data comparing five A cells and five B cells. While many genes are not detected, the aggregate signals across genes still provide sufficient information to differentiate between the two cell types, even when cells have perfectly mutually exclusive marker expression, as in the first two columns of cell type B. (C) A heatmap of single-cell data from unlabeled samples. Aggregate marker gene expression for each sample is plotted below the heatmap. Cell type identity can be inferred in noisy data because it is encoded in many genes.

signals can be modeled in low-dimensional space, the authors propose that expression data may also have this property. So just as images can be reconstructed from low dimensions thanks to high correlations between adjacent pixels, the authors argue that gene co-regulatory modules may be recoverable from high gene-gene correlations, effectively reducing the search space from 20 000 genes to a space of a few principal components that capture co-expressed gene modules. This is similar to the motivation behind the L1000 platform designed to measure the expression of 1000 'landmark' genes that recover a large fraction of information from the full transcriptome [34] and earlier work on expression imputation more broadly [35,36].

In a series of downsampling experiments, Heimberg *et al.* provide evidence that top principal components are robust to noise induced by signal loss, with robustness scaling with the proportion of variance explained. As such, they find that low-depth transcriptome coverage (~100 genes detected) is sufficient to characterize cell type differences that are represented within only a single principal component (glia versus neurons) but that higher depth (~1000 genes detected) is required to accurately recover cell types that differ along the top three principal components (between neuronal subtypes). These results nicely fit our intuitions about cell types and can be modeled to help researchers make decisions about experimental design. This low dimensionality also allows us to ground single-cell research in an area of expression analysis that has been of interest for decades: co-expression.

Linking Co-expression to Single-Cell Analysis

The observation that many genes jointly vary between cell types can be generalized to any source of conditional variation between samples, such as differences in age, treatment, or disease. Under any condition, genes that co-vary, or 'co-expressed' genes, can be identified by their significantly similar patterns of expression across samples, often assessed genome-wide between all possible gene pairs. Importantly, genes that are grouped by their expression profiles share molecular and biological functions, as shown in Eisen *et al.*'s seminal 1998 paper [37]. For example, members of the same protein complex, such as the proteasome, often have highly correlated gene expression. Co-expression links between genes are usefully visualized and analyzed as networks. While gene-gene networks often appear complex, they are motivated by the simple principle that genes with similar functions are preferentially connected [38]. As in single-cell analysis, the known functional groupings defined within these networks also imply lower dimensionality of transcriptional data relative to the number of genes. A natural question is whether the low dimensionality in single-cell data is directly linked to the low dimensionality implied by co-expression in bulk data.

This can be addressed by comparing co-expression networks built from bulk RNA-seq to those built from single-cell data. Here, the question is more precisely framed: Are the co-expression patterns observed in single-cell data unique, or do they overlap with the modules found in bulk RNA-seq networks? In an analysis of more than 400 bulk and single-cell co-expression networks our group found very similar results across the two data types [39], thus confirming a link between the low dimensionality of bulk and single-cell data. We also found that aggregating data across individual single-cell experiments strongly improved connectivity within known gene modules. Taken together, these two results support the strong conclusion that there exists a shared low-dimensional space that underlies cell identification across experiments and that these dimensions are observable in bulk co-expression. However, co-expression signals are strikingly weak in individual scRNA-seq datasets [40], regardless of the number of samples or the extent of cell type heterogeneity. This may be due to technical issues such as gene drop-out or to biological features such as transcriptional bursting that can

reduce the signal-to-noise ratio for cell type-relevant genes [41]. Even though cell type variation can be observed, a great deal of information is likely to be missing from any individual dataset.

To date, most co-expression applications in the single-cell field have been targeted toward correcting expression data for sample inference [42,43]. Eventually, gene-targeted evaluation will be the goal. Here again, prior experience from bulk analysis can help to guide us. We discuss this in more detail in the following section.

Co-expression Implies Collinearity: Lessons from Bulk Expression Analysis

As we have discussed, co-expression is inherent to expression data, provides low-dimensional properties, and makes characterization of samples robust, as in scRNA-seq. These are useful features, but it is important to understand their full implications to accurately model and interpret results. For this, we can learn from previous work to analyze gene expression data over the past decades, which highlighted the conceptual and statistical pitfalls that arise when co-expression is neglected, and genes are treated as independent variables. Gene collinearity complicates inference (Figure 2).

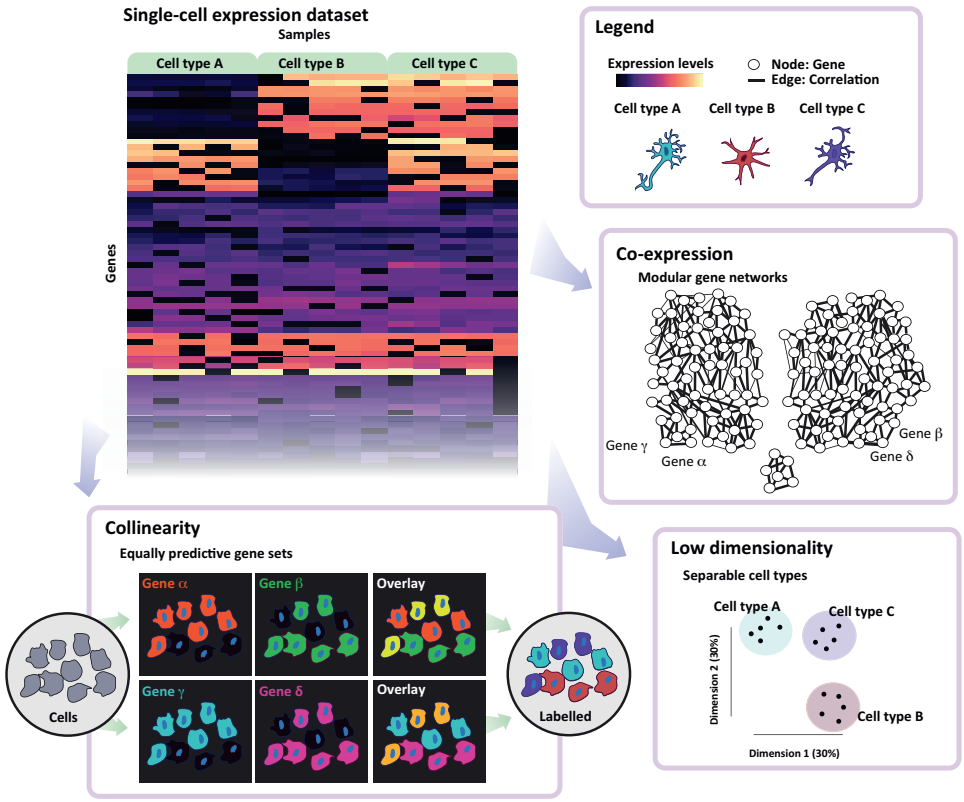

Figure 2. The Implications of Multi-gene Encoding of Cell Type. A single-cell expression dataset is shown in the top left of the schematic, where each row is a gene and each column is a cell. In this example, the three cell types (indicated by their distinct morphologies and colors) are distinguishable by their expression of two sets of genes: one set is expressed in cell types A and C and one set is expressed in cell types B and C. This has three corollaries (clockwise from right): (1) co-expression, where the genes that characterize cell types show correlated expression across samples, thus forming clustered modules; (2) low dimensionality, where cell types are easily separated in low-dimensional space; and (3) collinearity, where many gene pairs are equally predictive of cell type. This illustrates the importance of gene-gene relationships for cell identification and suggests caution when interpreting the significance of individual genes.

One area where this has been of particular importance is cancer subtyping. Similar to scRNA-seq applications, cancer researchers have aimed to define substructure among samples through unsupervised clustering approaches [44]. Rather than defining cell types, here the goal has been to define tumor subtypes and use these subtypes to predict patient-specific features such as prognosis or drug response. In breast cancer, for example, only a subset of patients respond to anti-estrogen therapy [45]. Could microarray technology determine the molecular characteristics of tumors that are unresponsive to known therapies? To an extent, the answer was yes: clustering did reveal structure among tumors, and the identified marker genes predicted patient survival [46,47]. Yet it came as a surprise when it was discovered that almost any set of genes could predict outcome [48]. This finding can readily be explained by co-expression. If co-expression is common, then each marker gene can be regarded as a stand-in for some module of genes with related expression profiles. If these modules are large (i.e., expression data are low dimensional), then a random set of genes will sample from these modules and therefore capture much of the performance historically observed by any marker set. Concretely, we might imagine that only two gene programs drive cancer progression: increased cell proliferation and reduced apoptosis. If 10% of genes were involved in each process, then randomly sampling even a hundred genes leaves one quite likely to obtain useful markers. In such a scenario, the exact genes chosen as markers are only a broad clue into the molecular processes important for phenotype.

These examples illustrate that it is critical to determine whether a gene is co-expressed or independent to correctly interpret its significance, and suggest caution for the interpretation of markers derived from single-cell data. We discuss this and related issues in more detail below.

Limitations of Co-expression for scRNA-Seq

There is a strong expectation that single-cell RNA-seq will continue to yield insights into the cellular composition of tissues and certainly in those that have yet to be profiled. The evidence from the single-cell literature supports the notion that many cell types can be identified on the basis of broad changes in gene co-expression and that this underlies the success of droplet-based, high-throughput sequencing approaches that only shallowly sample from the total RNA pool. Yet there are clear limitations that arise from relying on co-expression as our saving grace.

First, if cell types do not conform to expectations of characteristically broad changes in gene co-expression, then current approaches will fail. In the simplest case, where cell types differ from their nearest relatives via the expression of only a small number of genes, we will be hard pressed to find them if there are too few genes captured per cell or if too few cells are sampled. This has nicely been discussed by Torre *et al.* in their recent comparison of scRNA-seq and single-molecule RNA fluorescence *in situ* hybridization (FISH) [49]. When designing single-cell experiments, it is therefore important to have a clear hypothesis and goal in mind, as one size will not fit all.

Second, by contrast, if cell identification truly does require the co-expression of hundreds or thousands of genes, this creates an important conceptual problem, as we have discussed. In this case, nearly all genes are potential markers. Collinearity between genes makes it difficult at least, and meaningless at worst, to prioritize one gene as the most significant for cell function [50]. For this reason, although we may be able to identify cells from their co-expression patterns, we may not be characterizing them very much. Intuitively, if only a few randomly chosen genes are ever necessary to identify a cell, mechanistic understandings of cell identity are likely to remain out of reach using current data. However, what collinearity may lack in explanatory power on a per-gene basis, it may compensate for in practical utility. In addition to

enabling sparse representations of biological processes that describe cell types or cell states [51], taking advantage of co-expression would allow researchers to tap into the same information that is encoded in multiple genes by targeting the select few with pre-existing tools such as Cre-driver lines or monoclonal antibodies. As long as the notion of a marker is clearly defined as 'one of many equivalents', we should be safe from making conceptual errors when interpreting results.

And third, we must note that despite its utility for cell identification, gene co-expression cannot be assumed to make up for all of the missing data. Currently scRNA-seq is not a transcriptome-wide method but rather a method to sample the transcriptome. Estimates from single-molecule FISH suggest that only ~10–20% of the transcriptome is assayed with scRNA-seq [49,52]. The investigation of regulatory and network information from single cells is compromised by this inherent low coverage. If broad co-expression is exploited to impute missing data, as in the MAGIC pipeline [53] or with autoencoder approaches [54,55], we are unlikely to detect subtle changes from expectation, which may be necessary as the field progresses. Indeed, after imputation the assessment of gene-gene similarity is partially circular, exaggerating the similarity of gene expression profiles and the apparent significance of resulting co-expression relationships [42]. Unless transcriptome coverage increases dramatically, it will be necessary to return to pooled samples or targeted assays to make any statement about individual genes. Of course, this already occurs through cluster-based differential expression and with newer approaches to smooth single-cell expression profiles by averaging across nearest neighbors [56,57]. Whether it will be possible to define co-expression that is both genome-wide and driven by variation between single cells remains to be seen.

Concluding Remarks and Future Perspectives

Early single-cell experiments have been remarkably successful thanks to gene co-expression within individual cells and cell types: co-expression has been our saving grace. Yet the field's reliance on gene-gene covariation has been largely implicit, putting us at risk of misinterpreting results and transforming co-expression into our original sin. To move the field forward it will be important to evaluate covariation directly, as this will provide greater insight into the successes and the failures of scRNA-seq, and will contribute to our understanding of cell types and cell states (see Outstanding Questions). Indeed, over-reliance on global variance measures may cause us to miss rare events, such as dysregulation limited to a small total number of genes. As a start, we suggest that low-dimensional plots of single data should report what the dimensions represent. The trend of displaying cells with t-distributed stochastic neighbor embedding has been a roadblock to obtaining replicable features since the method does not consistently represent cluster variability or distances [58]. More clarity in underlying methods will improve our collective intuition about the quality of single-cell data and clustering solutions.

Understanding the mechanisms of cell identity, as opposed to just their correlates, will ultimately require controlled perturbation experiments to characterize gene drivers of cell phenotype. This is already beginning to be possible in a high-throughput way through Perturb-seq [59], CRISP-seq [60] and CROP-seq [61], and we are excited to see future application and refinement of these techniques. We note that while broad co-expression underlies the success and replicability of current cell clusters, new data may reveal that other classes of variation are important for cell type characterization. Combining scRNA-seq with other techniques such as multiplexed FISH, epitope barcoding [62,63], and for neurons, projection mapping [64,65] and patch-clamp recording [66–68], may show that expression is low dimensional even when cell identity is not, thus limiting the resolution of unsupervised techniques based on expression data alone. External validation, beyond expression data, will be the

Outstanding Questions

Are all cell type differences distributed across many genes? Most current expression protocols rely on this, but some important heterogeneity between cells might be reflected in the differential expression of just a few genes or from differential isoform usage.

What are the dimensions of cell identity? How well do these generalize across all cell types, conditions, or species? Under what conditions does differential co-expression occur? Individual studies may be low dimensional only because they each sample from a relatively narrow space of existing cell phenotypes.

How useful are discrete transcriptional types for explaining and predicting cell function? Continuous gradients naturally occur in tissues and will also be reflected by co-expression. Are cells better modeled by continuous processes? How would this help guide further experimentation?

What genes are most important for determining cell identity? Marker genes are necessary for targeting cell types but may not be required for cell phenotype or function. What genes are co-expressed with known markers? How much do known markers explain results?

ultimate arbiter as researchers begin to unravel whether the cell types identified through large-scale gene co-expression are useful for understanding the organization and function of biological systems.

Acknowledgments

We would like to thank Paul Pavlidis, Tony Zador, David Jackson, Shreejoy Tripathy, Anirban Paul and Sara Ballouz for their helpful feedback on this manuscript. We would also like to thank Sara Ballouz for drawing the figures for this article. M.C. is supported by a postdoctoral fellowship from the National Institute of Mental Health (F32-MH114501) and a NARSAD Young Investigator grant from the Brain & Behavior Research Foundation (26140). J.G. is supported by grants from the National Institute of Mental Health (R01-MH113005-02) and from the National Library of Medicine (R01-LM012736-01). Any opinions, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Brain & Behavior Research Foundation or the National Institutes of Health.

References

1. Svensson, V. *et al.* (2018) Exponential scaling of single-cell RNA-seq in the past decade. *Nat. Protoc.* 13, 599–604
2. Cao, J. *et al.* (2017) Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* 357, 661–667
3. Plass, M. *et al.* (2018) Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. *Science* 360
4. Fincher, C.T. *et al.* (2018) Cell type transcriptome atlas for the planarian *Schmidtea mediterranea*. *Science* 360, eaaq1736
5. McKenna, A. *et al.* (2016) Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science* 353, aaf7907
6. Zappia, L. *et al.* (2018) Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. *PLoS Comput. Biol.* 14, e1006245
7. Stegle, O. *et al.* (2015) Computational and analytical challenges in single-cell transcriptomics. *Nat. Rev. Genet.* 16, 133–145
8. Wagner, A. *et al.* (2016) Revealing the vectors of cellular identity with single-cell genomics. *Nat. Biotechnol.* 34, 1145–1160
9. Shapiro, E. *et al.* (2013) Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat. Rev. Genet.* 14, 618–630
10. Zeisel, A. *et al.* (2015) Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* 347, 1138–1142
11. Usoskin, D. *et al.* (2015) Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. *Nat. Neurosci.* 18, 145–153
12. Macosko, E.Z. *et al.* (2015) Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* 161, 1202–1214
13. Treutlein, B. *et al.* (2014) Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* 509, 371–375
14. Regev, A. *et al.* (2017) The Human Cell Atlas. *Elife* 6, e27041
15. Lin, P. *et al.* (2017) CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data. *Genome Biol.* 18, 59
16. Pierson, E. and Yau, C. (2015) ZIFA: dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biol.* 16, 241
17. Risso, D. *et al.* (2018) A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat. Commun.* 9, 284
18. Hicks, S.C. *et al.* (2017) Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics* Published online November 6, 2017. http://dx.doi.org/10.1093/biostatistics/kxx053
19. Buettner, F. *et al.* (2015) Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.* 33, 155–160
20. McDavid, A. *et al.* (2016) The contribution of cell cycle to heterogeneity in single-cell RNA-seq data. *Nat. Biotechnol.* 34, 591–593
21. Vallejos, C.A. *et al.* (2017) Normalizing single-cell RNA sequencing data: challenges and opportunities. *Nat. Methods* 14, 565–571
22. Habib, N. *et al.* (2016) Div-seq: single-nucleus RNA-seq reveals dynamics of rare adult newborn neurons. *Science* 353, 925–928
23. La Manno, G. *et al.* (2016) Molecular diversity of midbrain development in mouse, human, and stem cells. *Cell* 167, 566–580.e19
24. Shekhar, K. *et al.* (2016) Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics. *Cell* 166, 1308–1323.e30
25. Lin, C. *et al.* (2017) Using neural networks for reducing the dimensions of single-cell RNA-seq data. *Nucleic Acids Res.* 45, e156
26. Butler, A. *et al.* (2018) Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* 36, 411–420
27. Haghverdi, L. *et al.* (2018) Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* 36, 421–427
28. Han, X. *et al.* (2018) Mapping the mouse cell atlas by microwell-seq. *Cell* 172, 1091–1107.e17
29. Kiselev, V.Y. *et al.* (2018) scmap: projection of single-cell RNA-seq data across data sets. *Nat. Methods* 15, 359–362
30. Crow, M. *et al.* (2018) Characterizing the replicability of cell types defined by single cell RNA-sequencing data using MetaNeighbor. *Nat. Commun.* 9, 884
31. Jaitin, D.A. *et al.* (2014) Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* 343, 776–779
32. Pollen, A.A. *et al.* (2014) Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat. Biotechnol.* 32, 1053–1058
33. Heimberg, G. *et al.* (2016) Low dimensionality in gene expression data enables the accurate extraction of transcriptional programs from shallow sequencing. *Cell Syst.* 2, 239–250
34. Subramanian, A. *et al.* (2017) A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell* 171, 1437–1452.e17
35. Troyanskaya, O. *et al.* (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics* 17, 520–525
36. Donner, Y. *et al.* (2012) Imputing gene expression from selectively reduced probe sets. *Nat. Methods* 9, 1120–1125
37. Eisen, M.B. *et al.* (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U. S. A.* 95, 14863–14868

38. Oliver, S. (2000) Guilt-by-association goes global. *Nature* 403, 601–603
39. Crow, M. *et al.* (2016) Exploiting single-cell expression to characterize co-expression replicability. *Genome Biol.* 17, 101
40. Chen, S. and Mar, J.C. (2018) Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. *BMC Bioinform.* 19, 232
41. Raj, A. and van Oudenaarden, A. (2008) Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell* 135, 216–226
42. Huang, M. *et al.* (2018) SAVER: gene expression recovery for single-cell RNA sequencing. *Nat. Methods* 15, 539–542
43. Azizi, E. *et al.* (2018) Bayesian inference for single-cell clustering and imputing. *Genom. Comput. Biol.* 3, e46
44. Song, Q. *et al.* (2015) Cancer classification in the genomic era: five contemporary problems. *Hum. Genomics* 9, 27
45. Polyak, K. (2011) Heterogeneity in breast cancer. *J. Clin. Invest.* 121, 3786–3788
46. Perou, C.M. *et al.* (2000) Molecular portraits of human breast tumours. *Nature* 406, 747–752
47. Sørlie, T. *et al.* (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci. U. S. A.* 98, 10869–10874
48. Venet, D. *et al.* (2011) Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS Comput. Biol.* 7, e1002240
49. Torre, E. *et al.* (2018) Rare cell detection by single-cell RNA sequencing as guided by single-molecule RNA FISH. *Cell Syst.* 6, 171–179.e5
50. Tripathy, S.J. *et al.* (2017) Transcriptomic correlates of neuron electrophysiological diversity. *PLoS Comput. Biol.* 13, e1005814
51. Furchtgott, L.A. *et al.* (2017) Discovering sparse transcription factor codes for cell states and state transitions during development. *Elife* 6, e20488
52. Shalek, A.K. *et al.* (2013) Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* 498, 236–240
53. van Dijk, D. *et al.* (2017) MAGIC: a diffusion-based imputation method reveals gene-gene interactions in single-cell RNA-sequencing data. *bioRxiv* Published online Februry 25, 2017. http://dx.doi.org/10.1101/111591
54. Lopez, R. *et al.* (2017) A deep generative model for gene expression profiles from single-cell RNA sequencing. *arXiv* preprint arXiv:1709.02082
55. Eraslan, G. *et al.* (2018) Single cell RNA-seq denoising using a deep count autoencoder. *bioRxiv* Published online April 13, 2018. http://dx.doi.org/10.1101/300681
56. Wagner, F. *et al.* (2018) K-nearest neighbor smoothing for high-throughput single-cell RNA-Seq data. *bioRxiv* Published on February 25, 2017. http://dx.doi.org/10.1101/217737
57. La Manno, G. *et al.* (2017) RNA velocity in single cells. *bioRxiv* http://dx.doi.org/10.1101/206052 Publsihed on October 19, 2017
58. van der Maaten, L. and Hinton, G. (2008) Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9, 2579–2605
59. Dixit, A. *et al.* (2016) Perturb-seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell* 167, 1853–1866.e17
60. Jaitin, D.A. *et al.* (2016) Dissecting immune circuits by linking CRISPR-pooled screens with single-cell RNA-seq. *Cell* 167, 1883–1896.e15
61. Datlinger, P. *et al.* (2017) Pooled CRISPR screening with single-cell transcriptome readout. *Nat. Methods* 14, 297–301
62. Stoeckius, M. *et al.* (2017) Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* 14, 865–868
63. Peterson, V.M. *et al.* (2017) Multiplexed quantification of proteins and transcripts in single cells. *Nat. Biotechnol.* 35, 936–939
64. Tasic, B. *et al.* (2016) Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nat. Neurosci.* 19, 335–346
65. Chen, X. *et al.* (2018) Efficient *in situ* barcode sequencing using padlock probe-based BaristaSeq. *Nucleic Acids Res.* 46, e22
66. Fuzik, J. *et al.* (2016) Integration of electrophysiological recordings with single-cell RNA-seq data identifies neuronal subtypes. *Nat. Biotechnol.* 34, 175–183
67. Cadwell, C.R. *et al.* (2016) Electrophysiological, transcriptomic and morphologic profiling of single neurons using Patch-seq. *Nat. Biotechnol.* 34, 199–203
68. Foldy, C. *et al.* (2016) Single-cell RNAseq reveals cell adhesion molecule profiles in electrophysiologically defined neurons. *Proc. Natl. Acad. Sci. U. S. A.* 113, E5222–E5231
69. Grun, D. *et al.* (2015) Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature* 525, 251–255
70. Kiselev, V.Y. *et al.* (2017) SC3: consensus clustering of single-cell RNA-seq data. *Nat. Methods* 14, 483–486
71. Perraudeau, F. *et al.* (2017) Bioconductor workflow for single-cell RNA sequencing: normalization, dimensionality reduction, clustering, and lineage inference. *F1000Research* 6, 1158