Entering the era of single-cell transcriptomics in biology and medicine

Rickard Sandberg

Recent technical advances have enabled RNA sequencing (RNA-seq) in single cells. Exploratory studies have already led to insights into the dynamics of differentiation, cellular responses to stimulation and the stochastic nature of transcription. We are entering an era of single-cell transcriptomics that holds promise to substantially impact biology and medicine.

Our notion of transcriptomes has been forged mainly by population-level observations that have been the mainstream in biology over the last two decades. We are used to thinking about differences in expression in terms of graded or subtle fold changes when comparing data across entire tissues or conditions. But the actual differences between cells may be far larger. Subsets of cells may experience dramatic changes that are averaged out or diluted by the presence of a large number of nonresponsive cells. In fact, it was shown over 60 years ago that inductive cues often result in all-or-none responses in single cells but these responses are observed as a gradual increase when quantified across the population[1].

It is clear that assessing gene expression in single cells is critical to better understand cellular behaviors and compositions in developing, adult and pathological tissues. To this end, a long-standing goal has been to enable genome-wide RNA profiling, or transcriptomics, in single cells[2,3]. Only recently has the technology matured so that biologically meaningful differences can be robustly detected with single-cell RNA-seq. Detailed protocols[4–6] for sequencing library preparations and the introduction of commercial automation (for example, Fluidigm C1) have lowered the barriers for researchers to access these methods. Widespread adoption of these techniques will have a major impact on our understanding and appreciation of cellular states, the nature of transcription and gene regulation, and our ability to characterize pathological states in disease.

Rickard Sandberg is at the Ludwig Institute for Cancer Research, Stockholm, Sweden, and Department of Cell and Molecular Biology, Karolinska Institutet, Stockholm, Sweden.
e-mail: rickard.sandberg@ki.se

Above the noise

Single-cell transcriptomics relies on the reverse transcription of RNA to complementary DNA and subsequent amplification by PCR or *in vitro* transcription before deep sequencing—procedures prone to losses or biases. The biases are exaggerated by the need for very high amplification from the small amounts of RNA found in an individual cell. Although technical noise confounds precise measurements of low-abundance transcripts, modern protocols have progressed to the point that single-cell measurements are rich in biological information. For example, a recurrent theme in single-cell transcriptome studies is that cells reliably group by their cell type or state when subjected to unsupervised clustering[7–10]. Gene expression associated with cell identity or developmental stages thus has a stronger signal than technical noise or biological variability related to dynamic processes such as phase of the cell cycle. Moreover, the power to detect meaningful biological differences from single-cell data is demonstrated by the identification of hundreds to thousands of genes with differences in abundances between cell types[7,9]. Recent refinements will improve the signal-to-noise ratio even further by enhancing the efficiencies of reverse transcription and PCR[11] or applying molecular barcoding strategies that control for amplification bias[12].

Challenges in single-cell transcriptomics

Currently available single-cell RNA-seq methods were developed with several different objectives. Full-length transcripts can be profiled, such that sequence reads cover the entire gene to quantify both gene and transcript isoforms and also monitor single-nucleotide polymorphisms and mutations[9,11]. In contrast, tag-based sequencing of 5′ or 3′ ends[10,13] provides only an estimate of transcript abundance at the cost of coverage across gene structures but allows the assay to be scaled up and combined with molecule counting[12].

The unified goal in the field is to develop cost-effective, high-throughput methods that detect all RNA present inside the cell at full-length RNA coverage. Lowering RNA losses and enhancing the conversion of RNA to cDNA before amplification are areas where further development would boost RNA detection. Another important goal is to augment procedures for the dissociation, sorting and picking of individual cells[14] so that complex tissues can be dissociated into single-cell suspension without inducing changes in gene expression related to cell handling or picking. Finally, simultaneous detection of poly(A)$^+$ and poly(A)$^-$ RNA, irrespective of transcript length, and RNA modifications (for example, m^{6}A in ref. 15) are also desirable features for future development.

One of the mind-boggling features of transcription that only becomes apparent in single-cell analysis is that expression of a gene that is reliably detected in a population may be anywhere from absent, to low, to

high in a given cell because of random fluctuations. Such variability may be explained by models that describe transcription as occurring in discrete bursts[16] driven by stochastic molecular processes. The stochastic nature of transcription has been studied in greatest detail in prokaryotes and unicellular eukaryotes[16], but more and more lines of evidence point to similar phenomena in mammalian cells[17,18]. We must therefore take into account such transcriptional behavior in our strategies for analyzing single-cell transcriptome data and in our biological interpretation of the results. For example, standard differential expression tests might not be suitable for single-cell data that contain a fair number of cells with no detectable expression. Indeed, new tests have been proposed[19] that combine differences in transcript abundance with differences in the fraction of cells with expression.

Single-cell transcriptome studies to date require cells in suspension (for example, dissociated tissues or cultures) so that the spatial organization of the population is often lost, unless cells had been picked from defined areas. Spatial information can be recovered to some extent through RNA *in situ* hybridization analyses of marker genes for identified cell types, allowing cell type–specific expression profiles to be projected onto complex tissue structures. However, methods that simultaneously capture spatial structures and transcriptome-wide profiles at single-cell resolution are being developed but have yet to be described (for example, building on *in situ* sequencing or array-based multiplexing strategies). The ability to perform spatial single-cell transcriptomics on developing, adult or pathological tissues promises to dramatically elevate our understanding of life and disease, revealing the transcriptomes related to specific states of intercellular communication, polarity formation and local gradients.

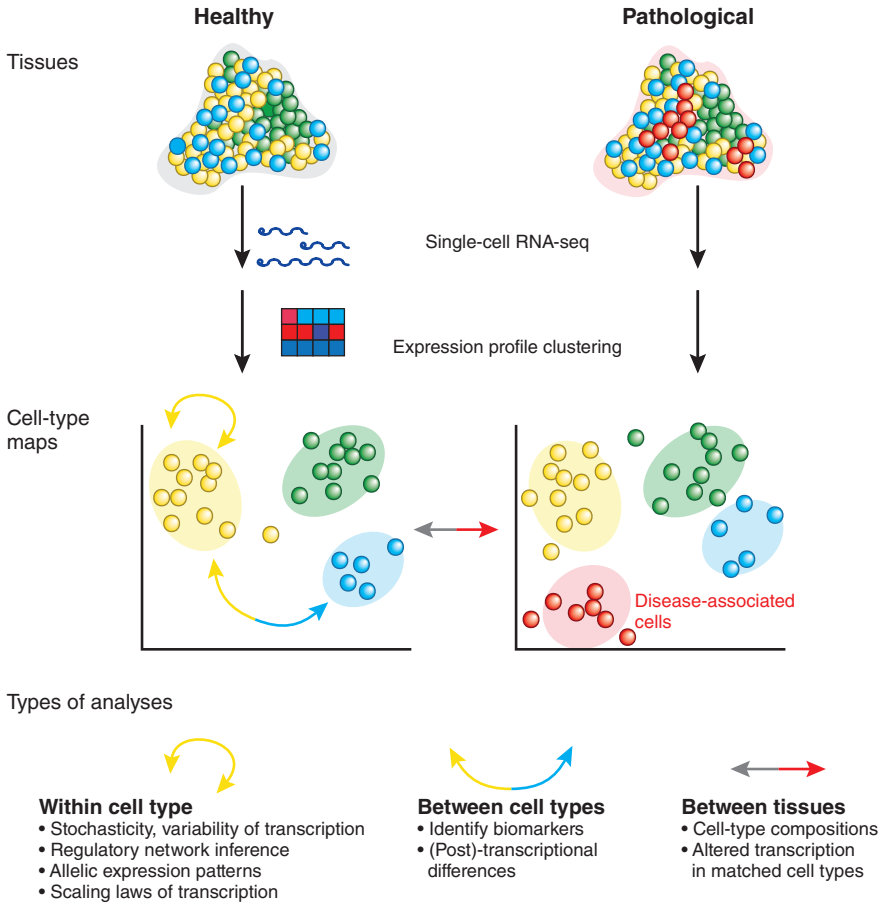

Figure 1 | Single-cell transcriptome analyses of tissues and cell types. Cells from a healthy or pathological tissue are dissociated, analyzed independently with single-cell RNA-seq and clustered based on their gene expression profiles. Clustering of cells reveals a cell-type map that can be used to assess the composition of the tissue including the identification of new cell types or subtypes. These rich data can be used to address many questions of gene expression and regulation within or between cell types and between tissues.

Implications for biology

The measurement of gene expression in single cells will revolutionize our understanding of gene regulation and resolve many longstanding debates in biology. Cells cluster by cell type or developmental state when grouped according to their expression profiles[7–10]. Thus, expression-based clustering allows for the unbiased reconstruction or 'reverse engineering' of cell types in any population or tissue after sequencing enough individual cells (**Fig. 1**). If the sampling of cells is extensive and sufficiently free from biases, such clustering can reveal all cell types present, including new ones. All cells in a cluster can also be used to derive robust cell-type expression profiles, again in a data-driven manner and without previous knowledge of which marker genes define a tissue or cell type. Single-cell profiling of RNAs is therefore the first method that could lay a foundation for a quantitative, data-driven classification of cell types.

Single-cell transcriptomics will also enable high-resolution transcriptional maps of both stable and transient cellular states during differentiation or reprogramming. Important for these aims is to sample sufficient individual cells that span the entire process, so that analyses can later zoom in on the subset of cells at critical bifurcation points of differentiation. The sample size should reflect how often cell types or events are expected to occur. Also, it is debated to what extent the human genome is transcribed, as several studies have identified very rare transcripts (for example, those present in one copy per 10,000 cells)[20]. These transcripts could either be expressed at high levels in rare cells (for example, ten copies in one of 100,000 cells) or have low (leaky) expression in a larger subset of cells. Analyses across hundreds or thousands of individual cells will likely resolve these questions and improve our understanding of cellular transcriptional landscapes and regulatory networks.

RNA-seq analyses across human tissues and cell populations have demonstrated the pervasive use of RNA processing to diversify the transcriptome and the proteome[21]. A large fraction of differences are subtle when comparing tissues, but it is possible that patterns of alternative splicing, polyadenylation and transcription start-site usage will have a more bimodal (on or off) distribution

at single-cell resolution, as suggested by a pioneering study on single cells[22]. Studies of the regulation of alternative polyadenylation have revealed a general shortening of 3′ untranslated regions in more highly proliferating cells[23] and in transformed cells *in vitro*[24]. Analyses of *in vivo* tumors would benefit greatly from single-cell RNA-seq to separately extract transcript abundance and isoform information from the mixture of transformed cells, stroma and other infiltrating cells. Single-cell transcriptomics of dissociated tumor and healthy tissues will enable the precise identification of mRNA isoforms that are important for the transformed state.

Implications for medicine

Transcriptomic approaches in medicine are often based on comparing pathological with matched healthy tissue[25] or analyzing a large number of pathological tissues to find subclassifications[26]. Cancer tissues are often characterized by changes in both cellular compositions (for example, infiltrating immune cells) and alterations in gene-expression programs in both the transformed cells and the surrounding stroma. Thus, observations at the tissue level contain several differential expression profiles superimposed on top of each other. High-throughput single-cell analysis of pathological tissues would simultaneously monitor changes in cellular composition (based on clustering) and associated gene expression profiles[27]. Comparisons could then be made between specific cell types observed in both the healthy and pathological tissues to reveal more precise gene expression programs of disease (**Fig. 1**). However, regional variations in cellular composition may necessitate sampling in multiple regions from the same tumor[28].

Areas of research that stand to benefit in particular from single-cell transcriptomics are those in which the clinically relevant cells are too rare to be studied using population-level techniques. For example, only a few circulating tumor cells (CTCs) are typically present in a milliliter of blood, which has precluded their genome-wide profiling. Two pioneering studies demonstrated the utility of single-cell RNA-seq analyses of CTCs of melanoma[9] or pancreatic[29] origin, as the transcriptome profiles both validated the cellular isolation procedure and were used to identify alterations in the gene expression programs. Single-cell RNA-seq with full-length transcript coverage[11] should enable simultaneous measurement of gene expression programs and detection of mutations that arise in the tumor through analyses of the CTCs. Transcriptome analyses of single CTCs is a noninvasive strategy to select treatment based on the inferred mutations[30] and also to monitor the development of drug resistance. It is time to determine to what extent CTC transcriptome profiling can be a future method for cancer diagnostics and treatment selection, and provide biomarkers for future therapies targeting CTCs.

Outlook

As we are just entering an era of single-cell transcriptomics, the near future will likely unravel many surprising and new characteristics of transcriptomes. It will be interesting to investigate whether certain scaling laws exist between RNA abundance profiles and cellular phenotypes such as cell or nucleus size. For example, to maintain protein concentrations inside membranes or subcellular compartments in cells of varying size, different abundances would be needed as volume and area scale differently with cell size. Sets of genes are likely to scale with characteristics such as plasma or nuclear membrane area, cytoplasmic volume and nuclear volume. Only with such knowledge at hand can we begin to resolve how cellular heterogeneity and cell type composition confound population-level transcriptome analyses. For example, comparisons of two tissues composed of cells of differing size might reveal differences in expression related to size, rather than the differences of interest. A better understanding of single-cell expression profiles will also provide a more rational basis for the design of future studies at the most appropriate level of resolution (for example, tissue, cell type, single cell or combinations of the three).

With the maturation of single-cell transcriptomics, I expect that studies of gene expression and regulation in single cells will boom in the coming years and the research community will soon obtain precise transcript-isoform quantifications across hundreds of thousands to even millions of individual cells. This information will answer many outstanding questions (**Fig. 1**) and lay the foundation for a quantitative definition of cell types and their variation in homogeneous as well as heterogeneous cell populations. Based on this knowledge it will become feasible to determine the transcriptome profiles of nearly all cell types in complex multicellular organisms. Single-cell profiling will also dramatically improve gene-regulatory network inferences[31], as the vast amounts of single-cell profiles are bona fide biological perturbations that should improve the power of inference.

ACKNOWLEDGMENTS

I am grateful to B. Reinius, J. Muhr, J. Holmberg and members of the Sandberg laboratory for comments on the manuscript. R.S. is supported by the European Research Council Starting grant 243066, the Swedish Foundation for Strategic Research FFL4 and the Swedish Research Council grant 2008-4562.

COMPETING FINANCIAL INTERESTS

The author declares no competing financial interests.

1. Novick, A. & Weiner, M. *Proc. Natl. Acad. Sci. USA* **43**, 553–566 (1957).
2. Brady, G., Barbara, M. & Iscove, N.N. *Methods Mol. Biol.* **2**, 17–25 (1990).
3. Eberwine, J. *et al. Proc. Natl. Acad. Sci. USA* **89**, 3010–3014 (1992).
4. Islam, S. *et al. Nat. Protoc.* **7**, 813–828 (2012).
5. Tang, F. *et al. Nat. Protoc.* **5**, 516–535 (2010).
6. Picelli, S. *et al. Nat. Protocols*, doi:10.1038/nprot.2014.006 (2 January 2014).
7. Yan, L. *et al. Nat. Struct. Mol. Biol.* **20**, 1131–1139 (2013).
8. Tang, F. *et al. Cell Stem Cell* **6**, 468–478 (2010).
9. Ramsköld, D. *et al. Nat. Biotechnol.* **30**, 777–782 (2012).
10. Islam, S. *et al. Genome Res.* **21**, 1160–1167 (2011).
11. Picelli, S. *et al. Nat. Methods* **10**, 1096–1098 (2013).
12. Kivioja, T. *et al. Nat. Methods* **9**, 72–74 (2012).
13. Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. *Cell Rep* **2**, 666–673 (2012).
14. Shapiro, E., Biezuner, T. & Linnarsson, S. *Nat. Rev. Genet.* **14**, 618–630 (2013).
15. Dominissini, D. *et al. Nature* **485**, 201–206 (2012).
16. Raj, A. & van Oudenaarden, A. *Cell* **135**, 216–226 (2008).
17. Chubb, J.R., Trcek, T., Shenoy, S.M. & Singer, R.H. *Curr. Biol.* **16**, 1018–1025 (2006).
18. Wills, Q.F. *et al. Nat. Biotechnol.* **31**, 748–752 (2013).
19. McDavid, A. *et al. Bioinformatics* **29**, 461–467 (2013).
20. Mercer, T.R. *et al. Nat. Biotechnol.* **30**, 99–104 (2012).
21. Wang, E.T. *et al. Nature* **456**, 470–476 (2008).
22. Shalek, A.K. *et al. Nature* **498**, 236–240 (2013).
23. Sandberg, R., Neilson, J.R., Sarma, A., Sharp, P.A. & Burge, C.B. *Science* **320**, 1643–1647 (2008).
24. Mayr, C. & Bartel, D.P. *Cell* **138**, 673–684 (2009).
25. Rhodes, D.R. *et al. Proc. Natl. Acad. Sci. USA* **101**, 9309–9314 (2004).
26. Golub, T.R. *et al. Science* **286**, 531–537 (1999).
27. Dalerba, P. *et al. Nat. Biotechnol.* **29**, 1120–1127 (2011).
28. Gerlinger, M. *et al. N. Engl. J. Med.* **366**, 883–892 (2012).
29. Yu, M. *et al. Nature* **487**, 510–513 (2012).
30. Vogelstein, B. *et al. Science* **339**, 1546–1558 (2013).
31. Kim, H.D., Shay, T., O'Shea, E.K. & Regev, A. *Science* **325**, 429–432 (2009).