Optimizing biological inferences from single-cell data

Two recent articles in *Nature Reviews Genetics* discuss the exciting opportunities of single-cell omics studies but also highlight the importance of appropriate data analysis strategies.

Over the past decade, several transformative technologies have enabled the profiling of increasing numbers of cells to generate omics data — primarily transcriptomes. Until recently, profiling a few hundred cells was considered a tour-de-force, yet the latest implementations of single-cell profiling now enable several hundreds of thousands (or even a few million)[1] cells to be characterized. As with various big-data fields, a key challenge is making sense of these catalogues of data, and two recent articles[2,3] in *Nature Reviews Genetics* discuss types of approaches that can be used.

Kiselev, Andrews and Hemberg[2] discuss clustering, which has rapidly established itself as a mainstay of single-cell analysis. Here, the rationale is that cells can be classified into groups that share similar profiles (such as transcriptomes), which represent the constituent cell types within the tissue being studied. The ubiquity of such analyses should not breed complacency in the way the methods are applied. As the authors point out, the range of available clustering methods have different algorithmic underpinnings, and different user parameters can skew the resulting clusters. Furthermore, there is no 'one-size-fits-all' solution: benchmarking studies have shown that no single tool outperforms all others across different biological applications and data types. Thus, clustering methods and their parameters should be carefully chosen to be suitable for the specific research question and data set.

When interpreting clustered single-cell data, the authors stress the need to consider the underlying biology of the system. With growing abilities to expand cellular throughput, there is now the opportunity to take an increasingly fractal view of cell types. Within traditionally defined cell types, single-cell profiling frequently reveals distinct subpopulations that can account for known functional heterogeneity. The deeper researchers look into cell populations at finer granularity, the more they can identify hierarchies of cell clusters within the subpopulations, and then within the sub-subpopulations, and so on. But when is deep enough? When does a new cluster represent a meaningfully distinct cell type? When might it merely represent heterogeneity within a larger cell type population? These are questions the field is currently grappling with.

“When does a new cluster represent a meaningfully distinct cell type?”

Beyond improvements in cellular throughput, there is also a drive to enhance the richness of the single-cell omics data produced. As Stuart and Satija[3] discuss, RNA sequencing (RNA-seq) data sets are made more informative by integrating them with other data from the cells, such as protein expression data, genome sequence or information on the spatial context of the cells in their original tissue. They describe a range of strategies for enabling multilayered data generation from the same cells. This can involve fractionating the single-cell lysates to capture and analyse the DNA, RNA or protein content separately. An alternative approach involves adopting poly(A)-labelled reagents — reporting on cell surface protein expression, CRISPR-based cellular perturbations or cell lineage relationships — that are analysed alongside regular poly(A) transcripts in RNA-seq workflows.

Single-cell omics data can also be enhanced retrospectively by integrating with other single-cell data sets collected on similar cell samples. Here, Stuart and Satija[3] describe the considerable and non-trivial analytical challenges involved in integrating single-cell data sets, not just in making data formats compatible — such as across omics types, or methodological protocols and platforms — but also in responsibly controlling for artefacts such as batch effects.

These integrative analyses are major components of current international Cell Atlas projects, which aim to characterize the cellular composition of organisms in unprecedented detail.

Single-cell omics technologies continue their awe-inspiring improvements in throughput and sensitivity, at ever-decreasing costs. Despite these distractingly impressive technological capabilities, these two articles serve as a reminder that thought and care still need to be invested in the appropriate analysis of the data, specifically that algorithmic tools are chosen wisely and data interpreted responsibly.

1. Cao, J. et al. *Nature* **566**, 496–502 (2019).
2. Kiselev, V. Y., Andrews, T. S. & Hemberg, M. *Nat. Rev. Genet.* https://doi.org/10.1038/s41576-018-0088-9 (2019).
3. Stuart, T. & Satija, R. *Nat. Rev. Genet.* https://doi.org/10.1038/s41576-019-0093-7 (2019).