https://doi.org/10.1038/s41592-024-02263-0

scRNA-seq: oh, the joys

To those who seek transcriptomic information at high resolution, scale and throughput, single-cell RNA sequencing brings the data. Scientists share tips and future plans as they reflect on the method's rise to stardom.

By Vivien Marx

"It has been incredible to watch the growth of this field over the last decade," says Rahul Satija, a researcher at the New York Genome Center. Thousands of labs are using single-cell sequencing to profile cells across a wide variety of organs and organisms, he says. It's now a routine method to measure what cells are doing. scRNA-seq experiments pose "a whole bunch of really exciting, conceptually new problems," says University of Washington researcher Cole Trapnell, who is also part of the Brotman Baty Institute, a collaboration between the University of Washington, Seattle Children's Hospital and the Fred Hutchinson Cancer Center.

The rapid rise of single-cell RNA sequencing (scRNA-seq) means researchers can find reviews[1–3] and evolving resources such as *Single-Cell Best Practices*. The book, co-authored by members of the Single-Cell Best Practices Consortium, addresses newcomers and advanced professionals alike. Book contributors can join through a Jupyter Notebook.

Satija, a Duke University alumnus and basketball fan, thinks back to a moment almost exactly 13 years ago. "I had the game playing on one monitor while mapping our very first sequencing results," he says. Duke was playing its Sweet Sixteen basketball game in the annual March Madness National Collegiate Athletic Association tournament. At the time, Satija was working with Alex Shalek and Joshua Levin in Aviv Regev's lab at the Broad Institute of MIT and Harvard. Duke lost the game, and it may be best to not linger on that upsetting memory, but that day Satija stayed in the lab to analyze his first scRNA-seq results. "I remember the thrill for the first time of realizing that the data we were looking at really came from a single cell," he says. The readout was from just 18 individual cells, but it felt like a beginning, especially in light of work from labs he calls the early scRNA-seq pioneers: those of Sten Linnarsson at Karolinska Institute and Fuchou Tang, then at the University of Cambridge.

His transformative moment concerning scRNA-seq, says Trapnell, happened as he watched transformation under the microscope. He was mid-postdoctoral fellowship and working with stem cells, which showed varied developmental trajectories. It was 2012 or 2013 when he and another postdoc did one

CREDIT: S. GLADWELL/GETTY IMAGES

At the Wellcome Sanger Institute, the Teichmann lab focuses on single-cell techniques and the insight they deliver.

of the first scRNA-seq experiments to study how muscle cells change during development. These cells transform into just one cell type, but what surprised him was to see them change at tremendously different rates. "It was so obvious that we could even build a whole algorithm that could put them in developmental order, just based on looking at how they were changing their genes up and down," he says. He sensed that soon researchers would have a pressing need for new software tools and new algorithms. Fast-forward to today and one finds more than 1,400 software tools from many labs, including the Trapnell team.

Many tools, sparse data

Indeed, the number of computational scRNA-seq tools is huge, say Sarah Teichmann, a researcher at the Wellcome Sanger Institute who co-founded and co-directs the Human Cell Atlas, and who responded jointly with colleagues Kerstin Meyer and Nick England. Given the rapid advance of technologies in this space, with sequencing instruments and barcoding techniques for spatial, proteomic and metabolic data, "there is still a great need for new tools," they note. Users want to analyze and combine results that are from different information modalities.

The previous software stack for gene expression analysis "just wasn't going to cut it," says Trapnell; tools need to address scRNA-seq data structures. For ideas, one can port methods from other fields. scRNA-seq experimenters face similar statistics issues to ones ecologists face when they count organisms at different locations. To resolve gene expression spatially, one can look at the way geology tools address an inference problem well-known in statistics to assess at what layer of rock an oil field begins and ends. "People are doing that, which is pretty cool," he says. It's not a conceptually new method, but "but it's conceptually new to biology in a lot of ways."

The sparseness of scRNA-seq data continues to be challenging, say Teichmann and her colleagues. Often, too few transcripts are captured. Imputing the missing values is a way to handle this[4] with methods that apply statistical models or deep learning, among others. While imputation can increase sensitivity for detecting differential expression, "it can also introduce false positives," they say, noting this is a point that colleagues make in a paper[5]. In that paper, Tallulah Andrews and Martin Hemberg, who are also at Wellcome Sanger Institute, evaluate different imputation methods for their risk of generating "false positive or irreproducible differential expression when imputing data." Those authors, the Teichmann team indicates, recommend that imputation can be useful in some instances, such as to visualize data, but for statistical tests, such as for differential expression analysis, unimputed data should be used.

"Imputation is widely used in human genetics where there are good known reference datasets," such as from the 1000 Genomes Project, say Teichmann and colleagues. As more scRNA-seq reference atlases become available, new methods could leverage them for more accurate imputation.

"Imputation, particularly at the level of the individual measurements, is kind of a fraught thing to do," says Trapnell. It's a way to express doubts about a value that, according to the assay, is zero but might be small. The doubts and the guessing have to be carried through the statistical analysis, and "you have to kind of propagate the uncertainty."

The alternative, he says, is measuring many more cells to "fill in the zeroes." To do so takes time and money, but with technology advances, he hopes this will become easier. The 'get more cells' strategy is better than spending time to develop an algorithm that can quickly become obsolete. His lab developed Census, a scRNA-seq imputation tool. Once barcoding techniques with unique molecular identifiers (UMIs) came into wide use, the algorithm was no longer needed.

Better now

Amplification bias was an issue, but UMIs "have solved this conundrum," say the Teichmann lab team. The method lends each gene a unique barcode. "After amplification, reads with the same UMI are collapsed into a single read, thus removing any amplification bias."

Karolinska Institute researcher Rickard Sandberg and colleagues note that errors within barcodes can occur, which must be corrected. Given the lack of experimental ground truths to help with that correction, they developed mRNA spike-ins that have highly diverse random sequences. They note[6] that UMIs shorter than eight nucleotides should be avoided except for shallow scRNA-seq experiments.

Some labs previously excluded the effect of cell cycle state on scRNA-seq data, as a confounder, but no longer do so, says applied statistician Kasper Hansen. He and colleagues at the Johns Hopkins Bloomberg School of Public Health developed Tricycle: Transferable Representation and Inference of Cell Cycle. It's a tool with which one can infer cell cycle state from scRNA-seq data.

During the cell cycle, many genes are differentially expressed, especially as the cell grows and increases its RNA content before cell division. Not only the genes regulating the cell cycle change. The team, says Hansen, amassed much evidence that Tricycle works well in mammalian cells. "If you use a more distant organism, we would love to hear your experience," he says.

Currently, users apply Tricycle to estimate cell cycle length, but it could reveal more about the interplay between cell cycle length, differentiation and cell fate. "We are actively working on this question," he says. In their method development work, the team confirmed Tricycle's results by comparing them to gold-standard cell cycle measurements. For technical reasons such datasets profile only a single proliferating cell type, but "what we need are methods which work on mixtures of cell types," he says.

As scRNA seq emerged, some labs used highly sensitive methods such as RNA fluorescence in situ hybridization (FISH) to probe and localize RNAs in cells or tissue sections. They counted individual mRNA molecules and came across "the bursty nature of gene expression regulation," says Trapnell, which was and is intriguing. FISH remains a great way to study the mechanics of transcription.

Over time, says Trapnell, scientists have learned that the cells contain much RNA, yet not many molecules are needed to "get a pretty clear idea of at least what kind of a cell a cell is, and maybe even where it is." Genes are correlated, and from an information theory or statistical perspective, a fraction of a cell's RNA content is telling. This has been a technical surprise to the scientific community, he says. scRNA-seq has enabled an era of cell atlas making.

Paths to scale-up

When prepping samples for scRNA-seq, tissue dissociation can bias the proportion of retrieved cell types, says the Teichmann lab team. But scRNA-seq delivers, in their view, the highest quality data, and experimenters can find subtle gene expression differences and characterize small subpopulations. Single-nucleus sequencing "is much less prone to bias," the researchers note, and the technique more faithfully represents the intact tissue's cell populations, "albeit with a slightly lower gene count," the team says.

Because every assay in biology has bias, says Trapnell, one needs internally consistent controls. Cells and nuclei are different beasts. And a cell has more RNA than just the nucleus alone. If a researcher wants to count how many cells of each type a sample contains, "nuclei are just fine for that," says Trapnell. Both scRNA-seq and single-nucleus RNA-seq have their place and limitations. To know which genes were expressed an hour ago, as opposed to yesterday, "nuclei are maybe even better for that."

The Teichmann lab and colleagues have combined scRNA-seq with recently improved spatial technologies that now offer "true single-cell resolution." They mention the 10x Genomics Xenium and Visium HD systems as examples. For many experiments[7], the Teichmann lab integrates data from scRNA-seq, single-nucleus RNA-seq and spatially resolved transcriptomics. They combined scRNA-seq and 10X Genomics Visium data in their work on limb development, for which they built a spatially resolved single-cell atlas.

Yuhan Hao, while a PhD student in the Satija lab, co-developed a bridge integration strategy for multimodal single-cell data analysis.

Together with colleagues in the UK, Germany and Australia, the team developed the WebAtlas pipeline[8] for sharing integrated single-cell datasets. One can query cell types and genes across single-cell data, as well as sequencing and imaging-based data. The datasets they applied their approach to include scRNA-seq data and datasets acquired with spatial technologies: 10x Genomics Visium CytAssist and Xenium, Vizgen's MERSCOPE and a mouse embryonic dataset generated with seqFISH developed in the Cai lab at California Institute of Technology.

In scRNA-seq, rapid scale-up is underway, say the Teichmann team. Companies such as 10x Genomics provide chips for loading 100,000 cells per inlet. Some companies use combinatorial indexing, which allows analysis of large numbers of cells at reasonable cost. Combinatorial indexing involves numerous rounds of barcoding of cells and analytes, and that scales up scRNA-seq experiments. One company the Teichman team mentions is Parse Biosciences. Another company in this space is Scale Biosciences, which Trapnell co-founded with University of Washington colleague Jay Shendure, Stanford University's Garry Nolan and Frank Steemers, who was previously at Illumina.

Integration in action

Says Satija, it's exciting how possible it is now to get different types of molecular information from individual cells – gene expression as well as protein abundance, chromatin accessibility and DNA methylation levels. "Measuring these other modalities provides a very different view of what the cell is doing, and can even be used to infer its past behavior or to predict its future state," he says.

The Satija lab has developed a bridge integration approach[9] to integrate single-cell datasets across modalities. In their paper, they consider two datasets generated from immune cells in people with COVID: RNA levels, which are cellular gene expression, and cellular protein level measurements. Each cell in a multiomic set of data is an 'element' in a dictionary. Dictionary learning is how translation between RNA and protein data is handled so the two sets can be integrated, says Satija. The method is implemented in the Seurat package, a widely used single-cell data analysis suite of tools from the lab.

The method focuses intensive processing and integration on a select group of representative cells, says Yuhan Hao, who heads data science at the biotech Neptune Bio and co-led this method's development as a PhD student in the Satija lab. These results are extended to represent the entire dataset. This shortens the otherwise extensive analysis time needed when integrating ten large scRNA-seq datasets with millions of cells that require much computational memory. "This process effectively brings datasets of different types into a common feature space, making integration straightforward," he says.

This method sits at the heart of the lab's Azimuth portal. It emerged as part of their activities in the Human Biomolecular Atlas Program (HuBMAP), which aims to map the human body at single-cell level. Azimuth offers annotated reference datasets that help with automated processing and analysis. There are reference scRNA-seq datasets of various types of human cells and scATAC-seq data. The method scATAC-seq, or transposase-accessible chromatin through sequencing, is a way to assess where transcription factors bind or where DNA methylation occurs.

New vistas

Especially because of plummeting sequencing costs, the overall costs of scRNA-seq experiments have dropped, says Trapnell. And whereas what used to be a typical experiment involved a few hundred cells from tissue from which cells could be readily disassociated, these days, experiments can be run with millions of cells from many specimens. Most early applications focused on regulation of individual genes. Such work continues, but scRNA-seq scale-up and throughput have opened up new experimental possibilities.

Instead of perturbing cells and asking "how does my favorite gene change?" or how

CREDIT: NEPTUNE BIO

scRNA-seq experiments pose "a whole bunch of really exciting, conceptually new problems," says University of Washington researcher Cole Trapnell.

individual genes are regulated, one can perturb model organism embryos and ask how a favorite cell type changes in proportion to other cell types. One gets a sense of the overall program because one can study how cell types depend on one another. One might be studying cancer cells to study residual disease. "It's like using single-cell RNA-seq like one would use flow cytometry," he says. "But on a much larger scale." These changes make it easier to design experiments.

The Trapnell and the Shendure lab have been applying scRNA-seq to developmental biology questions[10,11]. This work will be scaled up in their work at SeaHub, the new Seattle Hub for Synthetic Biology. Shendure is SeaHub's scientific director and Trapnell will co-lead.

When you can sequence millions of cells from different specimens — in their cases, embryos of model organisms that have been perturbed in various ways — one can study how a change affects all cell types across the development of the embryo, says Trapnell, and begin addressing problems in genetics and developmental genetics one could not address with more conventional tools. Along with new tools to address computational and statistical problems, such as for inferring which genes are required for which cell types, how the cell types depend on one another, or how the genes regulate one another, "I think it's going to provide us with a means of dissecting the genetic program that controls development."

Plenty of technical issues remain to be solved in scRNA-seq measurement and analysis, says Hao. "We need the scRNA data at the population level with genetic and curated clinical information," he says. Last year, the Chan Zuckerberg Initiative consolidated publicly available scRNA-seq data to build CZ CellxGene Discover Census, with which one can access, query and analyze scRNA-seq data. These data are invaluable for training AI models to learn the unified representation of all those cells. It would be useful, he says, to have data about the donors of these cells while also maintaining privacy.

Much exciting work is ongoing with scRNA-seq, says Hansen. Sequencing assays are tough to validate since measuring cells destroys them. He is glad to see methods that record the history of the cell before measurement, such as Phylotime, a retrospective lineage barcoding and analysis tool developed by his Hopkins colleagues Reza Kalhor and Hongkai Ji and their groups.

It's particularly useful to see the increase in spatial resolution that technologies are bringing, say Teichmann and her colleagues. This, along with new tools, will let scientists precisely map gene expression to individual cells at their exact location. For instance, understanding the exact cellular interactions between immune cells and their targets in cancer and autoimmune diseases holds much promise for treatment and drug discovery.

The research community, says Hao, has used scRNA-seq to identify and describe novel, rare and previously overlooked cell types. This opens a way to understand cell types and gene programs that lead to complex diseases and thus power therapeutics development. But one needs scRNA data not just from from tens to hundreds of individuals but thousands of individuals, says Hao. Lowering costs of scRNA increases its accessibility, and among the next technical challenges is determining how to collect information about the individuals who donated their cells and maintain privacy.

The next frontier, in Satija's view, is to move beyond observation and leverage these technologies "to understand not just what cells are doing, but why they are doing it." This is a new direction his lab is taking. One technique from this new direction is the team's Phospho-seq, to simultaneously profile proteins, quantify intracellular protein dynamics, use scATAC-seq in whole cells, and then integrate these data with scRNA-seq datasets using the bridge integration method.

One can track cell signaling during development and reconstruct gene-regulatory relationships this way. The lab has also begun large-scale experiments[12] to identify the regulators and targets of diverse cellular responses. This work involves pooled genetic screens, single-cell sequencing such as use of Perturb-seq combined with combinatorial indexing, and high-throughput sequencing to find targets of signaling regulators in different biological contexts. More than 1,500 individual perturbations are performed across six cell lines and five different biological signaling contexts.

With CaRPool-seq, the lab has combined use of CRISPR and single-cell genomics technologies to massively parallelize the measurement of cellular responses under high-throughput genetic perturbations, and those perturbations can involve either single genes or multiple genes.

Scale will keep rising and cost is likely to keep dropping. It helps that scRNA-seq can be performed with much less material than used to be needed and on a wider range of tissues, says Trapnell. One can do things, he says, that were "off limits before."

What trips people up the most in scRNA-seq work, says Trapnell, is study design. It has been too expensive for scientists to do the study they want, so they do a different experiment. "I think that what's really going to change in the next couple years is that now people will be able to do the study they want," he says. "And that's going to be really enabling for a lot of labs."

Vivien Marx ✉
Nature Methods.
✉e-mail: v.marx@us.nature.com

Published online: 23 April 2024

References

1. Kharchenko, P. V. *Nat. Methods* **18**, 723–732 (2021).
2. Heumos, L. et al. *Nat. Rev. Genet.* **24**, 550–572 (2023).
3. Vandereyken, K., Sifrim, A., Thienpont, B. & Voet, T. *Nat. Rev. Genet.* **24**, 494–515 (2023).
4. Cheng, Y., Ma, X., Yuan, L., Sun, Z. & Wang, P. *BMC Bioinformatics* **24**, 302 (2023).
5. Andrews, T. S. & Hemberg, M. *F1000 Res.* **7**, 1740 (2019).
6. Ziegenhain, C., Hendriks, G.-J., Hagemann-Jensen, M. & Sandberg, R. *Nat. Methods* **19**, 560–566 (2022).
7. Madissoon, E. et al. *Nat. Genet.* **55**, 66–77 (2023).
8. Li, T. et al. Preprint at *bioRxiv* https://doi.org/10.1101/2023.05.19.541329 (2023).
9. Hao, Y. et al. *Nat. Biotechnol.* **42**, 293–304 (2024).
10. Saunders, L. M. et al. *Nature* **623**, 782–791 (2023).
11. Qiu, C. et al. *Nature* **626**, 1084–1093 (2024).
12. Jiang, L. et al. Preprint at *bioRxiv* https://doi.org/10.1101/2024.01.29.576933 (2024)

CREDIT: BROTMAN BATY INST