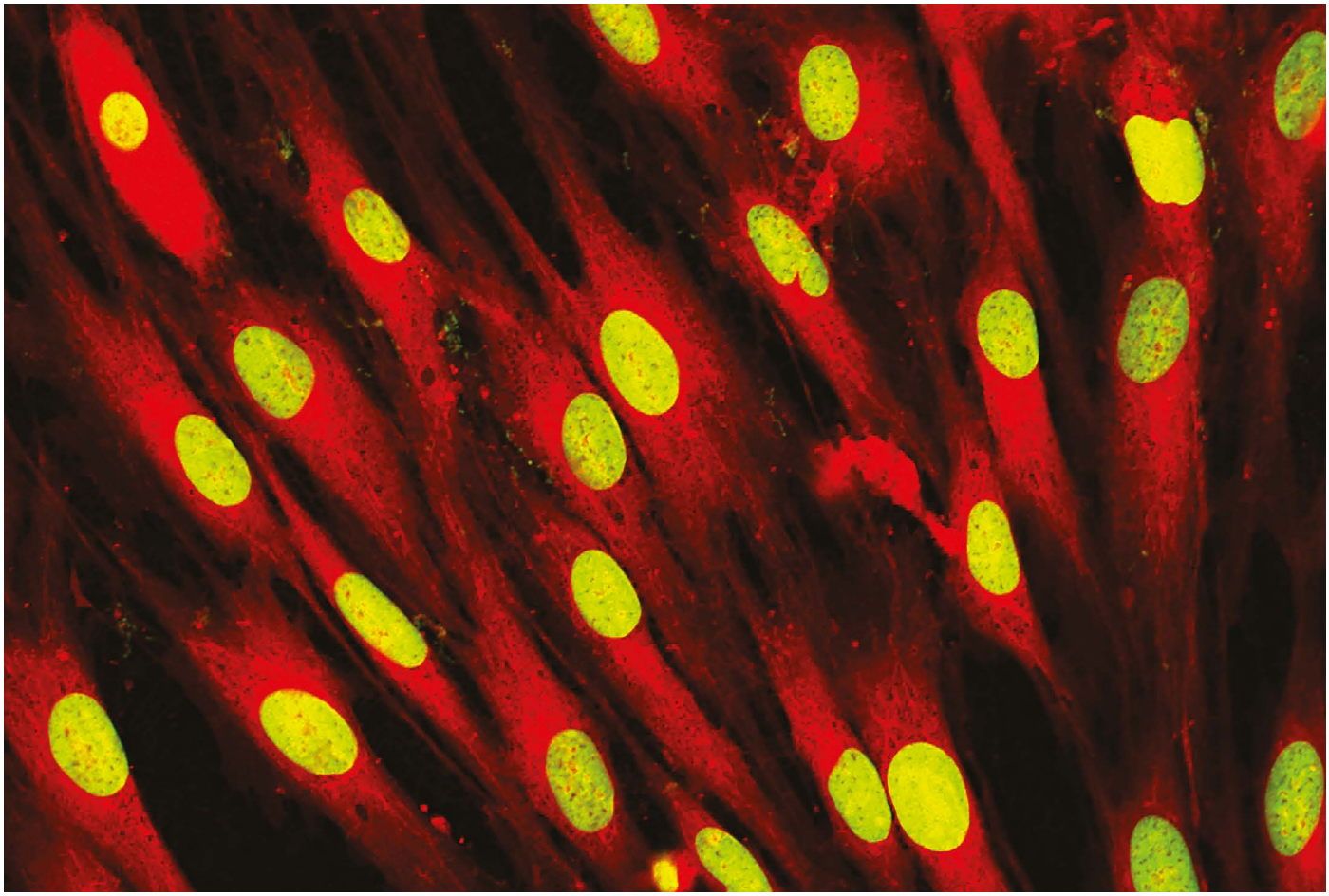


VSHIVKOVA/SHUTTERSTOCK

Researchers have developed an array of techniques to dissect the biology of individual cells, such as these skin cells.

SINGLE-CELL BENCHMARKING CLEARS UP CONFUSION

In the fast-paced field of single-cell biology, studies that compare methods help scientists to pick the right technique for their research. **By Amber Dance**

Single-cell biology has exploded over the past decade. Between 2015 and 2021, studies in the PubMed database of biomedical literature using the term 'single-cell' more than tripled, driven largely by technological innovations in isolating single cells and their molecular content.

Researchers use these and many other methods to probe individual cells in different ways, from characterizing their gene expression to documenting their epigenetic state, transcription factor activity and cell-to-cell communication.

"Single-cell biology has really been opened up to a much broader audience," says Samantha Morris, a developmental biologist at Washington University in St. Louis, Missouri.

But even though the proliferation of wet-lab techniques and analytical methods has expanded access to single-cell studies, it has also muddied the waters for researchers trying to pick the best approach. The online single-cell RNA tools catalogue, scRNA-tools, lists nearly 1,400 software packages that turn single-cell data into scientific insights. How are researchers supposed to choose?

"If you're going to do any kind of analysis, you are faced with a choice, and as a researcher, you are expected to make a justified choice," says Geir Kjetil Sandve, a bioinformatician at the University of Oslo.

And the choices scientists make matter, says Julio Saez-Rodriguez, a computational biologist at Heidelberg University in Germany. "Even a small change can lead to substantial differences in the results," he says.

The solution to the field's analytical abundance is benchmarking: the testing — ideally by neutral parties — of multiple methods, often applied to several kinds of data set, to determine which method works best for different purposes. Since the advent of single-cell research, scientists have conducted dozens of comparisons of wet-lab assays and analysis algorithms, which can guide researchers as they select methods for their own work. Although some researchers say these efforts are underappreciated by funding agencies, fresh initiatives are providing both support and credit to scientists who undertake benchmarking studies.

"The benchmarking papers are absolutely essential," says Morris. "This is such an important contribution to the community."

When embarking on single-cell research, the first choice that researchers must make is which technology to use to separate cells and analyse their molecules. For single-cell RNA (scRNA) sequencing, a common method is to first divide up the individual cells into wells or droplets, reverse transcribe the cell's RNA to generate complementary DNA (cDNA) and then use molecular barcodes to label the cDNA from each well with a different tag. Next, the cDNA from each cell is amplified to construct a library, then the cDNA strands and their tags are sequenced. Finally, the barcodes are used to ascertain which segments of RNA originated in which cell. But there are plenty of ways scientists can go about those tasks.

Putting tech to the test

In 2017, Joshua Levin, a geneticist and molecular biologist at the Broad Institute of MIT and Harvard in Cambridge, Massachusetts, worked with his colleagues to compare seven scRNA-sequencing methods. Two were low-throughput approaches, appropriate for analysis of hundreds of cells, which typically offer high sensitivity to capture rare RNA and cell types. Five were high-throughput approaches that could handle thousands of cells.

The researchers applied these methods to three samples[1]. One of these, human peripheral blood mononuclear cells (PBMCs), is a good test case, says Levin, because scientists already know which cell types should be present and what their molecular signatures look like. A mixture of human and mouse cell lines allowed the researchers to detect instances in which two cells landed in one well. This would appear as human and mouse genes sharing the same barcode, a clear indicator of a mistake. The team also tested four methods that were appropriate for single-nuclei RNA sequencing, which can be a good fit for certain tissues in which dissociating single cells is difficult. For those tests, they used mouse brain tissue because it is a common target in such studies.

The ideal approach depends on a scientist's data set and research questions. Smart-Seq2 and CEL-Seq2, both low-throughput methods, performed similarly, identifying the most genes per cell – although CEL-Seq2 sometimes assigned RNA sequences to the wrong cell. The downside of those two methods, says Levin, is that they're expensive. Among the high-throughput methods, the 10X Chromium system worked best, picking up the most genes per cell. Developed by 10X Genomics in Pleasanton, California, the system partitions samples into droplets.

One of the biggest challenges when performing benchmarking studies is ensuring that all the methods are run fairly, says Levin. To control for variation between experiments, the team prepared samples in one batch that they used with each method. When they got to the sequencing step, they ran each reaction in the same chamber of the machine. They also developed a computational pipeline to process all the data in as similar a manner as possible. Levin's team performed all the methods themselves, but that meant they weren't necessarily specialists in each technique. For methods that were new to them, the researchers asked the developers of each technology for assistance in getting it right.

A team led by biotechnologist Holger Heyn at the National Center for Genomic Analysis in Barcelona, Spain, took a different approach[2]. Heyn's team farmed out its testing to specialists in each single-cell and single-nucleus method, and sent samples from the same test tube to 13 centres across the world. "Single-cell technologies often require strong expertise and experience to be able to perform assays properly with the best possible result," Heyn says.

For their sample, the researchers chose a mixture of human PBMCs and mouse colon tissue. The former offered clearly defined types. The colon tissue included cells that were in the process of developing from stem cells to fully differentiated colon cells, providing a continuum of cell types and sizes.

To further assess how well the methods could detect rare cell types, the team added defined quantities of human, mouse and dog cell lines that were fluorescently labelled. This 'spiking in' of specific cell types, along with mixing distinct populations, allows scientists to check their observations against the exact quantities of cells they knew to be present in the sample. For example, one method tested in Heyn's study missed the canine cells that made up 1% of the cellular cocktail; another counted too many canine cells, putting them at 9% of the population.

> **"All of this dates pretty quickly as the technology changes."**

Heyn says methods that divide cells among wells on a microtitre plate often gave high-quality results, but limited throughput, compared with microfluidic methods that divide cells into droplets. CEL-Seq2 and another method, Quartz-Seq2, both excelled at finding genes – CEL-Seq2 in particular could detect even weakly expressed transcripts. But Quartz-Seq2 was the top performer overall, because it also scored well when grouping cells by expression of known biomarkers.

Assessing algorithms

Once scientists decide on an assay, they must select a method to turn its raw data into meaningful results. Again, options abound. Some algorithms take RNA-sequencing data and group cells by type, or detect a progression of gene expression that reflects the development of one cell type into another. Other algorithms evaluate DNA sequences to determine chromatin structure, analyse the action of transcription factors or assess molecules used in cell–cell communication. Some even combine data from different experiments, correcting for the differences between batches analysed at various times or by different researchers.

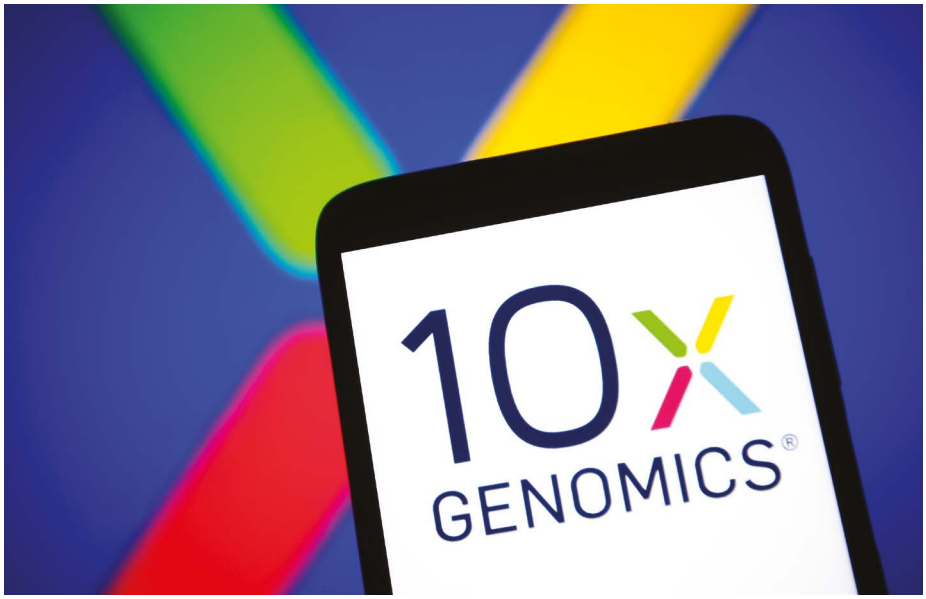
Benchmarkers evaluate analytical techniques on how accurately they perform those tasks, but specific metrics depend on the task at hand, says Mark Robinson, a computational biologist at the University of Zurich in Switzerland. For example, scientists benchmarking techniques that are meant to cluster similar cell types together might check the 'F1 score', which incorporates both the number of correct classifications made and how that number compares with the total number of correct classifications possible. By contrast, scientists aiming to quantify accuracy of differential gene expression might choose statistical power, a measure of the probability of detecting such differences. Benchmarkers might also evaluate metrics such as efficiency, computational demands, quality of documentation and whether the code behind an algorithm is open source.

Scientists can use two kinds of input data set: real data from cells, and artificial data generated by tools such as Splatter, a software package that can simulate scRNA-sequencing data. Simulated data offer a good starting point because they provide a sort of "sanity check", says Sandve – researchers know precisely what they put in the pipeline, so they know what should come out.

But simulated data can never stand in for true biological complexity, because researchers cannot simulate complexity that they do not fully understand, says Kim-Anh Lê Cao, a computational statistician at the University of Melbourne in Parkville, Australia. Genuine cells offer realistic complexity. "The disadvantage," says Saez-Rodriguez, "is we may not have a clear ground truth for what's happening."

Ideally, scientists test analytical methods against gold-standard data sets in which the cell types and biology are well characterized, so they can predict the results. The database of deeply integrated human single-cell omics information, known as DISCO, includes published data sets from a variety of tissue types, diseases and assay platforms.

More than 60 benchmarking studies of single-cell computational methods have been completed, according to a meta-analysis that Robinson posted on the preprint repository bioRxiv in September[3], so scientists looking for their ideal method have plenty of options. Look for neutral analyses, Sandve suggests, because developers of new algorithms naturally emphasize the advantages of their approaches. Then, specialists advise, scientists should look for methods that perform well with data sets similar to their own, and test-drive a few with their own data before



PAVLO GONCHAR/SOPA IMAGES/SHUTTERSTOCK

California biotechnology firm 10X Genomics makes gene-sequencing technology.

settling on the best approach.

The most complicated, advanced methods aren't always the best. For example, Luca Pinello, a computational biologist at Massachusetts General Hospital and Harvard Medical School in Boston, working with epigeneticist Jason Buenrostro at Harvard University in Cambridge, and their colleagues compared several methods to differentiate cell types on the basis of chromatin accessibility, a measure that correlates with gene expression[4]. When they plotted the computational run time against the quality of the results, they were surprised to find no relationship between the two. "Sometimes, doing really complex things doesn't boost your performance," says Pinello. That means scientists with limited computational resources can still find algorithms that will give good results.

And sometimes the answer isn't picking a single, best method, but using several. That's what Saez-Rodriguez and his colleagues concluded after they benchmarked methods to infer cell–cell communication pathways from scRNA-sequencing data and found poor agreement between them[5]. If multiple analysis pathways give the same result, he says, then the result is probably correct. His team developed an open-source framework called LIANA that allows users to run multiple algorithms on several data sets, allowing fair and unbiased comparisons, Saez-Rodriguez says. The system can also provide consensus results from all included methods.

It takes a village

Although the value of benchmarking is clear, scientists say such studies are neither particularly well funded nor highly esteemed. "Academia is not rewarded for these sorts of activities," laments Buenrostro.

But that is starting to change. The Chan Zuckerberg Initiative in Redwood City, California, has invested tens of millions of dollars in benchmarking and related studies for single-cell biology, says Ivana Jelic, the science programme manager for computational biology at the institute. And opportunities for recognition are growing, she adds.

Buenrostro, for instance, is excited about a paper format recently adopted by *Nature Methods* that seeks to reward benchmarking studies. Called Registered Reports, the format invites researchers to submit their plan before collecting data. If a submission meets the journal's criteria of novelty, scope and comprehensiveness, it is provisionally accepted ahead of time. Following data collection, if the study passes quality checks and interprets the findings in a manner the journal deems sensible, it will be published no matter what the results say. "That is super, super valuable in this space," says Buenrostro.

Researchers can also band together to benchmark. In 2019, several dozen researchers at the research institute Helmholtz Munich in Neuherberg, Germany, and the Ludwig Maximilian University of Munich gathered in the holiday resort of Schliersee in Germany. Their aim was to test various methods for integrating and normalizing data from different sources, which is crucial for researchers building large single-cell data sets. In teams, the group tackled simulated and real data on RNA sequences and chromatin accessibility from more than 1.2 million cells. Their work, after further efforts back at home, was published last December — more than two years after the data were collected[6]. That's a long time in the world of single-cell technology development. With tools coming out all the time, the shelf life of any single-cell benchmarking paper is only a few years.

"All of this dates pretty quickly as the technology changes, the protocols change, everything's improving, the software changes," says Matthew Ritchie, a bioinformatician at the Walter and Eliza Hall Institute of Medical Research in Parkville, Australia, who wasn't part of that study. "There is a need to have this as a process that is ongoing."

But it's not generally feasible for researchers to simply update benchmarking studies by adding a new method or two. Robinson, in his meta-analysis of 62 studies[3], found each study tended to use its own code, without any standards or systematization. "It's near-impossible to use that code and extend it and kind of build on it," he says.

To address this shortcoming, Robinson and others are building umbrella systems that will, similar to LIANA, allow users to compare algorithms using a variety of gold-standard data sets. One system, called Open Problems in Single-Cell Analysis, provides a "living benchmark platform", says site co-founder Daniel Burkhardt, a machine-learning scientist at Cellarity, a pharmaceutical company in Somerville, Massachusetts. Users can access leader boards that rank methods for their performance on different data sets; so far there are ten categories of functions, such as removing noise in data and batch integration, with more than a dozen data sets and about four-dozen methods included. The collaborators run all the analyses and provide code for users to download; methods developers can submit tools and offer tweaks or improvements to the code for their own approaches. The site has had 25,000 unique visitors since it launched in January 2021, says Burkhardt.

"Benchmarks done in a centralized fashion are much better than benchmarks that are done as papers are published," he says. Other options for systems include Robinson's OmniBenchmark, and OpenEBench from the ELIXIR Tools platform. But these have yet to take off; Robinson notes that of the 62 benchmarking papers in his meta-analysis, none of the studies used the systems.

With time, researchers predict, development will slow down and bioinformaticians will converge on more-standardized approaches, as often happens with new technologies.

"It is a mad race," says Lê Cao, "but you can see already that some people tend to use the same tools. Not necessarily because they are the best, but because they are easy to use." The field still needs time to settle on one or more winning strategies. For now, benchmarking can help to ease the confusion for scientists confronting a dizzying array of options.

Amber Dance is a freelance science journalist in Los Angeles, California.

1. Ding, J. *et al. Nature Biotechnol.* **38**, 737–746 (2020).
2. Mereu, E. *et al. Nature Biotechnol.* **38**, 747–755 (2020).
3. Sonrel, A. *et al.* Preprint at bioRxiv https://doi.org/10.1101/2022.09.22.508982 (2022).
4. Chen, H. *et al. Genome Biol.* **20**, 241 (2019).
5. Dimitrov, D. *et al. Nature Commun.* **13**, 3224 (2022).
6. Luecken, M. D. *et al. Nature Methods* **19**, 41–50 (2022).