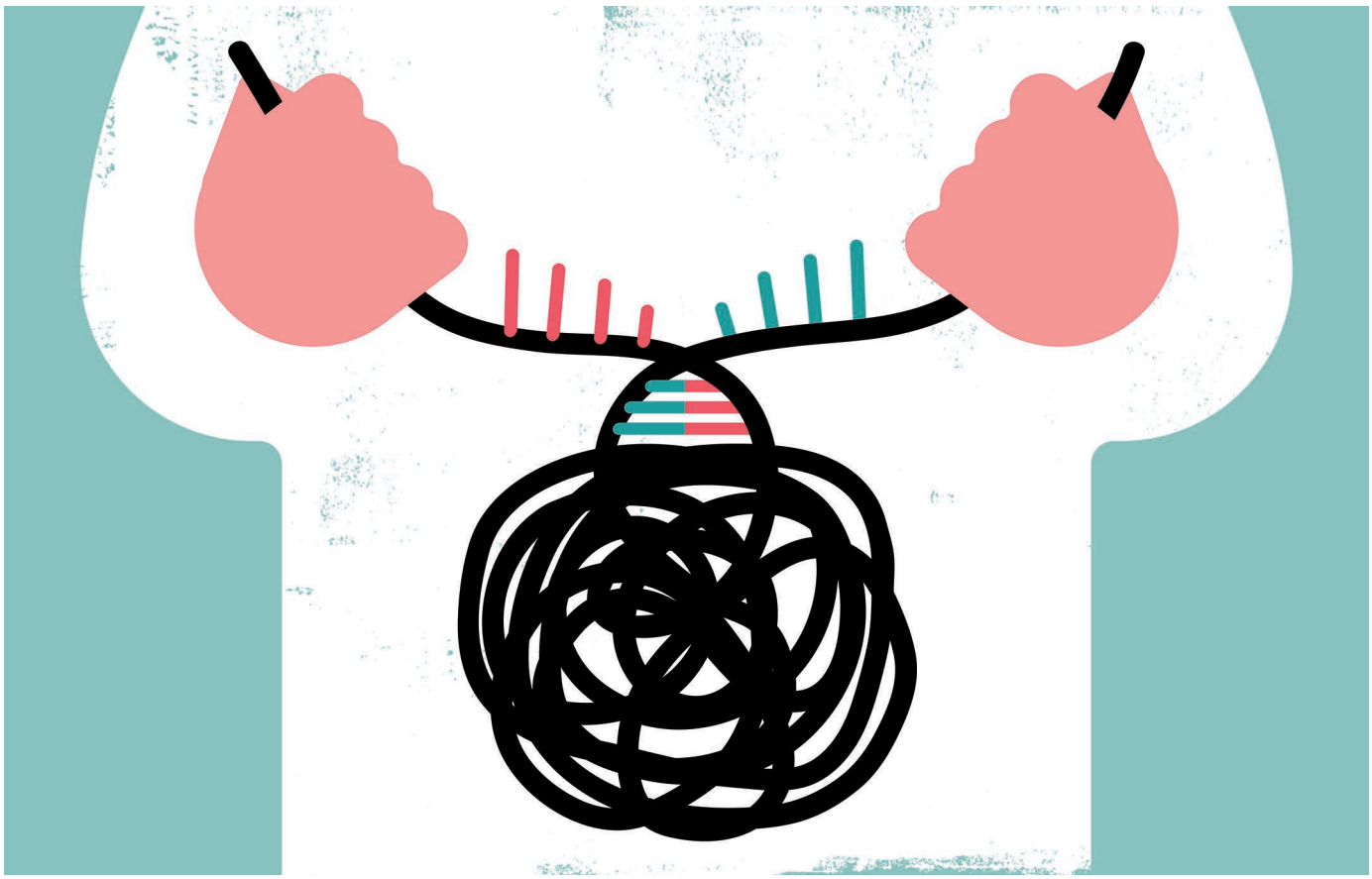
TOOLBOX

SINGLE-CELL SEQUENCING MADE SIMPLE

Data from thousands of single cells can be tricky to analyse, but software advances are making it easier.

ILLUSTRATION BY THE PROJECT TWINS



BY JEFFREY M. PERKEL

Single-cell biology is a hot topic these days. And at the cutting edge of the field is single-cell RNA sequencing (scRNA-seq).

Conventional 'bulk' methods of RNA sequencing (RNA-seq) process hundreds of thousands of cells at a time and average out the differences. But no two cells are exactly alike, and scRNA-seq can reveal the subtle changes that make each one unique. It can even reveal entirely new cell types.

For instance, after using scRNA-seq to probe some 2,400 immune-system cells, Aviv Regev of the Broad Institute in Cambridge, Massachusetts, and her colleagues came across some dendritic cells that had potent T-cell-stimulating activity (A.-C. Villani *et al. Science* **356,** eaah4573; 2017). Regev, who is profiled in a News Feature on page 24, says that a vaccine to stimulate these cells could potentially boost the immune system and protect against cancer.

But such discoveries are hard-won. It's much more difficult to manipulate individual cells than large populations, and because each cell yields only a tiny amount of RNA, there's no room for error. Another problem is analysing the enormous amounts of data that result — not least because the tools used can be unintuitive.

Typically, RNA-seq data is analysed by laboriously typing commands into a Unix operating system. Data files are passed from one software package to the next, with each tool tackling one step in the process: genome alignment, quality control, variant calling and so on.

The process is complicated. But for bulk RNA-seq, at least, a consensus has emerged as to which algorithms work best for each step and how they should be run. As a result, 'pipelines' now exist that are, if not exactly plug-and-play, at least tractable for non-experts. To analyse differences in gene expression, says Aaron Lun, a computational biologist at Cancer Research UK in Cambridge, bulk RNA-seq is "pretty much a solved problem". ▶

The same cannot be said for scRNA-seq: researchers are still working out what they can do with the data sets and which algorithms are the most useful.

But a range of online resources and tools are beginning to ease the process of scRNA-seq data analysis. One page at GitHub, called 'Awesome Single Cell' (go.nature.com/2rmb1hp), catalogues more than 70 tools and resources, covering every step of the analysis process. The field has spawned a cottage industry of computational-biology tools, says Cole Trapnell, a biologist at the University of Washington in Seattle.

BESPOKE TECHNIQUES

Lana Garmire, a bioinformatician at the University of Hawaii in Honolulu, laid out the basic steps of scRNA-seq data analysis (and some 48 tools to perform them) in a review published last year (O. B. Poirion *et al. Front. Genet.* **7**, 163; 2016). Although each experiment is unique, she says, most analysis pipelines follow the same steps to clean up and filter the sequencing data, work out which transcripts are expressed and correct for differences in amplification efficiency. Researchers then run one or more secondary analyses to detect subpopulations and other functions.

In many cases, says Christina Kendziorski, a biostatistician at the University of Wisconsin–Madison, the tools used in bulk RNA-seq can be applied to scRNA-seq. But fundamental differences in the data mean that this is not always possible. For one thing, single-cell data are noisier, says Lun. With so little RNA to work with, small changes in amplification and capture efficiencies can produce large differences from cell to cell and day to day that have nothing to do with biology. Researchers must therefore be vigilant for 'batch effects', in which seemingly identical cells prepared on different days differ for purely technical reasons, and for 'dropouts' — genes that are expressed in the cell but not picked up in the sequence data.

Another challenge is the scale, says Joshua Ho, a bioinformatician at the Victor Chang Cardiac Research Institute in Sydney, Australia. A typical bulk RNA-seq experiment involves a handful of samples, but scRNA-seq studies can involve thousands. Tools that can handle a dozen samples often slow to a crawl when confronted with ten or a hundred times as many. (Ho's Falco software taps on-demand cloud-computing resources to address that problem.)

Even the seemingly simple question of what constitutes a good cell preparation is complicated in the world of scRNA-seq. Lun's workflow assumes that most of the cells have approximately equivalent RNA abundances. But "that assumption isn't necessarily true", he says. For instance, he says, naive T cells, which have never been activated by an antigen and are relatively quiescent, tend to have less messenger RNA than other immune cells and could end up being removed during analysis because a program thinks there is insufficient RNA for processing.

Perhaps most significantly, researchers performing scRNA-seq tend to ask different questions from those analysing bulk RNA. Bulk analyses typically investigate how gene expression differs between two or more treatment conditions. But researchers working with single cells are often aiming to identify new cell types or states or reconstruct developmental cellular pathways. "Because the aims are different, that necessarily requires a different set of tools to analyse the data," says Lun.

One common type of single-cell analysis, for instance, is dimensionality reduction. This process simplifies data sets to facilitate the identification of similar cells. According to Martin Hemberg, a computational biologist at the Wellcome Trust Sanger Institute in Cambridge, UK, scRNA-seq data represent each cell as "a list of 20,000 gene-expression values". Dimensionality-reduction algorithms such as principal component analysis (PCA) and t-distributed stochastic neighbour embedding (t-SNE) effectively project those shapes into two or three dimensions, making clusters of similar cells apparent. Another popular application is pseudo-time analysis. Trapnell developed the first such tool, called Monocle, in 2014. The software uses machine learning to infer from an scRNA-seq experiment the sequence of gene-expression changes that accompany cellular differentiation, much like inferring the path of a foot race by photographing the runners from the air, Trapnell says.

Other tools address subpopulation detection (for instance, Pagoda, from Peter Kharchenko at Harvard Medical School in Boston, Massachusetts) and spatial positioning, which uses data on the distribution of gene expression in tissues to determine where in a tissue each transcriptome arose. Rahul Satija of the New York Genome Center in New York City, who developed one such tool, Seurat, as a postdoc with Regev, says that the software uses these data to position cells as points in 3D space. "That's why we named the package Seurat," he explains, "because the dots reminded us of points on a pointillist painting."

Although targeted to specific tasks, these tools often address multiple functions. Seurat, for instance, powered the cell-subpopulation analysis Regev's team performed to identify new classes of immune cells.

Most scRNA-seq tools exist as Unix programs or packages in the programming language R. But relatively few biologists are comfortable working in those environments, says Gene Yeo, a bioinformatician at the University of California, San Diego. Even if they are, they may lack the time required to download and configure everything to make such tools work.

Some ready-to-use pipelines have been developed. And there are end-to-end graphical tools too, including the commercial SeqGeq package from FlowJo, as well as a pair of open-source web tools: Granatum from Garmire's group, and ASAP (the Automated Single-cell Analysis Pipeline) from the lab of Bart Deplancke, a bioengineer at the Swiss Federal Institute of Technology in Lausanne.

ASAP and Granatum use a web browser to provide relatively simple, interactive workflows that allow researchers to explore their data graphically. Users upload their data and the software walks them through the steps one by one. For ASAP, that means taking data through preprocessing, visualization, clustering and differential gene-expression analysis; Granatum allows pseudo-time analyses and the integration of protein-interaction data as well.

According to both Garmire and Deplancke, ASAP and Granatum were designed to allow researchers and computational biologists to work together. Researchers "used to think of [bioinformaticians] as magical creatures who just get the data and magically generate the result", says Xun Zhu, a PhD student at the University of Hawaii at Manoa, and lead developer on Granatum. "Now they can participate a little bit in terms of tuning the parameters. And that's a good thing."

APPROACH WITH CAUTION

The tools aren't perfect for every situation, of course. A pipeline that excels at identifying cell types, for instance, might stumble with pseudo-time analysis. Plus, appropriate methods are "very data-set dependent", says Sandrine Dudoit, a biostatistician at the University of California, Berkeley. The methods and tuning parameters may need to be adjusted to account for variables such as sequencing length. But John Marioni at Cancer Research UK in Cambridge says it's important not to put complete faith in the pipeline. "Just because the satellite navigation tells you to drive into the river, you don't drive into the river," he says.

For beginners, caution is warranted. Bioinformatics tools can almost always yield an answer; the question is, does that answer mean anything? Dudoit's advice is do some exploratory analysis, and verify that the assumptions underlying your chosen algorithms make sense.

Some analytical tasks still remain challenging, says Satija, including comparing data sets across experimental conditions or organisms and integrating data from different 'omics. (A planned update to Seurat should address the former issue, he notes.)

But enough tools exist to keep most researchers occupied. Kendziorski suggests that people who are interested just dive in. Each new tool can unveil another facet of biology; just keep your eyes on the science, and be judicious in your choice. ■

Jeffrey M. Perkel *is the technology editor for* Nature.