There are amendments to doi:10.1038/s41587-020-0449-8

Go-ahead for first peanut allergy drug

The US Food and Drug Administration has approved a first food allergy drug, Aimmune Therapeutics' Palforzia, for children 4–17 years old with peanut allergies. Palforzia (AR101) consists of peanut (*Arachis hypogaea*) powder capsules to help patients build up tolerance to accidental peanut exposure. Patients start the oral desensitization course with 3-mg daily dose of peanut protein and gradually build up to a 300-mg daily maintenance dose. In a 496-patient pivotal trial, 67% of Palforzia recipients tolerated a 600-mg peanut protein challenge, after 6 months on maintenance treatment, with only mild allergic reactions. Only 4% of placebo recipients tolerated this challenge.

Peanut allergy affects around 1 million children in the United States, and accidental exposure can provoke life-threatening anaphylactic shock in some patients, leading Aimmune to predict the drug could exceed $1 billion in global annual sales. But others have their doubts. At over $10,000 per year, the drug's high cost could deter insurers, and its side effect profile — which mirrors the effects of peanut exposure — could put patients off, say some. While some patients might be tempted by a do-it-yourself peanut desensitization program, clinicians caution against it because of the increased need for epinephrine while on treatment. For Palforzia, treatment initiation and dose escalation must take place in a supervised medical setting.

Aimmune's closest competitor is DBV Technologies. The biotech, headquartered in Paris, first filed its transdermal peanut tolerance patch Viaskin Peanut for approval in 2018, but withdrew this submission months later pending more manufacturing and quality control data. It resubmitted the patch in 2019 and anticipates a decision by August. Aimmune and DBV Technologies are also working on other food allergy desensitization products, including products for egg allergy and milk allergy.

Published online: 9 March 2020
https://doi.org/10.1038/s41587-020-0458-7

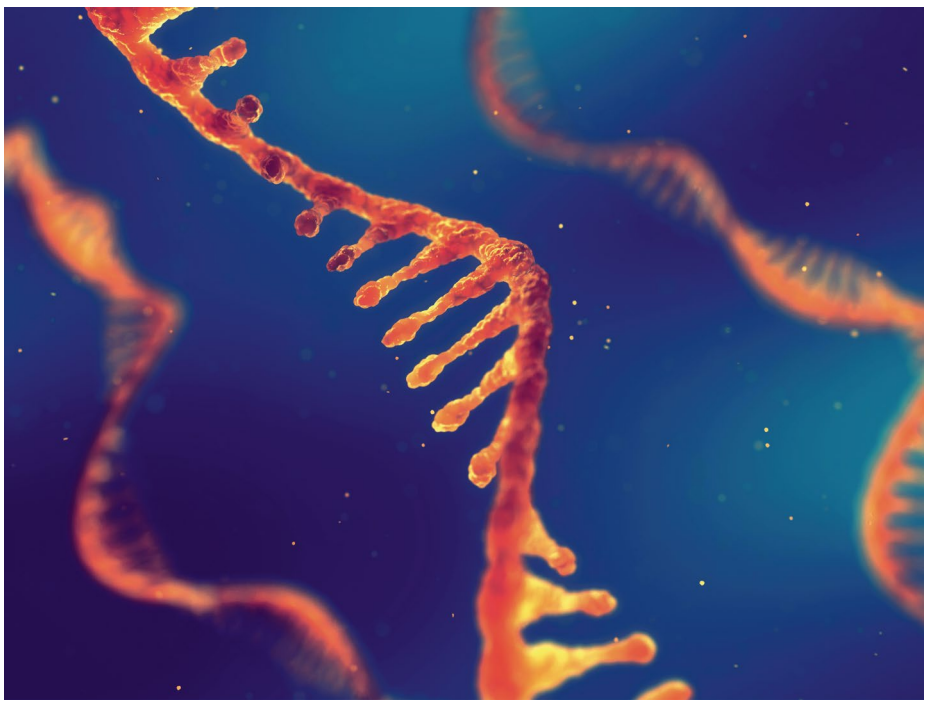
Single-cell RNA-seq analysis software providers scramble to offer solutions

A raft of tools have sprung up to help biologists work through the single-cell transcriptomic bottleneck, but integration remains elusive.

The market for single-cell analytical technologies is booming: from a value of $1.83 billion in 2018, it could triple by 2025, according to a report from ResearchAndMarkets.com published last year. The research community's enthusiastic embrace of single-cell RNA-seq has already transformed the fortunes of some instrument and kit manufacturers. 10X Genomics CEO Serge Saxonov, in a Q3 earnings call from November 2019, announced that the near-term market for his company's technologies for single-cell analysis could reach $13 billion. But this rapid uptake could also create headaches for new users as they learn how to come to grips with the complexities of single-cell data.

For over a decade, researchers have been developing RNA sequencing (RNA-seq) techniques to analyze changes in gene expression in individual cells rather than whole tissue samples. This technology is allowing researchers to tackle complex biological systems and phenomena — for example, profiling rare cell types within a tissue, analyzing tumor heterogeneity, or tracing lineages in differentiating cells — at single-cell resolution. Achieving such insights requires specialized algorithms that can account for the distinctive artifacts and biases associated with data arising from these experiments. "Single-cell data is much noisier than bulk RNA-seq," says biostatistician Stephanie Hicks, of the Johns Hopkins Bloomberg School of Public Health. Although bioinformaticians have devised computational tools to overcome that noise, as the scale of analysis expands, other challenges will grow too.

The scientific community has eagerly embraced single-cell RNA-seq (scRNA-seq). This interest has given impetus to academics'



Making sense of single-cell RNA-seq data can be problematic for labs without in-house bioinformatics capabilities. Credit: Nobeastsofierce Science / Alamy Stock Photo

Table 1 | Selected software providers for scRNA-seq analysis

Software name	Developer	Price structure	Platform-specific	Relevant stages of experiment
Cell Ranger	10X Genomics	Free download	10X Chromium	Raw read alignment, QC and matrix generation for scRNA-seq and ATAC-seq; data normalization; dimensionality reduction and clustering
Loupe Cell Browser	10X Genomics	Free download	10X Chromium	Visualization and analysis
Partek Flow	Partek	License	No	Complete data analysis and visualization pipeline for scRNA-seq data
Qlucore Omics Explorer	Qlucore	License	No	scRNA-seq data filtering, dimensionality reduction and clustering, visualization
mappa Analysis Pipeline	Takara Bio	Free download	Takara ICell8	Raw read alignment and matrix generation for scRNA-seq
hanta R kit	Takara Bio	Free download	Takara ICell8	Clustering and analysis of mappa data
Singular Analysis Toolset	Fluidigm	Free download	Fluidigm C1 or Biomark	Analysis and visualization of differential gene expression data for scRNA-seq
SeqGeq	FlowJo/BD Biosciences	License	No	Data normalization and QC, dimensionality reduction and clustering, analysis and visualization
Seven Bridges	Seven Bridges/ BD Biosciences	License	BD Rhapsody and Precise	Cloud-based raw read alignment, QC and matrix generation
Tapestri Pipeline/Insights	Mission Bio	Free download	Mission Bio Tapestri	Analysis of single-cell genomics data
BaseSpace SureCell	Illumina	License	Illumina SureCell libraries	Raw read alignment and matrix generation
OmicSoft Array Studio	Qiagen	License	No	Raw read alignment, QC and matrix generation, dimensionality reduction and clustering

QC, quality control; ATAC-seq, assay for transposase-accessible chromatin using sequencing.

in-house algorithms, as well as commercial development of both instruments and analytical software from startups like 10X, 1CellBio and Dolomite Bio alongside biotech giants like Illumina, Bio-Rad and Takara Bio. "The field has moved from users that do very cutting-edge methods on homemade systems to users that require complete, easy to use systems," says Juliane Fischer, senior post sales, applications and support specialist at Blacktrace Holdings, parent company of Dolomite Bio. Dolomite manufactures the Nadia scRNA-seq platform. And these turnkey systems are becoming popular — this January, Saxonov noted that 10X had placed its Chromium instrument for single-cell analysis at 96 of the world's top 100 research institutes.

With scRNA-seq taking academic labs by storm, it means many biologists are getting their first exposure to scRNA-seq data. Those labs with experienced bioinformaticians have developed open-source analysis tools in popular languages like R and Python. One of these, from Rahul Satija's group at the New York Genome Center, is Seurat; another is Scanpy, from Fabian Theis and colleagues; and contributors to the Bioconductor community have also generated an extensive toolbox for single-cell analysis. These algorithms may be intimidating to non-coders, but there is an ongoing effort to make these pipelines more user-friendly; concurrently, several commercial providers are also offering their own solutions for typical scRNA-seq experiments (Table 1).

Software developers include instrument makers themselves (e.g., 10X and Fluidigm) that have developed software specifically for their platforms, as well as companies that have an established track record in providing bioinformatics solutions. As an example of the latter, Partek has extended its Partek Flow software for end-to-end scRNA-seq analysis, taking users from raw reads to visualization. The SeqGeq pipeline, developed jointly by Illumina and FlowJo (now part of BD Biosciences), has also evolved from a general tool for next-generation sequencing analytics into a platform for scRNA-seq experiments. "I have been very encouraged by the development of tools that are designed to expose users to the standard workflows from the field without forcing them to write a single line of code," says Cole Trapnell, a genomics researcher at the University of Washington. Nevertheless, successful analysis still requires broad expertise and careful oversight. "We are not at the point where everything comes out and the results are just waiting for you," says Mirko Corselli, scientific marketing manager at BD Biosciences.

In a standard RNA-seq experiment, sequencing reads are computationally mapped to their gene of origin, and the resulting count matrix depicts the number of transcripts per gene. Single-cell experiments complicate this process. For example, many scRNA-seq protocols — including the workflows used by 10X and Dolomite instruments — capture individual cells in tiny droplets. All the RNAs captured in each droplet are tagged with a distinct barcode sequence indicating that they came from the same cell, and each transcript is also reverse-transcribed to include a unique molecular identifier (UMI) sequence to enable reliable counting of individual RNAs. Reads must therefore undergo careful quality control to ensure that both sequences are present before generating the matrix. This process must also account for droplets with multiple cells or no cells, as well as factors like innate differences in RNA content between cell types.

Furthermore, single-cell data can be sparse. If only a small fraction of a cell's RNA is captured, this means that genes that appear to be non-expressed may simply have eluded detection. Ian Taylor, Director of Product Innovation at BD Biosciences Informatics, notes that users expecting clean two-dimensional plots that clearly depict patterns of gene expression in their sample are "going to be somewhat shocked by the sparse data in front of them."

BIO's bias report

A survey by the Biotechnology Innovation Organization (BIO) of its member companies highlights a lack of both gender and racial diversity at the higher echelons of biotech. At the 98 companies who responded to the survey, women make up 45% of total employees, 30% of an executive team and 18% of the board, while people of color make up 32% of total employees, 15% of the executive team and 14% of the board.

BIO's report is the first to look at racial diversity in biotech. The findings have prompted the trade organization to recommend that companies should put a disproportionately high focus on recruiting diverse board members — by using, for example, targeted talent networks rather than word of mouth and personal connections.

BIO's gender-imbalance findings, however, aren't new to the industry. A 2017 report showed that women held only around one in ten of board positions. Another analysis by Massachusetts Institute of Technology (MIT) entrepreneur Sangeeta Bhatia and colleagues forming the Boston Biotech Working Group investigated why fewer female faculty at MIT set up companies as compared with their male counterparts. Their findings suggest that 40 to 50 more biotechs would exist if female MIT faculty had begun startups at an equal rate to that of their male colleagues. In response, several Boston venture capital firms are pledging that the boards of their portfolio companies will be 25% female by the end of 2022. The premise is that board participation gives women access to a network of investors, scientists and other contacts needed to start a business. BIO hopes to run the survey annually with increased participation to track progress.

Published online: 9 March 2020
https://doi.org/10.1038/s41587-020-0460-0

To make sense of the data, researchers must 'normalize' it first, a process that ensures they are comparing apples to apples when they attempt to identify gene expression differences between cells. Finally, these results must be subjected to dimensionality reduction, which simplifies the data in a way that enables visual representation and straightforward mathematical analysis, and clustering algorithms that can group the individual expression profiles into different cell types based on their similarity or dissimilarity to one another.

The initial stages of data processing are relatively straightforward. For example, 10X has developed the Cell Ranger pipeline, which is designed to perform efficient quality control — including barcode and UMI detection — and matrix generation on raw data from Chromium experiments. "They have been quite good about providing tools for low-level processing and viewing," says Harvard Medical School computational biologist Peter Kharchenko. "If you have a 10X machine and rely on their tools, you'll probably get a pretty good expression matrix." Ines Hellmann of the Ludwig-Maximilians University Munich notes that 10X software is designed to work with the specific idiosyncrasies of 10X data, whereas other pipelines might have to be customized to process the data appropriately. Takara Bio and Fluidigm have likewise opted to produce their own software suites for data processing, which are free to download but also platform-specific.

Some instrument makers, such as 1CellBio, made their analytical software freely available to researchers at the outset. But customers found the open-source software too complicated to navigate, says CEO Colin Brenan. 1CellBio therefore partnered with Partek to develop a pipeline that combines their inDrop platform with Partek Flow, which uses a far simpler, point-and-click user interface. This combination frees researchers from having to know coding, says Partek president Tom Downey, and allows them to process data generated from any single-cell RNA sequencing system. Dolomite made a similar decision to collaborate with Partek for its platform, although it also promotes the open-source dropSeqPipe software — developed by Patrick Roelli and colleagues at the Swiss Institute of Bioinformatics — for more expert users.

For novice scRNA-seq users, commercial software is easiest to use, and this is a major selling point. How much help researchers will need depends on the complexity of the analysis. "Finding new cell clusters is something most people could easily do," Dolomite's Fischer says of the Partek Flow software offered with her company's instruments. User-friendly software can also help in visualizing the data — a major focus of Qlucore's Omics Explorer software, and a step that could help or hinder interpretation. "We have an API [application programming interface] that brings in data from off of the Cell Ranger pipeline," says company president Carl-Johan Ivarsson, referring to the widely used 10X software tool. "Then you can move a slider or click a check box and the visualizations are updated instantly" to reflect the user's changes. Customer service and company onsite training programs often add to the appeal of commercial tools. "We try to put together workflows in presentations and webinars and education materials that guide people down a happy path," says Taylor of the Illumina/BD SeqGeq toolbox.

Enhancing the user-friendliness and accessibility of analytical tools is a natural sweet spot for companies, says Kharchenko. "[In academia] we're very good at developing algorithms, but we're not very good at polishing things and making them convenient," he says. This may be changing, however. For example, Satija's toolbox Seurat has won widespread praise as a powerful and relatively easy to use software and for its guided tutorials. And Hicks, who is part of the technical advisory board for Bioconductor, has worked hard to inform new users about how to use this initiative's software tools for scRNA-seq, including a recently published guide and an online e-book. "This is a way for people to sift through the enormous amount of rich resources," she says.

But there are dangers in too much simplification or blindly trusting the tools. BD's Corselli says, "The challenge is always: how do I know that what I'm seeing is true?" Many steps require careful tweaking depending on the experiment and the type of samples. For example, Hellmann recently embarked on a broad comparison of different analytical workflows to identify factors that can shape the success or failure of a given experiment. Normalization came out as a top factor. "In many big papers, normalization is not taken very seriously, but that can actually be very important," she says.

Before embarking on a deep analysis, scientists should also pay close attention to the dimensionality-reduction and clustering steps. One of the most popular mathematical approaches for dimensionality reduction is principal component analysis, but many groups have also been moving to implement newer, non-linear dimensionality reduction techniques such as *t*-distributed stochastic neighbor embedding (t-SNE) or the more recent uniform manifold approximation and projection (UMAP). However, all of these methods can potentially be confounded by the sparsity of scRNA-seq data, and at present there is still no clear consensus on which is the most robust 'gold standard' solution for simplifying and interpreting scRNA-seq data to identify biologically accurate cell clusters — or even whether such a single solution exists for all experiments.

As a consequence, scRNA-seq still requires a marriage of bioinformatics and wet-lab expertise. With this in mind, many academic centers set up hybrid workflows that incorporate user-friendly commercial algorithms and open-source software, which can be customized. Ivarsson notes that commercial package Qlucore aspires to offer a software for single-cell platforms that can adapt to diverse workflows. "Our vision is not to lock in users — rather the opposite, to make it easier to interact," he says. This is also true for commercially produced SeqGeq, which is designed to work with externally developed or custom-written R software packages. Dolomite's Fischer says that, in her experience, most researchers don't even think about the software aspect of the workflow until they begin wrestling with data — and at that point, flexibility and availability of support are the critical considerations.

Such interoperability is especially important for users looking to venture beyond routine applications like assessing differential gene expression. These cutting-edge applications remain works in progress and will almost certainly require open-source algorithms. "We're quickly moving past 'we did an scRNA-seq experiment and here are the cell types we see'," says Trapnell, whose group has developed computational tools for mapping the developmental trajectories of millions of individual cells in parallel. New tools are also enabling physical mapping of RNA-seq data in multicellular samples. For example, 10x Genomics' Visium Spatial Gene Expression Solution maps gene activity within tissue specimens, and although the company has developed software called Space Ranger to facilitate analysis, other algorithms will surely be needed to make the most of this still-novel experimental approach.

The ultimate challenge is integration. "The bottleneck for many researchers isn't necessarily the early stages of processing data, but the later, more in-depth analysis to integrate results," says Theis, who is Director of the Institute of Computational Biology at the Helmholtz Zentrum Munich. How to best combine data from many different studies using different samples — and, most likely, different experimental workflows — is still very much an unsolved problem. "Right now, it's sort of like a magic trick," says Kharchenko. "You put all your data in a pot and then say 'integrate!'" This problem becomes even more daunting when one begins to contemplate combining scRNA-seq data with other -omic data layers, such as chromatin accessibility, DNA methylation or protein expression.

The perfect software tool to deliver all answers is unlikely to emerge. "You can provide tools to make sure that you make no mistakes," says Corselli. "But ultimately it's still the responsibility of the scientist to look at the data and be rigorous." ❐

Michael Eisenstein
Philadelphia, PA, USA

Published online: 9 March 2020
https://doi.org/10.1038/s41587-020-0449-8

Acknowledgements
Additional reporting by Chris Lieu, Redwood City, CA, USA.

Around the world in a month

Credit: Map: © iStockphoto; Flags: pop_jop / DigitalVision Vectors / Getty

Published online: 9 March 2020
https://doi.org/10.1038/s41587-020-0451-1