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 in the lab of high-throughput scRNA-seq pioneer Steve McCarroll — 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.