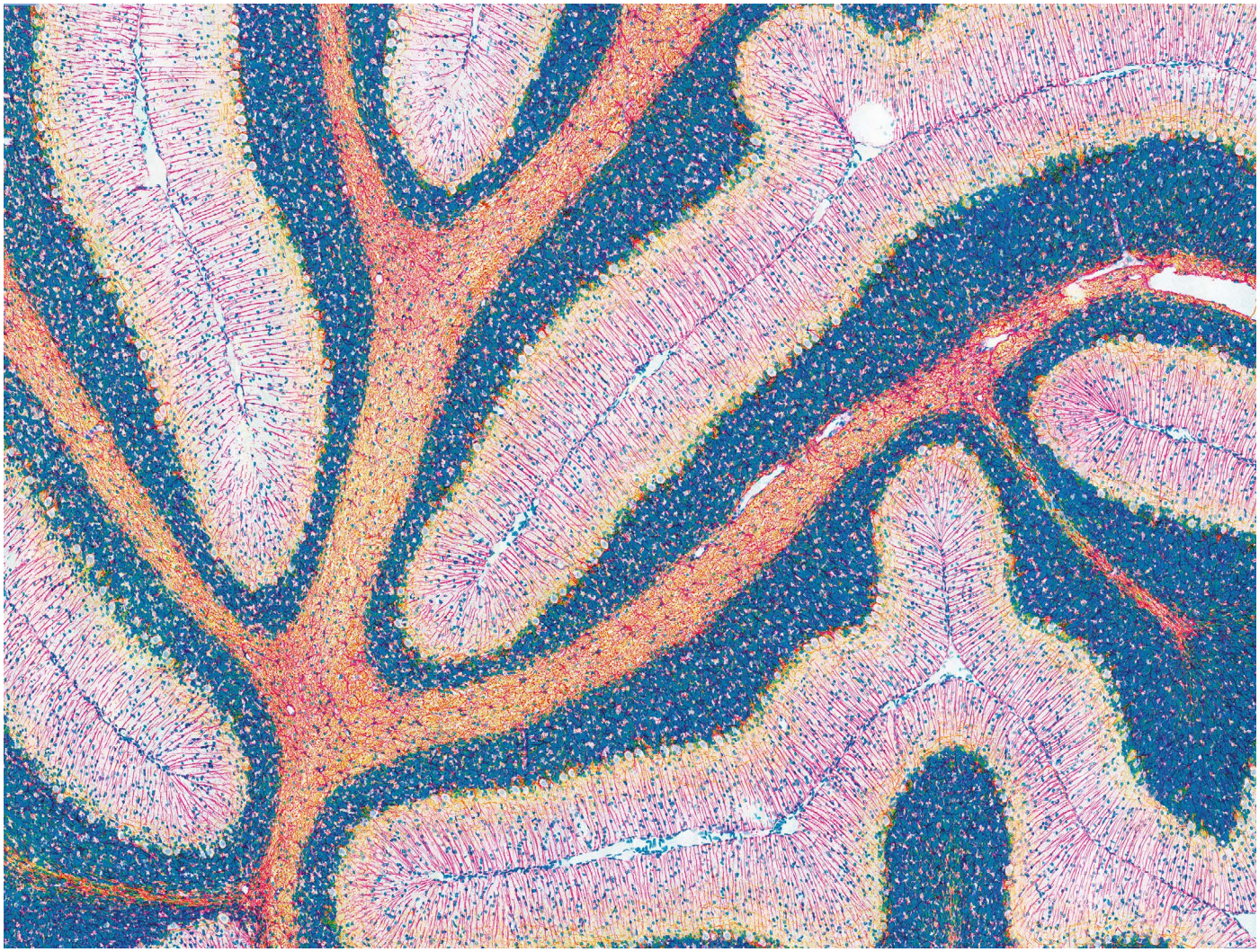

THOMAS DEERINCK, NCMIR/SPL

A fluorescent light micrograph of a section of the cerebellum, at the back of the brain, showing cell nuclei (red) and proteins (green and blue).

COMPUTATIONAL TECHNOLOGIES OF THE HUMAN CELL ATLAS

As the international effort reaches a 'critical mass' of achievements, *Nature* highlights seven tools poised to enable further discoveries. **By Amber Dance**

Single-cell technologies have shattered the fuzzy lenses through which researchers conventionally view biology. Instead of looking at the average behaviour of a swathe of cells, scientists can interrogate genes or other features cell by cell. But the technology also brings challenges: the data are expensive to collect and analyse, and typically force researchers to choose between resolution, throughput and physical location. When it comes to single-cell biology, researchers can learn a fair bit about any one cell, but it's harder to determine precisely where that cell came from.

At the forefront of the design – and use – of single-cell technologies is the Human Cell Atlas (HCA), which aims to catalogue every cell type in people. Launched in 2016, the project has profiled hundreds of millions of single cells, resulted in about 440 published studies and led to dozens of wet-lab and computational procedures.

Now, project co-chair Aviv Regev, head of research and early development at Genentech in South San Francisco, California, and the hundreds of scientists involved in the HCA say that they've hit a critical mass of accomplishments. To showcase this progress, the project is releasing more than two dozen papers this year across Nature Portfolio journals, including six in this issue of *Nature*. These papers highlight the project's accomplishments in cell-fate mapping, data integration and predictive modelling.

Here, *Nature* profiles some of the key technologies that made them possible. Available on the website GitHub, these computational tools include ways to catalogue cells and search atlas data; shortcuts for researchers to obtain spatial or multi-modal data at low cost; and *in silico* models that describe how cells interact and where and how diseased cells

might respond to treatment.

Such tools make massive data sets accessible, says Darcy Wagner, a biomedical engineer at McGill University in Montreal, Canada. "You just want to turn it in as many different ways as possible to look at it, because it's too complex for the human brain." Computational techniques, many that rely on forms of machine learning or artificial intelligence (AI), can step in and provide insights.

Search and annotate

As researchers collect single-cell data and refine them into cell atlases, one key task is to characterize and label, or annotate, each cell type. "This is normally a very time-consuming, onerous task reserved for a few experts in biology," says Evan Biederstedt, a computational biologist and head of the HCA Cell Annotation Platform at the Broad Institute of MIT and Harvard in Cambridge, Massachusetts.

Researchers have developed several programs to label cells automatically — but the tools don't always come up with the same answer. Enter popV. This tool does something simple but powerful: it incorporates eight automated cell-annotation tools into one platform, and more can be added as they become available[1]. "It's a speed-up tool," says co-developer Can Ergen, a computational biologist at the University of California, Berkeley. Researchers who have freshly generated single-cell RNA sequencing data can load them into popV, and each of the eight methods will 'vote' on cell identity — hence the tool's full name, popular Vote. For any given cell, users can check whether all eight annotations line up, or if there's a split vote on possible identities.

If the methods agree on a cell type, researchers can feel confident of its identity; if there is disagreement, maybe not so much. To quantify that, popV provides 'uncertainty scores' so that users will know how much trust to put in its identifications. "That's a really cool thing," says Regev. PopV was trained using data from Tabula Sapiens, a human cell atlas covering nearly 500,000 cells that represent 24 organs from 15 people. The researchers then tested it on a database from the Human Lung Cell Atlas[2]; popV's predictions agreed with most of the annotations and were more accurate than any single computational annotator, according to the resulting paper.

Biederstedt plans to incorporate popV into the HCA Cell Annotation Platform user interface, in which scientists will be able to view popV's predictions as they classify cell types. "It does get the community closer to the dream of automated cell annotations, and will help researchers tremendously," he says.

Once researchers have found an interesting cell type or state, they might wonder where else it occurs. Regev and her colleagues developed SCimilarity to answer this question. The software can take a cellular profile of interest and look for similar profiles[3], just as geneticists use the BLAST algorithm to find related genetic sequences.

G. HEIMBERG *ET AL.*/BIORXIV

The SCimilarity tool identifies related cells. Here, two million single-cell profiles have been clustered by the resulting annotations, including immune cells (purple and orange clusters).

"Figuring out if two cells are similar, that is a difficult problem," says Regev — the matching cell might be one among millions that are already in atlas databases. Fortunately, it's the kind of problem that has already been solved by image-processing algorithms, such as face-matching software. And that's the same approach her team used. The researchers fed a computer 50 million trios of cells, wherein each trio contained two similar cells and one outlier, until the software learnt the features that would distinguish matching cell types.

Each cell is initially defined by the expression of some 20,000 human genes, but the program compresses those into 128 key features for cell identity, explains co-developer Graham Heimberg, a computational biologist and AI scientist at Genentech; it is those features that drive the matching algorithm. Searches of the database, which covers more than 23 million cells from nearly 400 data sets, take just seconds.

To test SCimilarity, the researchers looked for data sets containing cells that are similar to certain immune cells found in fibrotic lung tissue, which would hint at techniques to produce and study those cells in the laboratory. Searching across 17 *in vitro* and *ex vivo* studies involving nearly 42,000 cells, the team unexpectedly found a hit among white blood cells grown in 3D hydrogel systems for the purposes of making blood stem cells[3,4]. Heimberg and his colleagues confirmed that the cultured cells, regrown in their lab, were similar to the lung cells. "We really wouldn't have expected that to come up as a hit," he says — but with SCimilarity, the link was clear.

Computer, enhance!

The expense of high-resolution or high-throughput single-cell experiments is prohibitive for many groups. But scientists are developing workarounds, using AI and machine learning to extrapolate single-cell or spatial data from much smaller or simpler data sets.

One example is scSemiProfiler. Suppose researchers want single-cell RNA profiles but can only afford bulk RNA sequencing. To help them make the most of their resources, scSemiProfiler uses bulk data and generative AI to produce the likely spread of single-cell profiles[5]. It's like taking a low-resolution digital photograph and then inferring the high-resolution equivalent, says developer Jun Ding, a computational biologist at McGill University Health Center.

The process does require some single-cell sequencing, but researchers can do that in small batches. Users add their handfuls of single-cell RNA profiles to the scSemiProfiler model until they and the program are satisfied with the output. The model will even advise researchers when more single-cell sequencing is necessary, and which cells to focus on.

Ding and colleagues test-drove scSemiProfiler on single-cell RNA profiles from the immune cells of 124 people with and without

COVID-19. The program was able to generate the correct single-cell profiles based on bulk sequencing of each sample and single-cell sequences from a representative subset: just 28 of the original panel. The researchers estimate that such an approach could save researchers nearly US$125,000 in a similar study, because it slashes the single-cell sequencing required by about 80%.

"It has the potential to really expand the applicability" of single-cell sequencing, which is currently limited by cost, says David Eidelman, a physician-scientist at McGill University Health Center.

Similarly, Regev and her colleagues are using machine learning as a shortcut to generate spatially resolved, single-cell RNA sequencing data from a readily available resource: tissue slices stained with haematoxylin and eosin (H&E). This pink-and-purple staining technique has been used for more than a century, and lab and hospital archives are stacked with the slides. Because those staining patterns must somehow be rooted in molecular features such as gene expression, Regev and her team wondered whether they could use the H&E information to generate what she calls "the fancy-schmancy stuff": spatial RNA data, which are otherwise laborious and expensive to acquire.

Indeed, the researchers could. Their program, SCHAF[6] (single-cell omics from histology analysis framework) comes in two versions, says co-developer Charles Comiter, a computer scientist at the Massachusetts Institute of Technology in Cambridge. The paired version is trained with H&E stains and limited spatial transcriptomics data from the same slice of tissue, and single-cell RNA profiles from an adjacent slice. Unpaired SCHAF, by contrast, is trained without any spatial RNA data. "You'll still get a good model, but maybe not as powerful," Comiter says.

The researchers tested SCHAF on data sets for which they had matched H&E, transcriptomic and spatial RNA data, two for breast cancer and one for small-cell lung cancer, and obtained "incredibly accurate" output of the spatial results, says Comiter.

Nicholas Krasnow, a protein engineer at Harvard University in Cambridge, Massachusetts, calls SCHAF "exciting". "I'm mainly just interested in seeing how it performs on new problems," he says. But Regev cautions that more training data are needed to ready the software for real-world clinical applications.

Integrate and predict

As well as using one type of data to predict another type, as SCHAF does, computer models can incorporate the data from multiple co-existing types from the same sample. That's the goal of multiDGD, which models biology using both RNA expression and chromatin-accessibility data from the same cells[7]. Using assays that measure the accessibility of the packaged DNA by determining which genomic segments are open and available for transcription, and which are tightly wound, along with information about which genes are being actively expressed, researchers can get a more complete picture of cell biology. Regev calls multiDGD "a nice generative model to learn these shared representations".

The input for multiDGD is based on expression levels for some 20,000 human genes as well as chromatin status (open or closed) for hundreds of thousands of segments across the genome — about 200,000 features per cell. These factors are reduced to a representative set of 20 or so features, which are then fed to the model.

This process of minimizing data "dimensions" makes it easier to identify similarities and differences, says co-author Emma Dann, a computational biologist at Stanford University in California. From there, researchers can move on to different tasks, such as clustering similar cell types or analysing developmental trajectories, she adds. MultiDGD outperformed other popular models in tasks such as cell-type clustering, particularly for small data sets, the team found.

"We always need to be cautious with these new tools."

Researchers can also ask questions of the model, perturbing a gene, say, or amplifying one gene's expression. In one example, the team tested how silencing 41 transcription factors *in silico* might alter chromatin accessibility of the target genes. Researchers can use such computational perturbations to generate hypotheses about how cells might react, says co-developer Viktoria Schuster, a data scientist at the University of Copenhagen.

In Montreal, Ding is also building models for *in silico* experimentation. One, called CellAgentChat, infers cell–cell interactions across a range of distances[8]. Unlike other methods that model cells at the population level, CellAgentChat treats individual cells as autonomous agents — each cell is doing its own thing in an environment of other autonomous cells. This might approximate the biological truth more accurately than do models that lump cells together, says Eidelman. Each cell-cum-agent has digital 'receptors' that can receive molecular 'signals' released by other cells. In response, cells activate new gene-expression patterns, just as real cells do[9].

Among other applications, such models can drive drug screens *in silico*, Ding says, for instance testing what happens if researchers block this or that receptor. His group tried that using a breast-cancer data set, and confirmed that the epidermal growth factor receptor, a known contributor and drug target, was a key interactor in its *in silico* interactions, too.

Ding's group has also developed a model, called UNAGI, that is dedicated to *in silico* drug testing and focuses on how cells change over time[9]. The team fed UNAGI data from four stages of the lung disease idiopathic pulmonary fibrosis to create a virtual disease progression "sandbox", as Ding puts it, with each cell represented by a couple of dozen features in a deep generative neural network. Using the model, the researchers could infer how gene expression changes as the disease progresses, and test whether different drugs would push cells back to an earlier gene-expression profile or towards a healthier one.

One drug that is already approved by the US Food and Drug Administration showed up in the researchers' screen: nintedanib, a growth factor inhibitor that prevents reproduction of fibroblasts. But the screen also flagged medications that might work even better, Ding says.

Wagner calls it "a really important new tool", and says that she's particularly excited about its potential to identify small molecules that could replace more expensive biological therapies such as antibodies.

But, she cautions, testing and validation are crucial. "We always need to be cautious with these new tools, and constantly benchmarking against something that's older," says Wagner. And that problem will only grow as single-cell data sets continue to expand. According to Ergen, future software will probably have to contend with ten million cells at once, and possibly even more.

But those tools are coming. The first version of the HCA should be released in the next year or two, according to the organizers, but further work is planned and more tools will surely emerge as the project's many collaborators continue to advance atlases and the technology to understand them.

"It's always just a work in progress; we're just trying to do better and better," says Heimberg. "The sky's the limit, and anything is in play."

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

1. Ergen, C. *et al. Nature Genet*. https://doi.org/10.1038/s41588-024-01993-3 (2024).
2. Travaglini, K. J. *et al. Nature* **587**, 619–625 (2020).
3. Heimberg, G. *et al. Nature* https://doi.org/10.1038/s41586-024-08411-y (2024).
4. Xu, Y. *et al. Protein Cell* **13**, 808–824 (2022).
5. Wang, J., Fonseca, G. J. & Ding, J. *Nature Commun*. **15**, 5989 (2024).
6. Comiter, C. *et al*. Preprint at bioRxiv https://doi.org/10.1101/2023.03.21.533680 (2023).
7. Schuster, V., Dann, E., Krogh, A. & Teichmann, S. *Nature Commun*. https://doi.org/10.1038/s41467-024-53340-z (2024).
8. Raghavan, V., Li, Y. & Ding, J. Preprint at bioRxiv https://doi.org/10.1101/2023.08.23.554489 (2024).
9. Zheng, Y. *et al.* Preprint at Research Square https://doi.org/10.21203/rs.3.rs-3676579/v1 (2023).