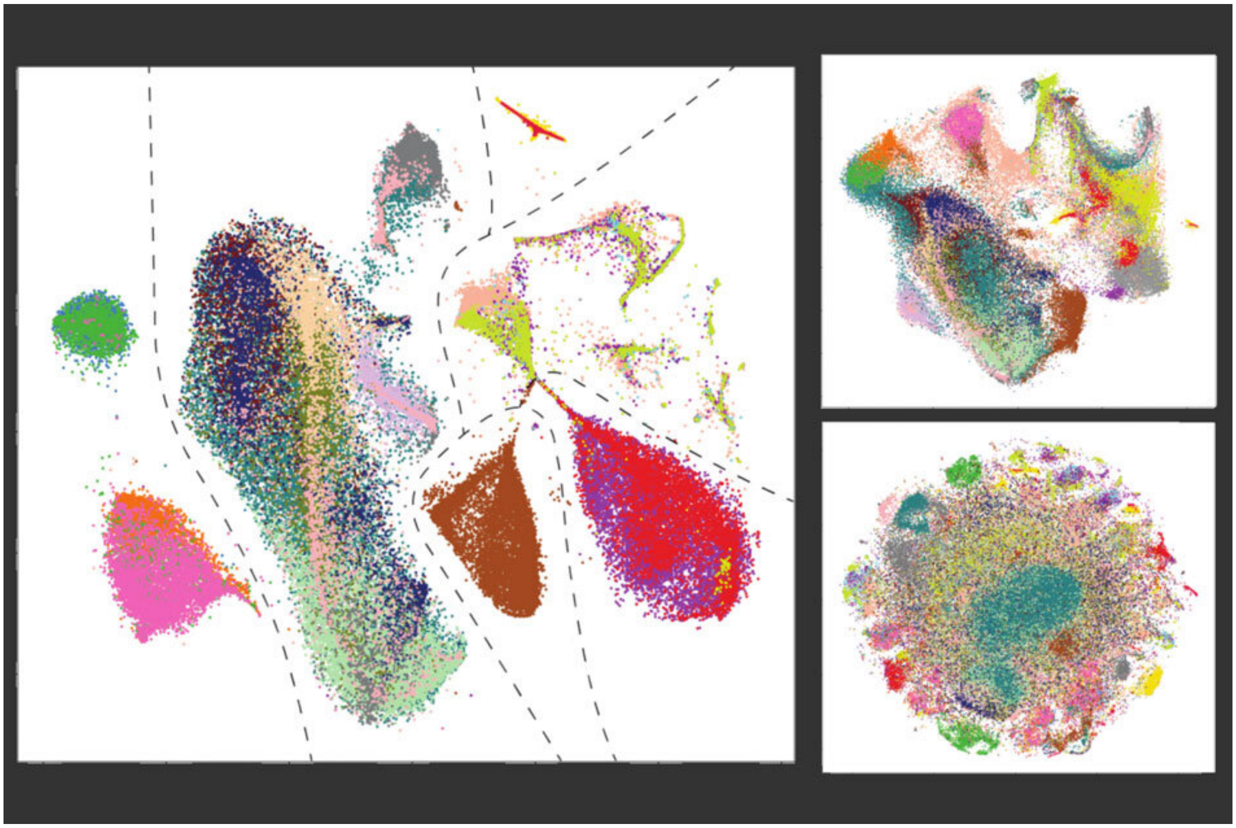
PHYS.ORG

Merging cell datasets, panorama style

May 7 2019, by Rob Matheson



A new algorithm developed by MIT researchers takes cues from panoramic photography to merge massive, diverse cell datasets into a single source that can be used for medical and biological studies. Credit: Massachusetts Institute of Technology

A new algorithm developed by MIT researchers takes cues from panoramic photography to merge massive, diverse cell datasets into a single source that can be used for medical and biological studies.

Single-cell datasets profile the gene expressions of human cells—such as a neurons, muscles, and immune cells—to gain insight into human health and treating disease. Datasets are produced by a range of labs and technologies, and contain extremely diverse cell types. Combining these datasets into a single data pool could open up new research possibilities, but that's difficult to do effectively and efficiently.

Traditional methods tend to cluster cells together based on nonbiological patterns—such as by lab or technologies used—or accidentally merge dissimilar cells that appear the same. Methods that correct these mistakes don't scale well to large datasets, and require all merged datasets share at least one common cell type.

In a paper published today in *Nature Biotechnology*, the MIT researchers describe an algorithm that can efficiently merge more than 20 datasets of vastly differing cell types into a larger "panorama." The algorithm, called "Scanorama," automatically finds and stitches together shared cell types between two datasets—like combining overlapping pixels in images to generate a panoramic photo.

As long as any other dataset shares one cell type with any one dataset in the final panorama, it can also be merged. But all of the datasets don't need to have a cell type in common. The algorithm preserves all cell types specific to every dataset.

"Traditional methods force cells to align, regardless of what the cell types are. They create a blob with no structure, and you lose all interesting biological differences," says Brian Hie, a Ph.D. student in the Computer Science and Artificial Intelligence Laboratory (CSAIL) and a researcher in the Computation and Biology group. "You can give Scanorama datasets that shouldn't align together, and the algorithm will separate the datasets according to biological differences."

In their paper, the researchers successfully merged more than 100,000 cells from 26 different datasets containing a wide range of human cells, creating a single, diverse source of data. With traditional methods, that would take roughly a day's worth of computation, but Scanorama completed the task in about 30 minutes. The researchers say the work represents the highest number of datasets ever merged together.

Joining Hie on the paper are: Bonnie Berger, the Simons Professor of Mathematics at MIT, a professor of electrical engineering and computer science, and head of the Computation and Biology group; and Bryan Bryson, an MIT assistant professor of biological engineering.

Linking "mutual neighbors"

Humans have hundreds of categories and subcategories of cells, and each cell expresses a diverse set of genes. Techniques such as RNA sequencing capture that information in sprawling multidimensional space. Cells are points scattered around the space, and each dimension corresponds to the expression of a different gene.

Scanorama runs a modified computer-vision algorithm, called "mutual nearest neighbors matching," which finds the closest (most similar) points in two computational spaces. Developed at CSAIL, the algorithm was initially used to find pixels with matching features—such as color levels—in dissimilar photos. That could help computers match a patch of pixels representing an object in one image to the same patch of pixels in another image where the object's position has been drastically altered. It could also be used for stitching vastly different images together in a panorama.

The researchers repurposed the algorithm to find cells with overlapping gene expression—instead of overlapping pixel features—and in multiple datasets instead of two. The level of gene expression in a cell determines its function and, in turn, its location in the computational space. If stacked on top of one another, cells with similar gene expression, even if they're from different datasets, will be roughly in the same locations.

For each dataset, Scanorama first links each cell in one dataset to its closest neighbor among all datasets, meaning they'll most likely share similar locations. But the algorithm only retains links where cells in both datasets are each other's nearest neighbor—a mutual link. For instance, if Cell A's nearest neighbor is Cell B, and Cell B's is Cell A, it's a keeper. If, however, Cell B's nearest neighbor is a separate Cell C, then the link between Cell A and B will be discarded.

Keeping mutual links increases the likelihood that the cells are, in fact, the same cell types. Breaking the nonmutual links, on the other hand, prevents cell types specific to each dataset from merging with incorrect cell types. Once all mutual links are found, the algorithm stitches all dataset sequences together. In doing so, it combines the same cell types but keeps cell types unique to any datasets separated from the merged cells. "The mutual links form anchors that enable [correct] cell alignment across datasets," Berger says.

Shrinking data, scaling up

To ensure Scanorama scales to large datasets, the researchers incorporated two optimization techniques. The first reduces the dataset dimensionality. Each cell in a dataset could potentially have up to 20,000 gene expression measurements and as many dimensions. The researchers leveraged a mathematical technique that summarizes high-dimensional data matrices with a small number of features while retaining vital information. Basically, this led to a 100-fold reduction in the dimensions.

They also used a popular hashing technique to find nearest mutual neighbors more quickly. Traditionally, computing on even the reduced samples would take hours. But the hashing technique basically creates buckets of nearest neighbors by their highest probabilities. The algorithm need only search the highest probability buckets to find mutual links, which reduces the search space and makes the process far less computationally intensive.

In separate work, the researchers combined Scanorama with another technique they developed that generates comprehensive samples—or "sketches"—of massive cell datasets that reduced the time of combining more than 500,000 cells from two hours down to eight minutes. To do so, they generated the "geometric sketches," ran Scanorama on them, and extrapolated what they learned about merging the geometric sketches to the larger datasets. This technique itself derives from compressive genomics, which was developed by Berger's group.

"Even if you need to sketch, integrate, and reapply that information to the full datasets, it was still an order of magnitude faster than combining entire datasets," Hie says.

More information: Efficient integration of heterogeneous single-cell transcriptomes using Scanorama, *Nature Biotechnology* (2019). DOI: 10.1038/s41587-019-0113-3 , www.nature.com/articles/s41587-019-0113-3

Provided by Massachusetts Institute of Technology

Citation: Merging cell datasets, panorama style (2019, May 7) retrieved 26 April 2024 from https://phys.org/news/2019-05-merging-cell-datasets-panorama-style.html