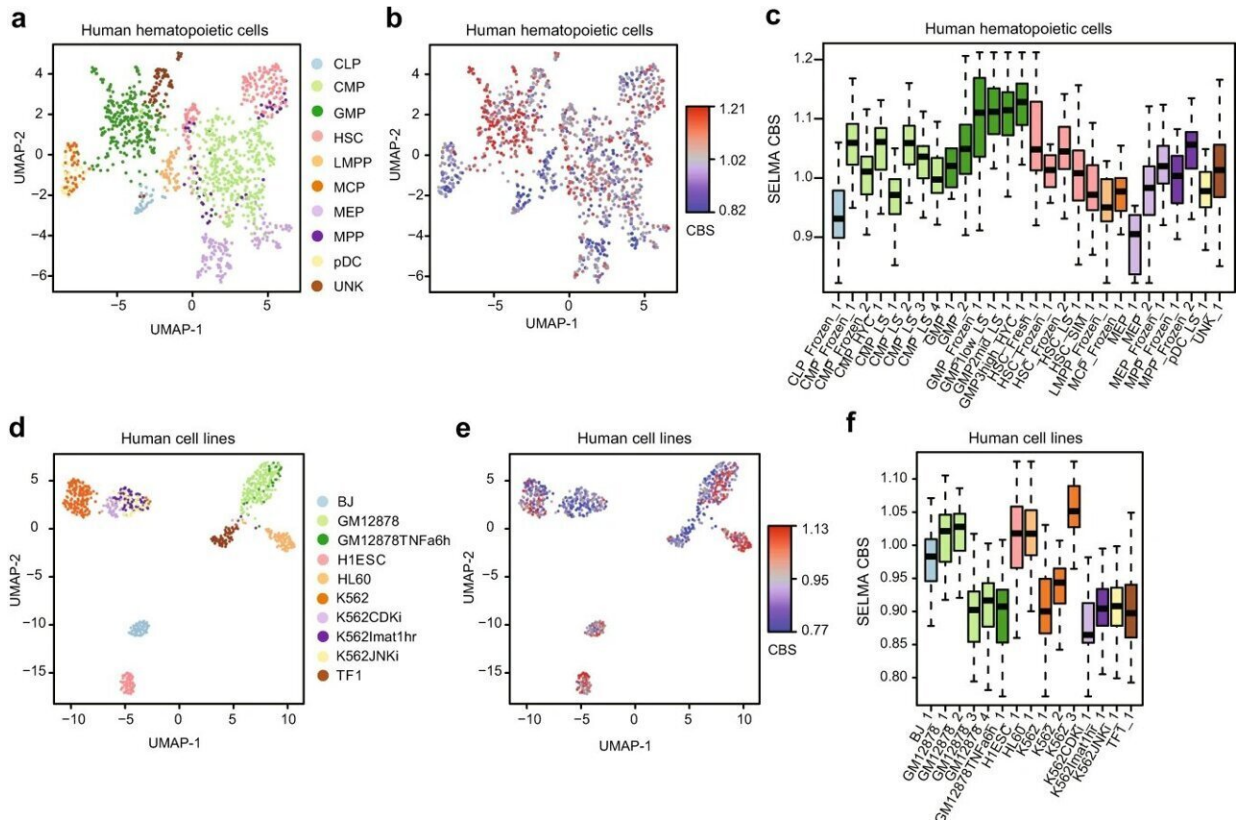
A powerful new tool to advance genomics, disease research

November 22 2022



Intrinsic cleavage biases affect single-cell ATAC-seq data analysis. Visualization of intrinsic cleavage bias effect in different cell clusters derived from scATAC-seq data for different biological samples and different experimental platforms: human hematopoietic cells (**a–c**), mixed human cell lines (**d–f**), mouse primitive gut tube (**g–i**), and 10× Single-Cell Multiome data for mouse embryonic brain (**j–l**), human peripheral blood mononuclear cells (PBMC) (**m–o**), and human lymph node (**p–r**). **a**, **d, g**, **j**, **m**, **p** UMAP visualization where cells are colored by cell type/labels/clusters. **b**, **e**, **h**, **k**, **n**, **q** Same UMAP visualization but cells

are colored by cell bias score (CBS). **c**, **f**, **i**, **l, o**, **r** CBS distributions of cells from different cell types/batches/clusters. Boxes are colored by cell clusters using the same color palette as the first column. The centerline, bounds of box, top line, and bottom line of the boxplots represent the median, 25th to 75th percentile range, 25th percentile – 1.5 × interquartile range (IQR), and 75th percentile + 1.5 × IQR, respectively. Credit: *Nature Communications* (2022). DOI: 10.1038/s41467-022-33194-z

UVA Health researchers have developed an important new tool to help scientists sort signal from noise as they probe the genetic causes of cancer and other diseases. In addition to advancing research and potentially accelerating new treatments, the new tool could help improve cancer diagnosis by making it easier for doctors to detect cancerous cells.

Developed by UVA's Chongzhi Zang, Ph.D., and his team and collaborators, the new tool is a mathematical model that will help ensure the integrity of "big data" about the building blocks of our chromosomes, genetic material called chromatin. Chromatin—a combination of DNA and protein—plays an important role in directing the activity of our genes. When chromatin goes wrong, it can turn a healthy cell into cancer or contribute to other diseases.

Scientists now can study chromatin within individual cells using a cutting-edge technology called "single-cell ATAC-seq," but this generates a tremendous amount of data, including much noise and bias. Zang's new tool cuts through that, saving scientists from false leads and wasted efforts.

As the best of times, large-scale, single-cell genomics research is like "hunting a needle in a haystack," Zang says. But his new tool will make it much easier by clearing away a lot of bad hay.

"Using the traditional way of analyzing the data, you might see some patterns that look like real signals of a particular chromatin state, but they are actually fake due to the bias of the experimental technology itself. Such fake signals can confuse scientists," said Zang, a computational biologist with UVA's Center for Public Health Genomics and UVA Health Cancer Center. "We developed a model to better capture and filter out such fake signals, so that the real needle we are looking for can more easily stand out of the hay."

About the genomics tool

Zang's new tool adapts a model from number theory and cryptology called "simplex encoding." He and his colleagues used that to code DNA sequences into mathematical forms and, ultimately, convert the complex genome sequence into a much simpler mathematical form. They can then compare different forms to detect bias and noise in the sequence data that cannot be found easily using conventional approaches.

"The DNA sequences' complexity increases exponentially when they get longer. They are difficult to model because a typical dataset has millions of sequences from thousands of cells," said Shengen Shawn Hu, Ph.D., a research scientist in Zang's lab and the lead author of this work. "But the simplex encoding model can give an accurate estimation of sequence biases because of its beautiful mathematical property."

Tests of the tool showed it was significantly better at analyzing complex single-cell data to characterize different cell types. This is important for both basic biology research and disease diagnosis, in which doctors must detect tiny numbers of disease cells within much larger specimens, ranging from tens of thousands to millions of cells.

"The biases were not easy to find because they were tangled with real signals and hidden in the big data. It might not be a big deal if people are

only going to pick the strongest signals from a large number of cells," said Zang, who recently co-led several other single-cell genomics research in studying coronary artery disease and gut development.

"But when you look at single-cell data, there are no low-hanging fruits anymore. The signals are always weak on the individual cell level, and the effect of noise and biases can be catastrophic. Bias correction is often ignored but can be vital in single-cell data analysis."

To make their new tool widely available, the researchers have created free, open-source software and posted it online. The software can be found on GitHub.

"We hope this tool can benefit the biomedical research community in studying chromatin biology and genomics, and eventually help disease research," Zang said. "It is always exciting to see our peers use the tools we developed to make important scientific discoveries in their own research."

The researchers have published their findings in *Nature Communications*.

More information: Shengen Shawn Hu et al, Intrinsic bias estimation for improved analysis of bulk and single-cell chromatin accessibility profiles using SELMA, *Nature Communications* (2022). DOI: 10.1038/s41467-022-33194-z

Software: github.com/zang-lab/SELMA and at doi.org/10.5281/zenodo.7048767

Provided by University of Virginia