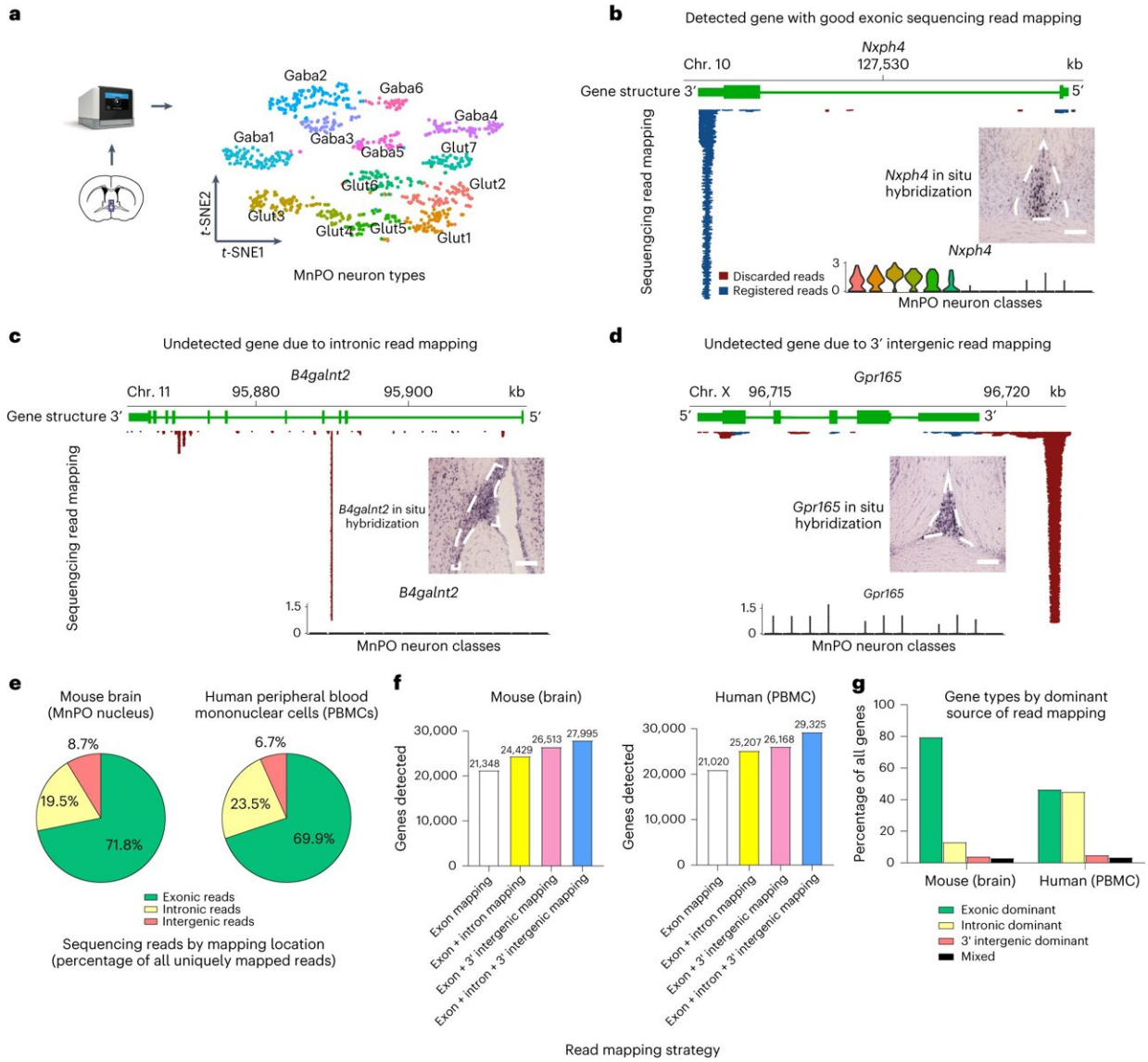
'Invisible' cell types and gene expression revealed with sequencing data analysis improvement

September 11 2023, by Lori Dajose



Missing genes and sequencing read registration in single-cell RNA-seq experiments. Credit: *Nature Methods* (2023). DOI: 10.1038/s41592-023-02003-w

In 2018, researchers in the Caltech laboratory of Yuki Oka, professor of biology and Heritage Medical Research Institute Investigator, made a major discovery: They identified a type of neuron, or brain cell, that mediates thirst satiation. But they were running into a problem: A state-of-the-art technique called single-cell RNA sequencing (scRNA-seq) could not find those thirst-related neurons in samples of brain tissue (specifically, from a region called the media preoptic nucleus) that were known to contain them.

"We knew that the gene labeling we added to our characterized neurons was being expressed in the median preoptic nucleus of the brain, but we didn't see the gene when we profiled that region of the brain with scRNA-seq," says Oka. "We heard this from many colleagues—scRNA-seq was missing cell types and gene expression that they knew should be there. We started wondering why that is."

Identifying different cell types is critical to understanding the vast number of functions performed by our bodies, from healthy processes like sensing thirst to cellular malfunction in disease states. For example, many researchers are currently looking for cell types that may be linked to specific diseases, such as Parkinson's Disease. Determining the precise cell types involved in such processes is critical for all of these studies.

Now, a collaboration between the Oka laboratory at Caltech and the laboratory of Allan-Hermann Pool at University of Texas Southwestern Medical Center has demonstrated how to optimize a key step in scRNA-

seq analysis to recover missing cell types and gene expression data that usually gets discarded. A paper describing the work appears in the journal *Nature Methods* on September 11.

"We've improved the analysis of existing state-of-the-art single-cell RNA sequencing data, revealing the expression of hundreds or sometimes thousands of genes for individual data sets," says Oka. "It is important to enable this type of precision because biological processes are rich and complicated. Recent research has identified over 5,000 distinct neuron types in the mouse brain, and the human brain is presumably more complex. We need our techniques to be as sensitive and comprehensive as possible."

Understanding gene expression

There are trillions of cells in your body, each carrying out the various functions that enable you to live your life—or in some cases, that lead to disease. Cells are differentiated from one another by their function. For example, the immune system's killer T cells seek out and destroy pathogens that cause illness, neurons fire electrical signals that underlie brain function, and skin cells pack together tightly to create a barrier against the outside world. Researchers have currently identified thousands of distinct cell types, but other unique varieties likely remain undiscovered.

Though cells can differ in shape and function, most cells in a given organism contain an identical genetic blueprint—the genome. The genome contains instructions on how to do any cellular task. The genes that comprise the genome are written in DNA, located in the cell's nucleus. Expressed genes are copied into RNA, which is transported out of the nucleus and into the rest of the cell to carry out functions.

In any given cell (and cell type), only a certain subset of genes are

expressed, or turned on, at a given time. These variations in gene expression rise to the differences in cell types.

As an analogy, think of a massive library with books sorted into different sections. If you want to build a plane, you might only check out the books about aviation and mechanics. If you are interested in other topics, you would browse a different set of books. The cells of an individual organism are no different: While each one contains the entire "library" of genes, only those genes that pertain to a specialized cell's unique functions are activated in the cell.

Improving techniques for gene expression estimation

scRNA-seq is a powerful technique to identify cell types. With this method, a cell is broken open and the genetic information expressed inside is labeled with a molecular tag that serves as a barcode. scRNA-seq can quickly do this for thousands of cells in a single tissue sample, with each cell receiving its own unique barcode. Computational analysis can then be performed to determine which sets of genes are expressed in individual cells, and computer models can evaluate that data to look for patterns and identify distinct cell types.

One problem with the technique, however, was that certain RNA sequencing data were commonly not included in gene-expression estimates, even though they represented expressed genes.

The reason, Oka and colleagues found, is related to an issue with the so-called reference transcriptome to which researchers map sequencing data. For example, researchers have extensively studied the mouse genome, and have labeled or annotated it in great detail, creating a digital reference, or "transcriptome," that maps out DNA sequences and their corresponding genes.

This annotation, the researchers found, must be optimized for scRNA-seq to prevent the loss of gene expression information—which can arise if the genes located at the tail ends of a DNA strand are poorly annotated, for example, or if there is extensive overlap between neighboring gene transcripts. Such complications can prevent the detection of thousands of genes. (These issues are particularly pronounced when using high-throughput forms of scRNA-seq, that to reduce cost, examine only the very tail end of genes; most of the atlases that have been created to describe the cellular complexity of our tissues rely on these methods.)

Precision and high resolution is incredibly important when identifying distinct cell types. For example, say that two cells each express genes "A," "B," "C," and "D, but only one cell expresses gene "E" while the other does not. If a sequencing technique does not capture the expression of "E," then the data would suggest that the two cells are identical when in fact they are not.

Led by Pool, a former Caltech postdoctoral scholar and the study's first author, the team optimized the reference transcriptome for the mouse and human genomes, and over the course of several years, built a computational framework to fix the reference transcriptomes of other organisms.

"Optimizing reference transcriptomes enables us to see cell types and states that otherwise we would be oblivious to," says Pool. "For example, with our optimized reference transcriptomes we are now able to observe the full repertoire of thirst-, satiety-, and temperature-sensing neural populations in our brain regions that we suspected would be there but were unable to detect. We expect our approach to also be highly useful in revealing new cellular and genetic diversity in existing and upcoming cell-type atlases for the brain and other organs."

The paper is titled "Recovery of missing single-cell RNA-sequencing data with optimized transcriptomic references." In addition to Pool and Oka, Caltech co-authors are former senior research scientist Sisi Chen and Matt Thomson, assistant professor of computational biology and Heritage Medical Research Institute Investigator. Helen Poldsam of the University of Texas Southwestern Medical Center is also a co-author.

More information: Allan-Hermann Pool et al, Recovery of missing single-cell RNA-sequencing data with optimized transcriptomic references, *Nature Methods* (2023). DOI: 10.1038/s41592-023-02003-w

Provided by California Institute of Technology