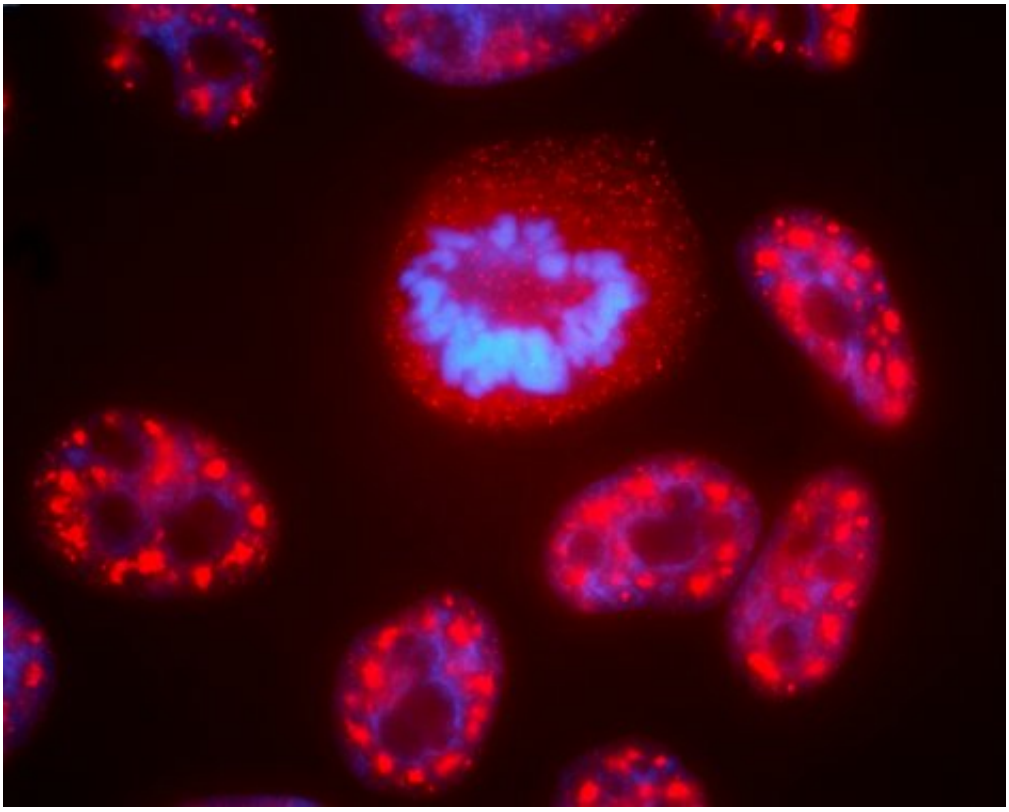
Math points to 100-times faster mapping of gene activity

April 28 2016, by Nicholas Weiler



Credit: National Institutes of Health

New research by UCSF scientists could accelerate – by 10 to 100-fold – the pace of many efforts to profile gene activity, ranging from basic research into how to build new tissues from stem cells to clinical efforts to detect cancer or auto-immune diseases by profiling single cells in a tiny drop of blood.

The study, published online April 27, 2016, in the journal *Cell Systems*, rigorously demonstrates how to extract high-quality information about the patterns of gene expression in individual cells without using expensive and time-consuming deep-sequencing technology. The paper's senior authors are Hana El-Samad, PhD, an associate professor of biochemistry and biophysics at UCSF, and Matt Thomson, PhD, a faculty fellow in UCSF's Center for Systems and Synthetic Biology.

"We believe the implications are huge because of the fundamental tradeoff between depth of sequencing and throughput, or cost," said El-Samad. "For example, suddenly, one can think of profiling a whole tumor at the single cell level."

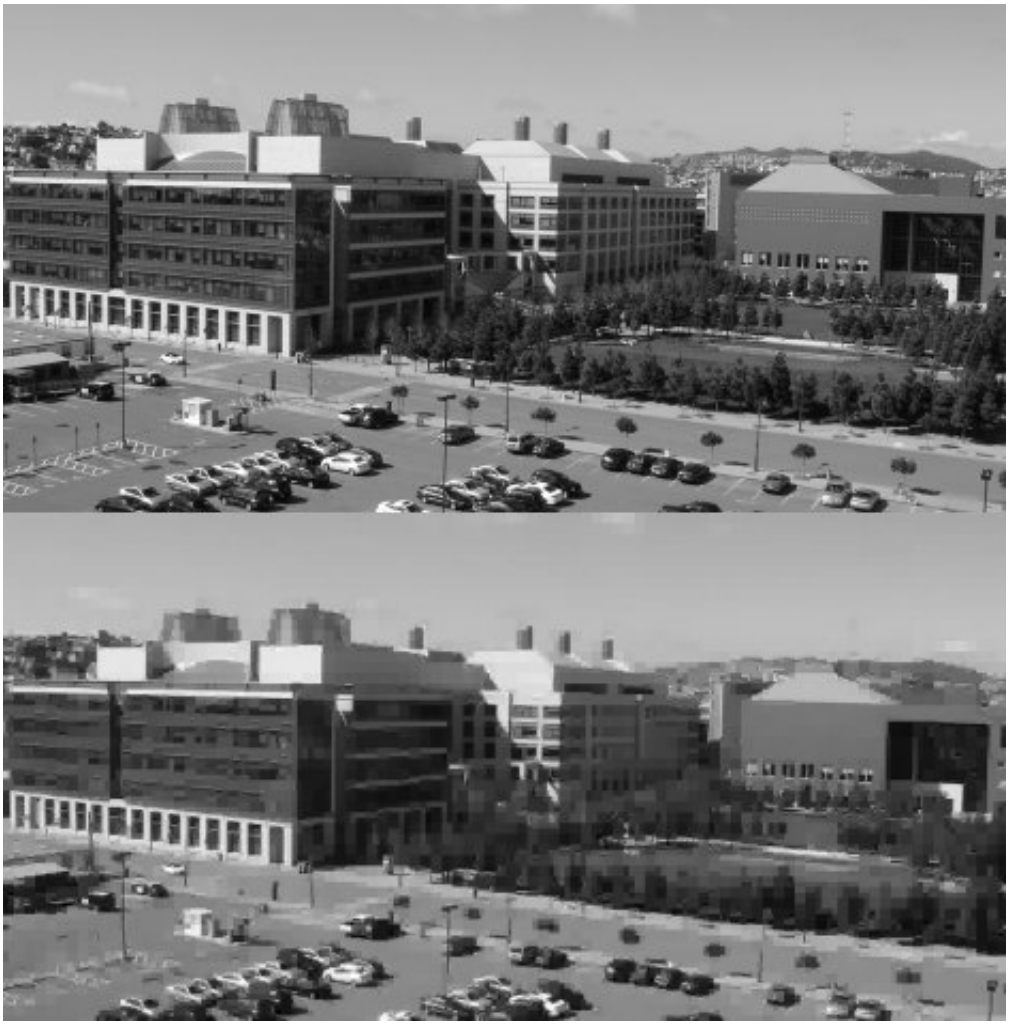
Research Brought Together Several Disciplines

The research stemmed from a collaboration between co-first authors Graham Heimberg, a PhD student in Thomson's lab with a background in bioinformatics, and Rajat Bhatnagar, PhD, a post-doctoral fellow in El-Samad's lab with a background in applied math and electrical engineering. Their combined expertise enabled them to apply engineering insights about how to extract key information from noisy signals to the pressing biological problem of how to more efficiently analyze large-scale gene activity datasets.

Heimberg and Bhatnagar had one fundamental insight: because each gene in a cell is typically part of one or more much larger gene programs – groups of dozens or hundreds of related genes that regularly get activated together—the complete readouts of gene activity made possible by deep sequencing are full of redundant information. For many modern applications of gene sequencing that care more about patterns of gene activity than the individual genes, they reasoned, the same results could be extracted from data with much lower resolution.

Think of it like looking at a fuzzy photograph of a city. You can probably pick out individual buildings and decide if you're looking at San Francisco or New York or Paris, but if you want to count the windows in the buildings or the number of cars on the streets, you're out of luck.

"We're not at all saying that high-depth sequencing is useless," said Thomson. "Far from it. Deep sequencing is an amazing tool for getting specific, molecular-level information about individual genes and gene mutations. We're just pointing out that with the right analysis, shallow sequencing can be much faster and cheaper for extracting cell-level gene expression information."



The researchers compare their approach to the image compression algorithms that reduced the bottom photograph of the UCSF Mission Bay campus by 100-fold, preserving most large-scale landmarks and many key details, while sacrificing the finer detail available in the original image.

To demonstrate their point, the researchers analyzed hundreds of publicly available gene expression databases derived from yeast, mice and humans. They showed that for common applications like detecting what tissue a cell comes from or picking out different types of neurons by their signature patterns of gene activity, the right mathematical analysis can pull the necessary information out of 10 to 100 thousand sequencing reads, rather than the millions of reads that constitute deep sequencing.

The researchers went further, deriving a theoretical framework that demonstrated exactly how deep sequencing needs to be to obtain a specific level of detail about the gene activity of a given cell or tissue. The basic conclusion, they say, is that the "dominance" of a given gene program within a dataset—that is how much of the dataset's spread that group of genes explains – determines the depth of sequencing needed to extract it. In other words, the big features are easy to pick out, but the details require higher resolution.

Low-Resolution Sequencing Could Speed Cancer Detection

The upshot of the new paper is that the sequencing pipeline could be made to flow tens to hundreds of times faster for the numerous genomic applications in which the big features of gene expression are probably the most important. This might include screening the blood for

individual cells on their way to becoming cancerous, identifying the genetic pathways that control stem cell growth, or building an atlas of the gene expression programs that build the human body.

This is crucial, Thomson and El-Samad say, because particularly for increasingly important techniques that rely on sequencing DNA from individual cells (such as the cancer liquid biopsy example above), the sequencing itself is now a major bottleneck.

For example, UCSF's Center for Advanced Technology (CAT) currently has a machine that can prepare 50,000 cells for sequencing in one long day of work, but even with the CAT's most advanced sequencing machine (which can do 5 billon reads in a day and a half) it would take more than two weeks for researchers to deep-sequence the full pattern of DNA activity in those 50,000 cells, at a million reads per cell. But if researchers can extract the relevant information from just 20,000 reads per cell, as the new research suggests, they could sequence 150,000 cells in just one day.

Speeding up the pipeline in this way could be transformative for many research and clinical applications of sequencing that are currently considered too costly or time consuming.

For example, El-Samad says, many cells have very redundant molecular pathways that all seem to do similar things within the cell, but which respond differently to different drugs.

"If you want to profile how different drugs affect these pathways – which are by definition composed of many different genes – it would be a huge waste of time and money to test every drug on every gene at full sequencing depth, not to mention probably logistically impossible," she said. "On the other hand, if you can quickly identify which pathways are activated from much lower resolution sequencing, it means you can

really go crazy on the number of drugs you can test in a reasonable amount of time."

The same principle holds in Thomson's work studying how stem cells differentiate, he says. There are huge numbers of genes involved, but a much smaller number of gene pathways, so low-resolution sequencing is enabling the lab to quickly identify which pathways transform the cells in particular ways, after which further experiments can be done to elucidate which genes can be controlled as key drivers of these pathways.

The researchers have used the equations derived in their paper to create a read-depth calculator to help other researchers determine the resolution of gene activity information required their specific application, which they posted online at Thomson's lab website.

"Our computational results from all these datasets demonstrate that this phenomenon holds across the board for all different kinds of genetic data, and the theoretical part explains why that is," El-Samad said. "But it's nice to just have a formula so you don't have to keep processing data over and over."

More information: *Cell Systems*, dx.doi.org/10.1016/j.cels.2016.04.001

Provided by University of California, San Francisco