Single-Cell Genomics: Can Bioinformatics Unlock its Potential?

The tools to sequence the genomes of individual cells yield data that’s noisy and somewhat unreliable. What bag of tricks can bioinformaticians use to address these challenges?

By Kristin Sainani, PhD

[*leave a comment*](http://biomedicalcomputationreview.org/content/single-cell-genomics-can-bioinformatics-unlock-its-potential#comments)

[*print*](http://biomedicalcomputationreview.org/print)

[*pdf*](http://biomedicalcomputationreview.org/sites/default/files/singlecell_bcr-spring-2016-web.pdf)

To study genomes, researchers have typically pooled the genetic material from thousands of cells together. But this approach can only get at “average genomes” or “average transcriptomes.” And sometimes this isn’t enough.

“Averages can be very misleading or not meaningful,” says **Cole Trapnell, PhD**, assistant professor of genome sciences at the University of Washington. “Plus, there are certain biological questions that you cannot answer unless you take single-cell measurements.”

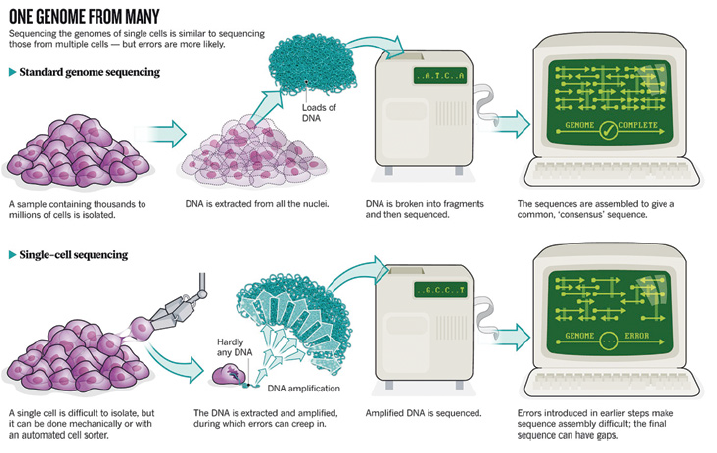
For example, the brain consists of multiple cell types enmeshed with one another; and cancerous tumors are a mix of genetically diverse cells, including some that drive invasion, metastasis, and treatment resistance. To understand what’s really going on in such heterogeneous groups of cells, researchers need to study the genomes of individual cells.

Now, using single-cell genomics, researchers have the tools to resolve individual clones within tumors, discover new cell types in the brain, find genetic abnormalities in embryos, and detect rare cancer cells in the blood, among other exciting applications.

But to unlock the full potential of single-cell data, advances in bioinformatics are needed. Many bioinformatics tools that were developed to process and analyze bulk data don’t work well when applied to single-cell data. Plus, novel algorithms will be needed to address the biological questions that only single-cell data can answer.

**Single Cell Genomics Defined**

With recent breakthroughs in isolating individual cells and in amplifying and sequencing DNA and RNA, researchers can now measure the genomes and transcriptomes of thousands of individual cells at once. A single cell carries only two copies of DNA and

*[](http://biomedicalcomputationreview.org/sites/default/files/imagecache/article-inline-image/u38/16srpingp10.png)*

*Reprinted with permission from MacMillan Publishers, Brian Owens, Sequencing DNA from individual cells is changing the way that researchers think of humans as a whole, Nature 491, 27–29 (2012).*

sometimes just a few messenger RNA (mRNA) transcripts. To sequence the DNA or make sense of its transcripts, researchers must first amplify those numbers as much as a billion-fold using powerful new amplification technologies such as Multiple Annealing and Looping Based Amplification or Multiple Displacement Amplification. This creates a “pool” of genetic material that can then be sequenced and analyzed just as pooled genetic material from multiple cells would be.

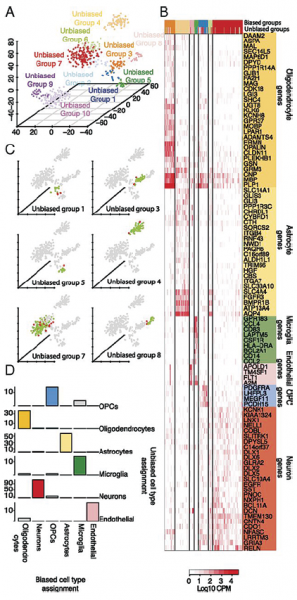
But there is a difference: The data are noisier and less complete than the bulk data garnered from multiple cells. Single-cell data are also larger in size and scale—often involving hundreds or thousands of samples rather than the tens of samples typical of bulk experiments. This is where bioinformaticians have their work cut out for them.

**Noisy Data**

When the small numbers of DNA or mRNA transcripts from a single cell are amplified, some regions of the genome or some transcripts are amplified better than others, leading to distortions. In addition, researchers cannot replicate their work on a single cell—because the same material can’t be amplified and measured twice. This makes it harder to separate experimental errors from real biological variation. In an October 2015 paper in *Nature Communications*, researchers in the UK estimated that of the observed variation in gene expression patterns in single-cell genomics experiments, only 18 percent was due to true biological variation. The remainder was due to technical noise.

 “Most of the computational efforts so far have been toward trying to separate the true signal from the noise. That’s where most of the field is now,” says **Peter Kharchenko, PhD**, assistant professor of biomedical informatics at Harvard Medical School.

Experimental tricks can help, Kharchenko notes. For example, researchers can tag each original mRNA transcript with a “unique molecular identifier”—a short random sequence that acts like a barcode—prior to amplification. Since each unique tag corresponds to only one transcript molecule in the original sample, this method can generate an accurate transcript count regardless of amplification errors. “The computational aspects of this are pretty straightforward, but it results in a drastic reduction of the noise in the data,” Kharchenko says. Also, researchers can add “spike-in RNAs” to each sample—control RNAs with known composition—to help detect experimental aberrations.

*[](http://biomedicalcomputationreview.org/sites/default/files/imagecache/article-inline-image/u38/16springp11.png)*

*When Steven Quake’s group at Stanford used an unbiased approach to sort single-cell genomic data for 466 individual cells into distinct groups defined by the entirety of their molecular signatures, they identified 10 distinct cell groups (8 adult [1-8] and 2 fetal [9-10]) as shown in the landscape in (A). These classifications compared favorably to a biased approach based on known markers for specific cell types (B-D). In C, for example, the classifications agreed for cells shown in green and did not agree for the small number of cells shown in red. And in D, agreement is shown in colored blocks and disagreement in gray, with the number of cells shown on the y-axis. Reprinted from S Darmanis, SA Sloan, Y Zhang, et al., A survey of human brain transcriptome diversity at the single cell level. PNAS 112(23): 7285-90 (2015).*

Computational tricks for reducing noise are also being introduced, but remain harder to implement. “Right now, the algorithms aren’t packaged in ways that average people can use, so you have to have a bioinformatics person to string it all together,” says **Robert C. Jones, MS**, executive vice president for research and development at Fluidigm, a company that makes tools for single-cell genomics. “Someday we hope to provide our customers with a regular pipeline so that people can just turn the crank.”

User-friendly software is beginning to emerge. A 2015 paper in *Nature Methods* introduced Ginkgo ([*http://qb.cshl.edu/ginkgo/*](http://qb.cshl.edu/ginkgo/)), an interactive Web-based program that automatically processes single-cell DNA sequence data, maps the sequences to a reference genome, and creates copy number variant profiles for every cell. The software—which was created in the lab of **Michael Schatz, PhD**, associate professor of quantitative biology at Cold Spring Harbor Laboratory—has built-in algorithms to correct amplification errors.

**Missing Data**

Amplification is fraught with another key problem: Some regions of the genome or mRNA transcripts may be completely missed. “Your body’s very good at copying entire chromosomes. To do so requires this amazingly beautiful orchestrated dance where you bring together many proteins, including for proof-reading and error-correction,” Schatz says. “We hijack some of those systems to make copies of the DNA through PCR (polymerase-chain reaction), but it’s not nearly as sophisticated as what goes on inside your body.” As much as 30 percent of the genome may be unamplified and missed; and as many as 60 percent of heterozygous alleles may be missed. With RNA, the problem is even worse—researchers estimate that some protocols miss as many as 60 to 90 percent of all transcripts present in the cell.

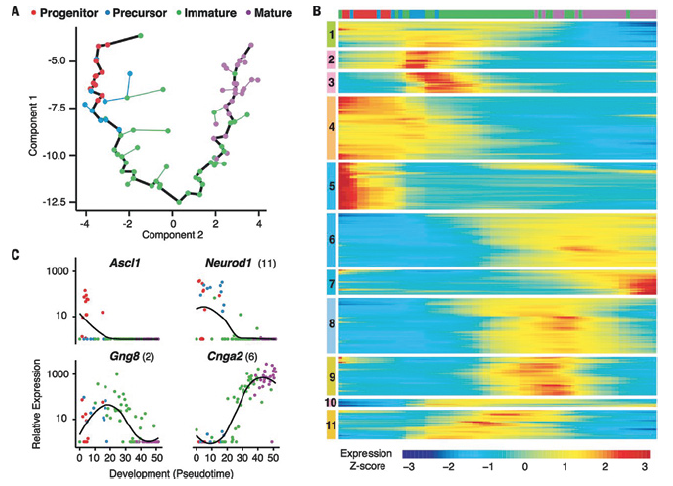
“There’s a big zero problem,” Trapnell says. “We have to find new ways to deal with that missing data. There’s discussion right now in the community about how best to do that: Do we want to fill it in based on our best guess? Do we want to build models that can tolerate a lot of zeros, and don’t have a problem with it? It’s not obvious what the right way to go forward is.”

Kharchenko’s group has developed software called SCDE (Single-Cell Differential Expression) to analyze single-cell RNAseq data ([*http://pklab.med.harvard.edu/scde/*](http://pklab.med.harvard.edu/scde/)). The model uses a Bayesian approach that accounts for the likelihood of dropout events. “We had to incorporate explicitly the probability of failing to observe a gene. By predicting the probability of not being able to see a gene in a given cell, then you can propagate that uncertainty further into other analyses,” Kharchenko says. “The computation becomes a little more complicated, but you’re better off taking into account the uncertainty of the measurement.”

Drawing on information gleaned from bulk data—such as the frequency of a particular mutant allele in a tumor—can also give clues as to the impact of dropouts. “So it’s not so much about sheer brute-force analytic methods in the single cell—it’s also knowing how to bring in different datasets to help you make better sense of everything,” says **Winston Koh**, a doctoral student in bioengineering in Stephen Quake’s lab at Stanford University. In a 2014 paper in *PNAS*, Koh and others combined bulk and single-cell genomic data to reconstruct the clonal architecture of childhood acute lymphoblastic leukemia.

**Making Sense of Single-Cell Data**

To gain biological insight, researchers start by grouping cells with similar gene or gene activity profiles. But clustering is tricky

*[](http://biomedicalcomputationreview.org/sites/default/files/imagecache/article-inline-image/u38/16springp12.png)*

*Using single cell gene expression, Trapnell and his colleagues determined that olfactory neurons exhibit large-scale shifts in gene expression during development. An unsupervised analysis of single-cell gene expression profiles using an algorithm called Monocle revealed a linear trajectory (black line in A) along which cells develop in pseudotime. Coloring of cells based on the expression of developmental markers shows that the trajectory corresponds to a stepwise development from olfactory progenitors to precursors to immature and ultimately mature olfactory sensory neurons. Global analysis of gene expression kinetics along the trajectory identified 3,830 genes that vary significantly during development (B) and can be hierarchically clustered into 11 nonredundant groups that covary over the trajectory. The bar on top shows the locations of individual cells, colored by stage of development, along this developmental trajectory. The Expression Z-score indicates changes in a gene relative to its dynamic range over pseudotime. Kinetic diagrams (C) show the expression of known markers of different developmental stages over the developmental progression. Reprinted with permission from NK Hanchate, K Kondoh, Z Lu, et al., Single-cell transcriptomics reveals receptor transformations during olfactory neurogenesis, Science 350:6265/1251-1255 (2015).*

because single-cell data are high-dimensional (involving thousands of sequences or expression profiles) and involve complex relationships. Traditional clustering algorithms such as Principal Components Analysis (PCA)—which assume a simple linear relationship between variables—aren’t optimal for these data. So, bioinformaticians are exploring alternatives, including t-SNE (t-distributed stochastic neighbor embedding), which is well-suited for high-dimensional, non-linear data.

“I’m pretty impressed with how it’s all coming together in the field. We are moving beyond simple PCA-type analyses to more sophisticated algorithms,” says **Stephen Quake, D.Phil**, professor of bioengineering at Stanford University and a founder of Fluidigm. “The whole ecosystem around that is looking very promising.” In a 2015 paper in *PNAS*, Quake’s team applied t-SNE to single-cell transcriptome data from 466 brain cells to identify 10 distinct cell groups in the adult and fetal brain—eight of which corresponded to known cell types in the brain. They then further classified cells into subpopulations based on their gene expression profiles—a first step towards building a comprehensive atlas of cell types in the brain.

Beyond grouping cells, researchers are also developing algorithms for ordering cells by temporal or developmental stage. This problem is challenging because it may require bioinformaticians to rethink their approach to cell classification, Trapnell says. Rather than trying to classify cells into clear-cut, discrete types, we should think of cells as lying more on a continuum, he says. “There’s a desire to put things into nice neat bins. And I think that’s not working well so far for single-cell data. So we might just need to be a little bit more flexible about how we analyze this stuff.”

Trapnell’s lab has developed Monocle, a toolkit for single-cell expression data that reconstructs the trajectory along which cells are presumed to travel, such as during development or differentiation ([*http://cole-trapnell-lab.github.io/monocle-release/*](http://cole-trapnell-lab.github.io/monocle-release/)). “Monocle is designed to put cells in continuous order by how differentiated they are—from undifferentiated stem cells to the fully differentiated state,” Trapnell says. In a recent paper in *Science*, Trapnell and others used Monocle to track the maturation of nasal olfactory cells in mice. “Because we capture the complete, continuous progression from neuronal progenitors to mature neurons, we can see the exact moment in development that these neurons select which member of a large family of sensory genes to express,” Trapnell says.

**Rapidly Changing Technology**

The technology for single-cell genomics is rapidly advancing—and the bioinformatics challenges may shift accordingly. “It’s quite frustrating because the approaches that you’ll use one day could be totally changed the next day,” Schatz says. Just as bioinformaticians will have to keep pace with the technology, biologists will need to stay abreast of the latest bioinformatics innovations, he says. “Researchers who want to use these technologies have to pay really close attention to the state-of-the-art in the field and make sure that they are using all the best practices available at the time.”

Tags:

* [*Feature*](http://biomedicalcomputationreview.org/site/search?keys=feature)

* [*Genomics*](http://biomedicalcomputationreview.org/site/search?keys=genomics)