Biology & big data

How computational biology is shaping the future of health and privacy.

by Lori Dajose

The study of living things is undergoing a revolution.

In the past few decades, cutting-edge biological tools have enabled the rapid collection of unprecedented amounts of data. Biologists and bioengineers have amassed countless terabytes of high-resolution videos of microscopic cells as they wiggle and grow and interact; sequenced millions of genomes, from *Escherichia coli* to mice to humans; and tweaked bacterial DNA to reengineer life. Within these vast data sets lie answers to fundamental questions of biology: What are the molecular rules that control development? How are stem cells "wired?" How many different types of cells make up the human brain? Can failures of single cells cause disease?

But manual analysis of all this information is virtually impossible. Fortunately, the fields of computer science and artificial intelligence are undergoing their own rapid development. These tools as applied to the biological sciences have given rise to the field of computational biology.

Lior Pachter (BS '94), **Matt Thomson**, and **David Van Valen** (PhD '11) all recently joined Caltech's faculty as part of an Institute initiative to focus attention on computational biology. *Caltech* magazine sat down with them to discuss the ongoing biological data explosion and its harmonious relationship with computational tools as well as how these intersecting revolutions will change the future of privacy, ethics, and what it means to be human.

What does it mean to study biology? What are the goals and challenges?

Matt Thomson: Doing biology is not only about measuring and observing. It's about actually changing and perturbing the biological systems, making predictions and models, tweaking them based on your observations and doing it all over again.

To give an analogy, let's say you throw a ball up in the air. If you know parameters like the ball's mass, the acceleration due to gravity, its initial velocity, and so on, you can accurately predict where the ball will land after a certain amount of time. We want to predict how biological systems will evolve, but it's difficult because there are so many parameters.

David Van Valen: The first set of parameters you think of are genes. For example, a simple *Escherichia coli* bacterium has 3,000 to 4,000 genes. What does each gene do? How do they interact? Imagine understanding an airplane and all of its component pieces ... an organism is at least 10 times harder.

Lior Pachter: In the past, some may have imagined that it was simple, that one gene encodes for, say, hair color, and another makes you happy or sad. It appears that biology doesn't work like that. It's a very complicated interwoven network of objects that interact in very complicated ways.

MT: Right. Living things are dynamic and heterogeneous, changing and evolving through time and space. Various genes can be expressed at different levels throughout an organism's lifetime.

DVV: You might think that you could just sort of take averages and glean insights that way, but you can't. You can't take a lung and blend it up and sequence all that matter and then understand a lung, because there are different types of cells (epithelial cells, endothelial cells, and so on) in different locations working with one another. We need techniques that can respect these heterogeneous differences in order to understand whole organisms.

But biology is really exciting right now, because for the first time, we're having solutions come up for all of these challenges. Simultaneously.

Can you talk about some of the new technologies that are impacting biology?

MT: In just the past decade, exciting and powerful technologies are emerging, like CRISPR-Cas9, the technique that allows us to edit genomes. The first full human genome was sequenced in 2003, after 13 years of work. Now, in 2019, sequencing all 20,000 genes in the entire human genome takes only a day or two, if not less.

LP: Matt and I work on developing techniques to identify all of the RNA molecules in individual cells within a sample of tissue. Knowing which RNA are present in a given cell can tell you which genes are activated and, therefore, what the cell is trying to do. The basic way RNA sequencing works is to flow cells, one by one, through a narrow pipeline and encapsulate each cell in its own water droplet. Within the droplet, the cell is broken open and all of the messenger RNA molecules inside are tagged with a barcode unique to that particular droplet. Then we gather up all of the messenger RNA from all of the cells and sequence it in one big batch. The barcoding enables us to know which genes came from which cells.

MT: We can profile 100,000 cells in a day and a whole mouse embryo in less than a week. Our colleague Long Cai is aiming to be able to profile 1 million cells in a day.

DVV: Genomic assays give us a sense of the composition of living systems in a way that we can respect their large "parts list." For understanding how things vary in space and time, we have imaging technologies. Our microscopes are now so good that we can look at whole tissues, we can look at single cells, we can look at single molecules. These technologies are starting to talk to one another, too. We basically repurpose machine-learning algorithms to identify individual cells, so that lets us look at things in a way that respects the important differences in datasets that have mixed information. These tools really are going to empower researchers to carry out a new generation of experiments.

In the 1960s, Gordon Moore (PhD '54) predicted that computational power would double every two years, which has turned out to be quite accurate. Is there a kind of Moore's law-like prediction for biology?

MT: Oh, it's actually super-Moore's law. It's faster.

LP: We've seen that in basically any given biological technology. I don't think there has been, in history, technological progression at this speed ever. Not even in computers.

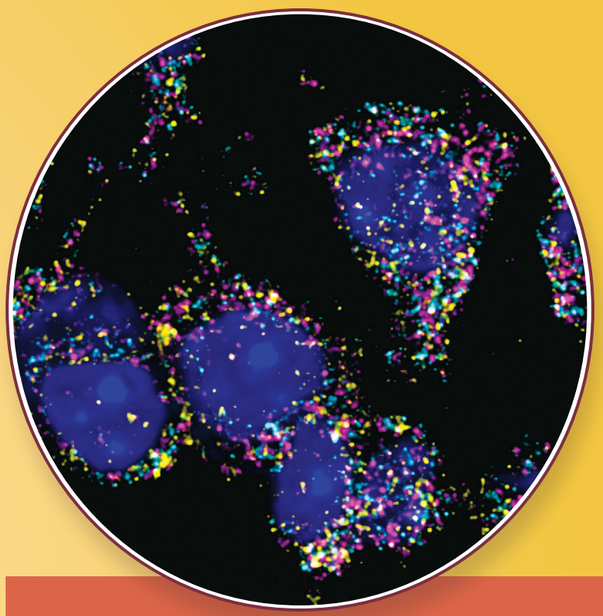
MT: Three years ago, a particular experiment to perturb a biological system and measure responses would require a massive national consortium of scientists, like 200 people, to collect and analyze data over years. Now, our lab can do these sorts of things in a week or so. We need automated and efficient ways to analyze all this data, and that's where machine learning comes in.

What exactly *is* machine learning?

LP: Machine learning is the process of using computational tools to predict and learn from data. These tools can be used in a variety of ways, from combing through telescope data to find planets outside our solar system to teaching a computer how to recognize moving objects in order to drive a car.

DVV: You can give a computer some example data sets and teach it how to look for insights. Then, once it has "learned," you can give it a totally new data set to analyze. It's a kind of artificial intelligence, and it has broad applications.

LP: Some of it even got its start here at Caltech, when



To understand complex biological systems, it is important to know how cells interact with and influence their neighbors. For example, a healthy cell located directly next to a cancerous cell will receive different chemical signals and behave differently than a cell elsewhere in the body. A newly developed technique from the laboratory of **Long Cai** colorfully illuminates every mRNA in every cell within a tissue sample with "super-resolution." The technique can be applied to study everything from embryos to cancers.

Santiago Lombeyda, a computational scientist with Caltech's Center for Data Driven Discovery, created the illustration featured in this article. In this visualization of data points from a large single-cell study, each dot represents a single cell, which consists of 20,000 independent gene expression measurements, that were then mathematically mapped into a 3-D space. Caltech senior research scientist Sisi Chen provided the data and analysis.

From left: Matt Thomson, David Van Valen, and Lior Pachter.

researchers in the 1980s were inspired by neuroscience to develop computational methods for data analysis called neural networks.

MT: Many of us biologists are now working with Caltech's AI4science Initiative, which brings together computer scientists and other researchers to use computing tools to get insights from our data. Machine-learning algorithms can be used for problems ranging from detecting fake news to classifying genetic sequences.

What kinds of things can we discover by applying machine-learning tools to biological data sets?

DVV: Image analysis is a big area where machine learning can help. In my lab, we work on repurposing machine-learning algorithms, like the ones used to do computer vision for self-driving cars or that Facebook uses to recognize and locate people in pictures. We refashion them to analyze imaging data from microscopes and make these tools available so that biologists everywhere can use them. For example, a pathologist can use these techniques to look at breast cancer cells and learn about the interactions between immune cells and cancer cells.

MT: We use machine learning to look at really high-dimensional data sets, including genomes and single-cell sequencing data. It can reveal relationships between networks of genes, such as which genes are controlling the expression of other genes in a network. This way we can find the "master genes" that control processes like brain development or that control the development of T cells in the immune system.

LP: The possibilities for what we can learn from all of this biological data are both exciting and sobering. I think we're really not that far from the unimaginable: changing who we are by changing our biology. What does it mean to be you? There are very profound fundamental changes and possible biomedical technology changes that are going to blow us away. It's going to come very fast.

DVV: Society is not prepared for it.

LP: Society is completely unprepared for it!

DVV: These technological developments will drastically reshape what the world looks like, but many of the people who are going to be affected by these tools are not aware of what's happening. There are questions about the medical industrial complex and healthcare system: Should we do basic genetic manipulations on embryos to get rid of disease-causing mutations? How about manipulations to determine eye color, height, skin color? We have to be having deep conversations, as a society, about what we actually want. How do we use these tools to create a just society?

LP: There is a law in the United States that prevents healthcare discrimination on the basis of genetics. But what about other kinds of discrimination? Can a university ask for a prospective student's genome and take genetics into consideration?

So, not only are computer science and biology meeting but philosophy and ethics are becoming part of the conversation as well.

LP: Yes. Here's one example. Private companies are offering to sequence anybody's genome because we have the ability to do that now. But the companies can collect this data, and we need to be having a conversation about the extent to which the data should or should not be private. Who can have access to it? It's complicated and subtle: if somebody makes their own data public, then they are implicitly releasing information about their relatives without their consent.

The genome is actually the most trivial thing that you can measure on a person these days. What if you could take a sample of someone's tissue and figure out aspects of their current state of health and well-being? This would be revolutionary in fighting diseases but also makes it very easy to get deeply personal information about a person.

MT: In parallel, there are all the internet companies that are figuring out your preferences and ideologies based on your search history, your social media friends, the things you post. Tons of data about your personality. Just imagine combining that with genetic data about you as a person. … You could get a full picture of society. Just imagine how advertisers could use this information. Right now, a vast majority of this information is in the hands of private companies.

DVV: Well, this got dark.

LP: On a more positive note, these hypothetical scenarios that we are imagining, the reason we're bringing this up is because we are all aware of how crazily exciting the development is in our field. It's just moving so fast. These are not conversations that are framed in some science fiction world, they're very real. There is so much data and information that has been collected already, and there is a lot to learn from it. I think we are very privileged to come to Caltech and do this now.

DVV: It's one of the unique things that makes Caltech so exciting. We get students who are fascinated by the biological questions but also have the quantitative backgrounds that are now necessary for doing this type of work.

Lior Pachter (BS '94) is a Bren Professor of Computational Biology and Computing and Mathematical Sciences, as well as an affiliated faculty member of the Tianqiao and Chrissy Chen Institute for Neuroscience at Caltech.

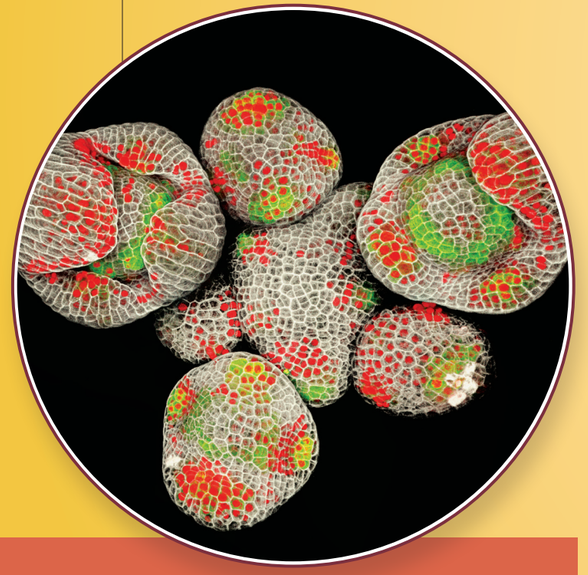
Matt Thomson is an assistant professor of computational biology and Heritage Medical Research Institute Investigator.

David Van Valen (PhD '11) is an assistant professor of biology and biological engineering.

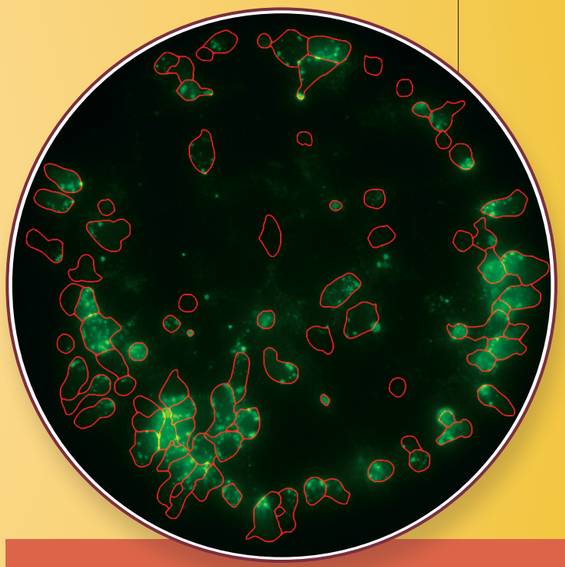
Long Cai is a professor of biology and biological engineering, as well as an affiliated faculty member of the Tianqiao and Chrissy Chen Institute for Neuroscience at Caltech.

Elliot Meyerowitz is the George W. Beadle Professor of Biology and a Howard Hughes Medical Institute Investigator.

Ellen Rothenberg is the Albert Billings Ruddock Professor of Biology.



How do plants grow? How do plant cells decide to be part of a flower or a stem? What genes do they express? In the laboratory of **Elliot Meyerowitz,** these questions are answered using machine-learning image- analysis tools to learn from detailed images and videos of plant growth and gene expression.



How do certain immune cells, called T cells, develop? How do they interact? What genes do they express, and when? Machine-learning image analysis enables researchers in the laboratory of **Ellen Rothenberg** to analyze T cell development and gene expression in real time.