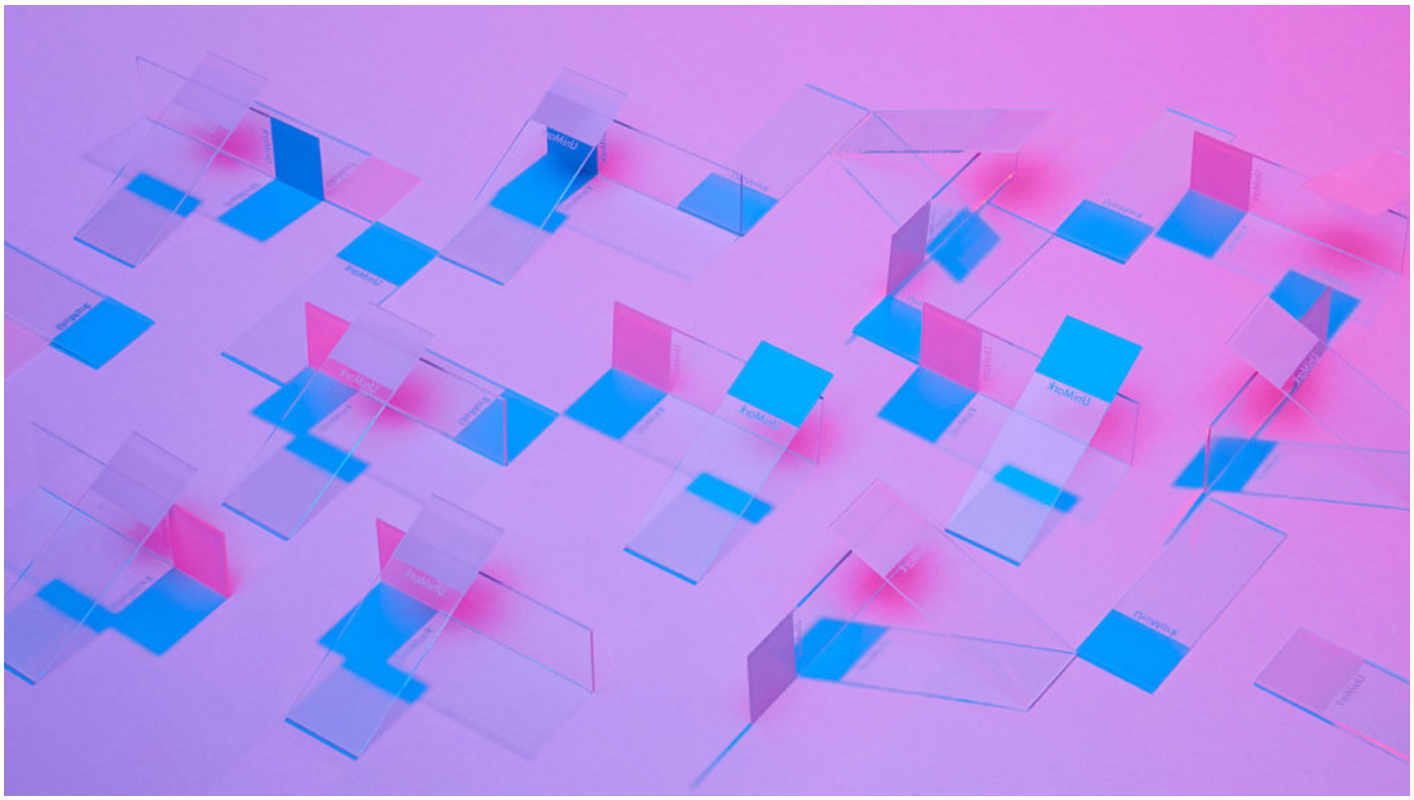
NEWS

Studying genetics in the age of big data

BY **TOM CHIVERS, MOSAIC**

4 OCTOBER 2018



New biomedical techniques, such as next-generation genome sequencing, are creating vast amounts of data and transforming the scientific landscape. They're leading to unimaginable breakthroughs -- but leaving researchers racing to keep up.

“This is when I start feeling my age,” **Anne Corcoran** says. She’s a scientist at the Babraham Institute, a human biology research center in Cambridge, England. Corcoran leads a group that looks at how our genomes -- the DNA coiled in almost every cell in our bodies -- relate to our immune systems, and specifically to the antibodies we make to defend against infection.

She is, in her own words, an “old-school biologist,” brought up on the skills of pipettes and Petri dishes and protective goggles, the science of experiments with glassware on benches -- what’s known as 'wet lab' work. “I knew what a gene looked like on a gel,” she says, thinking back to her early career.

These days, that skill set is not enough. “When I started hiring Ph.D. students 15 years ago, they were entirely wet lab,” Corcoran says. “Now when we recruit them, the first thing we look for is if they can cope with complex bioinformatic analysis.” To be a biologist nowadays, you need to be a statistician, or even a programmer. You need to be able to work with algorithms.

An algorithm, essentially, is a set of instructions -- a series of predefined steps. A recipe could be seen as an algorithm, although a more obvious example is a computer program. You take your input (ingredients, numbers or anything), run it through the algorithm’s steps -- which could be as simple as “add one to each number” or as complex as Google’s search algorithm -- and it provides an output: a cake, search results, or perhaps an Excel spreadsheet.

Researchers such as Corcoran need to use algorithms because, in the 17 years since she became a group leader, biology has changed. And the thing that has changed it is the vast -- the overwhelmingly, dizzyingly vast -- flood of data generated by new biomedical techniques, especially **next-generation sequencing**.

Not long ago, sequencing an entire genome -- determining the order of all 3 billion pairs of DNA letters in the helix -- took years. The Human Genome Project, the first completed sequence of an entire human genome, took around 13 years from conception to its completion in 2003, and cost more than 2 billion pounds. Today, next-generation sequencing can do the same thing in 24 hours for not much more than 1,000 pounds (about $1,300).

This has completely changed how scientists work. It’s not just that they get their hands dirty less often, nor simply that the required skills have changed. It’s that the whole process of science -- how you come by an idea and test it -- has been upended.

This has left a lot of senior scientists needing to understand and supervise techniques that didn’t exist when they trained. It’s left universities playing catch-up, with many degrees not teaching the skills that modern biologists need. But above all, it’s led to ground-breaking scientific discoveries -- breakthroughs that simply wouldn’t have been possible 20 or even 10 years ago.

From workbench to laptop:

A 10-minute drive from Babraham, in a village called Hinxton, there’s another major life-sciences center, the Wellcome Sanger Institute. It’s 25 years old this week, and the rapidly moving history of genomics is written in its very architecture.

“I did my postdoc at the Sanger,” says **Moritz Gerstung**, now a research group leader at the European Bioinformatics Institute next door. He chuckles at the memory. “You can almost sense when the building was conceived,” he says. “There’s so much space for laboratory work and not so much for where scientists can sit and analyze data on a computer.”

This is true everywhere, says **Gil McVean**, professor of statistical genetics at the University of Oxford's Big Data Institute. Genomic research has become something done mainly on a laptop, not a workbench. "If you look at any 15-year-old research lab, they're 90 percent wet lab," he says. "And if you go into one, almost all the people are sitting at computers. If you were to build a biomedical research center today, you'd build it 10 percent wet lab and 90 percent computing."

But that's not the only change. "One of the big changes in science," says McVean, "has been the move away from a very focused, targeted, hypothesis-driven approach, the 'I've got this idea, I design the experiment, I run the experiment and decide whether I was right or wrong' model."

It used to be that you had to have some plausible idea about *why* a gene might do something -- that you could imagine some sensible-sounding biochemical pathway that could link the gene to a disease or trait. The time it took to sequence genes and the limited computing power available meant you had to be quite sure you were going to find something before you dedicated all that expensive lab and analysis time.

Now you just collect a lot of data and let the data decide what the hypothesis should be, McVean says. If you look at 10,000 genomes of people with a disease and 10,000 without, you can use an algorithm to compare them, find the differences and then work out which genes are linked to the disease, without having to think in advance about which ones they might be.

© Dávid Biró for Mosaic

This approach is known as a genome-wide association study, a common form of analysis in the data-driven era. It’s a fairly simple idea. You take the genomes of a large number of people, sequence them and then use an algorithm to compare all of the DNA -- not just the 24,000 or so genes, that make up just 1 to 2 percent of the genome, but also all of the still-somewhat-mysterious non-coding DNA too. The algorithm can be quite simple: for instance, comparing how frequently a certain DNA variant appears in people with a certain trait or condition and people without it. If the variant appears alongside a trait or condition significantly more often than you’d expect by chance, then the algorithm flags it up as a possible cause.

Where it gets difficult is that diseases are almost all complex, and have tens or sometimes hundreds of genes or sections of non-coding DNA involved. This quickly leads to the need for complicated multidimensional analysis, and while the math involved isn't new, the sheer scale of the task means that algorithms are essential. Often they can be comparing tens or hundreds of parameters at a time.

It's a bit like the Google search algorithm. The process it uses to rank each web page isn't that complex -- for instance, measuring how frequently your search terms appear on a page, then where on the page it appears, then how many links there are to that page, and so on. But it combines hundreds of these measures and applies them to billions of web pages simultaneously. It would be impossible for a person to do.

Data-driven dividends:

The algorithmic approach has brought great dividends. Gerstung's field, the genomics of cancer, has had perhaps the most exciting developments, **for instance in relation to leukemia**.

This devastating and often fatal disease can -- in some cases -- be successfully treated with a full bone-marrow transplant. But that is a major procedure whose complications can sometimes be fatal themselves. You only want to give it to people with the most deadly forms of leukemia.

Predicting *which* leukemias will be the most deadly, though, is enormously difficult. The symptoms are complex and don't always tell you enough about the prognosis.

So what Gerstung's team did was sequence the genomes of 1,500 people's cancers to find the DNA mutations driving them, and then see which mutations correlated with which outcomes. There were 5,000 different mutations among the individuals, and around 1,000 different combinations, which the team divided up into 11 categories of greater or lesser risk. "It enables clinicians to make much more focused decisions," Gerstung says.

The influence of the data-driven approach extends much further. Sequencing the genomes of tumors has caused a "mind change" in our approach to cancer in general, says **Edd James**, associate professor of cancer immunology at the University of Southampton. "We're now much more appreciative that a cancer isn't just a mass of copied cells."

A single cancer may contain dozens of different kinds of cell, each with different combinations of DNA mutations and each vulnerable to different drugs. So sequencing allows clinicians to better target drugs at the individuals -- and tumors -- upon which they will work. "Before, people were treated as members of populations: 'X percent of people given this treatment will do well,'" James says. "But with this information, you can understand whether [individually] they're going to get the benefit."

As well as spotting differences, gene sequencing has revealed unexpected similarities between cancers too. Historically, James says, we've defined cancers by their anatomical site: as lung cancers, liver cancers, head-and-neck cancers and so on. "But using next-generation sequencing, you can see that there are cancers in different sites that share more in common with each other than with cancers in the same site. It's made us realize that some drugs that work for, say, breast cancer might work on others," he says.

Gerstung backs this up. "From a genetic perspective, there's substantial overlap between cancers from different anatomical sites. One even finds BRCA1 [a gene heavily involved in breast cancer] in some prostate cancers," he says.

This is going to become increasingly important. The U.S. Food and Drug Administration has recently licensed a cancer drug -- pembrolizumab -- for use in any cancer that shows signs of mismatch-repair deficiency, a form of DNA repair error. This is the beginning of drugs being licensed on the basis of a cancer's genetics rather than location.

And it's all because of the constant, gushing flow of data.

"We got so good at producing data," says **Nicole Wheeler**, a data scientist at the Sanger Institute who looks at the genomes of pathogenic bacteria, "that we ended up with too much of it."

McVean agrees. "In Moore's Law, the computing power you have doubles every 18 months," he says. "The growth of biomedical data capture -- through sequencing genomes, but also through medical imaging or digital pathology -- is much faster than that. We're super-Moore's-Law-ing in biomedical data."

The next generation:

It became completely impossible, in the early years of this century, for biological scientists to check their data themselves. And this meant that biologists had to recruit, or become, data scientists.

"We reached a bottleneck a few years ago," Corcoran says. "We had lots of data, but we didn't know what to do with it. So algorithms had to be invented on the fly, to deal with the data and maximize it," she says. "When you're looking at single genes, or a few, you can do it manually, but when you're looking at the expression of 20,000 genes, you can't even do the statistics by yourself."

© Dávid Biró for Mosaic

Biologists -- many of whom grew up, as Corcoran did, working on benches with glassware, not desks and laptops -- have had to learn to use these algorithms. “I think senior scientists are often intimidated by it,” she says, “and more reliant on their junior colleagues than they probably should be or would like to admit that they are.”

She’s evolved a “working knowledge” of how these algorithms function but admits that “it’s a slightly vulnerable period, where the people at the top don’t have the skills to check the work of the people beneath them.”

Wolf Reik, one of Corcoran's colleagues at the Babraham Institute who runs a research team looking at **epigenetics**, agrees. Older scientists have a completely different mindset, he says. "It's quite funny -- my staff in lab meetings think in terms of what the genome as a whole does. But I think about single genes and generalize from them -- that's how I learned to think."

It's important for people in his position, he says, to understand junior scientists' work and "most importantly develop an intuition about how to use the tools … because ultimately I put my name to the work," Reik says.

The younger scientists, on the other hand, have grown up with data. Some of them have come from that background -- Gerstung did a physics undergraduate degree -- although that's true of some group leaders as well, such as McVean. But others who came through a more biological route have ended up talking in terms of coding. "I did biology as an undergrad; that's my domain knowledge," says **Na Cai**, a postdoctoral researcher at the Sanger Institute who studies how genotypes relate to various human traits.

"Now I'm doing statistical analysis every day. It's been like learning another language, or several," she says. "I had to switch my brain from thinking in terms of biochemical pathways and flowcharts to a more structured kind of thinking in terms of code."

The senior scientists she works with have all been "quite good at keeping up with the latest developments," she says. "They might not be able to write the code, but they understand what the analysis does."

Wheeler, a colleague of Cai's, also came through the biology route and ended up coding. "I don't have a traditional software-engineering background," she says. "I learned to code on the side, during my Ph.D. [My coding] isn't the most efficient or glamorous, but it's about seeing what you have to do computationally and making it happen."

Making the shift:

In response to these needs, undergraduate degrees in the United Kingdom have been changing in the past few years. Newcastle University, for instance, now has a bioinformatics module in its biology undergraduate course, and the University of Reading's final-year research projects involve computational biology, although the earlier optional computing modules have a low take-up, so students in their final year are learning the skills last-minute. Imperial College London, which already has bioinformatics courses, is planning to add programming for first- and second-year students. "I think there's a recognition that biology involves more data than we used to have," says Wheeler, "so people need to have the skills to process it."

But the change is slow, and sometimes opposed by students, not all of whom got into biology to code. "I'd say some undergrad courses are catching up," Corcoran says. "But in general they

have not, as exemplified by the proliferation of post-degree Master's courses teaching these skills."

The change is necessary, though. Even the most wet-lab-oriented scientists interviewed said they spend less than 50 percent of their time doing experiments; some said it was as little as 10 percent or even, in Cai's case, none at all since she has become a full-time bioinformatician.

The shift toward being data-driven, says Wheeler, can be seen as a move from science that's hypothesis-*testing* to one that's hypothesis-*generating*. One scientist, who preferred not to put their name to the concern, worried that it had reduced the creativity in science, but according to Wheeler that's not the case. "It's moved the creativity around," she says. "In some ways there's more room for creativity. You can really try out some crazy ideas at relatively low cost."

This has other advantages. "You can become attached to hypotheses," says **Matt Bawn**, a bioinformatician at the Earlham Institute, a computational biology research center in Norfolk, England. "It's better to be a disinterested observer with no preconceptions, to look at the blank canvas and let the picture emerge."

But the greatest benefit is that data-driven studies are throwing up fascinating new findings all the time, in complex areas that were previously impossible to study.

Stefan Schoenfelder, another researcher at the Babraham Institute, studies the 3-D shapes of chromosomes and how they affect gene expression. When the Human Genome Project was completed, it was discovered that there were far fewer genes than previously expected -- about 24,000, roughly a quarter of what scientists thought was the minimum. The rest of the DNA doesn't code for proteins at all.

What has since been realized is that part of what those noncoding areas do is regulate the expression of the genes: They turn them on in some cells, off in others. And part of how they do that is by folding themselves into different shapes in different cells.

Chromosomes are usually depicted as X-shaped. But that's only when a cell is dividing. The rest of the time, the two meters of DNA inside almost every cell is coiled up in a complex tangle. So a length of DNA can be located a vast distance away from a gene on the chromosome but still be able to regulate it because in practice the two have close physical contact, Schoenfelder says. "That's why it's important to study this in 3-D context: If you just look at the sequences and assume they will regulate the gene next door, that's often incorrect."

On top of this, genomes fold differently, Schoenfelder says. "The same genome in a T cell will have a different conformation in a liver cell or in a brain cell, and that's linked to different genes being expressed and the cells acquiring different functions."

Working out the 3-D shape in each context is incredibly difficult. It involves sequencing cell types and seeing how they differ from other cell types, as well as which bits of DNA are interacting in that context. But the DNA first has to be treated using a complex technique known as cross-linking and ligation to allow the sequencing to see which bits are near each other. If two distant points are found together, it might be that they have been folded that way in order for one to affect the other. But -- much more often -- it's just the product of random jiggling.

Extraordinary times:

Finding the real correlations among the noise requires looking at billions of data points and seeing which links keep coming up slightly more often than others. It's then that the algorithms really come into play. Once you know which bits of the chromosome are regularly in contact with which other bits, you can use other algorithms to build 3-D models based on those points of contact.

"This whole field is only about 15 years old," Schoenfelder says. Before that, he says, "I didn't think of the genome's shape at all; I just thought of it as a ball of spaghetti crushed into the nucleus. I thought it was just a logistical problem, stuffing it into a nucleus that's maybe 5 microns across.

© Dávid Biró for Mosaic

“What’s blown me away is the fine level of regulation that exists, despite the extreme compaction, that still allows for this fine-tuning,” he says The 3-D shapes of chromosomes, and which regulatory elements interact with which genes on that shape, will be a large part of the story of how the 200 cell types in the human body arise.

Meanwhile, McVean says that genomic research has forced clinicians to reclassify the disease multiple sclerosis entirely. “We’ve found more than 250 bits of the genome which light up in terms of risk for the disease,” he says. “That’s let us make quite strong statements about the risk for the

individual. But it’s also allowed us to see overlaps with diseases like rheumatoid arthritis: Some of the genes that raise your risk of [multiple sclerosis] decrease your risk of arthritis.

“So we’ve learned it’s an autoimmune disease, even though it presents as a neurodegenerative disease,” McVean says. “There are four or five companies with new therapeutic programs coming out of this.”

And at the Babraham Institute, Reik has a thrilling, almost science-fiction story to tell. His work is in the field of epigenetics, looking at how the chemical environment of a cell affects the expression of genes; he sequences RNA, the messenger molecule that allows DNA to be read and proteins made, to see how it differs from cell to cell. His group is especially interested in aging.

Five years ago, it was discovered -- and Reik’s work has since confirmed -- that there is an aging clock in all our cells. It’s called DNA methylation. There are four letters in the DNA alphabet: C (cytosine), A (adenine), G (guanine) and T (thymine). As we get older, more and more of the Cs on our DNA gain a little chemical marker called a methyl group. To read this clock, the work is simple -- just counting the methyl groups up -- but, again, the sheer number of data points returned is so enormous that they absolutely have to be counted by algorithm.

“Reading that clock, we can predict your age, and my age, to within three years,” Reik says. “Which is surprisingly accurate: the most accurate biomarker of aging that we have.”

All of which is very interesting, of course: It’s “either a readout of an underlying aging process or our programmed life expectancy,” Reik says. But he says the implication is that we could interrupt it: “I’m sure there will be drugs and small molecules that can slow this aging clock down."

It may be too much to hope that big data will help us all live forever. But every scientist I spoke to agreed that the rise of algorithm-led, data-intensive genomic research has transformed the life sciences. It has left senior scientists sometimes unsure what their junior colleagues are doing, and left modern research centers with too much laboratory and not enough space for a laptop. The pace of change can be “disorienting,” Schoenfelder says.

“Life is a lot more complex now,” he says. “The skill set I had when I did my Ph.D., only 13 years ago, is absolutely not sufficient to keep up with today’s science.” But this change has brought an optimism back into genomic research. When the Human Genome Project neared completion, people were excited, believing that many diseases would fall quickly as their genetic components were revealed. But most of them turned out to be complex, polygenic, impossible to understand by looking at single genes. Now, though, it is possible to look at those diseases through the power of next-generation sequencing and tools that can sift the data it provides.

“Now when I run an experiment, I get 100 million, 200 million data points back,” says Schoenfelder. “I didn’t think that was possible in my lifetime, but it’s happened over the course of

a few years. We can address questions that were completely off-limits 10 years ago. It’s been an extraordinary revolution.”

Wellcome, the publisher of Mosaic*, founded the Wellcome Sanger Institute in 1993 and has funded it ever since. The Sanger Institute celebrates its 25th anniversary in October 2018.*

Gil McVean currently receives funding from Wellcome through an Investigator Award. Wolf Reik participates in a Sanger Institute resource collaboration that is funded by Wellcome.

This story originally appeared on **Mosaic***. It has been slightly modified to reflect* Spectrum*'s style.*