https://www.theatlantic.com/science/archive/2015/09/genome-big-data-disease-genes/404356/

How Data-Wranglers Are Building the Great Library of Genetic Variation

A huge project unexpectedly led to a way of finding disease genes without needing to know about diseases.

Let's say you have a patient with a severe inherited muscle disorder, the kind that Daniel MacArthur from the Broad Institute of Harvard and MIT specializes in. They're probably a child, with debilitating symptoms and perhaps no diagnosis. To discover the gene(s) that underlie the kid's condition, you sequence their genome, or perhaps just their exome: the 1 percent of their DNA that codes for proteins. The results come back, and you see tens of thousands of variants—sites where, say, the usual A has been replaced by a T, or the typical C is instead a G.

You'd then want to know if those variants have ever been associated with diseases, and how common they are in the general population. (The latter is especially important because most variants are so common that they can't possibly be plausible culprits behind rare genetic diseases.) "To make sense of a single patient's genome, you need to put it in the context of *many* people's genomes," says MacArthur. In an ideal world, you would compare all of a patient's variants against "every individual who has ever been sequenced in the history of sequencing."

This is not that world, at least not yet. When Macarthur launched his lab in 2012, he started by sequencing the exomes of some 300 patients with rare muscle diseases. But he quickly realized that he had nothing decent to compare them against. It has never been easier, cheaper, or quicker to

sequence a person's genome, but interpreting those sequences is tricky, absent a comprehensive reference library of human genetic variation. No such library existed, or at least nothing big or diverse enough. So, MacArthur started making one.

It was hard work, not because the data didn't exist, but because it was scattered. To date, scientists have probably sequenced at least 5,000 full genomes and some 500,000 exomes, but most are completely inaccessible to other researchers. There might be intellectual-property restrictions, or issues around consent. There's the logistical hassle of shipping huge volumes of data on hard drives. And some scientists are just plain competitive.

Fortunately, MacArthur's colleagues at the Broad Institute and beyond had deciphered so many exomes that he could gather thousands of sequences by personally popping into offices. Buoyed by that success, he started contacting people who were studying the genomes of people with cancer, heart disease, diabetes, schizophrenia, and more. "There's a big swath of human genetics where people have learned that you either fail by yourself or succeed together, so they're committed to sharing data," MacArthur says.

The most interesting variants turned out to be the ones that *weren't* there.

By 2014, he had amassed more than 90,000 exomes from around a dozen sources, collectively called the Exome Aggregation Consortium. Then, he had to munge them together.

That was the worst bit. Researchers use very different technologies to sequence and annotate genomes, so combining disparate data sets is like mushing together the dishes from separate restaurants and hoping that the results will be palatable. Often, they won't be.

Monkol Lek, a postdoc in MacArthur's lab who himself has a genetic muscle disease, solved this problem by essentially starting from scratch. He took the raw data from some 60,706 patients and analyzed their exomes, one position at a time. The raw sequences took up a petabyte of memory, and the final compressed file filled a three-terabyte hard disk.

The prize from all this data-wrangling was one of the most thorough portraits of human genetic variation ever produced. MacArthur went through the main results in the opening talk of this week's Genome Science 2015 conference, in Birmingham, U.K. His team had identified around 10 million genetic variants scattered throughout the exome, most of which had never been described before. And most turned up just once in the data, meaning that they lurk within just one in every 60,000 people. "Human variation is dominated by these extremely rare variants," says MacArthur. That's where the secrets of many rare genetic disorders reside.

But unexpectedly, the most interesting variants turned out to be the ones that *weren't* there.

The graduate student Kaitlin Samocha developed a mathematical model to predict how many variants you'd expect to find in a given gene, in a population of 60,000 people. The model was remarkably accurate at estimating neutral variants, which don't change the protein that's encoded by the gene, and so have minimal impact. But the model often wildly overestimated the number of "loss-of-function variants," which severely disrupt the gene in question. Repeatedly, the ExAc data revealed far fewer of these variants than Samocha's model predicted.

Why? Because many of these loss-of-function variants are so destructive that their carriers develop debilitating disorders, or die before they're even born. So, the difference between prediction and

reality reflects the brutal hand of natural selection. The variants are simply not around to be sequenced because they have long been expunged from the gene pool.

For example, the team expected to find 161 loss-of-function variants in a gene called DYNC1H1. By contrast, the ExAc data revealed only four—and indeed, DYNC1H1 is associated with several severe inherited neurodevelopmental disorders.

The model also predicted 125 loss-of-function variants in the UBR5 gene—and the data revealed just one. That's far more interesting because UBR5 has *never before been linked to a human disease*.

A full quarter of human genes are like this: They have a lower-than-expected number of loss-of-function variants. And while some of them are known "diseases genes," the rest have never been pinpointed as such. So, if you find one of these variants in a patient with a severe genetic disorder, the chances are good that you've found a genuine culprit.

That blew my mind. Here is a way of identifying potential disease-related genes, without needing to know anything about the diseases in question. Or, as MacArthur said in his talk, "We should soon be able to say, with high precision: If you have a mutation at this site, it will kill you. And we'll be able to say that without ever seeing a person with that mutation."

These results speak to one of the greatest challenges of modern genomics: weaving together existing sets of data in useful ways. They also vindicate the big, expensive studies that have searched for variants behind common diseases like type 2 diabetes, heart disease, and schizophrenia. These endeavors have indeed found several variants, but with such small effects that they explain just a tiny fraction of the risk of each condition. But "all this data can be re-purposed for analyzing rare diseases," says MacArthur. "Without those large-scale studies, we'd have no chance of doing something like ExAc."

"His talk really shows that you can't anticipate what these data sets will show you until you put them together," says Nick Loman from the University of Birmingham. "Our ability to interrogate biology if you can put hundreds of thousands, or millions, of genomes together is massive."