Scientists are on a path to sequencing 1 million human genomes and use big data to unlock genetic secrets

April 16 2021, by Xavier Bofill de Ros

A complete human genome, seen here in pairs of chromosomes, offers a wealth of information, but it is hard connect genetics to traits or disease. Credit: HYanWong/Wikimedia Comons

The first draft of the human genome was published 20 years ago in 2001 , took nearly three years and cost between US$500 million and $1 billion . The Human Genome Project has allowed scientists to read, almost end to end, the 3 billion pairs of DNA bases—or "letters"—that biologically

define a human being.

That project has allowed a new generation of researchers like me, currently a postdoctoral fellow at the National Cancer Institute, to identify novel targets for cancer treatments, engineer mice with human immune systems and even build a webpage where anyone can navigate the entire human genome with the same ease with which you use Google Maps.

The first complete genome was generated from a handful of anonymous donors to try to produce a reference genome that represented more than just one single individual. But this fell far short of encompassing the wide diversity of human populations in the world. No two people are the same and no two genomes are the same, either. If researchers wanted to understand humanity in all its diversity, it would take sequencing thousands or millions of complete genomes. Now, a project like that is underway.

Understanding genetic diversity

The wealth of genetic variation among people is what makes each person unique. But genetic changes also cause many disorders and make some groups of people more susceptible to certain diseases than others.

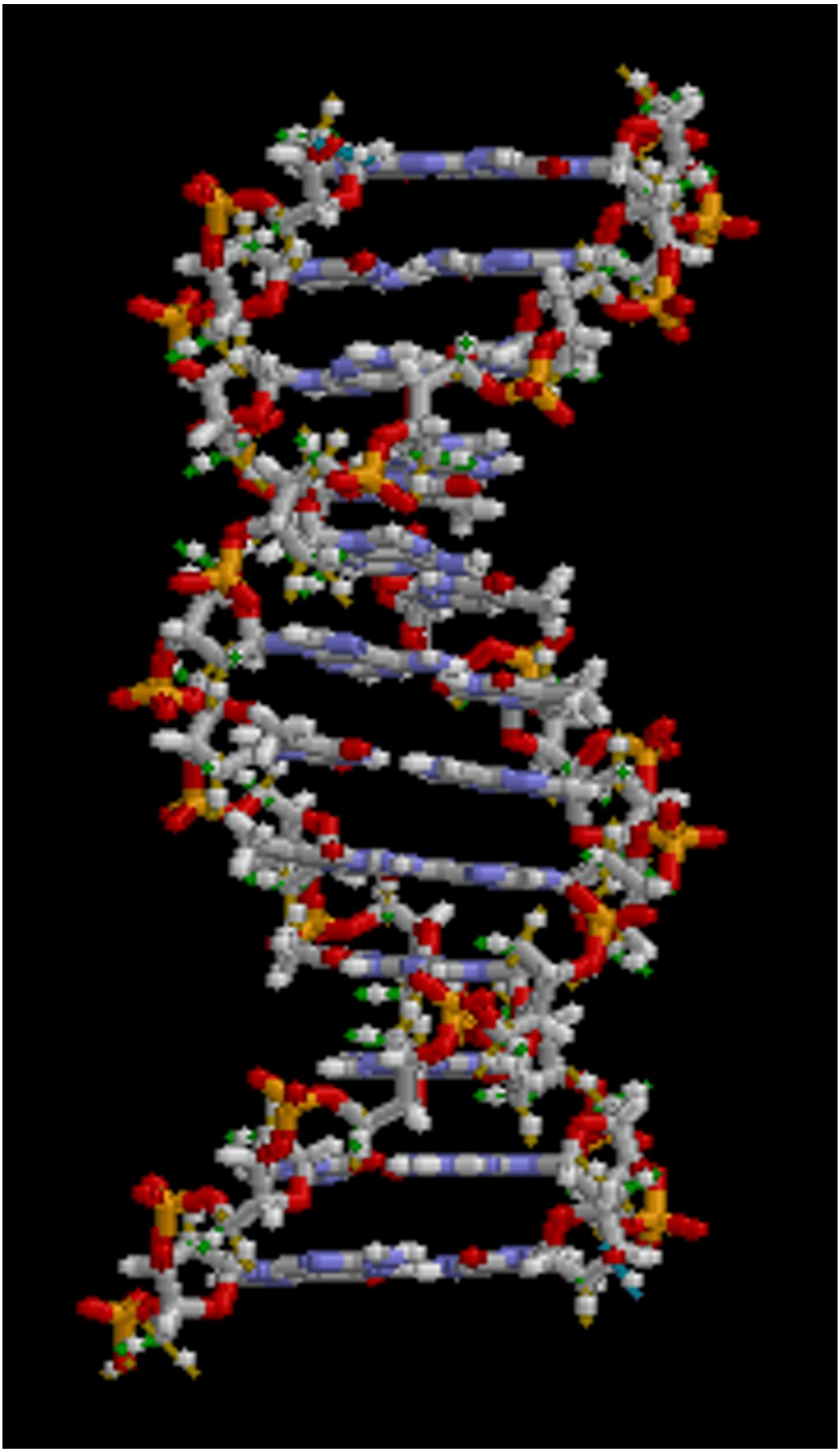
Around the time of the Human Genome Project, researchers were also sequencing the complete genomes of organisms such as mice, fruit flies, yeasts and some plants. The huge effort made to generate these first genomes led to a revolution in the technology required to read genomes. Thanks to these advances, instead of taking years and costing hundreds of millions of dollars to sequence a whole human genome, it now takes a few days and costs merely a thousand dollars. Genome sequencing is very different from genotyping services like 23 and Me or Ancestry, which look at only a tiny fraction of locations in a person's genome.

Advances in technology have allowed scientists to sequence the complete genomes of thousands of individuals from around the world. Initiatives such as the Genome Aggregation Consortia are currently making efforts to collect and organize this scattered data. So far, that group has been able to gather nearly 150,000 genomes that show an incredible amount of human genetic diversity. Within that set, researchers have found more than 241 million differences in people's genomes, with an average of one variant for every eight base pairs.

Most of these variations are very rare and will have no effect on a person. However, hidden among them are variants with important physiological and medical consequences. For example, certain variants in the BRCA1 gene predispose some groups of woman, like Ashkenazi Jews, to ovarian and breast cancer. Other variants in that gene lead some Nigerian women to experience higher-than-normal mortality from breast cancer.

The best way researchers can identify these types of population-level variants is through genomewide association studies that compare the genomes of large groups of people with a control group. But diseases are complicated. An individual's lifestyle, symptoms and time of onset can vary greatly, and the effect of genetics on many diseases is hard to distinguish. The predictive power of current genomic research is too low to tease out many of these effects because there isn't enough genomic data.

Understanding the genetics of complex diseases, especially those related to the genetic differences among ethnic groups, is essentially a big data problem. And researchers need more data.



The link between genetics and disease is nuanced, but the more genomes you can study, the easier it is to find those links. Credit: brian0918/Wikimedia Commons

1,000,000 genomes

To address the need for more data, the National Institutes of Health has started a program called All of Us. The project aims to collect genetic information, medical records and health habits from surveys and wearables of more than a million people in the U.S. over the course of 10 years. It also has a goal of gathering more data from underrepresented minority groups to facilitate the study of health disparities. The All of Us project opened to public enrollment in 2018, and more than 270,000 people have contributed samples since. The project is continuing to recruit participants from all 50 states. Participating in this effort are many academic laboratories and private companies.

This effort could benefit scientists from a wide range of fields. For instance, a neuroscientist could look for genetic variations associated with depression while taking into account exercise levels. An oncologist could search for variants that correlate with reduced risk of skin cancer while exploring the influence of ethnic background.

A million genomes and the accompanying health and lifestyle information will provide an extraordinary wealth of data that should allow researchers to discover the effects of genetic variation on diseases, not only for individuals, but also within different groups of people.

The dark matter of the human genome

Another benefit of this project is that it will allow scientists to learn about parts of the human genome that are currently very hard to study. Most genetic research has been on the parts of the genome that encode for proteins. However, these represent only 1.5% of the human genome.

My research focuses on RNA—a molecule that turns the messages encoded in a person's DNA into proteins. However, RNAs that come from the 98.5% of the human genome that doesn't make proteins have a myriad of functions by themselves. Some of these noncoding RNAs are involved in processes such as how cancer spreads, embryonic development or controlling the X chromosome in females. In particular, I study how genetic variations can influence the intricate folding that allows noncoding RNAs to do their jobs. Since the All of Us project includes all coding and noncoding parts of the genome, it is going to be by far the largest dataset relevant to my work and will hopefully shed light on these mysterious RNAs.

The first human genome sparked 20 years of incredible scientific progress. I think it is almost certain that a huge dataset of genomic variations will unlock clues about complex diseases. Thanks to large-scale population studies and big-data projects such as All of Us, researchers are paving the way to answering, in the next decade, how our individual genetics shape our health.

Provided by The Conversation