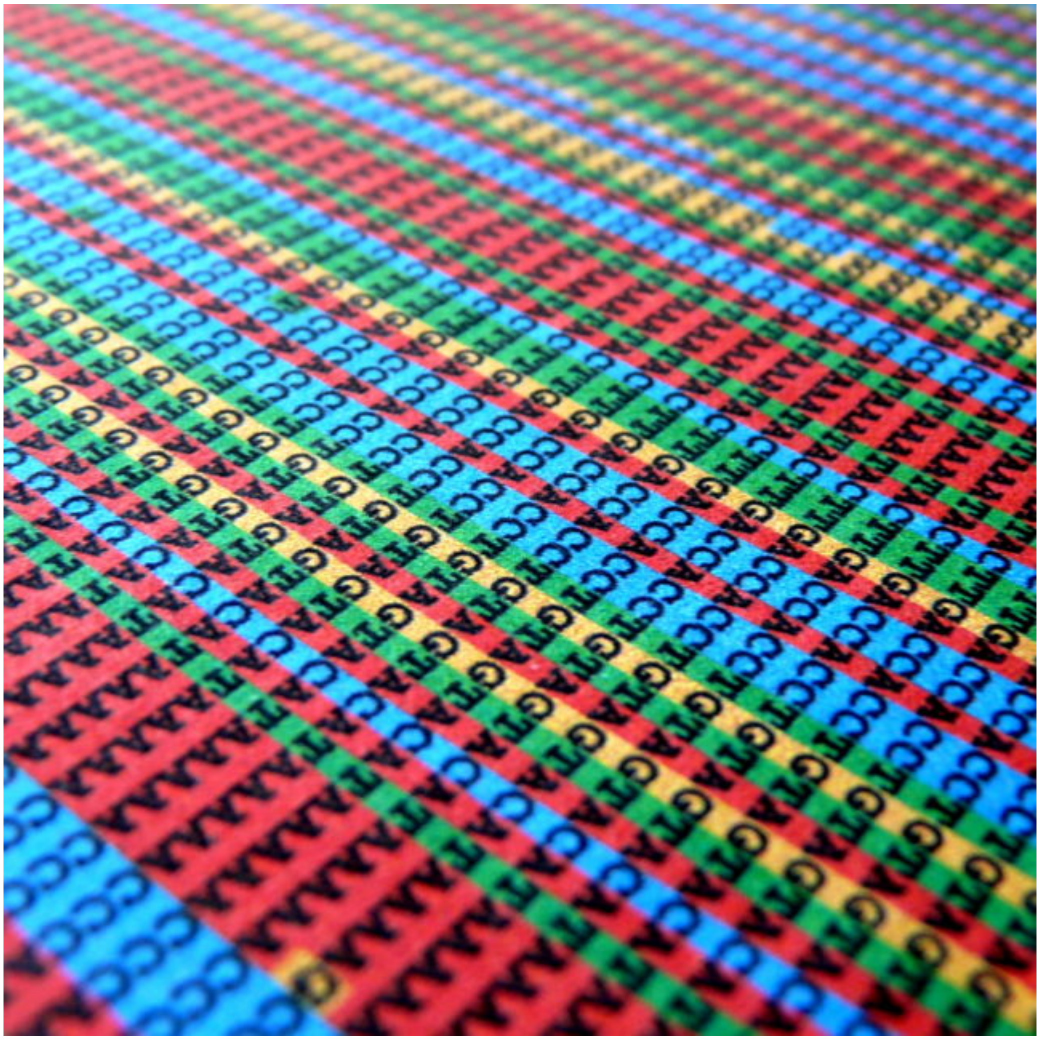
[Skip to Main Content](#)
STAT

Human ‘pangenome’ published, with goal of making genomics more useful for diverse populations

- 

By Megan Molteni[1] [2]May 10, 2023



Shaury Nash/Flickr

An international team of scientists has assembled the first human “pangenome” — an attempt to make a more representative reference genome, one that captures almost all the genetic variability residing in the DNA of humans around the globe.

The technological achievement, published Wednesday in Nature[4], is the result of years of work by more than 100 researchers behind The Human Pangenome Project[5], a $30 million effort launched in 2019 and funded by the U.S. National Human Genome Research Institute.

A reference genome is exactly what it sounds like. It serves as a guide for researchers building genetic tests and looking for new drug targets. The first one, released in 2001, was built from just 11 individuals, with 70% of the DNA coming from just one man[6] who answered an ad in a Buffalo, N.Y., newspaper. And though it has been improved upon over the years, the lack of diversity and resulting European bias has remained — hampering the diagnosis of rare disorders, resulting in genetic tests that work only for people of European ancestry, and undermining promises that personalized medicine would improve the health of everyone, not just a select few.

“Currently, when we map a sequence from a patient there’s always a fraction of the sequence, sometimes a significant section, that can’t be mapped,” Evan Eichler, a geneticist at the University of Washington and one of the leaders of the project, told reporters Tuesday. But the new reference, which adds 119 million bases to the existing reference and is composed of DNA from 47 people mostly from Africa, Asia, and the Americas, will allow for building a more complete understanding of how diseases are associated with all the unique remixes of DNA that have evolved throughout our species’ history. “That essentially means no sequence variations being left out when we map to a pangenome, at least in theory,” Eichler said.

In 2001, as scientists gathered in the East Room of the White House to declare victory in the decade-long quest to map the human genome, then-President Bill Clinton called the achievement the “most wondrous map ever produced by humankind.” Behind the fanfare, caveated in fine print, was the open secret that the “finished” sequence wasn’t actually finished.

In fact, it was littered with gaps untouchable by sequencing technologies available at the time. By some estimates, about 8% of the genome, mostly concentrated in areas that lie near where two parts of a chromosome connect, was made up of sequence-defying stutters of short DNA sequences known as repetitive elements.

This omission “wasn’t a cause of a great deal of agonizing at the time,” Eric Lander of the Broad Institute of MIT and Harvard recalled years later[9], because he and other Human Genome Project leaders expected the next generation of scientists to find a solution.

One of those scientists is Eichler, who started his career in the trenches of the Human

Genome Project and has spent the intervening decades worrying about the parts they left out. These highly repetitive regions, it turned out, are home to the vast majority of so-called structural variants, where large chunks of DNA shift around, making copies of themselves, flipping orientation, and recombining with other chromosomes.

These big changes can have big effects, like causing a cell to become cancerous. They also account[10] for the majority of the genome differences between one person and another. But as Eichler and his colleagues reported[11] in 2016, about 90% of structural variants had never been sequenced.

In a study[12] published in 2017, Eichler and colleagues estimated that there are up to 16 million single DNA letter differences between any random individual and the reference genome. Other researchers have shown that those numbers get worse the more an individual's ancestry differs from those in the reference genome. Scientists analyzing genomes from two dozen ethnic populations including Han Chinese, Yoruba, and Puerto Ricans found 60 million base pairs not present in the reference genome. In a 2018 Nature paper[13], researchers sequenced 910 individuals of African descent and discovered a single unfamiliar sequence spanning 300 million letters.

At a meeting convened that year by the National Human Genome Research Institute, experts concluded that the reference required a reboot to address these issues. They bet on an initiative proposed by Eichler and colleagues at the University of California, Santa Cruz, and Washington University, and The Human Pangenome Project was born.

What they published Wednesday is the first phase — consisting of DNA from 47 people who, years ago, gave blood to an older diversification effort called the 1000 Genomes Project, which launched in 2008. It was an easy way to jump-start the pangenome; cells from those samples were grown up and frozen for future use and the consent forms those individuals signed at the time would cover the use of their DNA data for new projects like this one. All the consortium had to do was get the cells out of cold storage and into long-read sequencing machines.

These technologies, developed by Oxford Nanopore and Pacific Biosciences, produce long, accurate stretches of DNA. They're more expensive than the workhorse sequencers produced by Illumina, which are currently the standard way to sequence a genome for medical purposes. But they have proven invaluable in recent efforts to map those finicky, repeat-riddled sections of the human genome. The Telomere 2 Telomere Consortium, of which Eichler is also a part, released the first fully complete[14] (actually, no caveats, no fine print) human genome sequence in 2021.

Leveraging this work, the new pangenome reference isn't just comprised of more individuals, but also more complete genomes. And more are expected to be added to it soon. The goal is to have 350 by the middle of next year. Why 350?

"Initially in the project there was a stated goal to get to representing human genetic diversity above 1% of global frequency," Eimear Kenny, founding director of the Institute for Genomic Health at Mount Sinai's Icahn School of Medicine, said at the press briefing. "Even though 47 is not a huge number it is actually getting us quite a long way along the path to that goal. With 350 we'll get even closer."

Because all those divergent sequences make storing and searching the reference genome increasingly space- and time-intensive, computational biologists with the Human Pangenome Project have also been working on ways of representing it as multiple branching paths more efficiently. "It's very much top of mind in terms of making this computable in a way that doesn't burden downstream pipelines," said Benedict Paten of the UC Santa Cruz Genomics Institute. "We're very aware that there's a cost benefit here and if the cost is higher than the benefit, people won't adopt it."

On Wednesday, he and his colleagues published a new software tool[16] in Nature Biotechnology for making pangenomes compatible with current methods for assembling and interrogating genomes.

Even with those efforts, scientists like Eichler expect it could be a decade or more before the benefits of a pangenome reference reaches patients and families anxious for an answer to their medical mysteries. But it's that promise that continues to push the project forward.

"I've been studying autism for the last two and a half decades, and still, for 70% of kids who come into the clinic, we can't explain why those children have autism," Eichler said. He believes that complete sequencing combined with a pangenome could finally provide answers for those families.

"If I can actually solve another 10, 20% of these cases and explain to parents why their children have autism, and hopefully from that will come better treatments, that for me is the holy grail of all this," he said. "It's not just the cataloging of structural variation of the genetic differences, it's not just discovering these dynamic regions, it's also making a difference in individuals' lives — families who suffer from a lot of the unknowns with respect to their kids' disabilities."

About the Author

Megan Molteni[1]

Science Writer

Megan Molteni is a science writer for STAT, covering genomic medicine, neuroscience, and reproductive tech.

megan.molteni@statnews.com [17]
@MeganMolteni [2]

Add a Comment

Create a display name to comment

This name will appear with your comment

There was an error saving your display name. Please check and try again.

Links

1. https://www.statnews.com/staff/megan-molteni/
2. https://twitter.com/MeganMolteni
3. https://www.parsintl.com/publication/stat/
4. https://www.nature.com/articles/s41586-023-05896-x
5. https://www.nature.com/articles/s41586-022-04601-8
6. https://www.statnews.com/2019/03/11/human-reference-genome-shortcomings/
7. https://www.statnews.com/signup/
8. https://www.statnews.com/privacy/
9. https://www.statnews.com/2017/06/20/human-genome-not-fully-sequenced/
10. https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.0050254
11. https://genome.cshlp.org/content/27/5/677.short
12. https://genome.cshlp.org/content/27/5/677
13. https://www.nature.com/articles/s41588-018-0273-y/
14. https://www.statnews.com/2021/06/01/researchers-claim-they-have-sequenced-the-entirety-of-the-human-genome-including-the-missing-parts/
15. https://www.statnews.com/2022/08/11/oxford-mexican-research-team-diversify-genomic-databases/
16. https://www.nature.com/articles/s41587-023-01793-w
17. https://www.statnews.com/2023/05/10/human-pangenome-published/mailto:megan.molteni@statnews.com
18. https://www.statnews.com/topic/diversity-and-inclusion/
19. https://www.statnews.com/topic/genetics/
20. https://www.statnews.com/topic/health-disparities/
21. https://www.statnews.com/topic/research/