THE QUEST FOR AN ALL-INCLUSIVE HUMAN GENOME

Efforts are under way to create a 'pangenome' that would catalogue almost all human genetic diversity. But not everyone is ready to sign on.

By Roxanne Khamsi

Several years ago, after an exhaustive search for uncharted variation in the human genome, Evan Eichler stumbled on something extraordinary. Eichler, a geneticist at the University of Washington in Seattle, and his colleagues struck on a massive stretch of DNA, about 400,000 letters long, that contained extra copies of genes — probably passed on from an ancient hominin group known as the Denisovans[1]. It appeared in about 80% of people living in Papua New Guinea, but practically nowhere else.

"We were shocked by the size," Eichler says. "We always knew there would be archaic segments in our genome." But the segment's length and its absence in much of the world, he says, "transformed our thinking".

This and other unexpected discoveries have made Eichler and other geneticists increasingly dissatisfied with the breadth and depth of the available maps of the human genome. The first draft genome from the US$2.7-billion Human Genome Project, released in 2001, was meant to become a reference point for future genetic research. But 93% of its sequence came from just 11 individuals, many of whom were recruited through a newspaper advertisement in Buffalo, New York; a whopping 70% of the DNA comes from just one man.

By 2003, that reference genome, known as GRCh38, would be deemed technically complete, but it still had hundreds of gaps and sections containing copious errors. These shortcomings came with consequences. Eichler worked with clinical geneticists at his university's medical centre and found that the reference genome lacks a region that has variants associated with Baratela-Scott syndrome, which can cause cognitive delays and skeletal malformations in children. Because that portion was missing, there was no quick way for the physicians to check for DNA errors there.

Genome maps have improved, but still don't adequately capture humanity's vast diversity. For example, in 2018, one group of researchers sequenced 910 individuals of African descent and discovered a sequence consisting of 300 million DNA letters, or bases, that was unfamiliar[2]. That's roughly 10% of the entire genome.

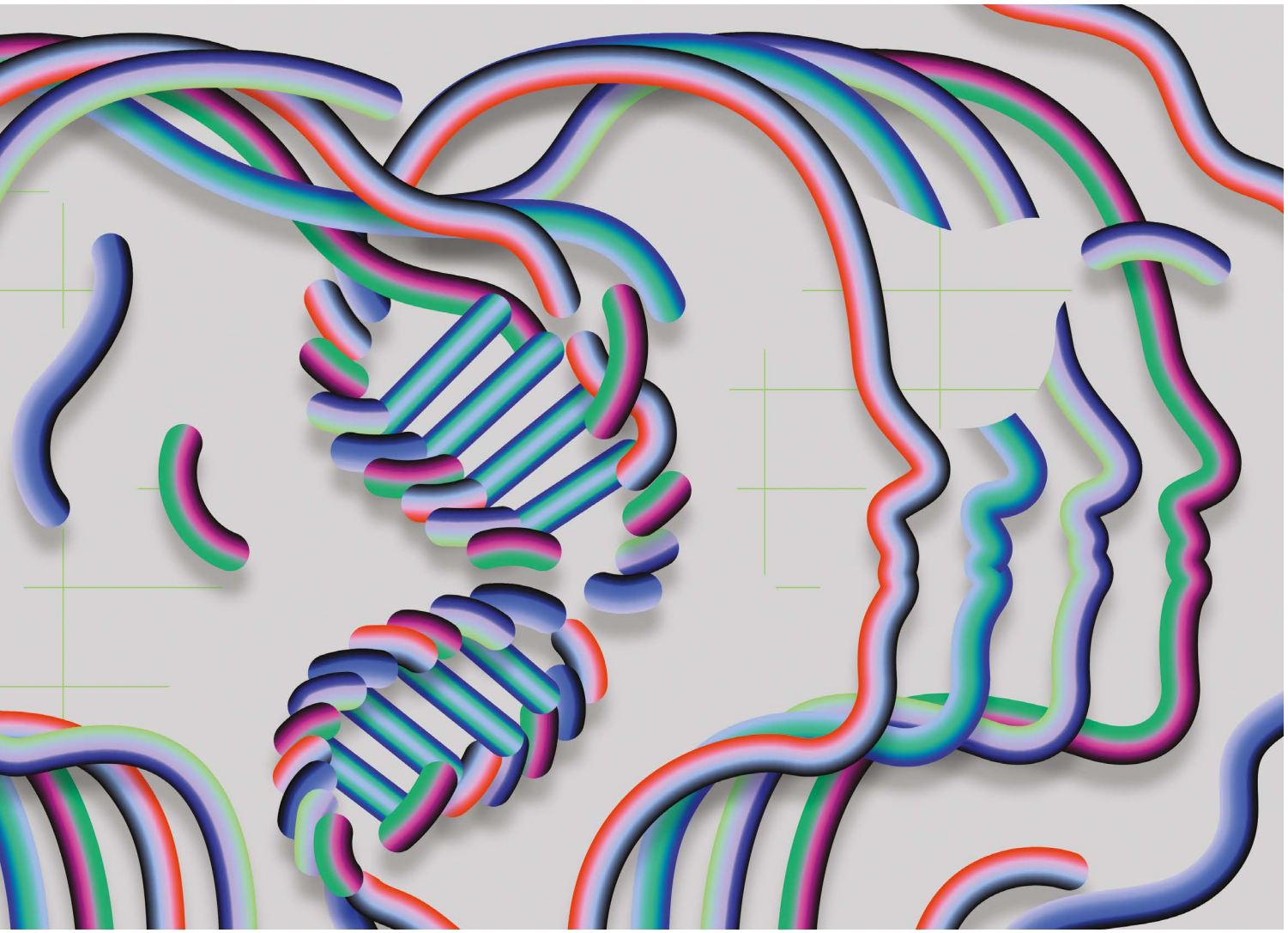
To create a reference that is more complete and more representative, Eichler has joined forces with a number of high-profile scientists, mostly in the United States. Their goal is to capture almost all human genetic variability — the dizzying number of genetic remixes in the human species, including additions, deletions and other types of mutation.

Rather than depicting the genome as a linear readout from a single individual, it would contain multiple paths branching in and out like the tangle of train lines on the map for the London Underground. These would represent the varieties of sequence that can be found in different populations, such as the long stretch of DNA found in many people from Papua New Guinea.

In 2019, Eichler and his colleagues started the Human Pangenome Project, a $30-million effort funded by the US National Human Genome Research Institute (NHGRI) in Bethesda, Maryland. The initial goal is to do detailed, reference-quality genome sequencing of about 350 people from different backgrounds and to share those data as freely as possible.

The effort will pose a significant technical challenge, but the scientists behind it, including Karen Miga at the University of California, Santa Cruz, and Ting Wang at the Washington University School of Medicine in St. Louis,

ILLUSTRATION BY ANA KOVA



> **MAKING HUMAN VARIATION INTUITIVE AND EASY TO UNDERSTAND IS PART OF OUR MISSION."**

argue it is worth it. They see it as crucial to making genomic medicine more equitable[3]. "To account for diversity is to better serve humanity," Wang says. "It is about both equity and equality. It is about building a more inclusive genomic resource for humankind."

The researchers in the pangenome effort are aware of the history of past missions to capture human genetic diversity, some of which were seen as 'vampire' projects that took data from marginalized populations and failed to respect their needs and wishes. In response to this, the pangenome effort engages bioethicists throughout the project, not just at periodic junctures, as was done by initiatives in the past. "They are not a separate entity working in silo, they are involved in every step of the project, including all the technical decisions," Wang says.

Nevertheless, some geneticists focused on the needs of Indigenous communities are wary of the initiative. They aren't calling for an end to the Human Pangenome Project per se, but they say that marginalized groups deserve control of their genetic data, and of the sequencers, too. "As we position ourselves to be in control of these technologies, we're empowering our communities," explains Keolu Fox, a geneticist at the University of California, San Diego, who is Native Hawaiian. "Nothing is as real deal as we are. We're from our communities."

Panning out

The concept of a pangenome traces back to the study of a bacterium known as *Streptococcus agalactiae*, or group B streptococcus, which can cause deadly infections in newborns. Scientists analysing six strains of the bacterium published a paper in 2005 trying to capture all of the microbe's genetic nuances[4]. What they produced was a core genome shared by all six strains and a "dispensable" genome of partially shared and strain-specific genes.

It was a tricky task, because bacteria swap and share bits of DNA, even with other species, mostly through a process known as horizontal gene transfer. "There's a lot of things that can happen in bacteria," says Candice Hirsch, a plant geneticist at the University of Minnesota in Saint Paul. As a result, biologists are continually updating the bacterial reference genomes. Humans, by contrast, do not add new variation as easily. That makes characterizing a human pangenome more feasible, Hirsch says.

But what it lacks in dynamics, the human genome makes up for in length and repetition. Chromosome 1, for example, the largest of the 24 different human chromosomes, stretches over about 250 million base pairs. That's more than 100 times the length of *S. agalactiae*. And it is riddled with long stretches of simple, repeated sequences and duplications of other, more complex segments. Until the past decade, scientists' main option for sequencing DNA involved breaking it into fragments and reading it in small chunks. This allows

them to detect single-letter changes in DNA relatively easily. But the short reads make it hard to recognize when a long stretch of DNA contains more than one copy of a gene. Eichler, who has specialized in identifying structural variants such as gene duplications and deletions, has opted for a newer approach, called 'long-read sequencing', which analyses bigger stretches of DNA at a time. This is what enabled him to find the previously unnoticed variant in people from Papua New Guinea.

In 2018, Eichler and other scientists gathered at the NHGRI to discuss a human pangenome effort. There, Eichler reconnected with a fellow scientist who shared his passion for long-read technology, Erich Jarvis, a neuroscientist and molecular biologist at Rockefeller University in New York City.

"We kept raising our hands and saying, 'You're not going to be able to do that unless you have high-quality reference genomes,'" Jarvis recalls. But long-read sequencing would require more money, and not everyone was keen to deploy it. Jarvis recalls feeling frustrated by some of the debates. "I even chipped a little bit of my front tooth on a fork at a restaurant. I was biting on it so hard," he says. Ultimately, he and others pushing for the long-read approach won.

Miga, who brings to the project a reputation for completing difficult-to-read sections of DNA, was already using long-read technology. She, along with Jarvis, Eichler and others, published the first-ever completely sequenced human genome, capturing all 3 billion letters, including the messy, highly repetitive sections that cap the ends of chromosomes — known as telomeres[5]. This first telomere-to-telomere genome sequence corrected numerous errors from previous references and uncovered around 100 unnoticed genes that probably code for proteins.

It was no simple feat, however. Typically, human cells contain two sets of 23 chromosomes — one from an egg and one from a sperm cell. But duplicated sequences and other structural DNA variations get jumbled up when machines try to read both sets at the same time. To circumvent this, the scientists analysed the DNA of a cell line derived from what's known as a molar pregnancy, in which a sperm fertilizes an egg with no nucleus. The DNA contained only one set of chromosomes.

The 350 genomes for the Human Pangenome Project, by contrast, will come from diploid cell lines, that is, cells that contain copies from both parents, so scientists will have to use complex computational tools to tease the genomes apart and make sure they capture the structural variation accurately.

The pangenome effort has already completed around 70 detailed genomes. It aims to finish telomere-to-telomere versions of all 350 by the end of the grant, in mid-2024.

And scientists are already working on ways to visualize the diversity and showcase the variations. Up until now, including for the GRCh38 reference genome, the convention has been to have a simple linear representation and a companion database with variations listed for different positions in the sequence, such as single-letter changes. "The community has used this convenient fiction of linear reference sequence for 20 years," says Benedict Paten, a computational biologist at the University of California, Santa Cruz. Paten, whose office is next to Miga's, is collaborating with a group to improve the sophistication of the pangenome visualization. In this new visualization, coloured lines represent distinct variants. More-frequent variations are indicated with thicker lines. "Making human variation intuitive and easy to understand is part of our mission in integrating the pangenome," Paten says (see 'Visualizing a pangenome').

Missteps and departures

Many of the 350 people whose genomes will be analysed in the Human Pangenome Project participated in the 1000 Genomes Project, an effort launched in 2008 to catalogue common and rare variants from 26 diverse populations. The DNA samples that were collected as part of that effort will be retrieved from cold storage and repurposed for the more detailed long-reads of the pangenome sequencing project. The consent forms that those individuals signed years ago also cover the use of their DNA data for the new project. But the Human Pangenome Project is taking further measures to ensure ethical collection and use of genetic data. In contrast to other major genetic sequencing efforts, in which scientists made decisions and then only had them vetted by an Institutional Review Board, for example, the Human Pangenome Project has social ethicists who are "embedded" in the decision-making process and continuously vetting the project, Eichler says.

As Wang puts it: "It's really about how to guide the nerdy scientists who may not think about social issues to do their science in the most appropriate manner."

In many ways, the leaders of the pangenome project are trying to overcome the ethically thorny legacy of past endeavours. The Human Genome Diversity Project, for example, launched in 1991 as an effort to collect DNA information from people around the globe, engendered staunch opposition from several communities. Indigenous groups, among others, felt they were being treated as living fossils, headed towards extinction[6].

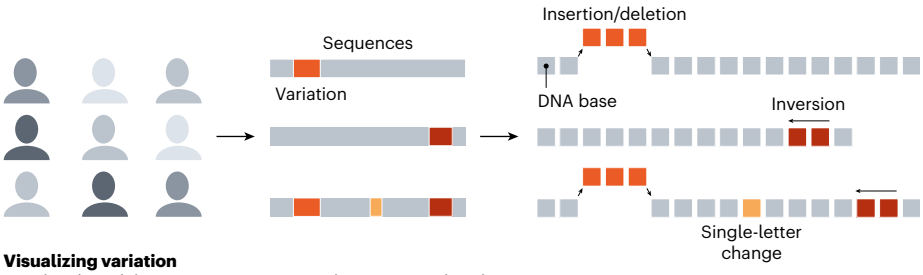
"Scientists were collecting Indigenous peoples' genomic data largely for the benefit of other, non-Indigenous peoples, which, when done without regards to Indigenous data

VISUALIZING A PANGENOME

The Human Pangenome Project aims to capture all of the variability in the human genome around the world. By analysing this variation and creating innovative ways to display it, the effort counters the assumption that there is a consensus of what a human genome looks like.

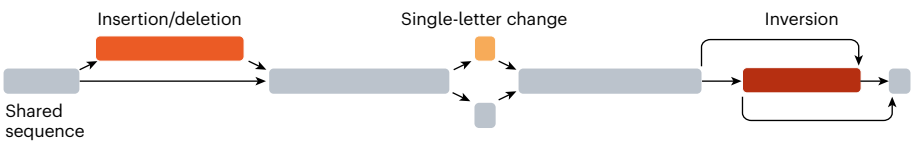
Gathering samples

Researchers will have to produce high-quality sequences for hundreds of individuals and catalogue the variants, including single-letter changes, insertions, deletions and inversions.



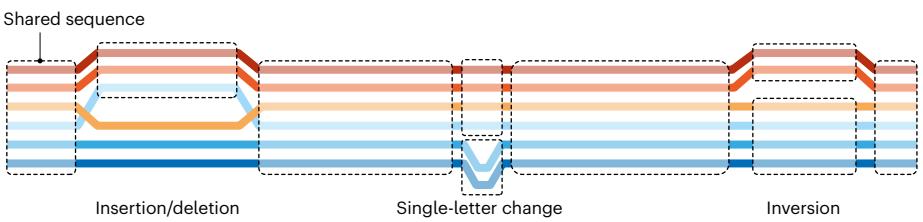
Visualizing variation

Graphical models can present variation data in a way that doesn't assume a standard, or default reference genome.



Exploring the pangenome

Representations that look like subway maps allow researchers to compare the variations in a population at a sequence level.



sovereignty, is a means of continued data extraction," says Krystal Tsosie, a geneticist and bioethicist at Vanderbilt University in Nashville, Tennessee, and a member of the Navajo Nation.

The next decade brought even more concern over ethical transgressions in genetic studies of under-represented groups, notably when the Havasupai Tribe filed a lawsuit against the Arizona Board of Regents and Arizona State University researchers in 2004. Members of the tribe had donated their DNA for genetic studies on type 2 diabetes, but discovered that it had been used without their consent for studies on schizophrenia and migration[7].

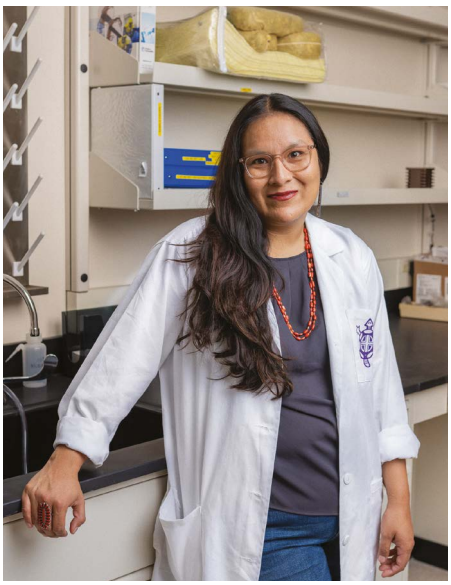
The researchers had also used stigmatizing words such as 'inbreeding' to explain genetic phenomena that were actually the consequence of population bottlenecks related to genocidal events, says Tsosie. She adds that, in the past, geneticists doing sequencing projects have often used racial language and failed to properly acknowledge the lasting legacy of colonialism in science, and the threat it poses to Indigenous people.

For several years, Fox and others have been calling for a massive departure from this approach. They say that Indigenous groups should have greater agency when it comes to the collection of their genetic data. Fox, who was a graduate student in Eichler's lab, says that he's not convinced that the pangenome project and others like it are involving the diverse groups they seek to sample in a way that truly empowers them. "I love Evan, man. When I have problems, I call him for advice," he says. "Despite that, you know, we don't agree on everything."

Fox advocates for an approach that puts sequencing power in the hands of the people. He and Tsosie are involved in the Native BioData Consortium, a non-profit research institute led by Indigenous scientists and tribal members in the United States that has been working to help Indigenous groups to acquire and run DNA sequencers on their own territory. The first sequencer was delivered to the Cheyenne River Sioux reservation in December 2020 says consortium co-founder Joseph Yracheta, a public-health geneticist at the Johns Hopkins Bloomberg School of Public Health in Baltimore. In February, Yracheta joined a Human Pangenome Project working group focused on ethical, legal and social implications of the project.

Fox is currently focused on genetic complexity found in the Pacific islands. He and his team mates are taking a holistic approach to sequence the genomes of agricultural species and other organisms in the environment in tandem, and are building a genomics institute to serve the community. Fox notes that the latest technologies, such as a 'distributed ledger' computer system that securely ties a person to their genetic data, can give people greater autonomy about whom they allow to access and use their information. "There are so many advancements in the data sciences right now that really allow for a new level of agency for participants," Fox says.

> "IF IT'S GOING TO HAPPEN, IT NEEDS TO HAPPEN IN THE BEST WAY THAT REPRESENTS INDIGENOUS PEOPLE."



TOMÁS KARMELO AMAYA/*NEW YORK TIMES*/REDUX/EYEVINE

Geneticist Krystal Tsosie.

Eichler is supportive of Fox's path. "I applaud his efforts to engage Indigenous scientists into genomics research – we need more of it," Eichler says. "It is not an either–or scenario, however, in my opinion." He adds that the Human Pangenome Project is encouraging Indigenous scientists to generate their own reference genomes. In those scenarios, "we will work together to make it happen by providing expertise and tools as needed".

No mutation without representation

Tsosie says that Indigenous groups might collaborate with big diversity projects in the future, but that it would have to happen in a way that would ensure that such communities can do their own sequencing. Moreover, although these major genome projects are often open-data efforts, Tsosie says it would be wise for there to be protections added for Indigenous people's deposited DNA sequences such that they be available only through access requests to avoid exploitation. "If it's going to happen, it needs to happen in the best way that represents Indigenous people," she says.

It's not just advocates from Indigenous communities in the United States who have voiced concerns about representation and data ownership. Others have argued that the pangenome project hasn't adequately involved researchers from regions outside the United States, according to Jarvis, who is on the project's sampling committee. He recognizes that some see the initiative as a largely US effort, but says that he and his collaborators are working to broaden it and involve scientists and participants from different parts of the world. For example, they have reached out to leaders of the Human Heredity and Health in Africa (H3Africa) programme to involve scientists in Africa who can do sequencing in countries there. (No sequencing effort seems immune from ethical challenges, however – even the H3Africa programme has had to straddle different countries' rules and norms governing the use of participant data, for example.)

Jarvis says he wants the Human Pangenome Project to achieve a better representation of human genetic diversity. "I'm a person of colour. I grew up as an African American. I grew up as an under-represented minority in the sciences," he says. "My diversity is not represented. So I have a personal motivation and a societal one to make sure that this pangenome really represents populations."

As they push forwards, the scientists also acknowledge that 350 genomes will not represent all human diversity. Ultimately, the true number of genomes needed to do this is difficult to pin down, and genetics often teaches us that rare differences can be important. "I don't think there is any magic number," says Adam Phillippy, head of the Genome Informatics Section at the NHGRI, and an investigator on the pangenome project.

Juggling the massive scientific undertaking while trying to avoid ethical pitfalls is something that weighs heavily on the pangenome researchers. "I'm sure there will be things that we will do that people will criticize five or ten years from now. I'm almost 100% sure of it," Eichler says. "But if we can go in with a clear conscience and say, we tried to do everything we possibly could to do it right, I feel that that's something."

Roxanne Khamsi is a science journalist based in Montreal.

1. Hsieh, P. *et al. Science* **366**, eaax2083 (2019).
2. Sherman, R. M. *et al. Nature Genet.* **51**, 30–35 (2019).
3. Miga, K. H. & Wang, T. et *al. Annu. Rev. Genom. Hum. Genet.* **22**, 81–102 (2021).
4. Tettelin, H. *et al. Proc. Natl Acad. Sci. USA* **102**, 13950–13955 (2005).
5. Nurk, S. *et al.* Preprint at bioRxiv https://doi.org/10.1101/2021.05.26.445798 (2021).
6. Dodson, M. & Williamson, R. *J. Med. Ethics* **25**, 204–208 (1999).
7. Garrison, N. A. *et al. Sci. Technol. Hum. Values* **38**, 201–223 (2013).