Diversity, inclusion and the Human Pangenome Project: Why capturing human genome diversity in our 4-letter language is such a big deal

“P

an” has several meanings.

As a noun, it refers to “a round metal container that often has a long handle and a lid.”

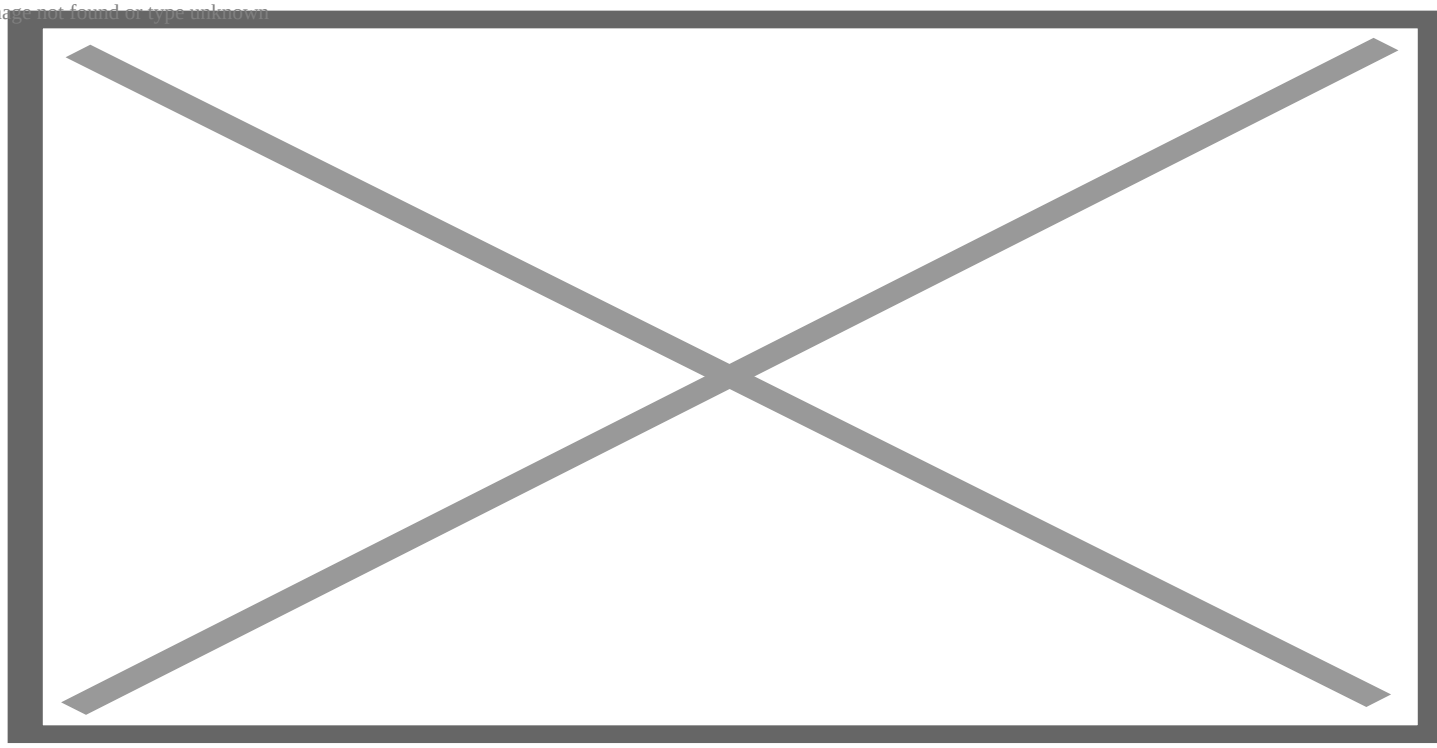
As a verb, it means criticism, like panning a film.

Peter Pan refers to an adult who doesn’t want to behave like one, from Sir James Barrie’s play about the boy who didn’t want to grow up.

As a prefix, “pan,” from the Greek, means “all, every, whole, and all-inclusive.”

Sigmund Freud reportedly used the term pan-sexualism in 1914, to mean “sex as a motivator of all things.”

Image not found or type unknown



Sigmund Freud. Credit: Toolshero

In genetics, the human pangenome is a complete reference of human genome diversity. It is envisioned as a new type of map that represents all of the ways that the sequence of 3,054,832 billion DNA base pairs – the building blocks of a genome – vary, plus or minus a few from short repeated sequences. The depiction is so densely packed that it resembles a map of the New York City subway system.

The Human Pangenome Reference Consortium is spearheading creation of a “genome reference

representation that can capture all human genome variation and support research on the full diversity of populations."

Such a resource is of course long overdue. Now that more than 30 million people have had their genomes sequenced, it's strange to think back about talk of "the" human genome, as if we are all identical for each of the 4 DNA nitrogenous bases – A, C, T, or G – occupying each of the 3 billion slots. We're not clones. But most biotechnologies take about 3 decades to mature, and since the human genome project got started in the early 1990s, things seem right about on schedule for a broader look.

Back in the mid 1980s, when I first attended meetings where the idea of sequencing "the" human genome surfaced, the task was expected to take at least a decade. About 93 percent of the first draft human genome sequence published in 2001 from the NHGRI and partners came from only 11 people, with 70 percent of the total from just one man, who was of 37 percent African ancestry and 57 percent European ancestry. The human genome published from Celera Genomics was reportedly Craig Venter's, head of that company.

After that, genome sequences began to trickle in, from celebrities, other rich folks, a handful of journalists who cranked out articles and books revealing their genetic selves, and a series of "firsts" – African, Han Chinese, and several modern peoples with ancient roots.

It's a little mind boggling to realize that today we can access our genome sequence data on our smartphones.

ma dna x e

Image not found or type unknown

Credit: Javier Jaen

Researchers began to catalog human genome diversity as the human genome project was winding down, by identifying single-base places in genomes that vary among individuals. These are the single nucleotide polymorphisms, or SNPs. As SNP collections peppered the chromosomes ever more densely, researchers quickly realized that new tools were needed to depict the unfurling diversity of our DNA.

Despite these sequencing advances, we still have a lot to learn about human genetic diversity, and that

calls for comparisons. Enter the human pangenome effort.

The diversity of our genome sequences is staggering. A study of whole genome sequences for 53,831 people found distinctions at 400 million places! Most were SNPs or an extra or missing DNA base. But it may be that much of our variability comes from only a few people. About 97 percent of the 400 million points of distinction came from less than one percent of the 53,831 participants, with 46 percent of them in only one person. We vary genetically in many ways, and some of us vary more than others, but we are all human.

For a few years, researchers compiled “reference” genome sequences to account for diversity in specific populations. These digital sequences displayed the most common DNA base found in many genomes from the group, at each point. But updating reference genomes took a long time, and it was a thankless task, never complete. By 2010, when more data from Asians and Africans had been added, still 5 million gaps in the reference sequences remained.

As the data swiftly outgrew attempts to capture genome diversity in a simple, clear visual tool, the idea emerged of the human pangenome: “a complete reference of human genome diversity.” The Human Pangenome Project officially began in 2019, and within a year, filled in the gaps remaining in genome sequences. The goal was to display the genome sequences of an initial 350 people from diverse ethnic groups, using “computational pangenomics” tools to create visuals called “genome graphs.”

In a genome graph, color-coded bases superimposed on the DNA depiction indicate how people vary, site-by-site. Like a geographical map with symbols denoting campgrounds, rest stops, and places of interest, genome graphs indicate SNPs and also missing parts of the genome sequence, extra hunks, and inverted regions. It also indicates meanings and context, such as distinguishing protein-encoding genes from control sequences, and highlighting places where the DNA sequence can be read from different starting points, which tells the cell to make different protein products.

The data pouring into the human pangenome project are coming from population biobanks and various genome sequencing projects. When all of this information is superimposed on the chromosome-length sketches, the genome graph indeed begins to resemble a subway map.

I grew up riding the New York City subways. Just as more train lines converge at the city’s center, Manhattan, with only a few lines extending into the boroughs, so too are the protein-encoding genes clustered toward each chromosome’s centromere, growing more sparse out towards the tips, the telomeres.

I think back in wonder at the first human genome meeting I attended, in 1986, I think in Boston. It’s been a long, strange trip, but we’re finally beginning to understand how a 4-letter language can spell out the astounding diversity of the human animal.

Ricki Lewis has a PhD in genetics and is a science writer and author of several human genetics books. She is an adjunct professor for the Alden March Bioethics Institute at Albany Medical College. Follow her at her website or Twitter @rickilewis

A version of this article originally appeared at PLOS and is reposted here with permission. Find PLOS on Twitter @PLOS