SciTechDaily

Global Genomes: Scientists Rewrite the Story of Human Genetics

https://scitechdaily.com/global-genomes-scientists-rewrite-the-story-of-human-genetics/?expand_article=1

By University of Washington School of Medicine August 7, 2023

The Human Pangenome Reference Consortium, a multi-institutional effort including UW Medicine, expands on the original Human Genome Project with data from 47 diverse individuals. It aims to improve understanding of genetic diversity and equity in human genome research, leading to transformative insights into genetic diseases.

University of Washington School of Medicine researchers played key roles in several aspects of a new genomic reference collection representing greater human population diversity.

UW Medicine genome experts made significant scientific contributions to a National Institutes of Health (NIH) Human Genome Research Institute reference collection that better represents the genetic diversity of the world's populations.

Called the Human Pangenome Reference Consortium, the multi-institutional effort expands and updates earlier work that started as the Human Genome Project. That original project, with drafts reported in 2001 and 2003, was based on a more limited sampling of human DNA. The goal then was to create an entire sequence of a human genome to use as a reference. It reflected data mostly from one person, with slight amounts of genetic information from about 20 others. That project was officially completed in 2022 with the release of the first telomere-to-telomere human genome.

Advancements in Human Genome Project

In contrast, the human pangenome reference contains nearly full genomic data from 47 people, representing different populations globally. This accounts for 94 human genomes, since each person carries two copies, one from each parent.

David Porubsky (left) and Mitchell Vollger (right) discuss the recent findings from the Human Pangenome Reference Consortium. Both led companion research studies published as part of the human pangenome reference collection report, May 2023. Credit: Randy Carnell/UW Medicine

“The pangenome approach represents a new way of thinking about human genetic variation,” said Evan Eichler, professor of genome sciences at the University of Washington School of Medicine in Seattle and one of the senior scientists in the Human Pangenome Reference Consortium. “It has the potential not only to improve discovery of genetic diseases but also transform our understanding of the genetic diversity of our species.”

Continued Expansion and Improved Equity

The current pangenome draft reference will continue to be expanded to include DNA sequencing and analysis from people from a variety of other ancestral and geographic roots. Eventually a cohort of more than 350 participants will enable researchers to capture the most common genetic variants, including ones that have been missed previously because they map to complex regions.

The latest research from the Human Pangenome Reference Consortium is reported in a series of papers in *Nature* and other scientific journals.

Impressive Research Outcomes

By reflecting variation across human populations, the pangenome reference collection is expected to improve equity in human genome research. Individuals and families from a wider range of backgrounds might benefit from new clinical advances based on knowledge of how genetic variation influences human health.

Researchers are already making discoveries that could not have been possible through previous human genome reference sequences.

The pangenome project studies in which University of Washington School of Medicine scientists made significant contributions were:

Drafting the Pangenome Reference

The overall project report, "A draft human pangenome reference," is published in *Nature*. Eichler, an expert in human genome evolution and variation, and their relation to disease, was among the senior authors. David Porubsky, Mitchell Vollger, William T. Harvey, Katherine M. Munson, Carl A. Baker, Kendra Hoekzema, Jennifer Kordusky and Alexandra P. Lewis, all from his department, were part of the project team.

This paper examines the diploid assemblies from 47 individuals. Diploid assemblies show a person's DNA sequence inherited from both parents, while only those from one parent appear in haploid assemblies. The assemblies were assessed to determine the extent of their coverage, accuracy, and reliability. The assemblies were found to be nearly complete (more than 99%) and highly accurate at the structural and base-pair levels. The researchers noted these assemblies outperformed earlier efforts at assembly quality, due to state-of-the-art sequencing technology and analytical innovations.

In addition to ascertaining known variants, the assemblies also captured new variants in structurally complex regions of the genome. These regions were previously inaccessible.

Challenges and Future Outlook

The authors also emphasized that the current pangenome reference is still a draft and that many challenges remain in building and refining this reference.

For example, the scientists plan to push towards a telomere-to-telomere or tip-to-tip sequencing of chromosomes to get a more complete picture of how people differ.

"That will give us a more comprehensive representation of all types of human variation," they noted. The researchers also would like to broaden subject recruitment because the present samples are insufficient to convey the extent of diversity in the human population.

Despite those and other limitations, the researchers anticipate that optimizing the pangenome reference collection will lead rapidly to a broad number of applications for scientists and clinicians.

Uncovering Variation Within Repetitive DNA

One of the related papers, a study led by UW Medicine researchers, is "Increased mutation and gene conversion within human segmental duplications," also appearing in *Nature*. The lead author is Mitchell R. Vollger, a postdoctoral fellow in genome sciences who collaborated with his colleagues as a student in the Eichler lab and with other Human Pangenome Reference Consortium scientists.

By overcoming previous obstacles in mapping areas of the genome containing large segments of repeated DNA code, they were able to spot more variants at the single-nucleotide level for many regions for the first time.

This is leading to a greater understanding of how, where, and to what degree mutations occur.

They discovered an elevated density of single-nucleotide variants within segmental duplications, compared to unique regions of the genome. They also found that almost a quarter of this increase was due to genes copying to new locations in a process called “interlocus gene conversion.”

The scientists created a map of hotpots that were prime locations for donating or receiving genetic material. They also observed that, from an evolutionary standpoint, areas of segmental duplication were slightly older than other parts of the genome containing unique sequences of DNA. However, this did not explain the increased density of single-nucleotide variants.

Interestingly, the nucleotide cytosine was more likely to convert to guanine, and vice versa, within duplicated sequences than were conversions among adenine and thymine. (A, T, C and G are the four chemicals that make up the alphabet for the DNA code.)

“These distinct mutational properties help maintain the higher cytosine and guanine content of segmental duplications of DNA, compared to unique DNA,” the researchers reported.

The scientists found more than 1.99 million single-nucleotide variants in these duplicated and gene-rich areas of the human genome—regions previously considered to be unreadable.

“A lot of this new sequence was uncovered last year [as part of the T2T Consortium] in copy number variable regions where there’s lots of differences between people,” Vollger said. “My focus in this latest work was looking at these variable regions and discovering the additional diversity that exists there and beginning to characterize it.”

He added, “Depending on how you choose to count, most human variation comes from these copy number variable regions that are only going to be unlocked using a pangenome reference. I think it’s absolutely critical that we continue to push the pangenome resource so that the scientific and clinical research community begins to adopt it.”

Closing the Gaps in Human Genome Assemblies

Another paper that is part of the series from the Human Pangenome Research Consortium appears in the journal Genome Research, under the title “Gaps and complex structurally variant loci in phased genome assemblies.” The lead author is David Porubsky, an acting instructor in genome sciences who conducts studies in the Eichler lab.

“Finishing multiple genomes is more difficult,” Porubsky said, “because human genomes are diploid. People carry two copies of a genome: the one inherited from the mom, and one inherited from the dad. So, the task is harder. That’s why there are gaps remaining. To resolve them, it will require more development in sequencing technology and more development in the underlying assembly algorithms, which we are using to put all these pieces together.”

Traditionally it has been challenging for scientists to separately reconstruct the DNA sequences for the two copies of our 23 chromosomes, but noteworthy progress has been made.

To do so, sequencing data usually is obtained from both parents, as well as from the child. However, in clinical settings, parental data is not always available.

Porubsky, Eichler, and their team are studying an approach that attempts to produce a complete genome assembly showing the set of genes from each parent—but without obtaining any parental data. They use a method called single-cell strand sequencing, or Strand-seq.

Either approach (trio-based or no parental data) can still result in gaps of missing information. The team analyzed gaps, assembly breaks, and misorientations from 77 phased and assembled human genomes from the Human Pangenome Reference Consortium. (A phased genome assembly tries to resolve the groups of variants in the chromosomes passed from each parent.)

The team learned several reasons for gaps arising in both methods, including areas where portions of DNA are incorrectly oriented. Many of these faulty orientations relate to large inversions, where things are figuratively turned upside down or inside out. Most of these occur between identical repeats of DNA code. There were also major assembly alignment discontinuities identified as regions of DNA that had undergone frequent expansions and contractions. Importantly, many of these areas overlapped with protein-coding genes, including areas with variations in copy number (how many times a section is repeated in one individual compared to another).

"My main task in this effort," Porubsky said, "was to better understand where we are coming short in the genome assembly, where the remaining gaps are, and how to close them. I was looking into where these gaps reside, their frequency, and the sequence properties. We found that many of these gaps are represented by these very long, highly repetitive sequences, which are difficult to assemble under the current technologies and algorithms."

Future Directions and Biomedical Relevance

"We are actually better positioned in the future to resolve them," Porubsky said, "and actually fill in these missing pieces of the puzzle and be able to better understand the human genome—even in these very complex parts of the human genome."

These regions contain biomedically relevant information, he noted.

"This is very important," he said, "because many of these complex parts of the genomes are associated with genetic disorders, such as certain forms of autism and Prader-Willi syndrome. Analyzing these regions may help in the future to better understand how to treat and diagnose these genetic disorders and identify perhaps new disorders which haven't been identified."

"A pangenomic representation [of these regions] would be most useful, yet more challenging, to realize," the researchers noted in their paper.

For more on this breakthrough, see:

- Human Pangenome Reference: A Deeper Understanding of Worldwide Genomic Diversity
- A Crystal Clear Image of Human Genomic Diversity
- Release of the New Human Pangenome Reference
- Piecing Together the Human Pangenome

"Increased mutation rate and gene conversion within human segmental duplications" by Mitchell R. Vollger, Philip C. Dishuck, William T. Harvey, William S. DeWitt, Xavi Guitart, Michael E. Goldberg, Allison N. Rozanski, Julian Lucas, Mobin Asri, Human Pangenome Reference Consortium, Katherine M. Munson, Alexandra P. Lewis, Kendra Hoekzema, Glennis A. Logsdon, David Porubsky, Benedict Paten, Kelley Harris, PingHsun Hsieh and Evan E. Eichler, 10 May 2023. *Nature*.
DOI: 10.1038/s41586-023-05895-y

The Human Pangenome Reference Consortium work at UW Medicine was supported in part by grants from the U.S. National Institutes of Health (5RO1HG002385, 5U01HG010971, R01HG010169, U24HG007497, and 1UO1HGO01973). Eichler is an investigator at the Howard Hughes Medical Institute.