Study addresses a long-standing diversity bias in human genetics

July 17 2024

The new gene expression dataset covers 26 populations across five continents. Credit: Will Kirk/Johns Hopkins University

Most research in human genetics has historically focused on people of European ancestries—a long-standing bias that may limit the accuracy of scientific predictions for people from other populations.

Now, a team of Johns Hopkins University scientists has generated a new

catalog of human gene expression data from around the world. The increased representation of understudied populations should empower researchers to attain more-accurate insights of genetic factors driving human diversity, including for traits such as height, hormone levels, and disease risk.

The work deepens the scientific field's understanding of gene expression in populations of Latin America, South and East Asia, and other regions for which limited data existed.

Published in *Nature*, the findings may improve future studies of human variation and evolution.

"We now have this global view of how gene expression contributes to the world's diversity, the broadest picture to date in populations that have been poorly represented in previous studies," said senior author Rajiv McCoy, a Johns Hopkins geneticist.

"We're trying to better understand the connection between variation at the level of our DNA and variation at the level of our traits, which previous genetic studies have looked at but with a really persistent bias that often excludes non-European ancestry populations."

While genetic research most often explores differences in DNA, the researchers set out to examine "gene expression"—the process by which genes in DNA are "transcribed" into RNA molecules.

RNA in turn serves as a blueprint to guide the assembly of amino acids into the proteins that provide structure and carry out various tasks within cells. But genetic mutations can affect how genes are expressed—changing how much RNA genes produce or the structure of the RNA itself. These mutations and associated effects on gene expression can critically impact the development of traits and diseases.

To identify mutations that change gene expression, the scientists measured RNA in cells from 731 people who had already participated in the 1000 Genomes Project, a previously established international collaboration that characterized the DNA sequence of the same individuals.

"We know not only their genome sequences, which were previously published, but we now have measurements of their gene expression. By combining these data, we can understand at a very basic level the genetic sources of gene expression differences between individuals," McCoy said.

"Ultimately that's what contributes most to the differences between you and me, even though at the level of DNA we are 99.9% identical."

Johns Hopkins geneticists Dylan Taylor (left) and Rajiv McCoy. Credit: Will Kirk/Johns Hopkins University

While the 731 individuals span 26 different groups across five continents, the scientists found that gene expression patterns are often shared between groups, a phenomenon also observed in patterns of DNA variation. Most of the differences in gene expression were seen within populations rather than between them.

"The distribution of our diversity is more complex than these geographically, politically, or socially defined labels," McCoy said.

The cohort's diversity allowed the scientists to spot possible connections between mutations and specific traits and health risks, including for mutations limited to subsets of populations that have previously gone unexamined, he said.

"We are demonstrating that by having this more-diverse cohort, we can really home in on specific mutations that could be driving these gene expression changes, and ultimately how they might be driving variation and how that affects traits or susceptibility to a disease," McCoy added.

The findings could also lead to better personalized therapies, said lead author Dylan Taylor, a Johns Hopkins doctoral candidate in biology.

"We can't really use these studies in a predictive fashion for personalized medicine equitably unless we have more diverse datasets," Taylor said.

"If you try to use results from a study using only European individuals to predict gene expression in individuals from an underrepresented

population—South Asians, for example—your results won't necessarily be very reliable."

Key gaps still exist. The 1000 Genomes dataset does not include many groups from the Middle East, Australia, and the Pacific Islands, and has limited samples from the Americas and Africa.

"The field is starting to move in this exciting direction to include diverse individuals in human genetic studies," Taylor said. "Our research is a proof of concept for other scientists. We are demonstrating we can really do this, and we should, and it's valuable."

  More information: Rajiv McCoy, Sources of gene expression variation in a globally diverse human cohort, *Nature* (2024). DOI: 10.1038/s41586-024-07708-2. www.nature.com/articles/s41586-024-07708-2

Provided by Johns Hopkins University