MedicalXpress

Widely used reference for the human genome is missing 300 million bits of DNA

November 19 2018

A depiction of the double helical structure of DNA. Its four coding units (A, T, C, G) are color-coded in pink, orange, purple and yellow. Credit: NHGRI

For the past 17 years, most scientists around the globe have been using the nucleic acid sequence, or genome, an assembly of DNA information, from primarily a single individual as a kind of "baseline" reference and human species representation for comparing genetic variety among groups of people.

Known as the GRCh38 reference genome, it is periodically updated with DNA sequences from other individuals, but in a new analysis, Johns Hopkins scientists now say that the collective genomes of 910 people of African descent have a large chunk—about 300 million bits—of genetic material that is missing from the basic reference genome.

"There's so much more human DNA than we originally thought," says Steven Salzberg, Ph.D., the Bloomberg Distinguished Professor of Biomedical Engineering, Computer Science, and Biostatistics at The Johns Hopkins University.

Knowing the variations in genomes across populations is essential to research design to reveal why certain people or groups of people may be more or less susceptible to common health conditions, such as heart disease, cancer and diabetes, and Salzberg says that scientists need to build more reference genomes that more closely reflect different populations.

"The whole world is relying on what is essentially a single reference genome, and when a particular DNA analysis doesn't match the reference and you throw away those non-matching sequences, those discarded bits may in fact hold the answers and clues you are seeking," says Salzberg.

Rachel Sherman, the first author on the report and a Ph.D. student in computer science at Johns Hopkins, says, "If you are a scientist looking for genome variations linked to a condition that is more prevalent in a

certain population, you'd want to compare the genomes to a reference genome more representative of that population."

Specifically, the world's reference genome was assembled from the nucleic acid sequences of a handful of anonymous volunteers. Other researchers later determined that 70 percent of the reference genome derives from a single individual who was half European and half African, and the rest derives from multiple individuals of European and Chinese descent, according to Salzberg.

"These results underscore the importance of research on populations from diverse backgrounds and ancestries to create a comprehensive and inclusive picture of the human genome," said James P. Kiley, Ph.D., director of the Division of Lung Diseases at the National Heart, Lung, and Blood Institute (NHLBI), which supported the study. "A more complete picture of the human genome may lead to a better understanding of variations in disease risk across different populations."

For the new analysis, described online Nov. 19 in *Nature Genetics*, Salzberg and Sherman began their project with DNA collected from 910 individuals of African descent who live in 20 regions around the globe, including the U.S., Central Africa and the Caribbean. Their DNA had been collected for an NHLBI-supported study at Johns Hopkins led by Kathleen Barnes, Ph.D., who is now at the University of Colorado and continues to lead this program on genetic factors that may contribute to asthma and allergy, conditions known to be overrepresented in this population.

Many researchers look for small differences between the reference genome and the genomes of the individuals they are studying—sometimes only a single change in chemical base pairs within the DNA. These small changes are called single nucleotide polymorphisms, or SNPs.

However, Salzberg's team focuses on larger variations in the genome. "SNPs correlate really well to figure out an individual's ancestry, but they haven't worked as well to determine genetic variations that may contribute to common conditions and diseases," says Salzberg. "Some conditions may be due to variations across larger sections of the genome."

Over a two-year period, Salzberg and Sherman analyzed the DNA sequences of the 910 people, looking for sections of DNA at least 1,000 base pairs long that did not align with or match the reference genome. "Within these DNA sequences are what makes one individual unique," says Sherman.

They assembled those sequences and looked for overlaps and redundancies, filtering out sequences shorter than 1,000 base pairs, and DNA likely linked to bacteria, which is found in all humans.

Then they compared the assembled sequences of all 910 individuals to the standard reference genome to find what Salzberg calls, "chunks of DNA that you may have and I don't."

In all, they found 300 million base pairs of DNA—which is about 10 percent of the estimated size of the entire human genome—that the reference genome did not account for. The largest section of unique DNA they found was 152,000 base pairs long, but most chunks were about 1,000-5,000 base pairs long.

A small portion of these DNA sequences may overlap with genes that encode proteins or other cellular functions, but, Salzberg says, they have not mapped the function of each sequence.

They also failed to find sequences that aligned with having asthma or not. But Salzberg isn't deterred: "Until you survey the landscape, you

can't figure out what's useful."

More information: Rachel M. Sherman et al. Assembly of a pan-genome from deep sequencing of 910 humans of African descent, *Nature Genetics* (2018). DOI: 10.1038/s41588-018-0273-y