inverse.com

There's a Huge Problem With the Core of the Human Genome Project

7-9 minutes

The Human Genome Project, which began in the 1990s, was *Homo sapiens*' successful attempt to map out the entirety of our species' DNA. It produced the human reference genome, a finely polished collection of human DNA that's crucial for genetics research and genetics testing services around the world. Integral as it has been to the science community, two researchers at Johns Hopkins University have discovered that the reference genome is missing a piece or two — well, 296,485,284 base pairs of DNA, to be exact.

The reference genome is an essential map of human genetic material that is used as a basis for comparison. When we sequence our own DNA for insight into health, family history, and future disease risk, we chop up the sequence into lots of little pieces and compare stretches of it to the reference genome, looking for areas where we differ. The fundamental problem with this, the scientists write in a recent paper in *Nature Genetics*, is that the reference genome is based largely on a single person. Considering the myriad genetic differences among the 7.7 billion people alive today, that's obviously not ideal.

Professor of computer science and biostatistics Steven Salzberg, Ph.D., and Rachel Sherman, a Ph.D. candidate, make the case that

this single reference genome doesn't capture the diversity of human genetics. Some populations, they add, differ *too much* from this reference genome. To make their case, they refer to the genomes of 910 individuals from twenty different countries, all of pan-African descent.

Locations where the pan-African genome differs greatly from the reference genome

In the DNA of these individuals, the team found 300 million pieces of DNA common that don't exist in our "reference" genome. If we disregard this much material, says Salzberg, we'd inevitably miss key insights into the health and history of specific populations. They too, are human, so shouldn't they be represented in the "human" reference genome?

"These regions are essentially invisible to the genetics community until we have a reference genome that includes those regions,"

Salzberg tells *Inverse.*

The Problem With The Reference Genome.

Over the years, we've continuously workshopped the reference genome. But recent analysis indicates that almost *seventy percent* of its material was gleaned from a single African-American individual, who is referred to only as RPCI-11, explains Salzburg.

That means that when scientists perform genetic analysis to identify differences between diverse populations from all over the world, most of the time, they compare those genomes to the genetic material from, mostly, one person. This leads us to often ignore material that might be too different from this reference, says Sherman. She calls them "missing pieces."

If you print out the entirety of the human genome, its material will fill a bookcase

"When you line things up, there are going to be pieces that don't line up at all because they're too different to match anything from the reference genome." Sherman says. "Then you ignore all the stuff that doesn't line up as not really relevant or not really worth looking, at when maybe these are actually the pieces that are of the most interest because they're the most different from the reference genome."

In the study, Sherman and Salzberg took big chunks of this "different" material (about 1,000 base-pairs long) and tried to determine whether they merely represented accidental strings of sequencing errors — or really did hold useful information about unexplored human DNA.

The team came to the conclusion that this "new" DNA is of high enough quality to warrant a second pass, even though they don't know what its significance to the human body might be just yet.

What Are The Consequences?

So far, Sherman says, we don't really know what we're missing by ignoring DNA that's not represented in the reference genome. But who knows what we might find there if we take a look?

Salzberg suggests we imagine a fictional population that has an extra chromosome — 24 instead of the usual 23 in each cell. Nothing in that extra chromosome from this population would line up with the reference genome. Maybe, she says, somewhere on that hidden chromosome, is the reason why the fictional population tends to develop a certain disease — and why the rest of the world doesn't. But because we don't have the right reference to compare it to, we would never know it's there.

> Original advertisement that brought in the donors for Human Genome Project (Buffalo News, 3/23/1997), h/t Pieter de Jong, who placed the ad pic.twitter.com/gNB7mMv3Yu
>
> — Jay Shendure (@JShendure) October 28, 2017

An ad recruiting donors for the human genome project

“If once in a while there were mutations in that chromosome that did cause problems, you’d never be able to study those,” Salzberg says. “You’d never be able to observe them if you relied exclusively on this reference genome.”

Let’s be clear: this research doesn’t provide evidence for some undiscovered chromosome. But it does suggest that we’re probably missing a lot when we use a single reference genome from some person called RPCI-11 as a basis for all of our analyses on the DNA of our entire species.

How Can We Fix It?

Instead of striving for a single universal reference genome, the team argues, we should have a *bunch* of reference genomes — perhaps one for each population of interest.

“What we’re kind of advocating here with this finding, is that we really need to be building reference genomes for each population,” says Sherman. “If there’s this much DNA missing from the reference in this population, the model needs to change.”

Some countries have taken it upon themselves to at least attempt to create their own reference genomes. Denmark, for example, is compiling genetic material from 150 Danes in an attempt to make a true “Danish” reference genome. A 2016 paper in *Nature* describes

an attempt to compile a reference for Korean individuals, though that paper too only describes research conducted on a single person. But other projects, like the 1,000 Genomes Project, are also trying to get this process started, but it's a lot of work to create a reference material as polished as the current version, known as GRCh38.

"You have to do more than just go and sequence another person from another population to create a reference genome," he adds. "You have to do quite a bit more."

Its not that researchers aren't aware that we need more reference genomes. Salzberg just hopes that there would be more of them by now, and that they'd at least be widely adopted as standard reference genomes. The paper laments that none of these attempts have achieved the same status and clout as GRCh38 — though that's the goal of the Danish project.

Going forward, Sherman and Salzberg are taking it upon themselves to get this project started, by building several additional reference genomes, which they hope to release in one to two years. They're looking to begin creating a library of reference genomes to help people get as much insight out of their genetic material, no matter how "different" it is.

"What we really need to have are hundreds of reference genomes," he adds. "That's going to happen one day."