Missing data? African study shows why we need to expand the human reference genome

The human genome sequence, first published in 2001, has some important information missing. The latest version of it, called GRCh38, has a monstrous 3.1 gigabases of information—but that's still not enough. A letter published in Nature Genetics [November 19] finds that the reference genome is missing a colossal 10 percent of the genetic information found in the genomes of hundreds of people with African ancestry—information that also appears in other human populations.

The "human genome" is in fact assembled from the genomes of just a handful of people, with the majority of GRCh38 coming from just one person.

…

[Author Rachel Sherman] set out to create a pan-genome for Africa, using DNA from 910 people of African descent.
…

Sherman and her colleagues looked for sequences more than 1,000 base pairs long that didn't match the reference and found a lot of them: nearly 300 million base pairs, which is about 10 percent of the size of the entire reference genome.

That's not to say this information is unique to African people: about 40 percent of this data matched either the Korean or Chinese genomes. This suggests that it's important genetic material that's present across a huge range of humans, but still not captured by the reference genome assembled from just a small number of people. There's a lot going on with humans that isn't reflected by the human reference genome.

Read full, original post: DNA data from Africans reveals sequences that we'd missed