MedicalXpress

Latino genomes point way to hidden DNA: 20 million missing base pairs mapped

August 8 2013, by Jake Miller

Hidden in the tangled, repetitious folds of DNA structures called centromeres, researchers from Harvard Medical School and the Broad Institute have discovered the hiding place of 20 million base pairs of genetic sequence, finding a home for 10 percent of the DNA that is thought to be missing from the standard reference map of the human genome.

Mathematician Giulio Genovese, a computational biologist in genetics at HMS and at the Broad Institute, working in the lab of geneticist Steven McCarroll, HMS assistant professor of genetics and director of genetics for the Stanley Center for Psychiatric Research at the Broad Institute, found a way to use the genomes of Latinos to interpolate the locations of these missing pieces. Their findings will be published in The *American Journal of Human Genetics* on August 8.

"In nature, polymerase, the molecular machinery that copies DNA within living cells, can sequence hundreds of millions of base pairs of DNA. The techniques we've developed to sequence DNA in the lab can only do relatively short segments, and we need to stitch those pieces together after the fact," Genovese said. "So while we wait for sequencing technology to catch up with nature, we wanted to see if we could use mathematical patterns to find a place for some of the missing pieces."

By using the genomes of admixed populations—populations, such as Latinos and African Americans that derive ancestry from more than one continent—the team developed a sophisticated mathematical method to

help fill in the uncharted regions on the human genome map. The map is a key tool that geneticists rely on to find disease genes and identify the functional genetic variations at the core of human diversity. The unmapped DNA also sometimes resembles known, mapped genes, which can interfere in attempts to study similar sequences.

Best known as the molecular hinges that help chromosomes divide, centromeres have been widely considered structural elements that were unlikely to harbor protein-coding genes, the researchers said. For this reason, their finding—that nearly half of the unmapped sequences contained in available genomic reference libraries, including many protein-coding genes, were located in the centromeres—was unexpected.

Insight from a diverse population

Surprisingly, the study also found that the genomes of Latino individuals are a uniquely powerful resource for assembling maps of the human genome. The study searched 242 Latino genomes from the 1000 Genomes Project Phase 1 for DNA sequences that have not yet been located on the reference human genome map.

"Throughout the history of genomic research, different populations have given unique gifts to genetic inquiry because of the history or structure of that population," said McCarroll.

The power of the Latino genome for Genovese's approach came from the contribution of the African ancestors that many Latino individuals have. Because of the long history of human evolution on the continent, the African genome is rich in genetic diversity. Other human populations evolved from subsets of that diverse population, as small groups migrated around the globe just a few tens of thousands of years ago. (Sometimes, however, the lack of diversity in a population can be an asset for researchers. There are island populations that have allowed the

discovery of recessive mutations that are rare in most of the world, but happen to be more common on a given island.)

"Latino populations have a relatively distinctive gift to give. Having some recent African ancestry, but just a little, can yield especially powerful information about what the structure of the human genome is in all populations," McCarroll said.

When chromosomes recombine with each other in each generation, they do so in relatively large segments or chunks. In the genomes of Latinos— many of whom trace ancestry to European, Native American and African populations—the mixed European, Native American and African sequences form a mosaic of large segments.

Imagined as separate colors, an admixed genome would look like a mosaic with large red, green and blue tiles, rather than a video screen with tiny, mixed-color pixels.

Genovese developed an algorithm that could use a missing sequence's proximity to known genetic markers to pinpoint where on the chromosome the missing pieces fit—a technique first reported in a related paper in February, which localized a smaller sample of genes.

The technique works best when individuals have some African DNA because the diversity among African genomes provides a high number of genetic markers. But Genovese discovered that his technique is most powerful when individuals have only a little African ancestry— because this genetic "signal" is then most localized to a small number of regions in their genomes. Because the sampled Latino genomes had low levels of African ancestry (on average, just a few percent, compared to around 80 percent in African Americans), it was more powerful for pinpointing where on the map the marker was.

The blank spots on the map that the researchers identified were the centromeres, the only places where the missing DNA could be hidden.

A new approach to mapping

Until this work, scientists have tended to assume that mapping the remaining patches of terra incognita in the human genome would require future improvements in sequencing technology.

"I think people have tended to assume that someone will invent some sequencing technology that can magically read chromosomes in sequence from end to end," McCarroll said. "Giulio approaches the problem as a mathematician, and his favorite genome technology is his own mind—he saw a way to answer this question using data that was already in front of us, looking for patterns and relationships in the data instead of trying to sequence everything."

The highly repetitive DNA that makes up much of the centromeres is especially challenging to sequence with current technology. Instead of trying to sequence all the way through the unknown regions, the researchers used known information on both sides of the gaps to show what fits in the middle.

The millions of base pairs of sequence that Genovese and McCarroll's team have located will be added to the next release of the reference human genome assembly—the "Google maps" of the human genome that geneticists use every day—providing a more comprehensive view of the genome and how the pieces all fit together.

Provided by Harvard Medical School