New methods identify thousands of new DNA sequences missing from reference map of human genome

April 20 2010, by Leila Gray

(PhysOrg.com) -- A person can have one or more copies, or no copy at all, of a particular DNA sequence, which may account for why these sequences were absent from the reference genome.

"A large portion of those sequences are either missing, fragmented or misaligned when compared to results from next-generation sequencing genome assemblies on the same samples,"said Dr. Evan Eichler, senior author on the findings published online in advance of print today, April 19, in Nature Methods. Eichler is a University of Washington (UW) professor of genome sciences and an investigator with the Howard Hughes Medical Institute. "These findings suggest that new genome assemblies based solely on next-generation sequencing might miss many of these sites."

Dr. Jeffrey M. Kidd was lead author of the article, which describes the new techniques the research team used to find some of the missing sequences. Kidd headed the study in Eichler's lab while earning his Ph.D. at the UW, and is now a postdoctoral fellow at Stanford University.

"Over the past several years, the extent to which the structure of the genome varies among humans has become clearer. This variation suggested that there must be portions of the human genome where DNA sequences had yet to be discovered, annotated and characterized,"he said

"We hope that these sequences ultimately will be included as part of future releases of the reference human genome sequence."

The reference genome assembly is a yardstick -- or standard for comparison -- for studies of human genetics.

The human reference genome was first created in 2001 and is updated every couple of years, Kidd explained. It's a mosaic of DNA sequences derived from several individuals. He went on to say that about 80 percent of the reference genome came from eight people. One of them actually accounts for more than 66 percent of the total.

Along with their collaborators at Agilent, the team designed ways to examine these newly identified sequences in a panel of people representing populations from around the world. The researchers found that, in some cases, the number of copies of these sequences varied from person to person.

The fact that a person can have one or more copies, or no copy at all, of a particular DNA sequence may account for why these sequences were missing from the reference genome. The researchers also found that some of these sequences were common or rare in different populations, depending on from which part of the globe their ancestors originated.

"Each segment of the reference genome is from a single person, and reflects the genome of that individual. If the donor sample was missing a sequence that many other people have, that sequence would not be represented in the reference genome."Kidd explained. "That is why some of the positions on the reference genome represent rare structural configurations or entirely omit sequences found in the majority of people."Kidd said that the study published in today's Nature Methods used information from nine individuals, representing various world populations, to search for and fill in some of the missing pieces.

By looking at genomes from seven kinds of animals, the researchers were also able to show that some of the newly identified DNA sequences appear to have been conserved during the evolution of mammals and man. The animals whose genomes were studied were chimpanzee, Bornean orangutan, Rhesus monkey, house mouse, Norway rat, dog, and horse.

"Some of the sequences were present in several different species, but were absent from the reference genome,"Kidd said. "Some of the sequences present in several mammals actually correspond to sites of variations in humans -- some people have retained a particular sequence, and others have lost it."

The researchers also developed a method to accurately genotype many of the newly found DNA sequences and created a way to look at variations in the number of copies of these sequences, thereby opening up regions of the human genome previously inaccessible to such studies.

"Scientists can now begin trying to understand the functional importance of these sequences and their variations,"Kidd said.

The 1,000 Genomes Project (an international effort to fully sequence the genomes of a thousand anonymous individuals) and other genome studies are amassing massive amounts of data on DNA sequences that are then mapped to the reference genome, he added. Any study, he continued, that improves the completeness and quality of the reference genome assembly will thereby benefit these projects and lead to a fuller picture of the extent of human genomic variation.

The findings are published in *Nature Methods* as "Characterization of missing human genome sequences and copy-number polymorphic insertions".

In addition to Kidd and Eichler, other researchers on the study were Nick Sampas, Paige Anderson, Anya Tsalenko, N. Alice Yamada, Peter Tsang, and Laurakay Bruhn, all of Agilent Laboratories, Santa Clara, Calif.; Francesca Antonacci, Hillary S. Hayden, Can Alkan, and Maika Malig, all of the University of Washington in Seattle; Tina Graves, Robert Fulton, Joelle Kallicki, and Richard K. Wilson, all of the Genome Sequencing Center at the Washington University School of Medicine in St. Louis; and Mario Ventura and Giuliana Giannuzzi of the Department of Genetics and Microbiology, University of Bari, Italy.

Kidd's work on this study was supported by a U.S. National Science Foundation Graduate Research Fellowship. The study was funded by a grant to Eichler from the National Institutes of Health entitled "Human Genome Structural Variation."

Provided by University of Washington