Researchers Develop Algorithm to Map Copy Number, Segmental Duplications Using Sequence Data

August 31, 2009

Byline: Andrea Anderson
Newsletter: GenomeWeb Daily News
GenomeWeb Daily News - August 31, 2009

NEW YORK (GenomeWeb News) – In a paper appearing in yesterday's advanced, online edition of Nature Genetics, a group of researchers from the University of Washington and elsewhere reported that they have developed and tested a read-mapping algorithm to map copy number variation in the genome from short read sequence data.

The team used their algorithm, dubbed micro-read fast alignment or mrFAST, to assess three human genomes, identifying segmental duplications that encompassed millions of bases of DNA as well as at least 113 copy number variable genes. Based on their findings, they estimate that there are an average of 73 to 87 copy number differences between any two individuals. Now, the researchers are gearing up to apply their method to data from the 1000 Genomes Project.

Although some approaches for finding CNVs are available, the researchers noted, commercial SNP arrays and array comparative genome hybridization approaches can miss some copy number changes — particularly those occurring in segmental duplications.

"Even sequence-based strategies such as paired-end mapping frequently cannot unambiguously assign end sequences in duplicated regions, making it impossible to distinguish allelic and paralogous variation" senior author Evan Eichler, a researcher with the University of Washington's Department of Genome Sciences, and his co-authors wrote.

The difference between the mrFAST algorithm and other methods for

translating high-throughput sequence data into copy number information is that mrFAST checks every location in the genome and matches these sites to reads with at least 94 percent identity, lead author Can Alkan, a post-doctoral researcher in Eichler's lab, told *GenomeWeb Daily News*.

For the current paper, the team then applied the algorithm to three human genomes: the Watson genome (sequenced by researchers at the Baylor College of Medicine and Roche 454 Life Sciences), the Yoruban genome (sequenced by Illumina researchers), and the Han Chinese genome (sequenced by a team from the Beijing Genomics Institute).

The Watson genome had been sequenced with Roche 454 GS-FLX technology and the African and Asian genomes were both sequenced using the Illumina platform.

Although a similar approach could be applied to longer reads, Alkan explained, the researchers calibrated the architecture of their algorithm to short reads for the current study. Consequently, they had to chop up the Watson genome into smaller bits so that the reads more closely resembled short reads available for the other two genomes.

Before looking for copy number changes, Alkan explained, the researchers tossed common repeats, such as LINE elements. And, he explained, in order to compensate for the GC bias present in the reads, they also increased the read depth in these regions slightly.

The researchers found that read depth did correspond to copy number patterns. For instance, when they focused in on a set of 961 known autosomal duplications, the team found that they could detect more than 90 percent of the segmental duplications larger than 20,000 bases with 20 times sequence coverage.

Overall, the team found 725 non-overlapping, large segmental duplications. Nearly all of these were present in the genomes of all three men. The team subsequently verified their results using array CGH, Alkan explained, finding fairly good matches between CNVs predicted by each method.

The researchers also found 68 gene families with full or partial copy number variation. And they predicted that, on average, nearly four percent of genes will likely show differences of at least one copy number. In the three genomes tested, for example, the team verified copy number changes involving at least 113 genes — including genes implicated in everything from psoriasis to color blindness.

The researchers did not attempt to look at any reads shorter than about 35

base pairs, Alkan said, though he believes the algorithm could use reads as short as about 25 base pairs.

Down the road, the team plans to try to use the algorithm to find CNVs in sequence data from the 1000 Genomes Project. If quality control studies indicate that the algorithm and sequence data are compatible, Alkan said, the researchers plan to start by applying mrFAST to about 100 genomes sequenced through the project.