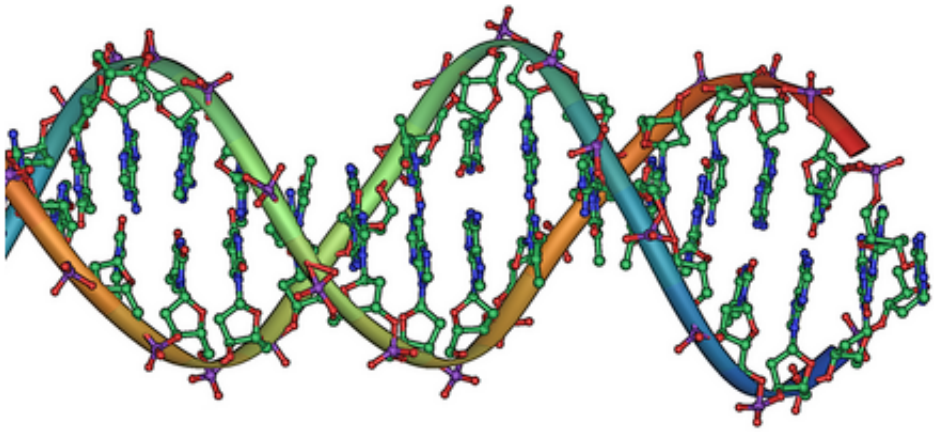
Genome sequencing of individual Korean offers opportunity to identify parts of sequence unique to Korean population

October 6 2016, by Bob Yirka



DNA double helix. Credit: public domain

(Phys.org)—A team of researchers with the Genomic Medicine Institute at Seoul National University in South Korea and Korean genetic sequencing firm Macrogen has conducted the most complete genome sequencing of a person from Korea to date. In their paper published in the journal *Nature*, the team describes how they sequenced the genome and the medical benefits they believe will come from their work.

As the researchers note, the genomes of many people have been sequenced—the first famously in back in 2000—but as they also note, people from Asia have been severely under-represented in such studies. Also, the researchers note, such sequences have gaps or incomplete sequence data in places that reflect the unique features of people from different parts of the world. After the first mapping was made, they explain, researchers began using it as sort of a template, plugging in information when it was found regarding people from different geographical regions.

In this new effort, the researchers went back to the beginning and sequenced the genome of a single individual Korean person from scratch. In so doing, they were able to fill in some of the gaps that have been in place since the first genome was fully sequenced. They also found a large number of structural differences between the genome of the Korean person and the original template—and several that were missing altogether. They suggest their findings will help in designing and testing drugs that might have different impacts on people from separate regions. The note, too, that their research will have another impact beyond Korea—they believe the sequenced genome will be a closer match to people of other Asian areas as well, offering scientists in such places a better base from which to begin their own research efforts.

The researchers also point out that the new method they used to sequence the genome will also make it easier to test for the best matches in organ donations and will help to speed up the process by which drugs are tested for adverse side effects. Fully sequencing the genomes of people from different parts of the world, the team claims, will create advancements in precision medicine, where health care overall is tailored to specific individuals based on their genetic makeup.

More information: Jeong-Sun Seo et al. De novo assembly and phasing of a Korean human genome, *Nature* (2016). DOI:

10.1038/nature20098

Abstract

Advances in genome assembly and phasing provide an opportunity to investigate the diploid architecture of the human genome and reveal the full range of structural variation across population groups. Here we report the de novo assembly and haplotype phasing of the Korean individual AK1 using single-molecule real-time sequencing, next-generation mapping, microfluidics-based linked reads4, and bacterial artificial chromosome (BAC) sequencing approaches. Single-molecule sequencing coupled with next-generation mapping generated a highly contiguous assembly, with a contig N50 size of 17.9 Mb and a scaffold N50 size of 44.8 Mb, resolving 8 chromosomal arms into single scaffolds. The de novo assembly, along with local assemblies and spanning long reads, closes 105 and extends into 72 out of 190 euchromatic gaps in the reference genome, adding 1.03 Mb of previously intractable sequence. High concordance between the assembly and paired-end sequences from 62,758 BAC clones provides strong support for the robustness of the assembly. We identify 18,210 structural variants by direct comparison of the assembly with the human reference, identifying thousands of breakpoints that, to our knowledge, have not been reported before. Many of the insertions are reflected in the transcriptome and are shared across the Asian population. We performed haplotype phasing of the assembly with short reads, long reads and linked reads from whole-genome sequencing and with short reads from 31,719 BAC clones, thereby achieving phased blocks with an N50 size of 11.6 Mb. Haplotigs assembled from single-molecule real-time reads assigned to haplotypes on phased blocks covered 89% of genes. The haplotigs accurately characterized the hypervariable major histocompatability complex region as well as demonstrating allele configuration in clinically relevant genes such as CYP2D6. This work presents the most contiguous diploid human genome assembly so far, with extensive investigation of unreported and Asian-specific structural

variants, and high-quality haplotyping of clinically relevant alleles for precision medicine.