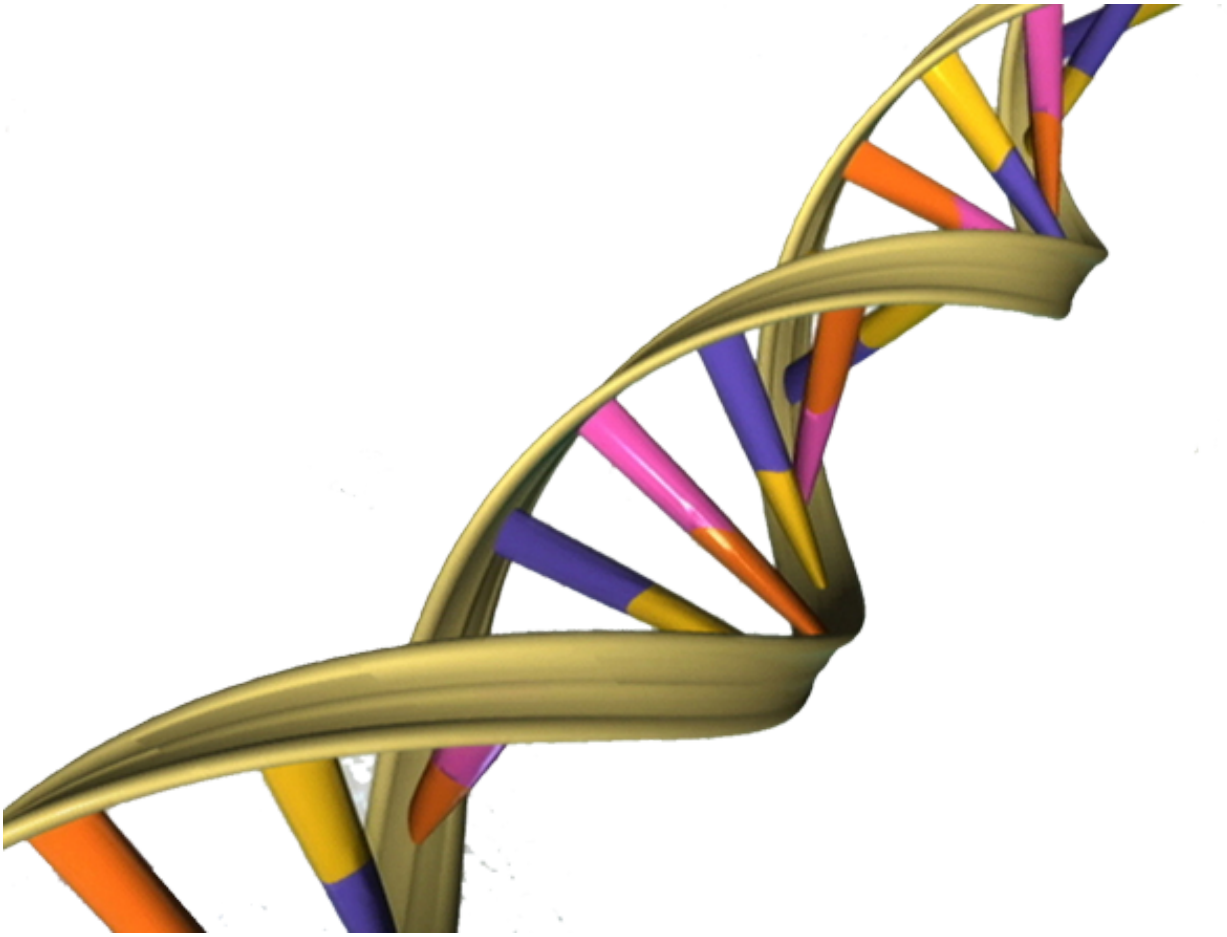
Researchers conduct sequencing and de novo assembly of 150 genomes in Denmark

July 27 2017, by Bob Yirka



A depiction of the double helical structure of DNA. Its four coding units (A, T, C, G) are color-coded in pink, orange, purple and yellow. Credit: NHGRI

(Phys.org)—A large international team of researchers has developed a Danish reference genome catalog based on the de novo assembly of 150 genomes sequenced from 50 family trios. In their paper published in the journal *Nature*, the group describes the multi-year effort, its purpose, and where they believe such efforts are leading.

One of the ways that scientists are learning about diseases, particularly those that are hereditary in nature, is by sequencing the genomes of large groups of people—doing so enables searching for variants that cause or contribute to a given disease—the ultimate goal would be to sequence every single person on Earth. In this new effort, the researchers have sequenced the genomes of 150 people in Denmark (from 50 family trios) of Danish descent—native Inuits and immigrants were screened out.

The effort, the researchers report, was carried out partly to create a Danish reference catalog and partly to learn how to conduct large-scale genome sequencing. They note that such a project required the combined efforts of multiple people and organizations working in coordinated fashion. Ultimately, the project cost approximately $10 million.

One major aspect of the project was selecting technology and determining how many people to sequence. The team wound up using samples from 50 families and did the sequencing using both combinations of paired-end and mate-pair libraries with the Illumina HiSeq2000. They note also that de novo assemblies were used because the researchers believed other approaches left out pertinent information. De novo refers to deriving a peptide sequence from a mass spectrum without the use of a sequence database. It is the preferred approach to sequencing when the aim is to identify novel peptides in organisms that have not been previously sequenced. The researchers report very few gaps in the data. They also note that three assemblies were used: Allpaths-

LG14, SOAPdenovo2 and SGA. Accuracy was measured by comparing their results with the human reference genome.

The researchers hope that their effort will lead to improved medical interpretation of genetics in Denmark.

More information: Lasse Maretty et al. Sequencing and de novo assembly of 150 genomes from Denmark as a population reference, *Nature* (2017). DOI: 10.1038/nature23264

Abstract
Hundreds of thousands of human genomes are now being sequenced to characterize genetic variation and use this information to augment association mapping studies of complex disorders and other phenotypic traits. Genetic variation is identified mainly by mapping short reads to the reference genome or by performing local assembly. However, these approaches are biased against discovery of structural variants and variation in the more complex parts of the genome. Hence, large-scale de novo assembly is needed. Here we show that it is possible to construct excellent de novo assemblies from high-coverage sequencing with mate-pair libraries extending up to 20 kilobases. We report de novo assemblies of 150 individuals (50 trios) from the GenomeDenmark project. The quality of these assemblies is similar to those obtained using the more expensive long-read technology. We use the assemblies to identify a rich set of structural variants including many novel insertions and demonstrate how this variant catalogue enables further deciphering of known association mapping signals. We leverage the assemblies to provide 100 completely resolved major histocompatibility complex haplotypes and to resolve major parts of the Y chromosome. Our study provides a regional reference genome that we expect will improve the power of future association mapping studies and hence pave the way for precision medicine initiatives, which now are being launched in many countries including Denmark.