SHAPEIT4: An algorithm for large-scale genomic analysis

December 20 2019

Credit: CC0 Public Domain

Haplotypes are a set of genetic variations that, located side by side on the same chromosome, are transmitted in a single group to the next generation. Their examination makes it possible to understand the heritability of certain complex traits, such as the risk of developing a disease. However, to carry out this analysis, genome analysis of family

members (parents and their child) is usually necessary, a tedious and expensive process.

To overcome this problem, researchers from the Universities of Geneva (UNIGE) and Lausanne (UNIL) and the SIB Swiss Institute of Bioinformatics have developed SHAPEIT4, a powerful computer algorithm that allows the haplotypes of hundreds of thousands of unrelated individuals to be identified very quickly. Results are as detailed as when family analysis is performed, a process that cannot be conducted on such a large scale. Their tool is now online under an open source license, freely available to the entire research community. Details can be found in *Nature Communications*.

Nowadays, the analysis of genetic data is becoming increasingly important, particularly in the field of personalized medicine. The number of human genomes sequenced each year is growing exponentially and the largest databases account for more than one million individuals. This wealth of data is extremely valuable for better understanding the genetic destiny of humanity, whether to determine the genetic weight in a particular disease or to better understand the history of human migration. To be meaningful, however, these big data must be processed electronically. "However, the processing power of computers remains relatively stable, unlike the ultra-fast growth of genomic Big Data," says Olivier Delaneau, SNSF professor in the Department of Computational Biology at UNIL Faculty of Biology and Medicine and at SIB, which led this work. "Our algorithm thus aims to optimize the processing of genetic data in order to absorb this amount of information and make it usable by scientists, despite the gap between its quantity and the comparatively limited power of computers."

Better understand the role of haplotypes

Genotyping makes it possible to know an individual's alleles, i.e. the

genetic variations received from his or her parents. However, without knowing the parental genome, we do not know which alleles are simultaneously transmitted to children, and in which combinations. "This information—haplotypes—is crucial if we really want to understand the genetic basis of human variation, explains Emmanouil Dermitzakis, a professor at the Department of Genetic Medicine and Development at UNIGE Faculty of Medicine and SIB, who co-supervised this work. This is true for both population genetics and in the perspective of precision medicine."

To determine the genetic risk of disease, for example, scientists assess whether a genetic variation is more or less present in individuals who have developed the disease in order to determine the role of this variation in the disease being studied. "By knowing the haplotypes, we conduct the same type of analysis, says Emmanouil Dermitzakis. However, we are moving from a single variant to a combination of many variants, which allows us to determine which allelic combinations on the same chromosome have the greatest impact on disease risk. It is much more accurate!"

The method developed by the researchers makes it possible to process an extremely large number of genomes, about 500,000 to 1,000,000 individuals, and to determine their haplotypes without knowing their ancestry or progeny, while using standard computing power. The SHAPEIT4 tool has been successfully tested on the 500,000 individual genomes present in the UK Biobank, a scientific database developed in the United Kingdom. "We have here a typical example of what Big Data is, says Olivier Delaneau. Such a large amount of data makes it possible to build very high-precision statistical models, as long as they can be interpreted without drowning in them."

An open source license for transparency

The researchers have decided to make their tool accessible to all under an open source MIT license: the entire code is available and can be modified at will, according to the needs of researchers. This decision was made mainly for the sake of transparency and reproducibility, as well as to stimulate researchers from all over the world. "But we only give access to the analysis tool, under no circumstances to a corpus of data," Olivier Delaneau explains. "It is then up to each individual to use it on the data he or she has."

This tool is much more efficient than older tools, as well as faster and cheaper. It also makes it possible to limit the digital environmental impact. The very powerful computers used to process Big Data are indeed very energy-intensive; reducing their use also helps to minimize their negative impact.

More information: Olivier Delaneau et al, Accurate, scalable and integrative haplotype estimation, *Nature Communications* (2019). DOI: 10.1038/s41467-019-13225-y

Provided by University of Geneva