In his Master's Thesis titled “From SNPs to Signals: Automatic Result Filtering and Novelty identification for Genome-Wide Association Studies”, Arto Lehisto, a student of Life Science Technologies at Aalto University, developed a tool for identifying significant results of genome-wide association studies (GWASs). The tool, which was produced for the needs of the FinnGen study, can be used to identify biologically interesting genomic locations from the wealth of information produced by the FinnGen study.

Used widely in the study of common diseases such as type II diabetes or cardiovascular diseases, GWASs produce a summary statistic of the significance of association between a trait, such as type II diabetes, and locations in the genome. However, due to the large amount of genomic locations that are measured, as well as a phenomenon called linkage disequilibrium, in which genomic locations correlate with each other, it is difficult and time-consuming to identify locations in the genome which could have a biological effect on the trait in question. Finding biologically important associations is important for developing new treatments for the traits, as they can imply new biological mechanisms behind the disease that could be targeted using novel drugs.

The FinnGen study, started in 2017, aims to combine the genome information and digital health care data of 500,000 Finns. This wealth of information can then be used to produce a large amount of GWASs, each showing possible links between a phenotype and the human genome. In the thesis, a software tool and a data processing pipeline were developed to automatically process the GWASs. The tool automatically filters the associations down to significant ones, and compares the found associations to a database of disease-genome associations found in earlier studies. This way, new links between the genome and a disease can be identified with much less manual labour. In his thesis, Lehisto assesses the performance of the pipeline using a set of GWAS results from the UK Biobank, a data source gathered using the genomic data and health records of 500,000 UK individuals. The data consisted of GWAS scans for multiple diseases, and the analysis was performed on Google’s cloud computing resources.

Using the dataset, the tool was found to successfully identify the regions of interest in the GWAS scans. The performance of the tool allows it to be run in reasonable time for the datasets that FinnGen will produce. The few problems identified during the run were found to be a product of using data from different sources, which will not be the case in the FinnGen study. Using a snapshot of the FinnGen data, the tool was successfully run for the entirety of the FinnGen GWAS scans, proving that it will scale up to the data sizes required by the study. The data pipeline will make it possible to run the tool on the dataset consistently in different phases of the project.

More information:
Arto Lehisto
arto.lehisto(at)gmail.com
Two-sentence summary for social media.

In his master’s thesis, Arto Lehisto created a tool for automatically filtering and identifying novel associations between diseases and genomic loci from genome-wide association study (GWAS) results. The tool can be used to identify biologically interesting genome-disease associations for the FinnGen study, e.g. for discovering new therapeutic targets.