A million European genomes

The Finnish Institute for Health and Welfare (THL), together with CSC, has simulated the genomes of one million Europeans. The data used for the simulation was publicly available full-genome sequences, but in the simulations they were formed into synthetic genomes, no longer representing real people. The simulation was done with CSC’s LUMI supercomputer. This was one of the most comprehensive human genome simulations in the world. The simulation was done for the EU’s 1+MG Initiative.

In 2018, the EU published the 1+Million Genomes (1+MG) Initiative, with an ambitious goal of collecting data covering the genomes of one million Europeans. The project was one of the biggest of its type in the world, with 27 countries participating. The secure use of European genomic data enables personalized healthcare and better diagnostics. This improves the treatment prognosis of especially cancers and neurological disorders.

The dataset is anonymized, so there is no unique and identifiable data. The goal is to create federated management that transcends national boundaries and gives access to national genomic archives.

"1+MG synthetic data project, from my viewpoint, the unique challenge was how we can create an effective simulation of a population which in its final generation includes a million people and corresponds in all its properties – in terms of genome, data formats and size – to genuine genomic data which, when simulated, is freely shareable without data security issues. In the end, we simulated a population of some 25 million people, of which just over a million was assigned synthetic genomes. This type of dataset enables a wide range of research, training and development project, such as 1+MG, without having to consider any ethical or legal issues or data security problems," says associate professor Tero Hiekkalinna of THL.

At this point, a simulation of the synthetic material of a million people was conducted, with dozens of phenotypes. This means that environmental effects on the phenotypes of individuals are included in the dataset.

The simulation of a million genomes was funded in Finland by the Ministry of Social Affairs and Health as well as the Ministry of Education and Culture. Hiekkalinna says that there were huge challenges in creating and managing the material.

"Owing to the size of the material, we needed dozens of terabytes of disk space."

The 1+MG Initiative was followed by B1MG (Beyond 1 Million Genomes) that ran from 2020 to January 2024. The B1MG project specified the guidelines and recommendations for the federated management of genomic data obtained from various European countries. CSC, the ELIXIR node in Finland, was one of the project’s managers and coordinators. The aim is to make the operation of biobanks compatible with cross-boundary data infrastructure. In the B1MG project, CSC was in charge of the technical infrastructure work.

The million-genome data simulated by THL and CSC will be made available in the Federated European Genome-phenome Ar-
chive (FEGA). The FEGA is designed for the storage and publishing of biomedical data for research purposes, but not to be available for the general public. The Finnish database is maintained by CSC. The FEGA is connected to the European Genome-phenome Archive (EGA). The EGA is one of the world’s most extensive data storage facilities.

The same simulated data will in future be available also for the Genomic Data Infrastructure (GDI) project. Started in 2022, the Genomic Data Infrastructure is coordinated by ELIXIR. The goal of the GDI is to create the final infrastructure, giving access to genomic data and clinical data collected of Europeans.

In future, Europeans can expect to have faster and more accurate diagnoses. Collected and analyzed genomic data enables better drug design and preventive drug treatment. All this will lead to better health and a higher life expectancy. This requires data preprocessing and harmonization, and secure, scalable and flexible technical solutions.

Five use cases, from cancer to rare diseases

These three related projects make use of five use cases. These use cases are relevant for the construction of the final GDI infrastructure. The Genome of Europe will create a collection of reference data for health programs utilizing genomics in European countries: each country submits genomic data in proportion to its population. A data model will be developed from clinical cancer information and metadata obtained from genomics. A polygenic risk score (PRS) is created for the purpose of deciding on the patient’s treatment: the individual risk score will take into account millions of genetic variations. In rare diseases, the key aspect lies in the presence of gene variants in different populations, and discovering the link between gene mutation and illness. The sharing of COVID-19 data collected by each country in Europe will also be tested.

Hiekkalmona, Tero; Heikkiläinen, Vilho; Perola, Markus; Terwilliger, Joseph (2023): Simulated European Genome-phenome Dataset of 1,000,000 Individuals for 1+Million Genomes Initiative.

1+MG
https://framework.onemilliongenomes.eu

Beyond 1 Million Genomes (B1MG)
https://b1mg-project.eu

The Finnish Institute for Health and Welfare (THL)
https://thl.fi/

CSC – Tieteen tietotekniikan keskus Oy
https://research.csc.fi/cloud-computing

Hiekkalmona, Tero; Heikkiläinen, Vilho; Perola, Markus; Terwilliger, Joseph (2023): Simulated European Genome-phenome Dataset of 1,000,000 Individuals for 1+Million Genomes Initiative.

1+MG
https://framework.onemilliongenomes.eu

Beyond 1 Million Genomes (B1MG)
https://b1mg-project.eu

The Finnish Institute for Health and Welfare (THL)
https://thl.fi/