The aim is to improve diagnostics and pharmacogenomics – in other words, to improve the impact of individual differences in hereditary factors on drug response. Another aim is to support the secondary use of these data for research. Valuable data will been collected from patients with cancers and rare and polygenic diseases. Data has also been collected on disease-causing pathogens as well as infectious diseases such as COVID-19.

This data can provide the basis for personalised drug treatments using multi-gene risk assessment. The risk is calculated using a personal polygenic risk score (PRS), which covers millions of genetic variations.

The three-year B1MG project ended in October 2023. The Finnish ELIXIR Node CSC led the technical infrastructure work in the project. The Genomic Data Infrastructure project, launched in 2022, is coordinated by ELIXIR, a cross-European life-science infrastructure for biological information. The aim of the GDI is to create the final infrastructure to provide access to genomic and clinical data collected from Europeans. GDI is a consortium of partners from 20 European countries. The B1MG project made recommendations to GDI regarding data and metadata management.

“B1MG was a coordination and support action grant, that was tasked with determining the roadmap and best practices for the deployment of the required infrastructure to support the 1+ Million Genomes ambition. As work package co-lead for the technical infrastructure CSC was able to drive forward the decisions on the roadmap ensuring that these aligned with existing and future CSC requirements, such as Sensitive Data Services,” says senior coordinator Dr Dylan Spalding from CSC.

The project made recommendations to GDI regarding data and metadata management.

“This data can provide the basis for personalised drug treatments using multi-gene risk assessment. The risk is calculated using a personal polygenic risk score (PRS), which covers millions of genetic variations.

The three-year B1MG project ended in October 2023. The Finnish ELIXIR Node CSC led the technical infrastructure work in the project. The Genomic Data Infrastructure project, launched in 2022, is coordinated by ELIXIR, a cross-European life-science infrastructure for biological information. The aim of the GDI is to create the final infrastructure to provide access to genomic and clinical data collected from Europeans. GDI is a consortium of partners from 20 European countries. The B1MG project made recommendations to GDI regarding data and metadata management.

“B1MG was a coordination and support action grant, that was tasked with determining the roadmap and best practices for the deployment of the required infrastructure to support the 1+ Million Genomes ambition. As work package co-lead for the technical infrastructure CSC was able to drive forward the decisions on the roadmap ensuring that these aligned with existing and future CSC requirements, such as Sensitive Data Services,” says senior coordinator Dr Dylan Spalding from CSC.

CSC has supported deploying federated sensitive data nodes. In this role CSC has used its experience in federated sensitive data services.”

Spalding worked for B1MG-project as a work package co-leader, which focused on the personalised medicine.

“The real benefit of B1MG is how it has set the direction for the GDI project, which will deploy a federated infrastructure across Europe to support cross-border access to over 1 million genomes. This has the potential to help democratise research, and drive personalised medicine across the EU.”

CSC as co-leads of the infrastructure pillar has a leading role in this work. Also the Life Science AAI (Authentication and Authorization Infrastructure) & REMS (Resource Entitlement Management System) are applications already in use to support access management to data. According to
Spalding this should align well with the existing Federated EGA node and Sensitive Data Services. The Federated EGA (European Genome-phenome Archive) is a distributed solution for sharing and exchange of human -omics data across national borders.

"GDI is very important for rare disease and personalised medicine, but also cancer, infectious disease, and common and complex diseases. However, the infrastructure isn’t specialised for any particular disease, but should support research into all disease types. The development is driven by the 1+ Million Genomes use cases, as well as Genome of Europe which is aiming to build reference cohorts of 500,000 citizens across Europe."

According to Spalding B1MG demonstrated a proof of concept version of the Starter Kit for both rare disease and cancer use cases. The Starter Kit is a set of software applications and components co-developed by the 20 GDI nodes.

Starter Kit
The Starter Kit has been created for the basis of GDI. There are five functionalities which were defined in B1MG that need to be supported – data reception, data discovery, data access management, storage and interfaces, and processing.

The Starter Kit includes more than 2,500 synthetic genomics and phenotypic data on cancer and rare diseases. It is a first step towards a production infrastructure.

"Starter kit contains all necessary functionality to deploy a demonstration system that allows the discovery, access, and analysis of sensitive genomic and phenotypic data. A set of synthetic data is included that can demonstrate these functionalities without risk of leaking real genomic or phenotypic data.”
An evolved version of Starter Kit will be integrated to the GDI portal.

**Personalised treatment through AI**

Spalding believes that the huge amount of data in the GDI project will enable a better-personalised treatment.

“GDI has the potential to support new machine learning and AI methods, speeding up the transition to personalised medicine across Europe.”

Professor **Arto Mannermaa**’s group from the University of Eastern Finland is developing learning algorithms based on genomic and clinical data to identify and predict risk factors for breast cancer. Genomic and clinical data are combined to form an artificial intelligence model that not only helps to determine the risk of illness, but also in drawing up individual treatment plans.

Mannermaa’s group creates AI models from image data. What other data should be combined with image data to improve healthcare?

“We have now incorporated genomic data into our imaging data. The more data modalities we can combine, the better we can identify factors related to successful cancer treatment, and the more likely we are to identify factors influencing disease risk.”

Factors influencing disease risk include data on treatment response or other clinically relevant information related to treatment.

“The more data there is, the more demanding the computing environment becomes. Ancillary data can be obtained from various sources, such as electronic health record systems through biobanks.”

21.2.2023 | **Ari Turunen**

---

**MORE INFORMATION:**

- **Genomic Data Infrastructure**
  https://gdi.onemilliongenomes.eu

- **Beyond One Million Genomes**
  https://b1mg-project.eu/1mg/genome-europe

- **University of Eastern Finland**
  https://www.uef.fi/en/

---

**CSC – IT Center for Science**

is a non-profit, state-owned company administered by the Ministry of Education and Culture. CSC maintains and develops the state-owned, centralised IT infrastructure.

http://www.csc.fi

https://research.csc.fi/cloud-computing

---

**ELIXIR**

builds infrastructure in support of the biological sector. It brings together the leading organisations of 21 European countries and the EMBL European Molecular Biology Laboratory to form a common infrastructure for biological information. CSC – IT Center for Science is the Finnish centre within this infrastructure.

http://www.elixir-finland.org

http://www.elixir-europe.org