Quick DNA analysis of patient samples with artificial intelligence

The human genome contains millions of genetic variants that make each individual unique. Some variants affect eye colour or blood type and others affect hereditary diseases. The DNA sequence may also include a pathogenic sequence variant that causes various disruptions in the function of the gene. The disruptions manifest themselves as hereditary diseases. Blueprint Genetics from Finland classifies genetic variants found in the genome from patient samples and analyses their connection to the described symptoms of the patients.

Blueprint Genetics started its operations focusing on the diagnostics of cardiovascular diseases. The company is now able to analyse majority of hereditary diseases based on the patient samples it receives. More than 6,000 disorders resulting from a defect in a single gene are known in humans. On average, one in two hundred will inherit a genetic defect from their parents. There are also many multi-factor disorders in which the combination of multiple genetic variants causes the disease or increases the risk of illness. These include, for example, Alzheimer's, diabetes, rheumatoid arthritis or cancer.

Jussi Paananen, Director of Data Science at Blueprint Genetics and researcher at the University of Eastern Finland, has a background in computer science with data science as his field of specialisation. Paananen became interested in biomedicine at an early stage because it utilises technologies that produce a lot of data. In recent years, he has been interested in machine learning and artificial intelligence, which are on their way to becoming research methods in bioinformatics thanks to increasing computing power.

"I am interested in how artificial intelligence can help geneticists in decision-making as well as processing large amounts of data."

Artificial intelligence helps identify variants

Research into artificial intelligence is on the rise and the methods are changing. In machine learning, the computer learns to arrive at a particular outcome independently. Machine learning algorithms find patterns that people are not able to detect from large data sets. Machine learning utilises neural network research, which has a long tradition in Finland. The neural network learns the non-linear dependen-

When part of a chromosome disappears, it is called deletion. In such cases, the chromosome often breaks at two different points, whereupon the part that broke away disappears. This results in some of the genes also disappearing, which causes developmental disorders. A view from the IGV (Integrative Genomics Viewer) software in which a geneticist is examining a deletion in the ORF15 region of the RPGR gene. ORF15 is one part of the RPGR gene. In practice, it is one exon that controls the protein production of the RPGR gene. Mutations in the RPGR gene cause two thirds of all cases of retinal degeneration linked to the X chromosome. The coloured bars shown are nucleotide sequences sequenced from a patient sample. The colour indicates the direction from which the DNA molecule has been read. A deletion of two nucleotides is visible in the middle of the sequences read from the patient sample.

cies of the variables directly from the observation data. It is able to classify the ears from animal-themed images, for example.

"Neural networks are at their best in solving classification problems", says Paananen.

"In image analysis, images or parts of images are identified and classified. A machine can identify objects and things: this is a human, this is a car, this is a cancerous tumour. What we do is classify DNA variants. From patient samples, we try to find which DNA variants cause diseases as well as which genetic variants are a part of our normal genome."

Genetic variant is identified by screening different sources

The customers of Blueprint Genetics are doctors treating patients. The doctors want to find out whether the illnesses of their patients are due to hereditary factors or not. Doctors from around the world send Blueprint Genetics their patients' blood or saliva samples, the isolated DNA of which is then sequenced. Sequencing generates a huge amount of data from which the interesting variants are drawn. In practice, this means that the patient's genetic variants are compared to the average human reference DNA.

Blueprint Genetics employs top professionals, geneticists and doctors who classify the variants. They go over the data mass that has already been processed and divided into smaller parts. The experts practically sieve through existing scientific literature and databases.

"We are trying to figure out which of these variants explains the disease or its symptoms."

Since similar information has been collected around the world, a single DNA variant that explains the disease can often be found in scientific articles and databases.

"We issue a clinical statement based on the results. The clinical statement is typically a few pages long document, that is delivered to the customer physician. The physician uses the statement as an aid in diagnosis and planning of treatment."

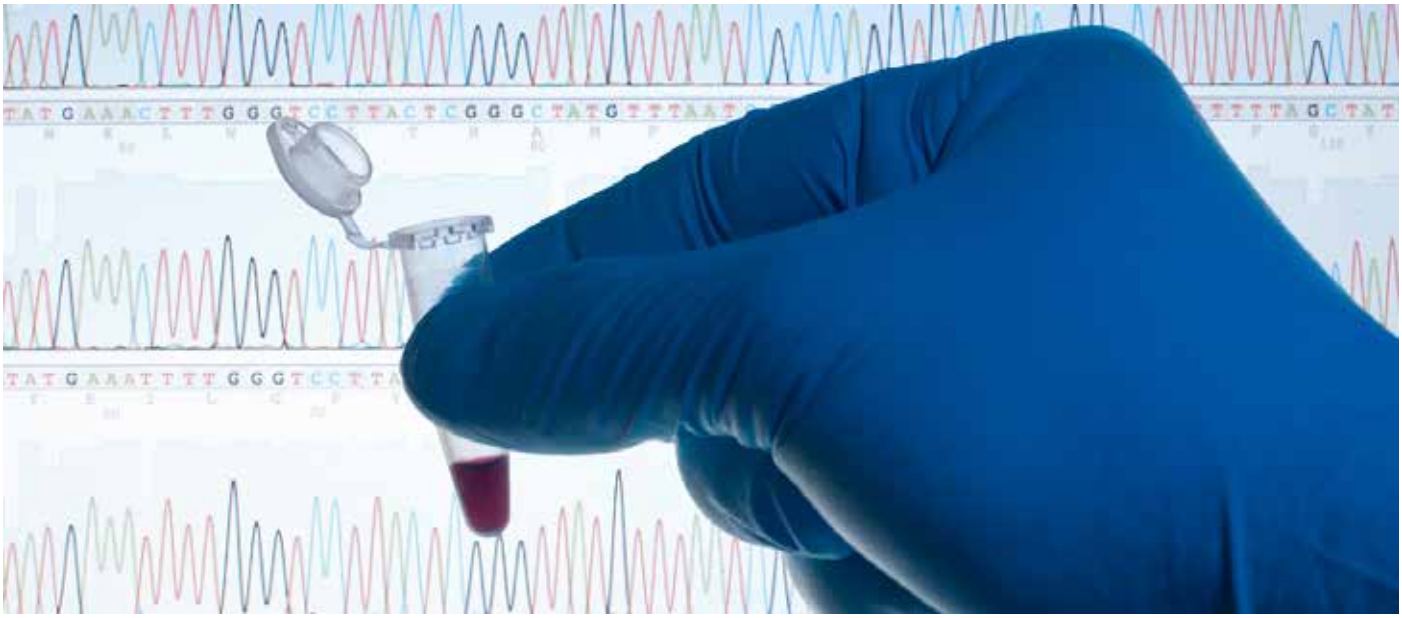
Blueprint Genetics utilises a variety of data sources. Where possible, the analysis of the data is automated. Software analyses the data and performs complex data processing. The field is under constant development. Software is updated several times a year, data volumes and computing power are increasing. Methods evolve and change rapidly.

"We have our own software production combining different data sources and facilitating literature searches. However, the final interpretation is always carried out by a geneticist."

Analysis and interpretation of patient data is demanding work because it involves a lot of legislation and regulation. Blueprint Genetics provides medical doctors with processed information, but the doctors always make the actual decision.

Blueprint Genetics is also interested in cooperation between the public and private sectors.

"The utilisation of genetic data is an enormous challenge that concerns the whole human race. The solution requires



Blueprint Genetics receives a blood or saliva sample, and the genetic variant caused by a possible disease is sought from the DNA obtained from the sample. The analysis takes about three weeks.

cooperation from companies, academic research groups as well as publicly funded organisations. Blueprint Genetics strives to contribute to the development of open science solutions and is constantly looking for new partners."

Databases listing genetic variants are important

Initially, Blueprint Genetics focused on certain interesting genes, or gene panels, based on the patient's symptoms. A panel typically includes about a hundred known genes associated with a particular disease. A team of geneticists sieves through the approx. 2,000 variants studied using the panel. The company has now shifted to exome sequencing, meaning that it sequences all protein-encoding genes, of which there are approx. 21,000 in our genome.

The human exome is the part of DNA with which all human proteins are produced. The part of the gene that encodes and directly guides protein production is called the exon. All the human exons in our genome together are called the exome. The human exome is approx. 1.5% of the entire genome.

"When our analysis focused on gene panels, we obtained, for example, 2,000 variants that a team of geneticists went through. Now, there may be 200,000 variants. As we advance to sequencing the entire genome, the number of variants will be 5 million. This amount of data cannot be sieved through manually."

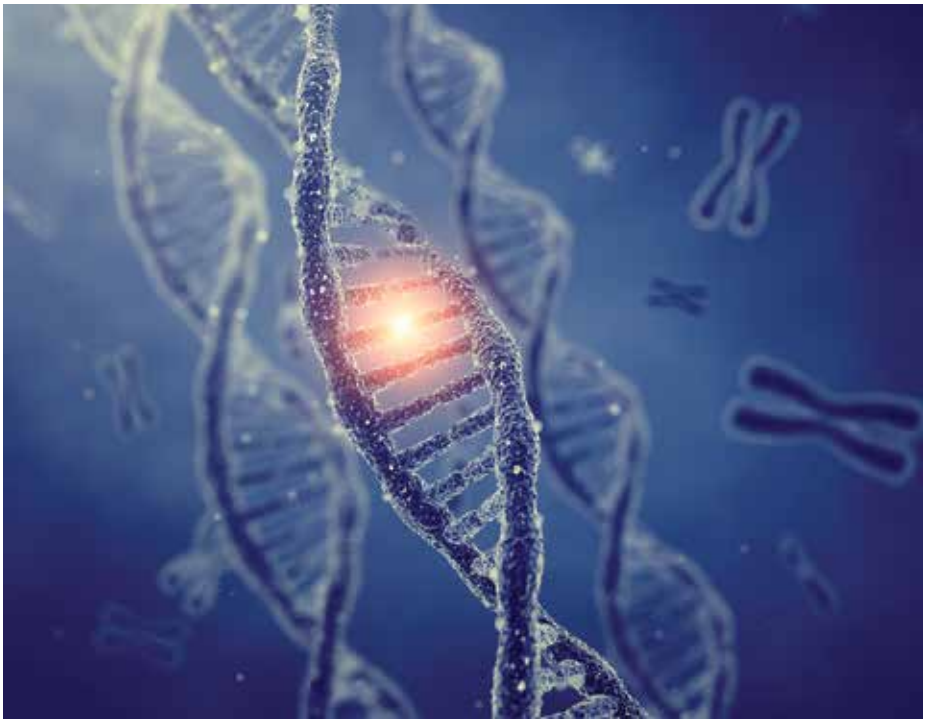
External databases are important in interpreting the data collected from patient samples. Genomic variants have been catalogued in various international databases, the most important of which are located in the organisations of EMBL-EBI in Europe and NCBI (The National Center for Biotechnology Information) in the US. In addition, ELIXIR coordinates the public biomedical infrastructure in Europe, enabling genetic variants to be mined from these international databases.

Variant databases provide useful lists that can be used to find correlations between genetic variants and phenotypic data. EMBL-EBI classifies, stores and distributes information on genetic variants. The most important databases include the European Genome-phenome Archive (EGA) where patient data from biomedical research is stored, the European Variation Archive (EVA) that includes genetic variants, Ensembl that provides interpretation for these variants, the gnomAD service for population-level variant occurrence data and the ClinVar archive for clinically significant variants. Therefore, the doctor often needs information from more than one service in order to produce the correct interpretation of the genomic variant for the patient. For this reason, European and American services regularly exchange information on the latest research results so that the services would always provide the latest information on our genome for research and medicine.

"Genetic variant databases are important because they have information on the prevalence of the variants in healthy people. This information can be utilised, for example, when it is known that only 1% of people have a certain rare hereditary disease. When we see that there is a variant that 5% of people have, it can be concluded that this cannot be the variant causing the disease. Thus, it is possible to filter out major, common DNA variants that cannot be associated with the rare disease."

Public sector data services offered by ELIXIR are important.

"We utilise our own local copies of different data sources. Physical distance and communications links require the sources to be in the same place. From public services, I would like to see more measures related to the versioning of databases. Old



versions should not be discarded. Long-term storage should be available for different versions."

Standardisation of metadata is challenging

A major challenge in both public research organisations and the private sector is the standardisation of the data used for interpretation. Data notations can vary greatly. The big challenge for Blueprint Genetics is the so-called phenotypic data.

"In one sense, it is metadata in itself, i.e. information accompanying a patient sample: symptoms, diagnosis and other background information. A sample may be accompanied by a lot of metadata or none at all."

The standardisation of phenotypic data has the same problem as patient data in health care, where the challenge is different notations.

"We obtain information from different countries that has been recorded in different ways. The background information varies."

Jussi Paananen thinks that firms like Blueprint Genetics find it difficult to utilise data produced and managed by publicly funded and research-focused organisations.

"Research organisations and joint infrastructures are interested in large population cohorts, in which case we are talking about a huge amount of data being collected and harmonised. We process information in different ways than cohorts which, for example, compile the information of tens of thousands of people living in the same geographical area. We, however, always deal with individuals."

Blueprint Genetics seeks to use internationally consistent classification, terminology and standards in its operations.

"We produce the DNA data ourselves and can decide what form it is in and which standards it conforms to. However, we utilise guidelines provided by others when interpreting the results."

The first attempt at such a standard was made a few years ago. The American College of Medical Genetics and Genomics (ACMG) has issued guidelines on how sequence variants should be classified.

ACMG has proposed the following common terminology for single-gene disorders: pathogenic, likely pathogenic, uncertain significance, likely benign and benign.

"We have our own modified version of ACMG's classification."

The challenge for companies like Blueprint Genetics is the ability to utilise data. There is a lot of information in peer-reviewed publications, and the aim is to develop good text mining tools in order to automate the screening of articles.

Ari Turunen

MORE INFORMATION:

https://blueprintgenetics.com
https://www.elixir-europe.org/platforms/data/core-data-resources
https://www.ebi.ac.uk/ena/
https://www.ebi.ac.uk/eva/
https://www.ensembl.org/
https://www.ncbi.nlm.nih.gov/clinvar/
http://gnomad.broadinstitute.org
https://www.ebi.ac.uk/dgva

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