

*eli*ir ***Finnish User Cases***



## Finnish User Cases

Ordered and secured. REMS is a data management software that provides security and only grants access to authorised material.

PAGE 14

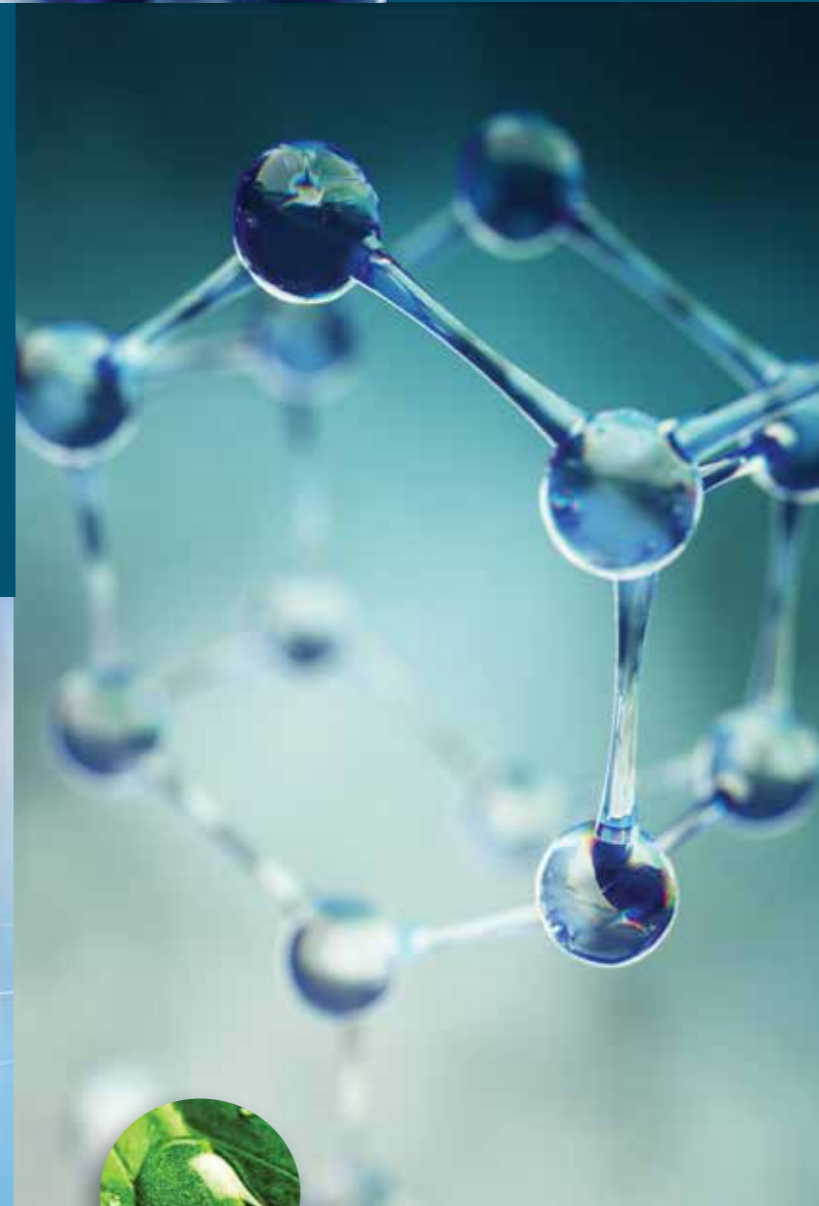


## ALGORITHM DETERMINES THE APPROPRIATE DRUG

The goal of Professor Mikko Niemi is to devise an interpretation algorithm that helps doctors determine the appropriate drug and correct dosage for a patient. Treatments become more effective and side effects are reduced, thereby decreasing the costs. PAGE 40

## Quick DNA analysis of patient samples with artificial intelligence

PAGE 56



## INDEX

- 4 Preface: Open data of life
- 6 Mapping the genomes of all organisms: new vaccines and medicines
- 10 "Smart life insurances" offered: human biological data is only useful when interpreted correctly
- 14 Ordered and secured
- 18 Disease prediction models are becoming more accurate
- 26 Looking for a good drug
- 30 Half of all drug ingredients affect only three protein families
- 36 Fighting cancer with mathematics
- 40 Algorithm determines the appropriate drug
- 44 Striving for a national service to utilise genomic data in health care
- 46 BBMRI.fi: an IT infrastructure for shared biobanks
- 48 Bank of million patient samples
- 52 Storing the whole genome of the Finnish population? The data will benefit disease research
- 56 Quick DNA analysis of patient samples with artificial intelligence
- 60 Webmicroscope stores tissue samples in the cloud
- 62 Secrets of the intestines
- 66 Microbes and climate change
- 70 Pups and pooches behind genetic discoveries in human diseases
- 74 Saimaa ringed seal aids the study of population genomes
- 78 Better harvests on the horizon? Data will also be collected in the future

### ELIXIR FINLAND

Tel. +358 9 457 2821  
e-mail: servicedesk@csc.fi

[www.elixir-finland.org](http://www.elixir-finland.org)

[www.elixir-europe.org/about-us/who-we-are/nodes/finland](http://www.elixir-europe.org/about-us/who-we-are/nodes/finland)



[www.elixir-europe.org](http://www.elixir-europe.org)

Realisation:  
Ari Turunen and  
Paula Winter  
Up-to-Point Ltd.  
2018

Extensive data sets and databases are increasingly being used in cancer research. The research group of Sampsa Hautaniemi develops methods that can be used to integrate data from various sources, such as DNA, gene expression and protein function. PAGE 36





# Open data of life



The size of a human cell is astronomical compared to the size of an atom; a single carbon atom in a cell measures one tenth of a nanometre. In drug design, atoms and their types define how a drug interacts with a human cell. The message chain conducted by interactions of the atoms in biological molecules affects the machinery inside a living cell. These messages are transmitted flawlessly right now in your cells and in your gut microbes.

Imagine that the atom is a human being. This human being could affect the functioning of a whole solar system and change events that take place dozens of millions of kilometres away.

It is possible to collect data about molecular biology anywhere within the reach of the Internet and availability of lab equipment, from urban jungle to the tundra. Labs can be miniaturised, and research has created massive amounts of data. Life science data accumulation has been a prerequisite for innovating new drug molecules or developing better diagnostics for diseases. Collecting this data is relatively easy. Reliable interpretation, however, requires an understanding of whole cells and biological systems. Understanding the cell is

similar to understanding a solar system at a one-metre resolution – with an increased complexity. The cell imagined to the size of a solar system does not have any vacuum, but it is filled with at least human-size interacting objects.

ELIXIR, European infrastructure for life science information, is committed to organising and sustaining life science data to facilitate its interpretation. ELIXIR allows, for example, secure processing of human data by using information technology innovations that prevent illegal use of data.

ELIXIR has 22 members, 21 countries and the EMBL – European Molecular Biology Laboratory. ELIXIR consists of nearly 200 organisations, which form a federation of trusted parties. In the beginning of 2017, with the help of the ELIXIR infrastructure, 21,000 scientific articles have been published and 8,500 patents have been granted. These include patents for vaccines, biomarkers, enzymes and prevention of the Ebola virus.

This book showcases Finnish user cases of ELIXIR. The ELIXIR node in Finland is operated by the CSC – IT Center for Science. In 2018, ELIXIR Finland supports more than 300 life science

research projects, many of them publicly funded.

### International collaboration is the key

The human genome has approximately 20,000 genes that guide all functions of the body. Sometimes, genetic information becomes corrupted, which can lead to, for example, breast cancer. International research has shown that there are exactly 93 genes in the human genome that may upon a change turn a healthy cell into a breast cancer cell. Detailed information on the gene is crucial when designing new treatments, because proteins produced by a cell from a mutated gene are targets for individualised healthcare treatments.

Denying access to this type of life science data would be wrong. Healthcare will need data services to facilitate understanding lab tests in the light of accumulated open knowledge, and so does research. Only open data services, with carefully planned data structure and reliable user policy, can lead to better interpretation of the life science data created by global research.

To meet data protection regulations without compromising openness,

ELIXIR operates the European Genome-Phenome Archive (EGA). The EGA services allow only identified and authorised individuals to analyse restricted life science data contents. The EGA and its technologies are used in the Nordic countries for publishing long time series data and research data regarding the entire population.

Storing and analysing genomic data at the scale of a million people and connecting this data with other resources – such as data on lifestyle and disease history – will transform our understanding of how diseases are born, how they progress, and most importantly, how they can be cured through individualised treatments. Bringing together massive life science data sets in Europe and making them reliably accessible for

both healthcare and research is a core mission of ELIXIR. The costs related to sustaining life science information are shared between ELIXIR members. Through its data resources, ELIXIR also supports research in countries that cannot create large data resources on their own.

The impact of life science data integration will first be seen, for instance, in the way rare diseases, cancer, and brain-related diseases are diagnosed and treated. The fruits of life science data integration will ripen by aligning national and international expertise into a long-term, standard-based infrastructure, which operates at an international scale.

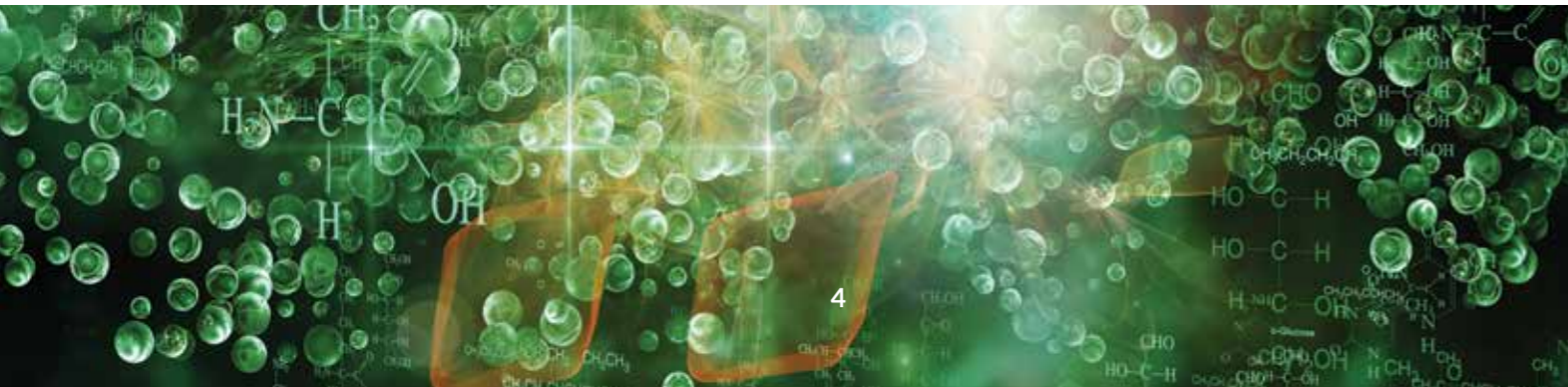
In the future, data will influence how society is organised. Enlightened


citizens will demand new kinds of health services. The private services sector in the field of data interpretation will inevitably grow. We need international data infrastructure resources as well as standards that the private sector, universities, and research groups can leverage to become part of the development. These data users need guarantees about the quality of data interpretations. This ecosystem enables a journey towards understanding the microcosmos of a cell, a cluster of cells, and eventually entire living organisms. Correlation between genetics, molecules, and other data of life has only begun.

**Tommi Nyrönen**  
Head of Node, ELIXIR Finland



“Storing and analysing genomic data will transform our understanding of how diseases are born, how they progress, and most importantly, how they can be cured through individualised treatments.”





# Mapping the genomes of all organisms: New vaccines and medicines

With the development of the methods used in bioinformatics, also the costs have lowered. It has become faster and cheaper to find out the genome of various organisms. However, we have a formidable task to be done to understand the information contained in the genome of various organisms and humans. It will require cooperation between various research organisations and well organised databases.

The mapping of the whole human genome was completed in 2003. Owing to the internet, the Human Genome Project was completed earlier than anticipated, since it enabled efficient cooperation between various laboratories. The entire human DNA was sequenced. The human genes have been packed into three billion base pairs. Now, the next step is to find out how these genes work. Through the analysis of the base pairs of the genome we will begin to understand the pathogenetic mechanism of various illnesses and effective forms of treatment.

Today, research is generating quite versatile genome data. The aim is, for example, to use the information to evaluate the status of the environment and effects on health by analysing microbes, to cultivate edible plants into plants that will better withstand draught to alleviate the crises caused by climate change, or to develop drugs against diseases for which there is no cure at the moment. To do this, new kind of linking and analysis of the sources of data will be needed.

## All the genomes of the known species will be mapped

It is becoming faster and cheaper to find out the genome of various organisms. Now, as part of the Earth Bio-Genome Project (EBP), the aim is to map the genome of all eukaryotic organisms. Eukaryotic archaea and eubacteria, i.e. prokaryotes,

are cells the DNA of which is constituted of only one chromosome. The group of eukaryotes consists of unicellular protozoans and three groups of multicellular organisms: plants, fungi and animals.

By means of bioinformatics, we can map the remaining 80 to 90 per cent of those organisms, whose genome still remains unknown. In 2011, Census of Marine Life estimated the number of animal species to be approximately 8,7 million, 6,5 million of which are terrestrial and 2,2 million are marine animals. According to the estimates based on high-performance sequencing methods, there may be as many as 5,1 million species of fungi. There are approximately 400,000 plant species.

For the first time in human history, we will have the opportunity to efficiently sequence the genome of all known eukaryotic organisms. EPB's aim is to sequence all of the known 1.5 million eukaryotes. Samples are being gathered all around the world. Part of them, probably around half a million, will be derived from botanical gardens. The rest will need to be directly collected from the nature. One of the most significant collection sites is the Amazon. In January 2018, EPB launched cooperation with a Brazilian gene bank project which concentrates on the organisms of the Amazon area.

The Amazon area has a richer variety of plant and animal species than any



Probably one third of all terrestrial species are found in the Amazon area.

where else in the world. Probably one third of all species are found there. Rain forests are the home of a huge potential of new drugs.

For example, ACE inhibitor, i.e. the angiotensin-converting enzyme, was discovered from the venom of the jacaraca viper in the Amazon. The enzyme generates angiotensin, which helps lower blood pressure and lighten the pumping of the heart. In the 1970s, researchers developed a synthetic version of the venom of this snake.

### Massive data archives

The oceans are the largest continuous ecosystem in the world. The signifi-

cance of planktons for the global climate is at least as important as that of the rain forests. However, only a fraction of those organisms which create this ecosystem, have been classified and analysed. The ecosystems constituted by planktons contain a vast amount of life: in every litre of ocean water there are more than 10 million organisms, containing viruses, prokaryotes, unicellular eukaryotes and cnidarians. These genuine organisms contain bioactive compounds, which can be used in the pharmaceutical industry, foods, cosmetics, bioenergy and nanotechnology. In 2009-2013, the researchers of Tara Oceans, an international expedition, collected 35,000 biological samples in 210 different measurement locations from oceans around the world. This is the largest plankton collection until this day. Another campaign in which samples were collected from the sea, was Ocean Sampling Day. In that campaign, research stations were asked to collect samples and to generate data. BioSamples collects descriptions and metadata from biological samples that have been used in research. The samples are references or have been used in various databases.

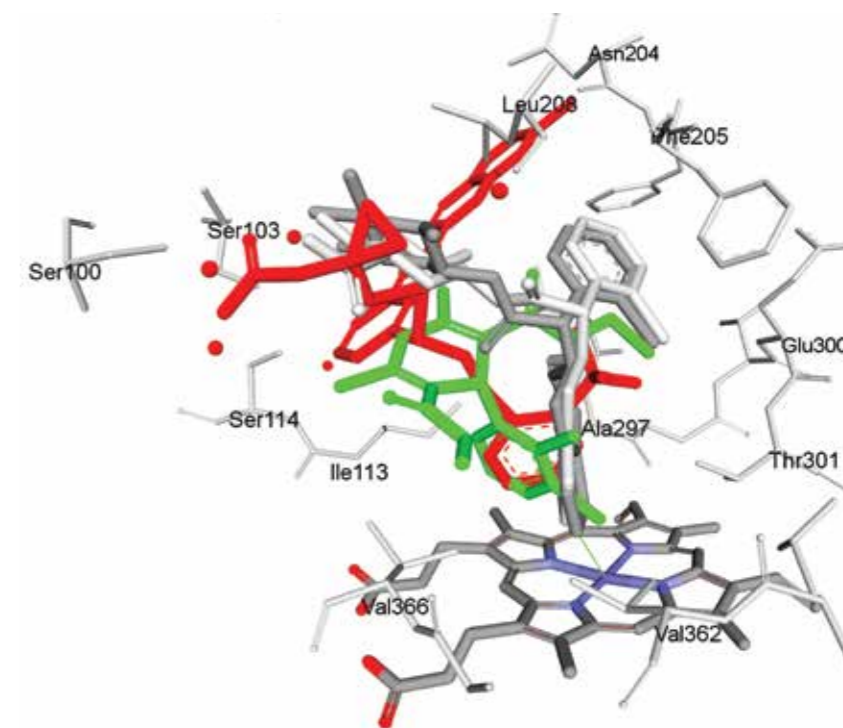
Analysing genomes and the proteins that determine their operation is a huge task, which would not be possible without cooperation. The European life science infrastructure for biological information ELIXIR provides an ef-

ficient platform for cooperation with members from nearly 200 research organisations, and an infrastructure which is used by almost half a million researchers. ELIXIR enables access to various data archives.

Massive sequencing of cultivated plants and forest vegetation allows us to do research on what is causing plant diseases. EURISCO (European Search Catalogue for Plant Genetic Resources) contains information on 1.9 million cultivated plants and their wild cousins. The samples have been collected by nearly 400 different organisations. A total of 43 countries are involved, and the aim is to preserve the agrobiological diversity of the world.

Uniprot (Universal Protein Resource) is collecting protein sequences and annotation data. An annotation means the determination of the functioning of the protein on the basis of the sequence. Owing to Uniprot's data, we can learn more about the functioning of proteins and their interaction with other molecules as well as their location in cells and organisms. The aim is to collect all publicly available protein sequence data. Uniprot is the largest publicly available protein sequence database.

The European Nucleotide Archive ENA is a collection which offers free access to all published nucleotide sequences and annotated DNA and RNA sequences. The International Nucleotide Sequence Database is a collabora-



tion forum between DNA Data Bank of Japan (Japan), GenBank (USA) and ENA. New data is synchronised between these three databases every day. Already in 2012, these databases contained the entire genomes of 5,682 organisms. The amount of data is doubled every ten months.

The European Genome Archive EGA is one of the largest public data storages in the world with patient data from biomedical projects. EGA stores the genotype and phenotype data collected from humans on the basis of a separate consent for research use of the sample and

**“To receive the best benefit from the data, genotype data should be linked to other health data.”**

the data. Thanks to EGA, many of the ELIXIR research projects have become possible.

### Biomedical data to the health records

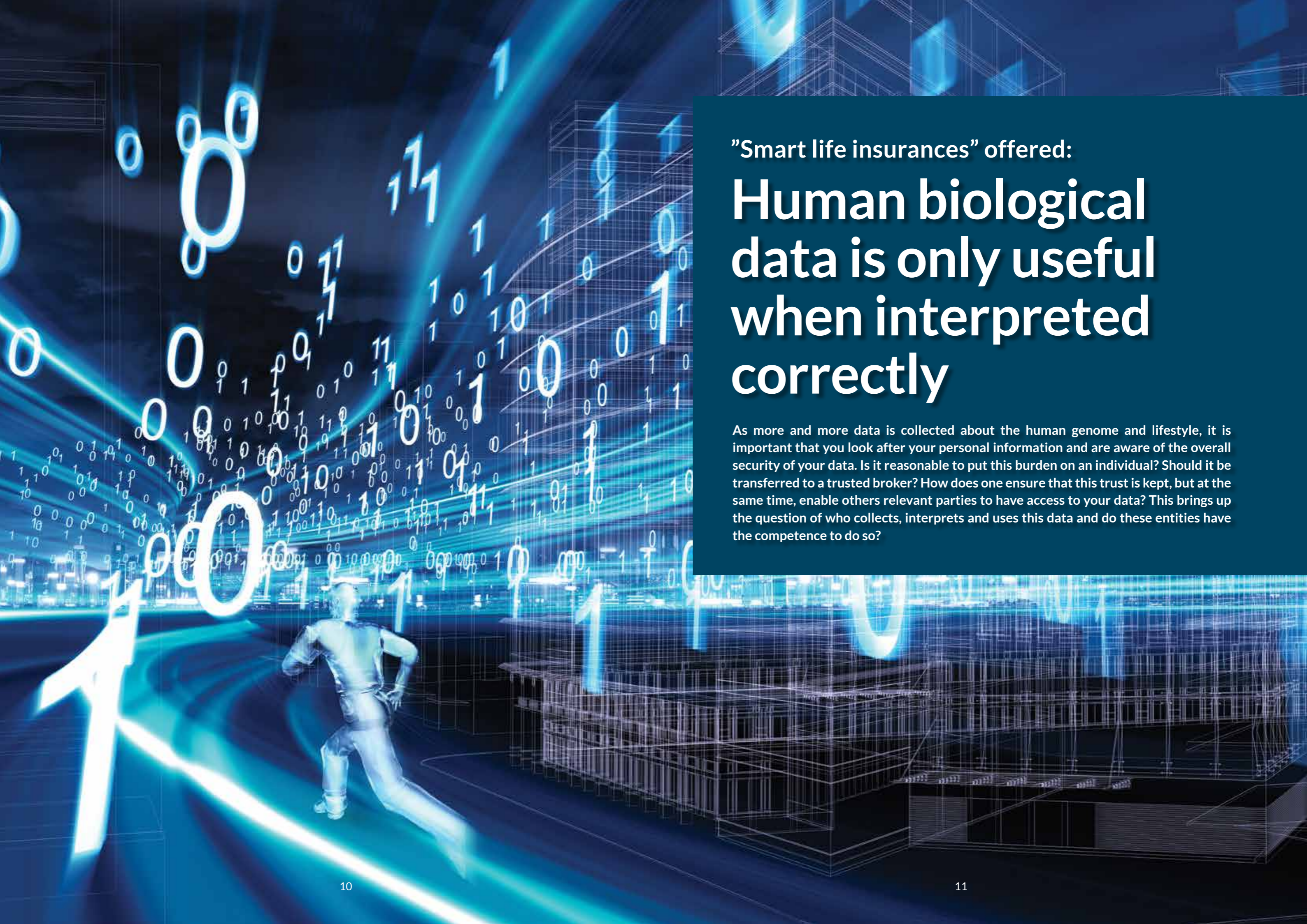
The ELIXIR infrastructure has more than 20 member states. Biomedical data is offered for use by researchers through the national centres in the member states. The benefits are indisputable. The genes of dogs and cats have proven useful in the analysis of rare human diseases. Through the Finnish centre the researchers have had access to a DNA bank of dogs and cats, the data of which has allowed us to discover the gene of a nerve degeneration disease, for example. The aim is now to develop a drug for the disease. Canine genes have proven useful in the research of human diseases, because the canine and human genomes are 95 per cent identical. The canine gene bank contains more than 70,000 samples from 60,000 dogs and 300 breeds of dogs. It is probably the largest of its kind in the world.



Researches developed an antihypertensive drug from the venom of jacaraca viper.



Ocean Sampling Day.



”Smart life insurances” offered:

## Human biological data is only useful when interpreted correctly

As more and more data is collected about the human genome and lifestyle, it is important that you look after your personal information and are aware of the overall security of your data. Is it reasonable to put this burden on an individual? Should it be transferred to a trusted broker? How does one ensure that this trust is kept, but at the same time, enable others relevant parties to have access to your data? This brings up the question of who collects, interprets and uses this data and do these entities have the competence to do so?



The exponential growth in biological information has an impact on both individuals and communities. It will become possible to predict a person's entire lifespan with certain genetic premises and lifestyle factors. As this information is increasing, also the possibilities to use this data for purposes other than what it was originally intended for will increase. Do we dare consume unhealthy foods anymore in the future if information is collected about it that may impact our insurance terms?

**The Finnish insurance company cooperates with Polar, who produces biomonitors and collects, heart rate and lifestyle data for the application.**

Economic and societal impacts will be felt after the next five to ten years when bioinformatics will be applied in preventive health care. For example, if a person has a genetic disposition to fall ill as a result of liver disease, which can

be treated by lifestyle changes, revealing the issue to them at an early stage will probably influence the person's lifestyle choices. Health care professionals can justify their recommendations by presenting well-known examples of long treatment histories from the health care system or biobank.

Open questions still remain: How and to what extent is modern biological information interpreted and used in public health care? How will/should the legislation evolve? As the need for better health care for the aging increases, so does the cost, and therefore such issues must be clarified quickly. The legislative aspects is particularly important as many insurance companies and giants of data processing, such as Google, are interested in the opportunities that are opening up.

#### Biomedical data is valuable

The American 23 & me provides genetic tests for anyone, which then provides information about hundreds of medical risks related to one's own genome. There are already many illnesses that can be analysed at the molecular level so it is therefore possible to diagnose, for example, one's propensity for can-

cer, which in turn can radicalise and tailor treatments to reduce side effects related to generic "heavy" treatments. It is envisaged that such new technologies will also be able to predict changes in the state of health of an individual.

Who can, who is allowed to, and who is able to participate in the continuous observation of one's own health? Who interprets whether a person is drifting towards a serious illness, and can this diagnosis be trusted? Whose rare disease can be cured and should it be done using public resources? Which ethical boundary conditions are used to coordinate access to the latest treatments?

Technology provides increasing opportunities for observing health and lifestyle on an individual level in real time. Different kinds of technological devices for monitoring our own health are becoming cheaper and integrated into devices that we already carry with us - mobile phones, clothes or watches. The Finnish insurance company Lähtapiola is conducting a new experiment where the company provides "Intellectual Life Insurance". The insurance company cooperates with Polar, who produces biomonitors and collects, for instance, heart rate and

lifestyle data for the application that help doctors make predictions about the person's state of health. It is possible for the client to lower their insurance fees if certain healthy lifestyle options are met in the data given to the insurance company. Individuals would therefore benefit from lower insurance payments that encourage a healthier lifestyle. To return the favor, the insurance company accepts data as a "currency" that it utilises.

This type of data is valuable. Reliable and well organised data sources that are used in interpreting an individual's health are currency in international commerce. In the UK, the National Health Service NHS has decided to open up the health care history of more than one million Londoners to Google. They are hoping that access to the data would enable Google's experts to help prevent kidney diseases that are the source of great costs in public health care. It is estimated that as many as every fourth of these cases could be prevented if the risks were detected earlier and the people would change their lifestyle. This would bring about considerable savings in the public sector and improve public health.

#### Who owns the data and its interpretations?

The data that people accumulate about themselves regarding their lifestyle, e.g. engagement in sports, food and alcohol consumption, currently ends up online in very different services, or is deleted within one year. The aim of the services collecting this data is usually to gain profit by "encouraging people to be engaged with their technological ecosystem". Connecting this type of accumulated data with third-party data sources is usually not possible. Further, using this type of data to support reliable diagnosis requires access to vast studies, so that an individual's data retrieved from the sample can be interpreted correctly. This kind of data integration is still in its early phases.

However, the development pace is fast. Examining data collected from dogs, for example, is legally less restrictive than data from humans, and there are many

services combining genetics and lifestyle for advancing dogs' health already available (MyDogDNA). The next great favour by man's best friend may be showing how genetic biological information should be used in health care.

Health care organisations collect data and samples from people in connection with treatments for research purposes. A medical professional is always responsible for the confidential collection of data and samples. Permission from the collector is requested if these are used for new purposes.

The prevailing practice significantly facilitates conducting studies to improve health. In the Nordic countries, centralised health care has been in use for decades, which has also been able to organise and provide high-quality data for research purposes. For example, 30 percent of Norwegian citizens have a sample in the biobank. In Finland, more than 150 million medical histories have been collected for archives from 4,3 million citizens.

There are 5,4 million people in Finland, and in 2016 nearly all medicine prescriptions go to the same archive. The Biobank Act that came into force

in Finland a few years ago also ensures that responsible research use of the data is allowed without informing every citizen about the issue separately. The collection provides an excellent starting point for interpreting the connections between genetic premises and factors that happen during a person's life, if safe and sufficiently open access to the data can be created for a large group of international, skilled analysts.

But what can we read in the data now and, above all, in the future? In the UK, Google has been given access to all patient data because it is not possible to know in advance which factors predict and explain the development of a kidney disease. But what if, when trying to predict this, it turns out that the person has an acute risk to have a heart attack? Should the person be informed about this? Nordic biobanks have studied that approximately 60 percent of people want to know about random discoveries. The rest 40 percent do not want to know. Who owns the data and samples collected from people, and who has the right to control it for example for research purposes?



# Ordered and secured

To start with, personal medical data is private and strictly protected. However, progress cannot be made in medicine without human data. The solution is a data management software program that provides security and only grants access to authorised material.

Data on the human genome should be treated with the utmost care and complying with information security protocols. In order to ensure information security, ELIXIR provides a service in which researchers log into a system that identifies their electronic identity while also distributing access rights to the biomedical data stored in the cloud. In this way, the researcher creates a secure analysis environment for the data to be analysed. This is made possible by the REMS tool.

“I have not heard of any other general resource entitlement software like REMS.”

ELIXIR strictly adheres to the EU law on information security. When researchers utilise data, the REMS tool can be used to ensure that the shared data is subject to authorisation.

CSC, the Finnish ELIXIR node, develops and maintains the open source REMS tool that can be used to manage access to datasets containing confidential material. REMS (Resource Entitle-

ment Management System) is an access management tool that, where necessary, prevents the illegal use of data. With the REMS tool, it is possible to order a specific file from a large amount of data and have it delivered to the ordering party locked in a secure manner.

“There may be various tools within an organisation handling similar things. Although there are many ready-made tools and services available for identity and role management, I have not heard of any other general resource entitlement software like REMS”, says the REMS tool’s product owner **Tommi Jalkanen** from CSC.

## ELIXIR AAI: a federation of 200 organisations

REMS is part of a federated system formed by the ELIXIR community comprising nearly 200 organisations. Becoming a federation has required agreements between the different organisations regarding information security, personal data law, rights and obligations. This has resulted in ELIXIR’s own trust network, ELIXIR AAI, the rules of which each member organisation has committed themselves to follow.

In practice, ELIXIR AAI is a community that uses federated authentication and identity management. This

federation (HAKA) has been developed based on the trust network of Finnish universities and research institutes. The ELIXIR federation enables Single Sign-On (SSO) to joint services.

ELIXIR’s member organisations maintain basic user information that shows the role of the user in addition to the name and contact details. Determining the role is important because the REMS tool distributes access rights based on it. That is to say, REMS decides what kind of a view opens for the user in the service on the basis of personal details. This is so-called entitlement-based REMS.

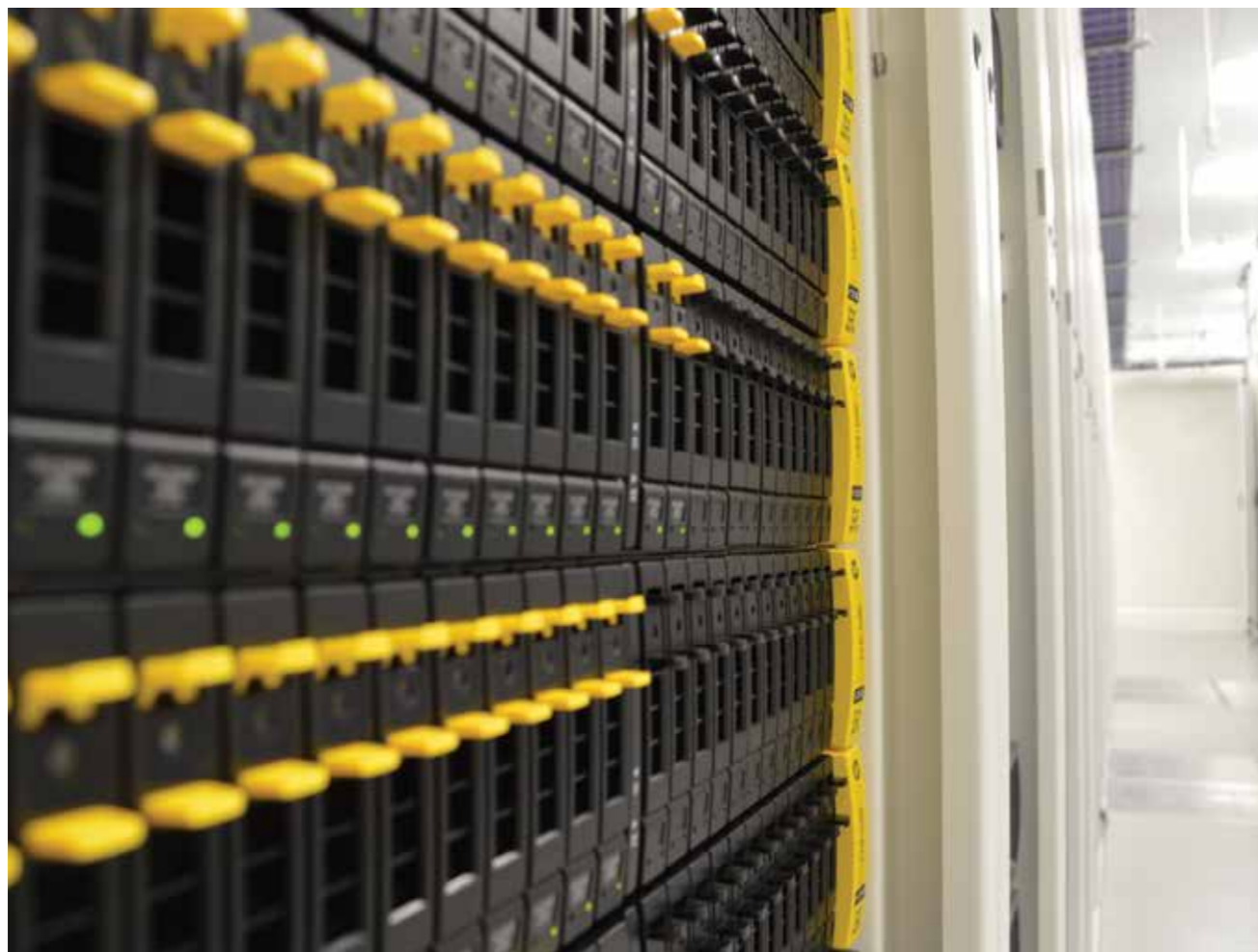
Despite the high level of information security, REMS is still easy to use. No separate sign-on is required to use the tool. Logging in to the service is done with the user name and password of the ELIXIR home organisation. So no service-specific user name/password pair is required. It is this federated management that ensures the use of data resources can be monitored. At the same time, it is possible to ensure that the materials are not used for wrongful purposes. The use of the service can be monitored and reported.

The way the service works is that a researcher applies for permission to

use the data with the REMS tool. The researcher logs in to REMS with their federated identity and then fills in an application for data use and agrees to comply with the terms of use. ELIXIR’s Data Access Committee (DAC) receives the application through REMS and approves or prohibits the use of the data. The applicant is notified of this by e-mail. If approval is granted, the applicant is provided with instructions on what happens next. REMS directs the data request to CSC’s Data Access Service. It provides the researcher with a view of the entitled data in the ePouta cloud service.

A federated user ID can be easily closed by the responsible organisation if the user switches workplaces, for example. The use of strong identification facilitates traceability and reporting. Fumbling with user name/password pairs is also reduced, as are password resets. Single sign-on reduces the need for separate user IDs and saves time, effort and money. Overlapping data management is reduced and data quality is improved. The service owner can focus on the service as the data administration of the ELIXIR organisation manages the IDs. These new practices support, for example, the use of ELIXIR’s many software services.





### Interface support for utilities

A new feature of the REMS software is a programming interface support for utility programs. A modern and widely-used web technology that enables the joint use of services, such as databases, is now available for researchers. This makes it possible to easily and safely build ecosystems and grant third-party access to the service. REST (Representational State Transfer) is a

well-known and frequently used application architecture for decentralised systems. The REST interface allows different software programs from different platforms to use the same resource.

“Creating an all-encompassing interface is currently in the works, providing extensive opportunities for the building of third-party utilities”, says Tommi Jalkanen.

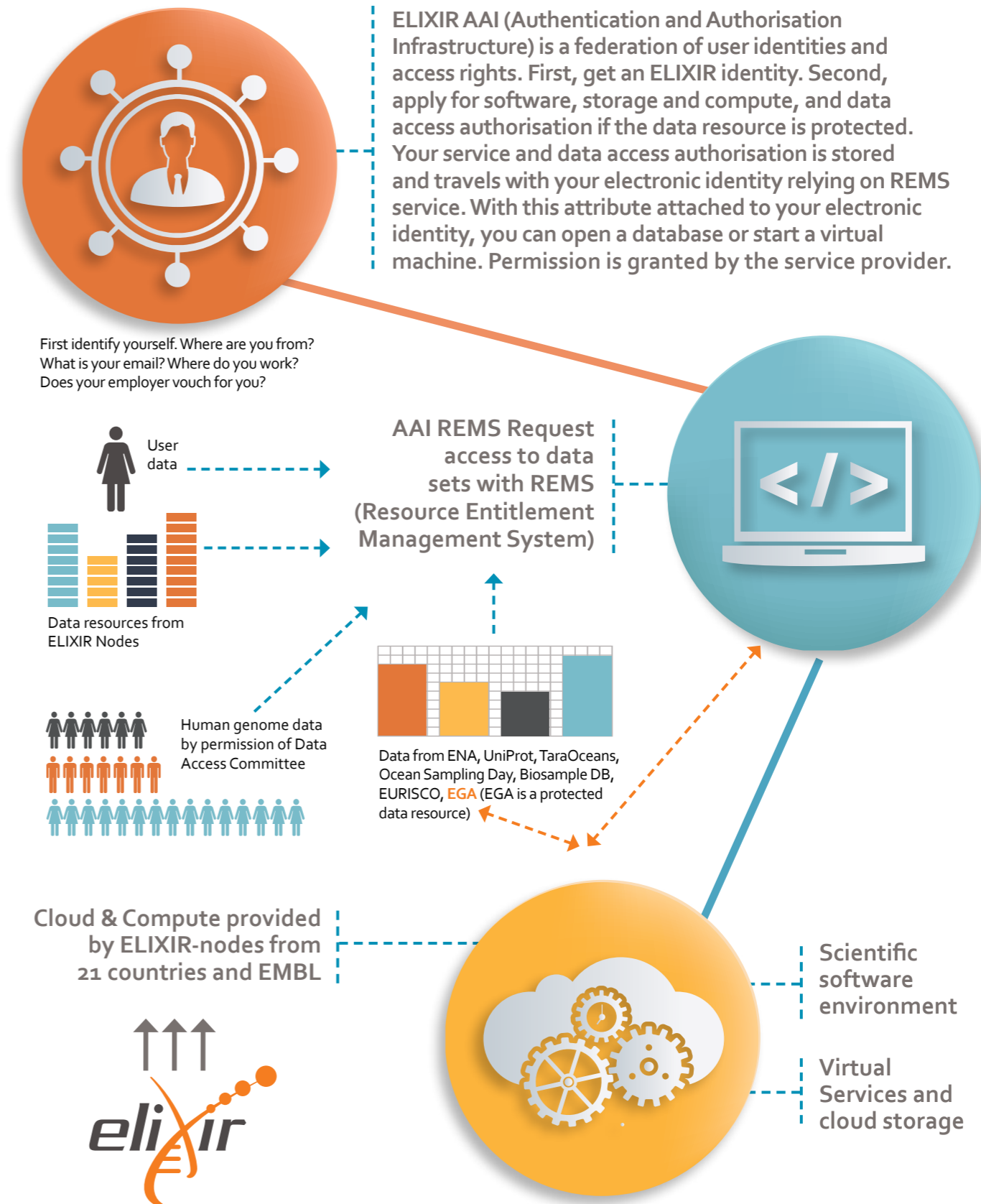
### Why monitoring access rights is important

Using statistical methods, it is possible to identify a person with sufficient probability from anonymised material if genomic information is available on the subject. Therefore, this issue must be approached through information security, the usage agreements of the service providing genomic data as well as national and international legislation.

If the anonymised material is linked with additional information, such as

**A modern and widely-used web technology that enables the joint use of services, such as databases, is now available for researchers.**

year of birth or the name of the disease, the researcher must be reliably authenticated in the service and accept the service’s terms of use, which prohibit the identification of the persons included in the materials. It is also possible to profile users, in which case each profile can be provided with an appropriate view of the material. The access rights and legislation define how the materials should be, for example, stored and analysed.

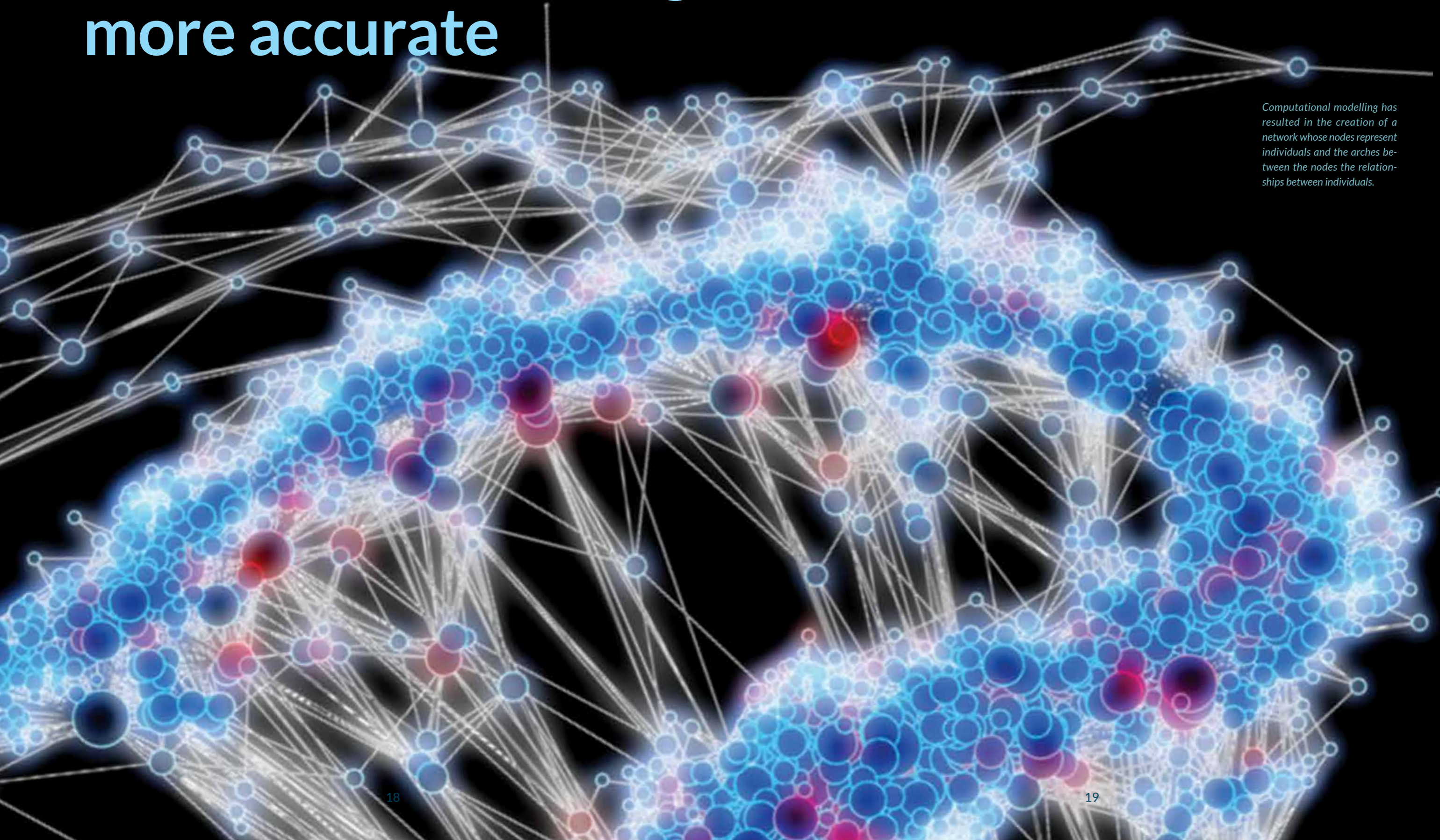


*The ELIXIR compute platform provides a seamless workflow for users: the researchers may use their electronic identity to securely create a scientific software analysis environment, and gain access to large sensitive biological data resources stored on a cloud. The platform also helps research groups to create scalable services.*

# Disease prediction models are becoming more accurate

Computational methods can now be used to deduce from data sets as to who is at risk of developing, for example, diabetes or cancer. Laura Elo and her research group develop methods which are used to find different predictive markers for diseases. Combining clinical data with molecular data can also provide valuable information about suitable drug treatment.

*Computational modelling has resulted in the creation of a network whose nodes represent individuals and the arches between the nodes the relationships between individuals.*



**R**esearch conducted on human biology produces a lot of new data for researchers to study. DNA sequencing generates an individual's genetic profile. RNA sequencing, in turn, provides measurement data on the activity of genes. It tells which genes are expressed and possibly produce proteins in the cells at any given time.

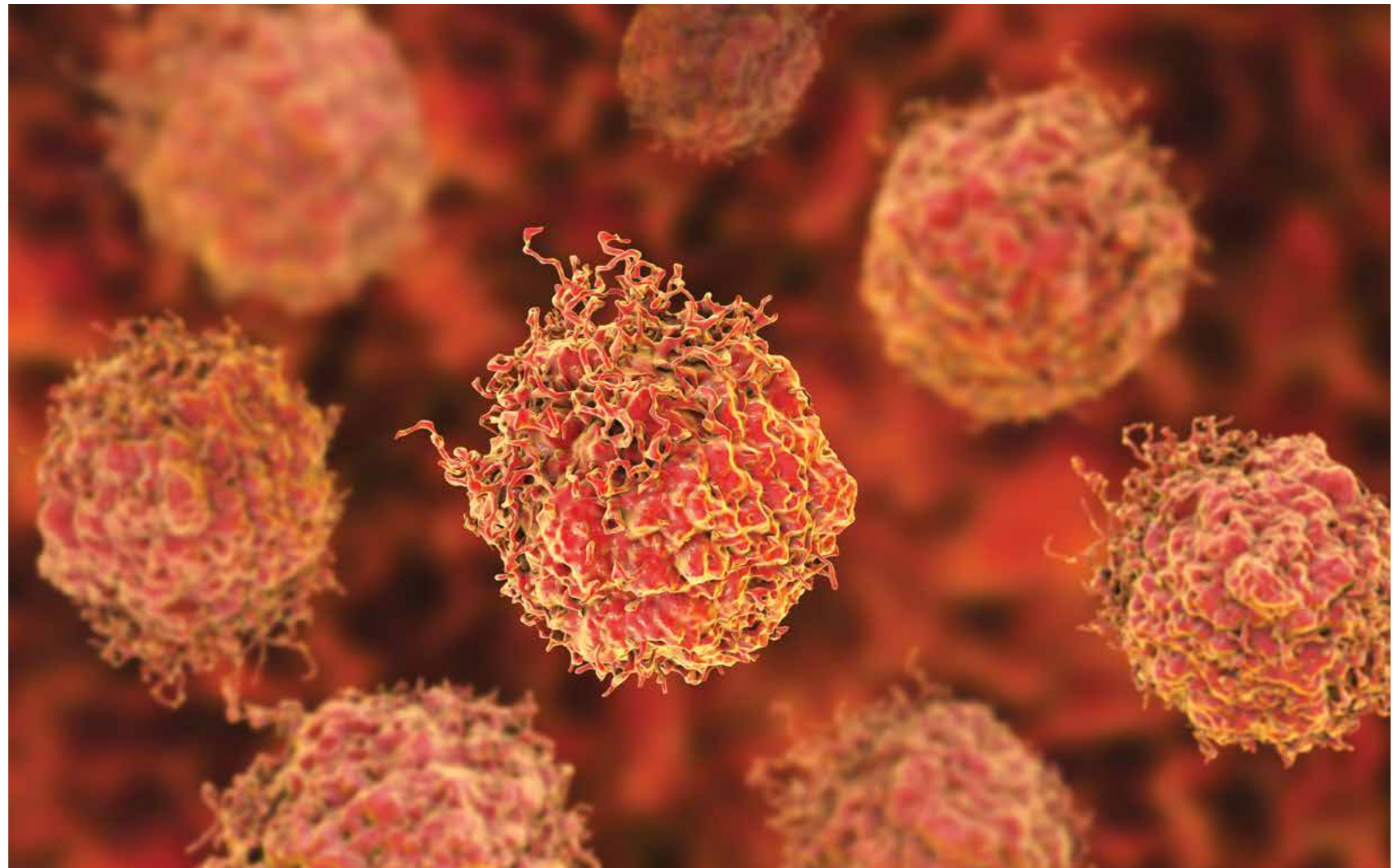
Thousands of different molecules and their interactions can be measured from a tissue sample. For example, it is possible to study different active forms, or transcripts, of a gene. When the goal is to determine the function of proteins or their deviations in connection with diseases, it is called proteomics. Mass spectrometers are used as aids to measure molecular mass.

**Laura Elo**, Research Director of Bioinformatics at the Turku Centre for Biotechnology, and her research group develop modelling methods that allow the measurement data collected in follow-up studies to be utilised to determine disease risk on an individual basis.

"I started my career as a mathematician at a time when bioinformatics was still a marginal field. I became excited about biology and medicine back then", Elo says.

One important material for the researchers is the data collected from different populations. The studies use data stored in Auria Biobank in the Turku region as well as data obtained from elsewhere in Finland and from other countries. The electronic medical records also have a lot of data collected from patient care that can be used in research, subject to consent. However, the data alone is not enough to determine the emergence and development of diseases. Computational methods and models are required to make comprehensible interpretations from data masses. The aim is to develop functional models for use by doctors.

"Almost all of our research is related to medicine and the needs of doctors. One of our major goals is to provide practical tools for doctors. The data alone is not useful, unless it can be modelled and interpreted. In the future, our work will hopefully al-



low patients to be offered treatments that are increasingly individually targeted."

Effective treatment is always personal because drugs and treatment methods work in different ways for different individuals. A patient's treatment response is affected by a number of factors, the information about which is obtained, for example, through laborato-

ry measurements. In addition to clinical variables related to the patient's health, there are many factors at the gene and protein level, which affect the efficacy of treatment methods. Mathematics helps with the analysis of data obtained about an individual.

"Biology is complex. One disease can actually present itself in many different ways at the molecular level, and differ-

ent treatments can be effective for different people. A specific drug can cause serious adverse effects for some while being ineffective for others. Computational methods allow us to predict, who will suffer from the adverse effects and who are likely to benefit from the treatment. We mathematicians can help medical scientists to identify the key predictive factors", Elo says.

*Prostate cancer that spreads metastases and is resistant to hormonal treatment is a malignant disease leading to the patient's death. The cytostatic drug docetaxel was introduced over a decade ago. However, approx. 10–20% of patients have side effects that force them to stop the treatment. International research groups created mathematical models that predict the side effects of cytostatic prostate cancer treatment for the Prostate Cancer DREAM 9.5 Challenge. The researchers developed a total of 61 models for the challenge, seven of which turned out to work and were awarded in the competition. A model developed by the joint research group of the University of Turku and the Turku University Hospital was one of the winning models. More information: Journal of Clinical Oncology Clinical Cancer Informatics: <http://ascopubs.org/doi/abs/10.1200/CCI.17.00018>*

### The model must also be suitable for new data

Development of mathematical models requires large volumes of data as their raw material. For example, some of the predictive models have been developed using clinical patient data from the US, but they are also suitable for the patient data of the Turku University Hospital.

“When a sufficiently large amount of genomic and clinical data is obtained, they can be combined and the modelling phase can start. This is only possible if the description of the data, metadata, is in order.”

Many things have to be taken into account in the development of models. It is important to assess the prediction ability of the model in advance. Models easily become overfitted for the data that is used to create them. This means that the model is too well-suited for the data. Therefore, the predictive model works with one data, but the prediction is no longer good with new data. Validation is required to verify the model. This can be achieved, for example, by using patient cohorts from another hospital or country. Checking the model by using other patient data is important to allow for the general adoption of the model. Data from different biobanks helps with this.

“If the model is built and tested using the same data, you may get it to work almost perfectly in that data. However, it may not work on new individuals. Therefore, we strive to build models which predict the outcome as closely as possible but can still be generalised to new data.”

The work of Laura Elo and her research group with modelling involves continuous experimentation and change.

“After developing a model and showing that it works with certain data, the validation process is continued. We aim to find as many new data sets as possible to test the accuracy of the predictions produced by the model. You can always develop a model that works in one data. However, it is only after it has been verified in several data sets that the predictive model can be considered reliable enough to be given to doctors to support de-

cision-making. The more widely the model can be tested, the better we can assess whether it only works for a specific population or if it is more universally applicable.”

New factors are added to models and their effect on predictions is analysed. For example, linear, simplifying models are easy to understand and interpret in hospitals. However, sometimes the interactions between molecules are so complex that linear models do not work and, therefore, other solutions are needed.

“The more new variables are added to a model, the more critical its validation becomes. An important question is understanding which variables are most significant for prediction and how their combinations provide the best predictions. You need to find balance for the model: it must be complex enough for prediction, but the model must not be overfitted to the data.”

### Predictive model for renal cell carcinoma

Laura Elo and her team have been involved in the development of predictive models for renal cell carcinoma. Renal cell carcinoma originates in the epithelial cells of the renal cortex. The prognosis of renal cell carcinoma is poor as 40% of patients die within five years.

A new computational method can be used to find predictive markers from patient samples. The study found that the expression of 152 genes can predict the life expectancy of patients with renal cell carcinoma after surgery.

“The prognosis of renal cell carcinoma is usually good if the cancer is localized. On average, however, 50% of patients develop metastases after surgery. The goal is to predict as early as possible whether the patient’s prognosis is good or bad in order to select the best treatment strategy.”

Two different sets of data were utilised in the development of the predictive model. The gene expression data of more than 400 renal cell carcinoma patients were obtained from the international Cancer Genome Atlas (TCGA) database. The model was then validated using an independent Japanese data set of 100 patients.

### Identifying the underlying mechanisms of type 1 diabetes at the cellular level

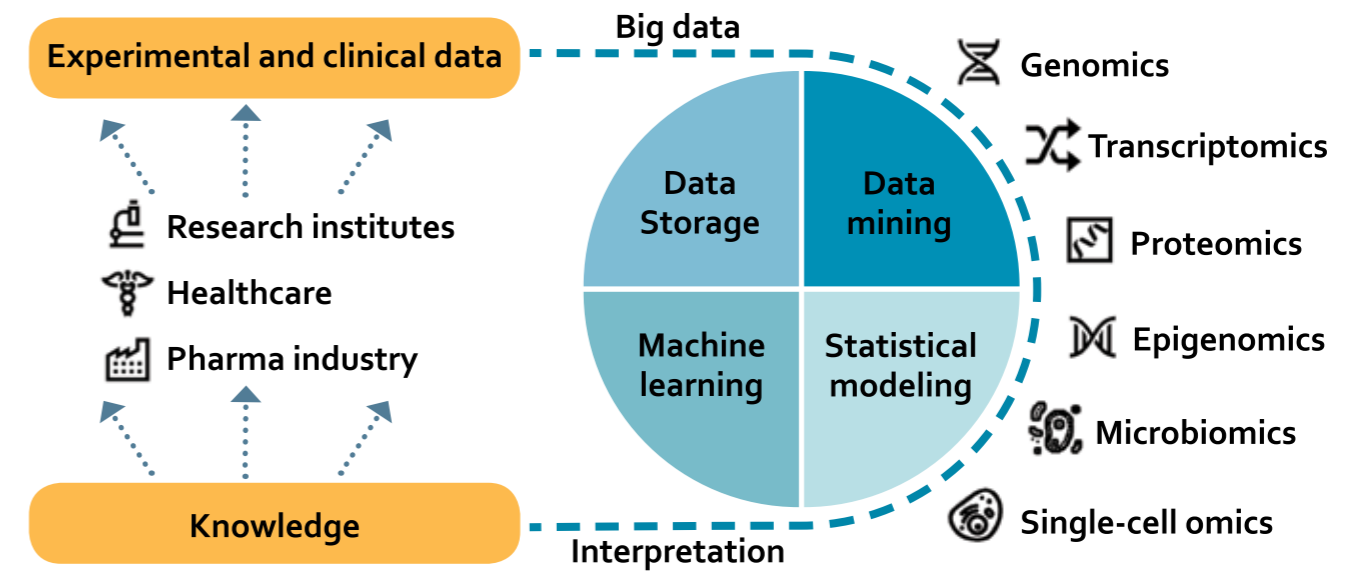
Laura Elo studies patient data to search for different biomarkers that can predict disease onset or treatment responses. A biomarker is a factor or characteristic that indicates a change in biological status, for example, in genes or proteins. In Finland, researchers have aimed at determining the underlying mechanisms of type 1 diabetes for a long time. Type 1 diabetes is caused by the destruction of insulin-producing cells. The pancreas does not produce the insulin hormone needed by the body, thereby causing blood sugar to rise.

“Finland has the highest incidence of type 1 diabetes in the world relative to the size of the population. Both genetic and environmental factors play a role in the development of the disease. We look for biomarkers that could predict the development of the disease as early as possible.”

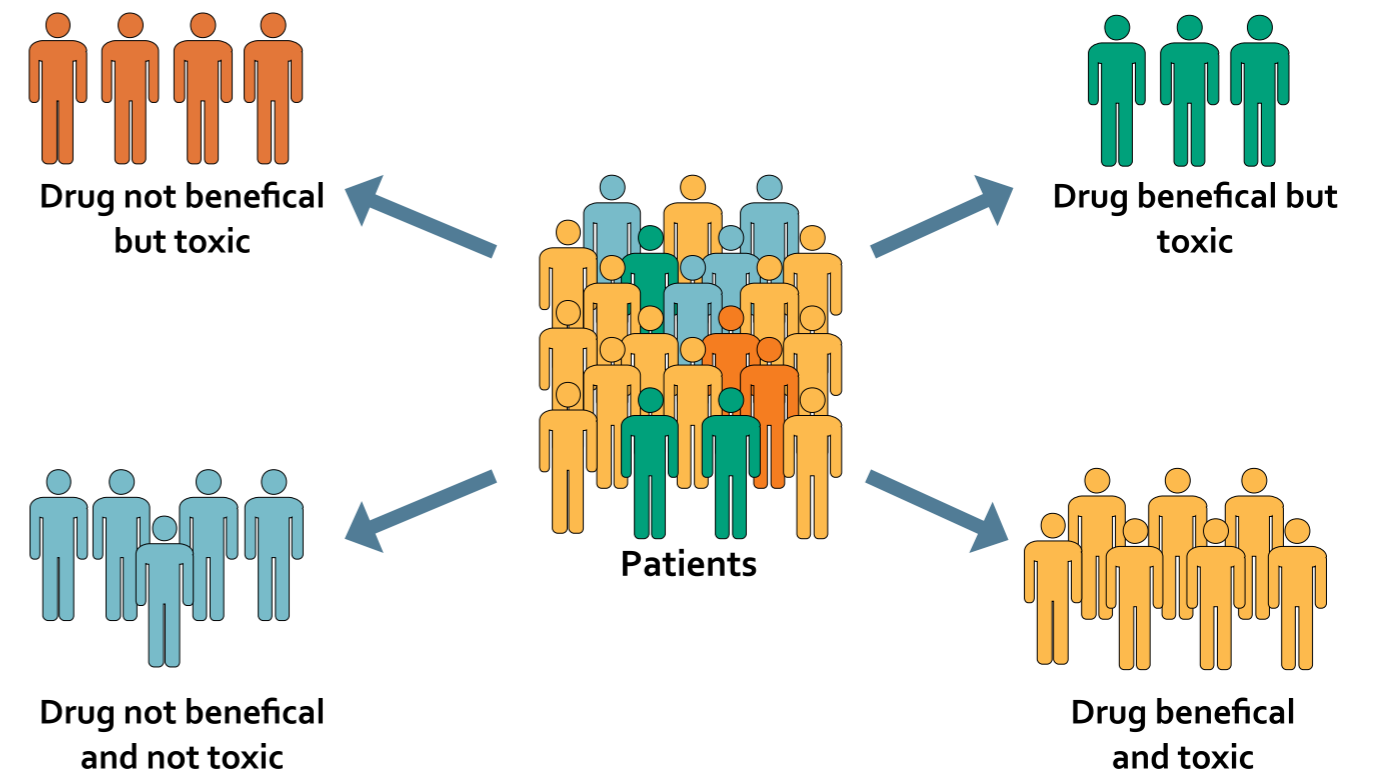
Because Finland has the highest levels of type 1 diabetes relative to the population in the world, diabetes research here is also significant. As early as in 1994, the ambitious and extensive research project DIPP (Diabetes Prediction and Prevention) was started in Finland. Genes that predispose you to type 1 diabetes are being sought in the blood samples of newborns. Children who are found to have a genetic risk to develop diabetes are invited to a follow-up study. Samples are taken every three months and, from age 2 onwards, every six or twelve months. The screening participants include the university hospitals in Turku, Tampere and Oulu.

“The children with a genetic risk of developing type 1 diabetes have been monitored until the age of 15. The goal is to identify the factors affecting the onset of the disease at the cellular level even before it can be diagnosed with the current methods.”

Laura Elo collaborates with Professor Riitta Lahesmaa, whose research group studies leucocytes and aims to understand what factors make cells cause diabetes. In the future, this could lead to preventing the onset of diabetes and curing current patients.



The Medical Bioinformatics Centre develops computational data analysis tools and mathematical modelling methods for the needs of biomedical research. Special focus is put on the analysis and interpretation of the extensive measurement data produced by modern biotechnology (e.g. deep sequencing and mass spectrometry). The goal is to improve the diagnostics, prognoses and treatment of complex diseases, such as diabetes and cancer, in close cooperation with doctors and medical researchers.



The aim of personalised medicine is to identify factors that can be used to find the most suitable treatment strategy for each individual.

## New tools

Going forward, Laura Elo wants to focus on the underlying mechanisms of diseases and the risk factors for falling ill. The statistical modelling of the complex interactions between different factors requires many new meth-

**“The more new variables are added to a model, the more critical its validation becomes.”**

ods and measurement technologies developed and tested by researchers.

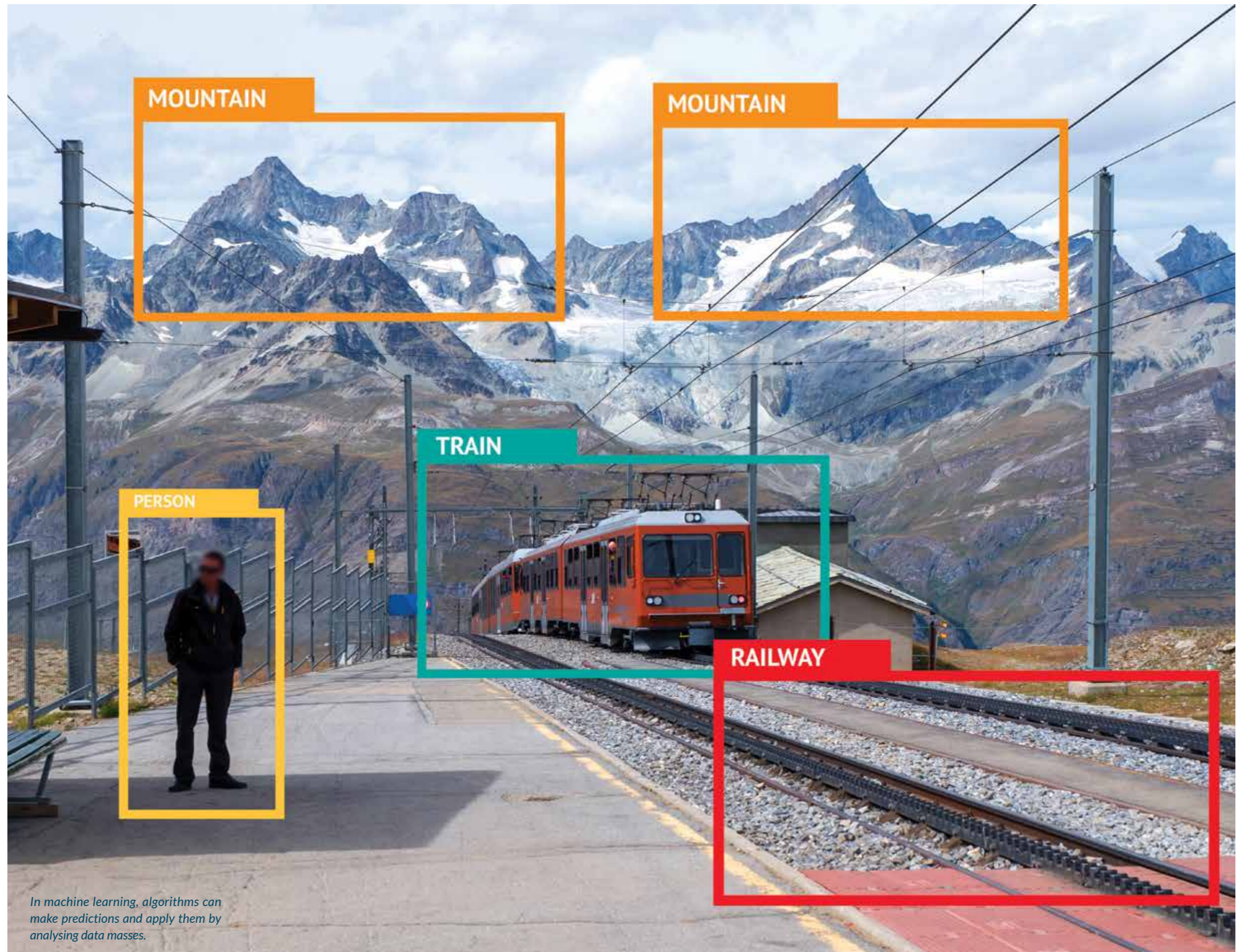
In addition to statistical modelling, Elo and her team applies different machine learning techniques to create predictive models. The machine is taught to learn the essential factors from the data. For example, the machine can learn to provide binary predictions of the consequences of treating an illness with medication: good response/bad response.

“New tools and methods must be brought as close to the patient as possible. We are constantly thinking what is required so that the model can be used in treating patients. What should be measured and how? Is there anything that could be done better? The model must be sufficiently simple and easy to use in order to end up at a clinic to be used by a doctor in their everyday work. It is important to know how doctors use them.”

“The essential thing about this work is that it is interdisciplinary. Just how much more information can be obtained using computational methods than sieving through the data only manually. Computation has become part of medicine.”

The Turku Centre for Biotechnology has its own computer cluster whose computing capacity is supplemented by a connection to the ePouta cloud service of the Finnish ELIXIR node CSC.

“The computing capacity and tools provided by ELIXIR facilitate the utilisation of data produced by other organisations. Utilising European data is important, but the data should be standardised. Making data compatible is a job for a large infrastructure.”



*In machine learning, algorithms can make predictions and apply them by analysing data masses.*

# Looking for a good drug

A good drug molecule will not be created unless it is known which proteins it affects in our body. That is why, in drug design, it is important to utilise massive databases with all the discovered protein structures and protein families as well as knowledge about how proteins function in our cells.

The majority of the drugs in use are designed so that their target molecules are the body's biomolecules, i.e. proteins.

Most drugs take effect in the body by binding themselves to these targets like receptors of signal molecules. Receptors are natural targets for example for signal molecules such as neurotransmitters and hormones. They are specialised triggers of the cell associated with cellular signalling mechanisms.

**Approximately 2,500 drug molecules are available for medicine.**

The idea for drug design is to build small synthetic molecules that selectively affect the desired proteins. Most of the target proteins of drugs belong to only ten protein families, and up to half belong to only three families. Small mol-

ecules are able to absorb well into the bloodstream, allowing the drug to take effect. Depending on the location of the protein, the drug molecule has to penetrate the cell or transmit a signal outside the cell that affects the processes within the cell. The aim is to design the molecules, for example, in such a way that they slow down or accelerate the functioning of a particular protein.

In the past, little was known about in which part of the cell the drug takes

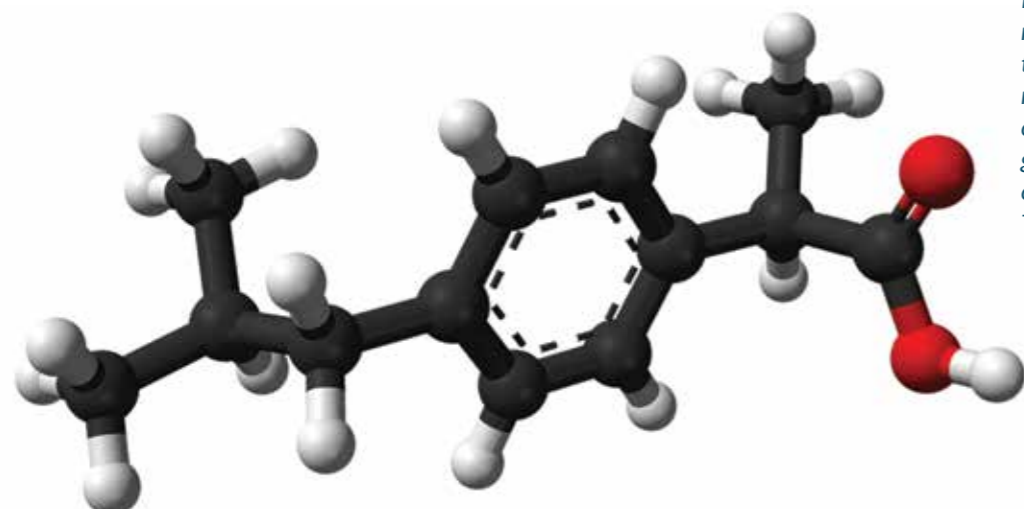
effect. In 1980, 150 of these target areas of effect were known. However, that number has grown enormously with the determination of the genomes of organisms since, currently, already more than 5,000 possible target areas of effect are known. Approximately 2,500 drug molecules are available for medicine. The function of the human genome is being investigated more and more closely and, in the next few years, the number of known possible

target areas of effect for drug ingredients may rise to 10,000. According to current estimates, our body has 2,000-3,000 proteins that are possible target proteins for a drug. Existing drugs have been shown to work through only about 450 drug targets on a limited number of diseases. Thus, drug designers have two major goals - to build new safe molecules that can be used to safely affect known targets and, on the other hand, to study the use of known, safe

drugs for new illnesses for which there is currently no drug approved by the authorities.

The goal of researchers is, among other things, to understand which structural and chemical characteristics of a drug molecule play a key role as they modify the function of proteins at the cellular level.

An effective drug can be developed once a three-dimensional structure of the target protein, which allows inter-



*Ibuprofen, used in, for example, many painkillers, inhibits the function of the cyclooxygenase enzyme, reducing the production of chemicals and hormones called prostaglandins which are needed in the communication of pain receptors. This reduces the sensation of pain.*

action with the drug molecule, is found. Chemical counterparts that recognise the amino acids at the protein's binding site are built into the drug molecule. When this kind of molecule encounters the target protein in the body, it automatically finds its way to the binding site of the protein because attaching itself there is energetically advantageous for it.

The binding of a well-designed drug molecule to the target protein could be compared with putting on a wool glove. It fits firmly on the hand with precisely five fingers: it would be very uncomfortable for one with six or seven fingers. In addition, a left-hand glove fits poorly on the right hand.

speeds up research because hundreds of times more protein amino acid sequences are known than protein structures that have already been determined by testing. It can roughly be said that the task of genomics is to determine the sequence of nucleotides. This sequence is translated into an amino acid polymer in the cell, but it starts to function only after the protein folds up into its three-dimensional shape. This function is investigated through proteomics. Thus, cooperation between experts in genomics, proteomics and drug molecule modelling supports one another.

#### Protein structures and locations in databases

Even though there is much information, the development of new drugs is quite challenging. Only 5% of drug ingredient candidates progress through laboratory testing even to treatment tests on animals. Of those, only a few per cent will ultimately be suitable as medications. It has been estimated that up to 75% of the price of drugs is due to the costs of failed pharmaceutical development projects.

One major challenge is minimising side effects. With the development of genomics, drug molecule have been found to have an individual effect. Historically, drugs have been developed assuming that people are similar in terms of their biochemistry but, in reality, we are unique at the cellular level in the same way as people are slightly dif-

ferent physically. When small drug molecules are used to try and influence the situation of a diseased body in a healing way, these individual differences at the molecular level may affect the performance of the drug.

By collecting and storing human biological data, it will be possible in the future to target drug molecules for treatment purposes that do exactly what they should in that exact situation, and tailored to the person who needs the medication. This is called personalised medicine.

A particular gene produces a specific protein affected by the drugs. When the DNA base sequence of a person's genome is known, it is also possible to deduce the basic structure of the corresponding protein in that person. Like DNA, a protein is also a string consisting of successive building blocks, and a specific block of a gene always corresponds to a specific block of a protein.

One person may have – inherited or caused by the environment – a change in one DNA nucleotide that is reflected in a protein through this chain. That change may be just where the protein should receive signals from elsewhere in the body or interact with a drug molecule. By storing protein structures and sharing them to be used by researchers, this phenomenon can be controlled and understood. The shapes of the drug molecule and the protein molecule can be matched to each other so that the drug is adapted to the situ-

ation, allowing the drug to adhere and take effect as effectively as possible. Many cancer treatments are based on this. The genome of a tumour changes over time. Tumours at different stages can be affected through drugs, but the shape of the drug molecules must take into account the changes in the shape of growth-stimulating proteins.

That is why especially proteins whose three-dimensional structure can be determined by tests or predicted through modelling are studied in drug design. The adherence of the drug molecule can be studied using modern computer modelling software in which the three-dimensional protein and

Now, the entire arsenal of proteins and drug molecules can be screened and the best candidates selected. This is due to the advancement of molecular biology, computer computing power and databases. It is now possible to screen the entire protein range of the body.

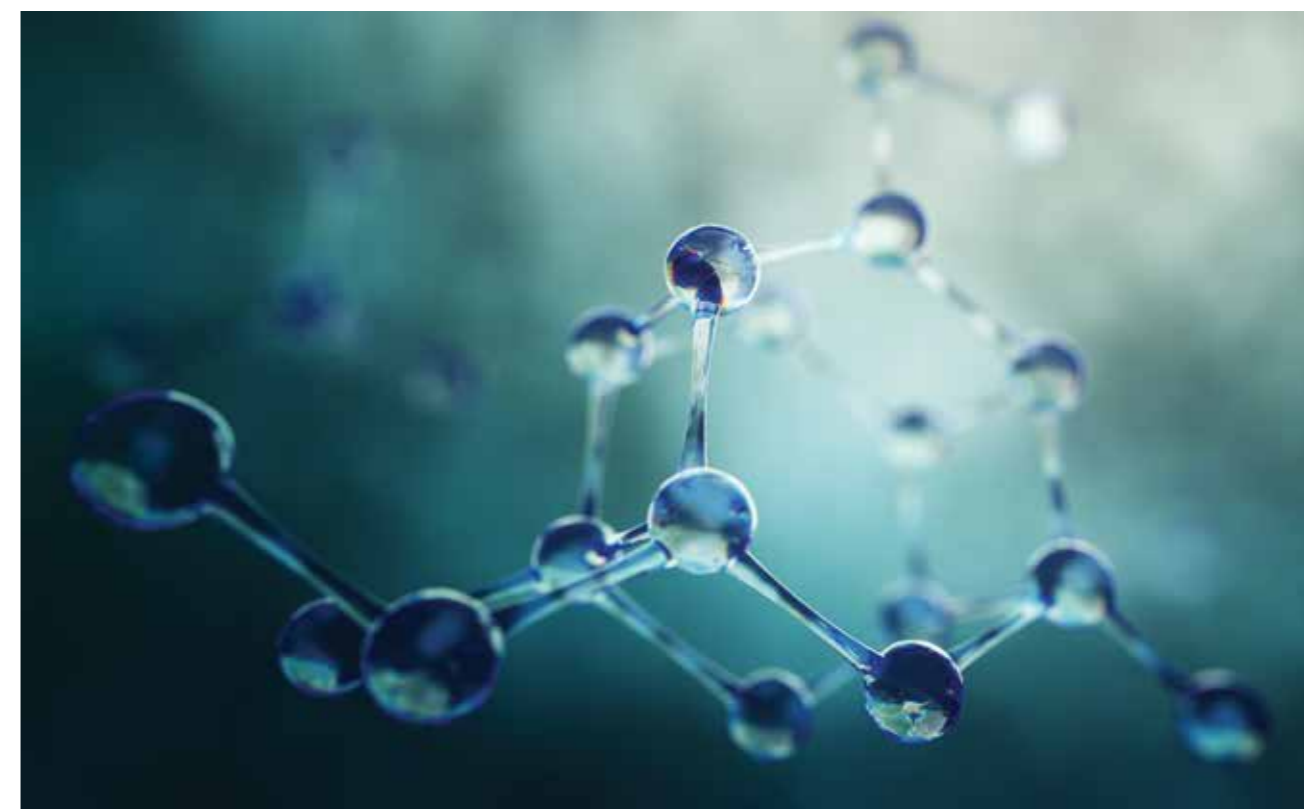
The Protein Data Bank, i.e. the PDB protein database, includes more than 100,000 protein structures divided into protein families. The members of a protein family are usually similar in terms of their three-dimensional structure, which is why they also function in a similar manner.

The PDB database is maintained by the international consor-

**The shape of proteins tells more about the function of the molecule than the amino acid sequence.**

ously. These include antibody-based imaging, mass spectrometry-based proteomics, transcriptomics and systems biology. All the data collected is open to researchers.

In January 2015, the Human Protein Atlas published a map showing the locations of 17,000 different proteins in the human body, providing valuable information for drug design. The map



**The shape of proteins tells more about the function of the molecule than the amino acid sequence.**

The shape of proteins tells more about the function of the molecule than the amino acid sequence. Proteins with the same shape can function similarly biochemically even if their amino acid sequences differ from each other by more than 80%.

Once the structure of one member of a protein family has been determined, the structure of other proteins belonging to the same family can be predicted by modelling. Modelling which is carried out using a computer

drug models are matched to each other. This also enables the tailoring of the ideal drug shape.

Usually, a drug takes effect by adhering to a defective protein in the body and altering its function. An ideal drug does only this; it does not interfere with healthy proteins or cause other side effects. Up to the present, we have been happy to find one protein affecting a disease and a drug molecule that is moderately effective against it.

tium Worldwide Protein Data Bank (wwPDB). It is tasked with maintaining individual macromolecular structural data that is freely available to researchers.

The Human Protein Atlas is a Swedish-based programme started in 2003 with the aim to map all the human proteins in cells, tissues and organs. Various omics technologies are used in the mapping, meaning technologies in which all genes or the proteins produced by them are studied simultane-

included the locations of proteins that were the target proteins of approved drugs. Researchers can view proteins in 32 different tissues, representing all of the most significant tissues and organs in the body.

In December 2017, the Human Protein Atlas released version 18. At that time, the database contained 26,000 antibodies targeting proteins encoded by almost 17,000 genes. It corresponded to 87% of protein-encoding human genes.

# Half of all drug ingredients affect only three protein families

Up to 50% of all the approved drugs affect only three protein families: nuclear receptors, G protein receptors and ion channels.

**D**rugs usually affect the cell receptors or enzymes in the body, both of which are proteins. Many drug molecules also bind themselves to enzyme receptors and transport proteins on the cell membrane. The drug may, for example, bind itself to the active site of an enzyme, thereby inhibiting the chemical reaction controlled by the enzyme. Most often the enzymes that catalyse the chemical reaction caused by the drug are metabolised by cytochrome P450 enzymes before they are in active form.

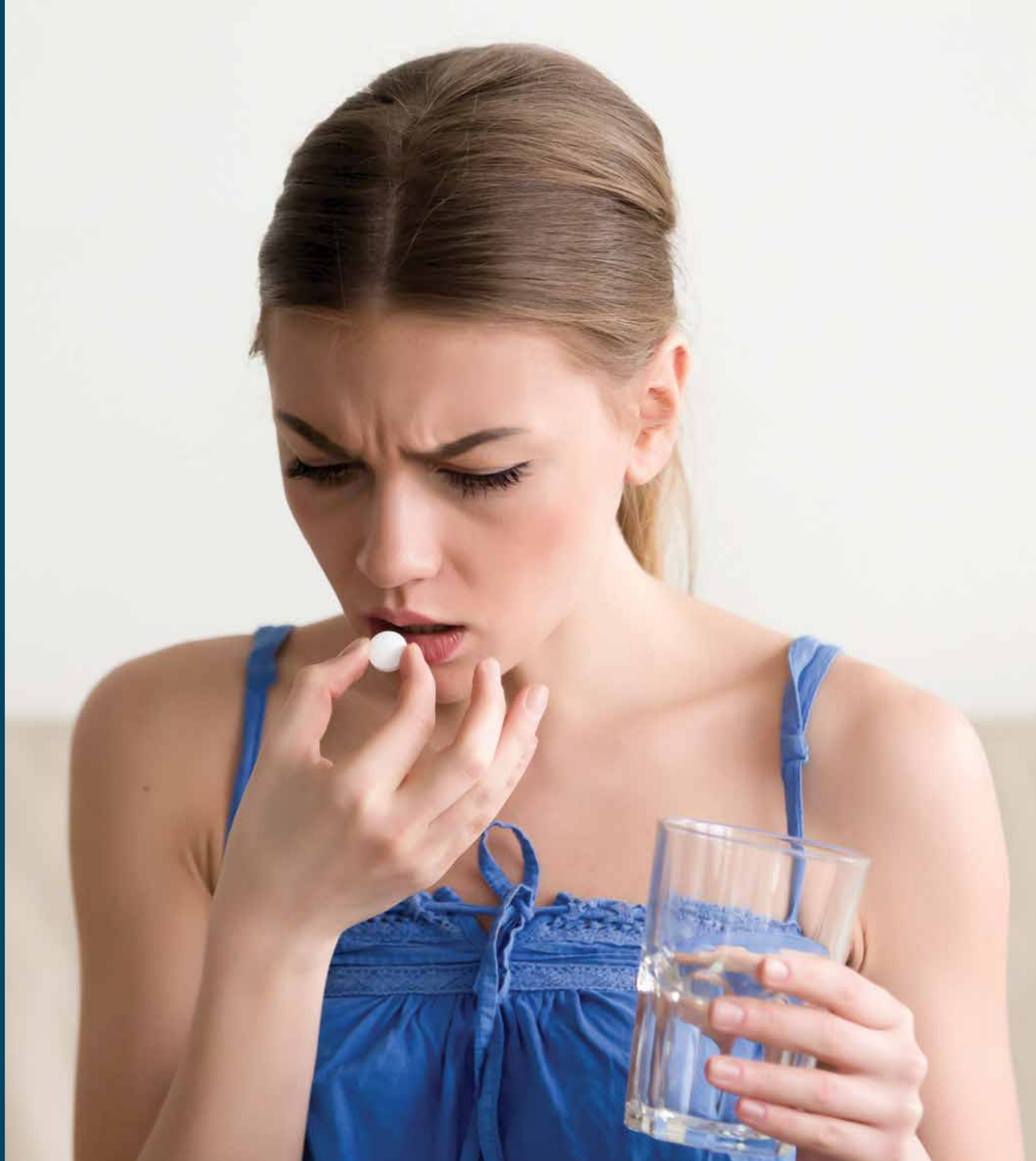
Most of the target proteins of drugs belong to only ten protein families, and up to half to only three families. Proteins belonging to a specific family have a similarly folded three-dimensional structure, function and significant similarity in amino acid sequences, which usually indicates a common ancient history. Proteins of the same family are derived from a single original form which, through evolution, has adapted and specialised due to envi-

ronmental pressures as well as functional roles that differ from their original role in cellular processes.

## Nuclear receptors and endocrine diseases

Protein families were discovered when the structure and amino acid sequences of a few proteins began to be known. It was then found that proteins consist of several independent, structurally distinct areas with a special task. These became known as domains.

New protein families have been discovered when studying the underlying mechanisms of various diseases. Nuclear receptors, for example, were discovered while studying breast cancer. It has been known for a long time that tumour growth ceased in one third of women with breast cancer whose ovaries or adrenal glands had been removed. However, the molecular basis of breast cancer was still a mystery. In 1947,



medical researcher **Elwood Jensen** began to investigate this. Jensen discovered the estrogen receptor and found that the estrogen receptor is activated when its natural estrogen, estradiol, binds itself to it. After this, the activated estrogen receptor travels to the nucleus of the cell where it participates in regulating the function of genes.

The estrogen receptor, a protein molecule belonging to the nuclear receptor family, is very important to humans. If changes occur in its function, they have a great significance for cell health. The estrogen receptor plays an important role in the emergence of breast cancer. Normally, estrogens reg-

in breast cancer samples, had become a standard test for breast cancer patients.

The discoveries revealed a protein superfamily functioning in the cells, the nuclear receptors, which includes the estrogen receptor. The nuclear receptor family includes, among others, the estrogen receptors alpha and beta, androgen receptor, progesterone receptor and vitamin D receptor. What nuclear receptors have in common is that they are activated when a cell membrane-penetrating the signal molecule, i.e. a ligand, a nuclear receptor hormone, binds itself to them, after which they travel to the nucleus to in-

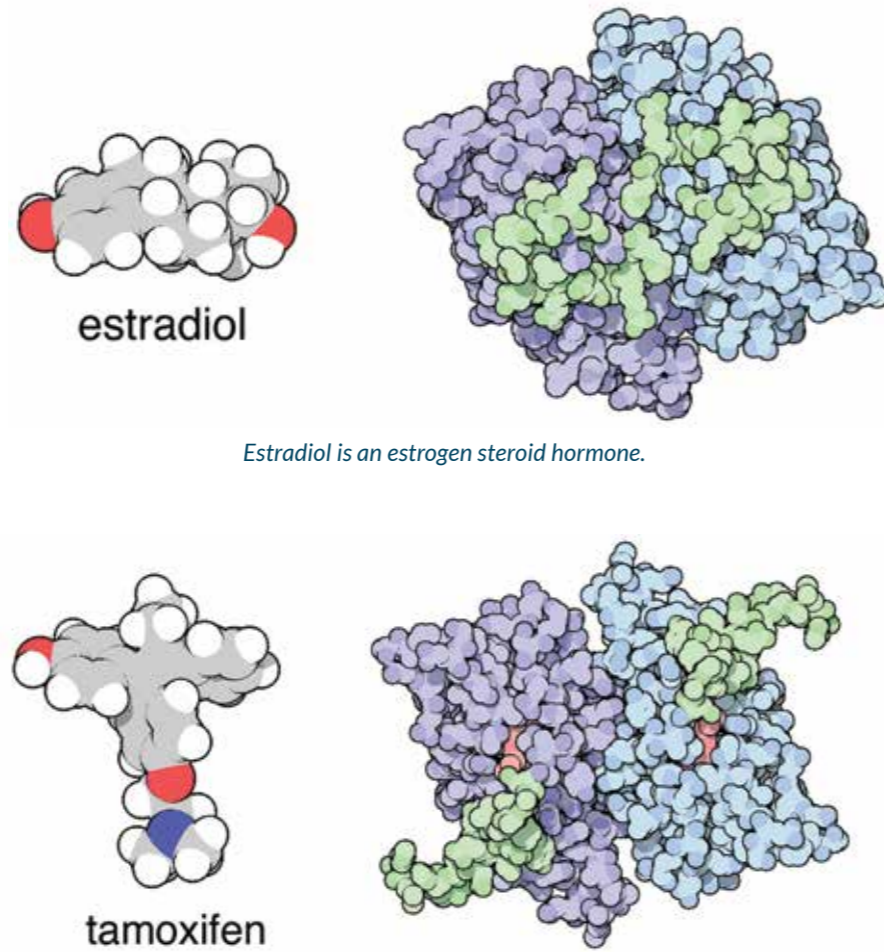
fluence cellular processes. Hormones that activate the members of the nuclear receptor superfamily include, among others, testosterone, estradiol, progesterone, glucocorticoids, mineralocorticoids and vitamin D as well as molecules created through drug design that mimic the structure of natural ligands. For example, the lip balm of **Therese Johaug** of Norway's 2016 national skiing team may have contained clostebol, a ligand of the androgen receptor. Clostebol functions as an anabolic factor, meaning that it promotes muscle cell protein growth.

Small molecules introduced to the human body through drugs or other

**New protein families have been discovered when studying the underlying mechanisms of various diseases.**

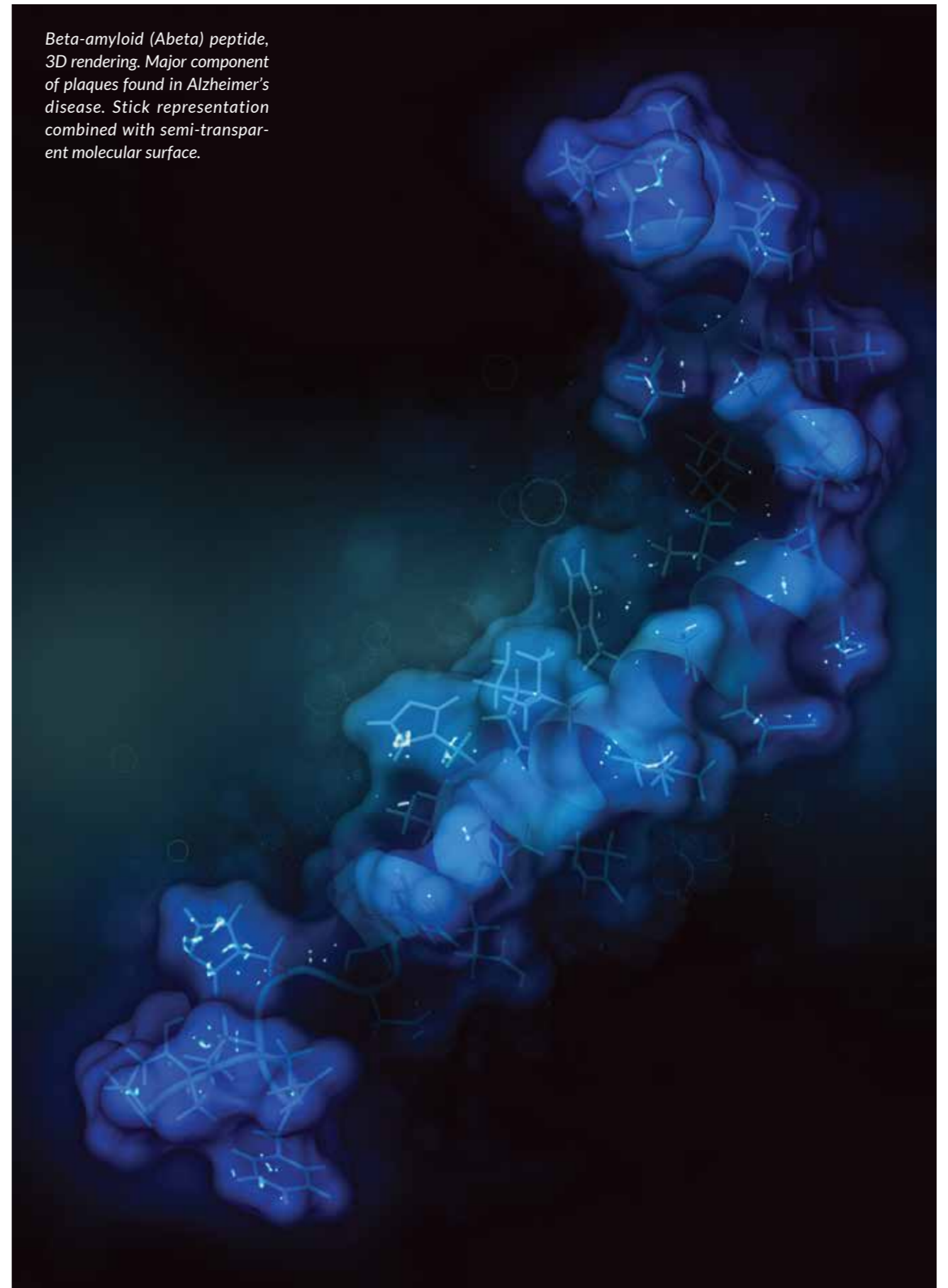
ulate the activity of the estrogen receptor in the cell. The changed shape of the estrogen receptor is active all the time, meaning that the normal regulation mechanisms of the cell based on the level of estrogen do not function correctly. This can lead to cancer, i.e. uncontrolled growth of abnormal cells.

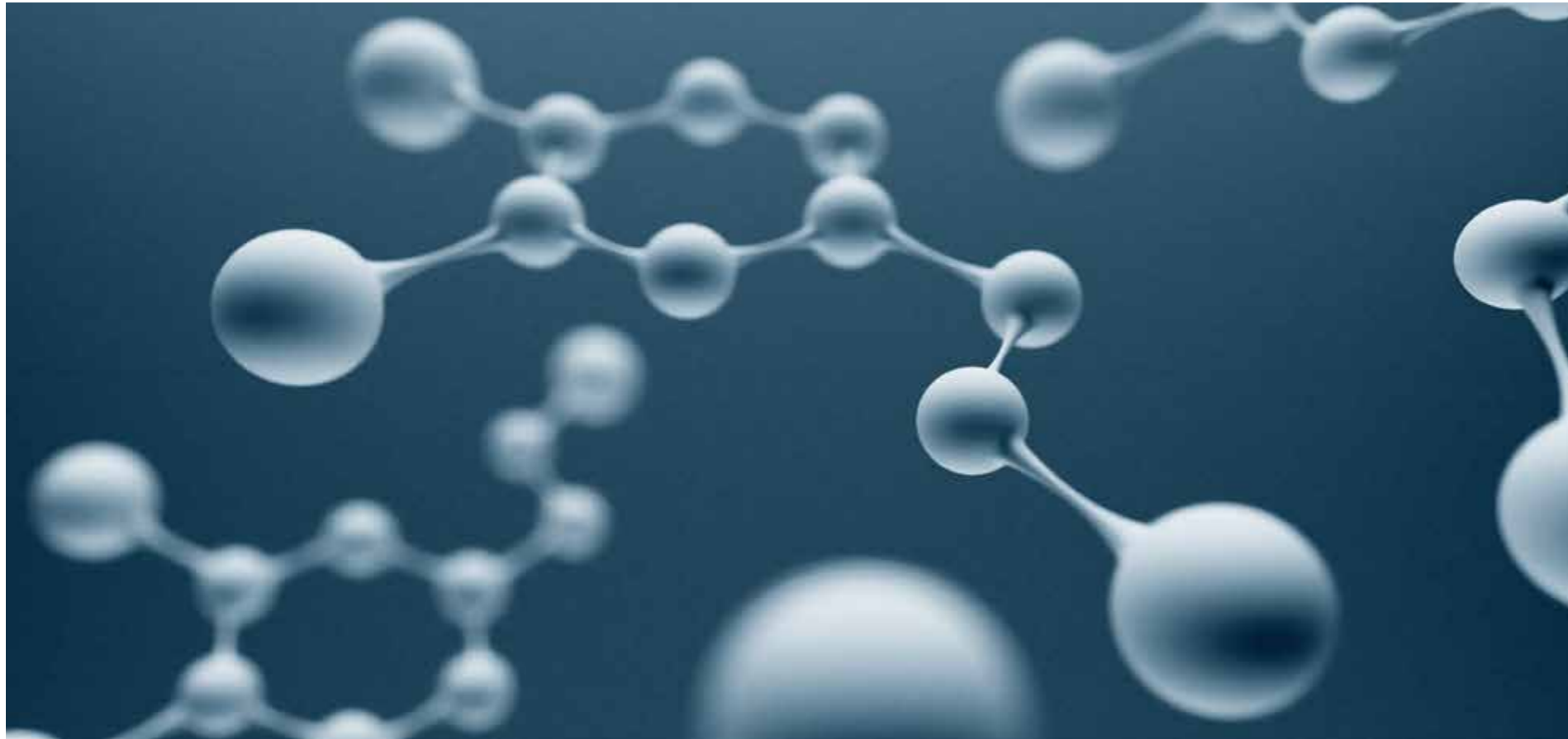
Elwood Jensen proved that breast cancer patients with a low estrogen receptor concentration in their cancer cells did not benefit from the removal of the ovaries. The ovaries produce a large proportion of women's active estrogen. The receptor concentration indicates who should have surgery and who should skip it. In the mid-1970s, Jensen and his colleague **Craig Jordan** discovered that cancer patients whose mutated tumour cells had a large number of estrogen receptors were also likely to benefit from tamoxifen. It is an estrogen, meaning that it overrides the effect of estrogen in cells. The patients with low numbers of receptors, in turn, could immediately be transferred to other treatments. By 1980, the test developed by Jensen, which was used to measure the number of receptors



*Estrogen receptor and tamoxifen. Tamoxifen is used to treat breast cancer. In the case of breast cancer, estrogen can enhance unnatural growth and make the disease worse. The drug tamoxifen is used to treat cancer by blocking the action of estrogen. Tamoxifen is a small drug that mimics the shape of estrogen and binds tightly to the estrogen receptor. When it binds, it changes the shape of a signaling loop on the surface the receptor, colored green here. The lower structure has the drug bound. Since the drug is larger than the hormone, it forces the activation loop out into an inactive conformation, blocking the signal to grow. Image: Protein Data Base*

*Beta-amyloid (Aβ) peptide, 3D rendering. Major component of plaques found in Alzheimer's disease. Stick representation combined with semi-transparent molecular surface.*





Dopamine.

routes can thus affect nuclear receptors by activating or deactivating them, thereby affecting the functioning of the cell's genes. The discovery of nuclear receptors has revolutionised biochemical endocrinology research. Endocrinology is a speciality that studies and treats diseases of hormone-producing organs. The diseases can result from the excessive production of hormones or lack thereof; furthermore, both benign and malignant tumours can occur in endocrine tissues. Prior to the discovery of nuclear receptors, the functioning of the hormones in the human body was a complete mystery. Now it can already be slightly modified.

### G proteins and cell signal transduction

In order for an organism to function, signals must be transmitted in the body's cells and the organs they form. The body as a whole sends and receives signals through electrical currents and certain molecules. **Martin Rodbell** and **Alfred Gilman** determined how signal

transduction occurs through the cell membrane via cooperation of molecules. In 1970, Martin Rodbell proved that the signal transmission takes place in three stages: signal reception, transmission and amplification. The transmission takes place so that a cell sur-

### G proteins are perhaps the most important molecules involved in signal transduction.

face protein transmits a command to exchange the guanosine diphosphate (GDP) bound to the protein located on the other side of the cell membrane for guanosine triphosphate (GTP). This phenomenon is data transfer at the molecular level.

In 1980, Alfred Gilman studied leukaemia cells and found that they did not respond to the external signals transmitted by hormones. This was due to a mutation of the receptor protein which caused the signal transduction of hor-

mones to be inhibited. Gilman isolated the protein from normal cells, and, with these proteins, he was able to repair the damaged cell. The molecules involved in the signal transduction are a large family of proteins that bind themselves to guanosine triphosphate. When they are bound to GTP, they are 'on', and, when they are bound to GDP, they are 'off'. He called them G proteins (also known as guanine nucleotide-binding proteins).

G proteins are perhaps the most important molecules involved in signal transduction. In addition to some forms of cancer, they are associated with diabetes, alcoholism and the underlying molecular mechanisms of many other diseases.

The protein family of the G protein-coupled receptors in the cell membrane conveys signals to the G proteins inside the cell membrane. G proteins, in turn, respond by exchanging GDP for GTP. The consequence of this activation is, for example, an enzyme released for the breakdown work in the cytoplasm inside the cell, opening or

### By understanding the functioning of ion channel receptors, researchers can develop, for example, treatments for addiction by changing the activity of the receptors.

ample. One third of all the known drug ingredients affect G protein-coupled receptors. Catecholamines (e.g. adrenaline, noradrenaline and dopamine), peptides, glycoprotein hormones and rhodopsin are examples of ligands that bind themselves to these receptors. Alfred Gilman and Martin Rodbell received the 1994 Nobel Prize in Medicine for the discovery of G proteins. In 2012, the Nobel Prize for Chemistry was awarded to **Robert Lefkowitz** and **Brian Kobilka** for explaining the workings of G protein-coupled receptors.

Proteins perform their work cyclically and accurately by exchanging shapes and molecules based on signals. The shapes of the G protein and G protein-coupled receptors are in a continuous, dynamically changing biochemical equilibrium reaction with each other. Changes in the sensitive balance of G proteins may result in illness. For example, the toxin of the cholera bacterium locks G proteins into one shape and affects the nerves

that control the absorption of salt and liquid in the intestines.

### Activity of ion channel receptors in the treatment of addiction

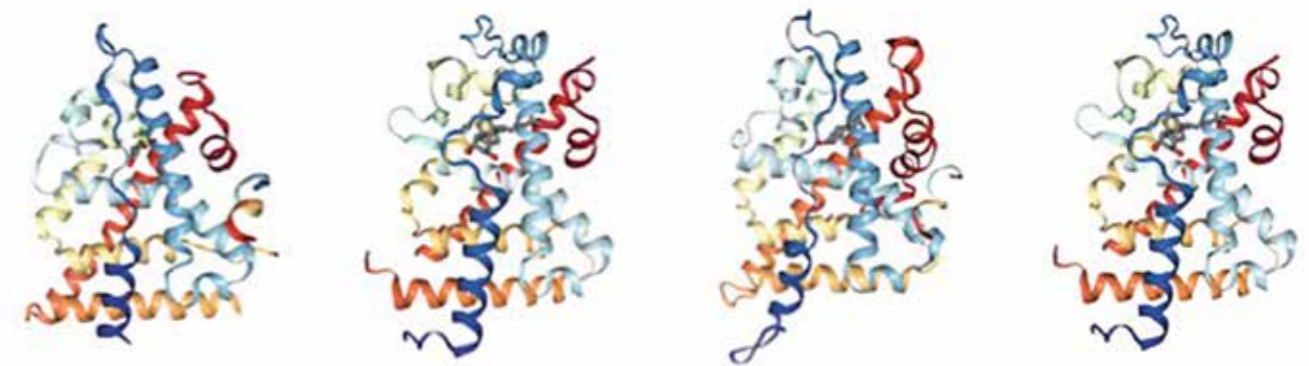
Some ion channels are complex, multi-part structures that are huge in terms of molecule size. Huge ion channel-coupled receptors react directly to tiny ligand molecules, such as the ionotropic receptor located in the brain that reacts to the amino acid glutamate. The protein, consisting of four domains, readily changes its shape as the thousands of times smaller glutamate binds itself to the domain for signal transduction and opens the ion channel permeating the cell membrane. Memantine is used for Alzheimer's disease. It protects brain neurons from destruction by blocking the excessive glutamate transmitter effect.

Receptor-gated ion channels open when a particular chemical compound binds itself to them. The chemical compound may be an extracellular molecule, such as a hormone, a neurotransmitter, a drug ingredient or a toxin, or an intracellular molecule.

By understanding the functioning of ion channel receptors, researchers can develop, for example, treatments for addiction by changing the activity of the receptors.

closing of an ion channel located in the cell membrane.

This mechanism is how rhodopsin, with which we detect light, or see, with our eyes, works in the eye, for ex-



From left to right: the estrogen receptor, androgen receptor, vitamin D receptor and progesterone receptor. These proteins are similar in structure and their molecular mechanism, although they affect entirely different functions in the body. The receptors belong to the same protein family of nuclear receptors. This means, among other things, that the receptors share the same prehistoric origin. The small molecule interacting with each receptor, such as testosterone in the androgen receptor, is visible at the top of the structure as a grey ball-and-stick model. The Protein Data Bank, i.e. the PDB protein database, includes more than 100,000 protein structures divided into protein families. The members of a protein family are usually similar in terms of their three-dimensional structure, which is why they also function in a similar manner.

A woman wearing a light blue hijab and a matching short-sleeved top is sitting on a white sofa. She is wearing white pants and white socks. She has a small tattoo on her left forearm. She is smiling and looking upwards and to the right. The background is a plain, light-colored wall.

# Fighting cancer with mathematics

Extensive data sets and databases are increasingly being used in cancer research. The research group of Sampsa Hautaniemi, Professor of Systems Biology at the Faculty of Medicine of the University of Helsinki, develops methods that can be used to integrate data from various sources, such as DNA, gene expression and protein function. When the analysis results are combined with biomedical databases, it becomes possible to generate experimentally testable predictions. This is useful in the diagnostics and design of treatment methods.



**S**ampsa Hautaniemi worked at the Massachusetts Institute of Technology (MIT) before setting up his own research group at the University of Helsinki in 2006. Hautaniemi's laboratory analyses complex, disease-related biological systems using mathematical methods. The analysis of data masses is not possible without computational assistance.

"Biomedical research requires databases and computational methods, especially in the interpretation phase of the results", says Hautaniemi.

The objective of the systems biology group is to apply computational methods to medical research questions. For example, which genetic profiles affect cancer risk or what is the prognosis of a patient with a particular genetic profile? The aim is to find a unique treatment in accordance with the genomic profile for the patient.

"Our goal is to understand the behaviour of the cancer cell and look for targets that, when their activity is modified, allow cancer cells to be destroyed with minimal side effects. When want-

ing to treat a cancer patient, you must first understand how the tumour cells make decisions on how they grow, multiply and move. We pursue this through genome-wide measurement and mathematical methods."

In the treatment of breast cancer, for example, it is important to be able to predict the probability of metastases emerging. Even though the treatment prognosis for breast cancer is improving all the time, metastases greatly increase the risk of disease.

"The problem is that we do not know how and which cells detach from the tumour, where they go and how they function there."

The aim is to deduce who has a high probability of forming metastases by studying gene activity and combining data. Current measurement methods generate enormous amounts of data.

"At this time, we do not yet know the main internal cell factors that affect the treatment response of cancer. That is why we use methods from different levels that measure the whole genome in research."

In addition to DNA and RNA sequencing, such methods include, for example, epigenetics, or analysing the impact of lifestyle on gene function. Proteomics, which determines the function of proteins and their structure, is also important.

#### Suitable medication based on data

More than four billion observation points can be measured from one cancer tumour. From this mass of observations, you should be able to identify the most characteristic factors for cancer development and drug response.

**"We are striving to find the factors for each tumour type and individual tumour."**

According to Hautaniemi, there has been quite a change compared to the situation 10–20 years ago when the usual number of observations to be processed was a few dozen or hundred.

"In addition, databases have genomewide data available on thousands of cancer patients. Utilising this data alongside Finnish material is important, but challenging."

In addition to prognosis, Hautaniemi's group also looks for suitable treatment methods based on computational analysis. Hautaniemi's group is mapping, for example, the impact of genetic modifications on drug response. Cytostatics, which destroy cancer cells, are used in the treatment of cancer. It is important to find a suitable cytostatic because the patient does not always respond well to the given drug.

In cooperation with the group of Professor **Olli Carpén**, Hautaniemi's laboratory has used genome-wide data on hundreds of ovarian cancer patients in their research. The researchers have been looking for subgroups of patients that have developed a resist-

ance to conventional chemotherapy in which platinum derivatives and taxoids are used as cytostatics.

The research project uses hundreds of thousands of processor hours of supercomputer computing time and dozens of terabytes of storage capacity.

"For a person with a certain type of genetic profile, some medications may even be harmful, while others provide the optimal benefit."

#### How data becomes knowledge

Hautaniemi and his group have developed methods by using data related to lymphoma together with the group of Professor **Sirpa Leppä**. The challenge is how to convert the data collected from genes and proteins into knowledge. Observations from clinical samples are always rather noisy and multidimensional, meaning that there are thousands of genes, proteins and potentially interesting areas of DNA. Therefore, it is essential to answer the correct and necessary medical questions so that the results are useful. The research questions can then be solved by mathematical methods.

When analysing lymphoma and ovarian cancer data, Hautaniemi's group used the so-called deep sequencing method. The method involves DNA or RNA being divided and sequenced, after which the base sequence of the molecules is converted into a format understood by a comput-

er. There may be hundreds of millions of short sequences converted into a computer format. According to Hautaniemi, when converting medical data into knowledge, the most significant bottleneck that is faced is the comprehension of medical questions so that they can be modified into computational problems.

**"The group have developed a software program called Genomic Region Operation Kit. It allows questions to be converted into computational problems and solved based on the data."**

To solve this problem, Hautaniemi and his group have developed a software program called GROK (Genomic Region Operation Kit). It allows questions to be converted into computational problems and solved based on the data. GROK is a universal tool and it has been used to understand the progression of prostate cancer. The study was conducted in cooperation with the laboratory of Professor **Olli Jänne**. The cooperation resulted in a better understanding of the function of the FoxA1 protein with the AR protein, which is the main protein affecting prostate cancer. Furthermore, the study found that a large number of

FoxA1 proteins provide a poor prognosis and a small number provides a good prognosis. In future, the results can be used to prepare a treatment prognosis and for planning treatment. According to Hautaniemi, the methods developed can be applied to any kind of cancer.

"We have used the methods we have developed to study, for example, breast, prostate and ovarian cancers. Although the tumours are found in different organs, they have a significant number of similarities at the molecular level. Therefore, in future, it might be possible to use a breast cancer drug for certain subtypes of ovarian cancer, for example. Prior to this, it must be possible to characterise the subtypes of each cancer. This means that, in future, we will be able to reliably find similar cancers regardless of their location and then recommend effective medication suitable for them."

Hautaniemi believes that cancer cell sequencing will be part of routine cancer diagnostics in future.

"We are striving to find the factors for each tumour type and individual tumour, and it is only a matter of time before we understand the biology of tumours so well that we can quickly calculate a prognosis and combinations of drugs that are likely to be effective based on their genome. Computational sciences play a key role in achieving this and utilising technology."



# Algorithm determines the appropriate drug

The goal of Professor Mikko Niemi is to devise an interpretation algorithm that helps doctors determine the appropriate drug and correct dosage for a patient. Treatments become more effective and side effects are reduced, thereby decreasing the costs.

People react differently to medications; the efficacy of drug treatment remains insufficient for some, while others suffer from adverse effects. The reason for the atypical responses may be our physical characteristics, other medication and each person's genetic makeup. An algorithm could be used to help predict the necessary dose or adverse effects of a drug when data on the patient's genome is also available in addition to physiological information from the patient. A genetic test can be performed through a simple blood sample.

New information about the human genome is obtained all the time. At the same time, the costs of genetic research and bioinformatics have fallen significantly. Data is accumulated and there are many new opportunities for utilising it. Pharmacogenetics is the study of the effect of genes on the efficacy and safety of drug ingredients. If the data on patient genomes was available to doctors, medication costs and significant adverse effects would often be reduced. The number of days in hospital care would also decrease.

**“The information on the results of the genetic test should be available when a medication is prescribed.”**

“If the genomes of patients were tested systematically, drug treatments could be better tailored and their dosages measured more individually”, says

Professor of Pharmacogenetics and Chief Physician **Mikko Niemi**.

Niemi is leading a research group at the University of Helsinki studying how genes affect the concentrations, safety and efficacy of drug ingredients. He is also investigating when genetic tests should be considered in drug selection.

“The information on the results of the genetic test should be available when a medication is prescribed, but generally you have to wait a week or two for the result. It could, therefore, be sensible to proactively test for the most important genetic variants affecting drug treatments. Through our research, we seek to identify those patients who would benefit the most from such proactive testing.”

Niemi's research group is also developing decision-making support systems related to pharmacogenetics. The aim is to devise an interpretation algorithm for doctors treating patients with cardiovascular disease to help find the most effective and safe cholesterol medication for each patient. The algorithm uses data on the patient's characteristics, illnesses, other medications and genome.

Statin drugs intended for cardiovascular disease reduce the level of LDL cholesterol and increase the level of good HDL cholesterol in the blood. However, they cause muscle pain in some patients. The predisposition for muscle symptoms is partly hereditary.

## Drug metabolism is individual

The dosage requirement of individual drug ingredients may vary by more

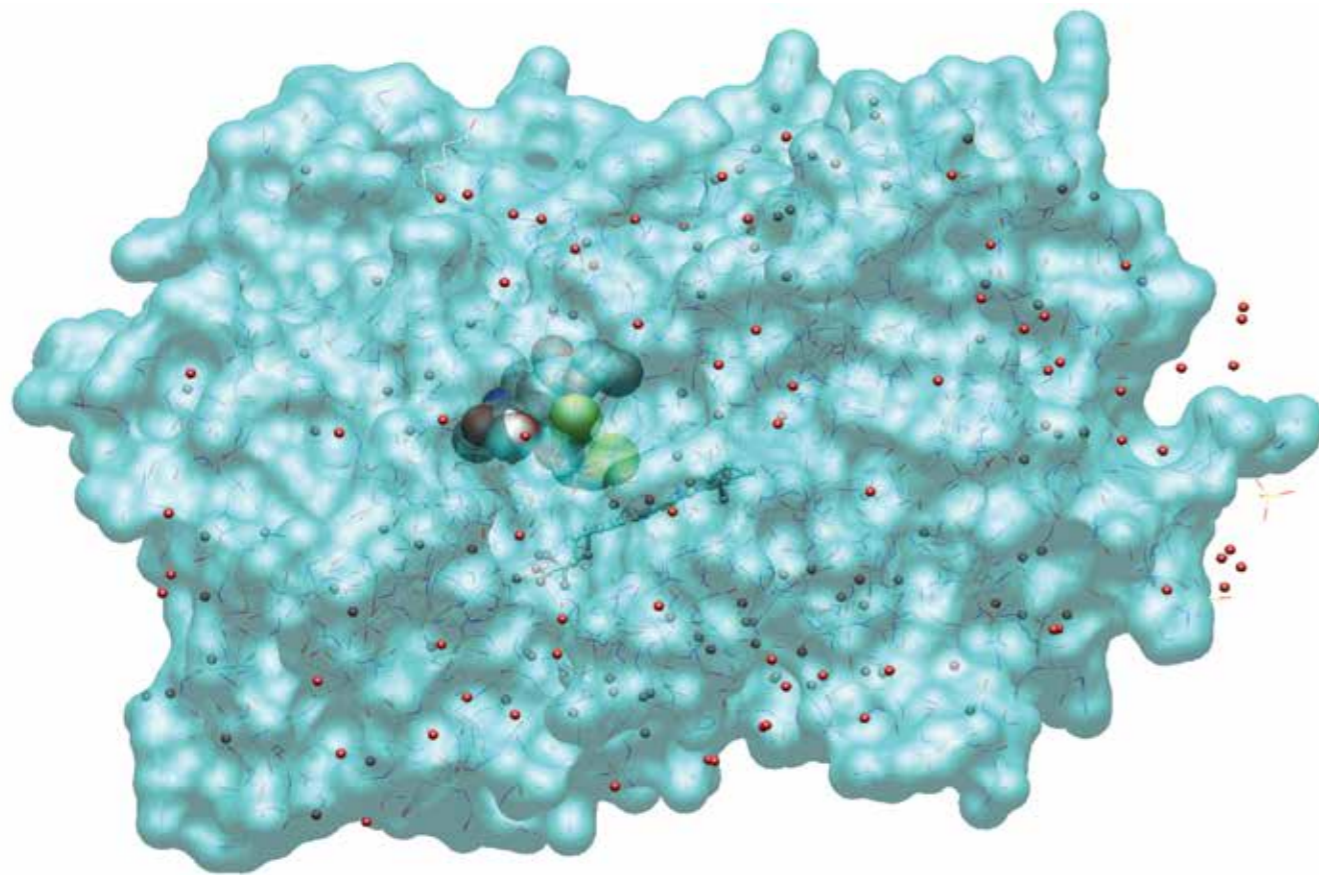
than tenfold between different individuals. This may result from how rapidly or slowly the drug leaves the body. Cytochrome enzymes (CYP) are central to the breakdown and removal from the body of many foreign substances, such as drugs. CYP enzymes are present especially in the liver.

**The algorithm uses data on the patient's characteristics, illnesses, other medications and genome.**

When Mikko Niemi was working on his doctoral dissertation on the synergistic effects of diabetes drugs, he suspected that the variation in drug metabolism in different individuals was hereditary. Of particular interest are the three CYP enzymes CYP2D6, CYP2C9 and CYP2C19, as they affect up to one third of all drug ingredients in clinical use. Genetic variation in the activity of the CYP enzymes is high. This variation may lead to manifold differences in the concentrations of different drug ingredients and the responses to them in different individuals.

Genetic tests allow people to be classified into up to four different groups, depending on the drug, based on how quickly the body eliminates certain drug ingredients: very fast, normal, slowed down and slow. This so-called metabolic rate can affect the dosage requirement, efficacy and adverse effect risk of a drug.





Cytochrome P450 (CYP) enzymes are some of the most important enzymes that break down drug ingredients. Pictured is the three-dimensional structure of the CYP2C8 enzyme.

In very fast metabolisers, the drug ingredient leaves the body faster than normal and its effect can be insufficient. In slow metabolisers, the drug exits slower than normal and its effects may be intensified. Consequently,

CYP2C19 metabolism. It is, therefore, advisable to opt for alternative medication with such patients.

Variation in the CYP2D6 enzyme, in turn, has a significant effect on, for example, codeine. Codeine is a common

patient would not need to suffer from inadequate pain management.”

Other enzymes besides CYPs are also relevant. TPMT, for example, is an enzyme that affects the metabolism of thiopurine drugs. Thiopurines are used to treat, for instance, autoimmune diseases, inflammatory bowel diseases and leukaemia.

“A hereditary TPMT deficiency predisposes you to the severe adverse effects of thiopurine drugs on blood cells. A genetic test to identify this hereditary deficiency has been in clinical use in Finland already since 2005”, says Mikko Niemi.

Around a dozen genetic tests related to drug treatments are currently available in Finland.

#### Decision-making support algorithm for doctors

The suitability of a drug ingredient for each individual depends on many factors. It is not solely affected by en-

prescription painkiller, part of which usually turns into morphine in the liver via the CYP2D6 enzyme. In slow metabolisers, the effect of codeine may be inadequate. In very fast metabolisers, the amount of morphine in the body may run too high.

“Were the doctor to already know at the start of treatment that the patient’s CYP2D6 metabolism is slow, the

**Genetic tests allow people to be classified into up to four different groups, depending on the drug, based on how quickly the body eliminates certain drug ingredients.**

the same drug dose may be too low for some and too high for others.

Some drugs become active by means of CYP enzymes. With such drugs, the effect of the hereditary metabolic rate is reversed. For example, in one third of the population, the effect of clopidogrel, a drug that inhibits blood coagulation, is weaker than normal due to hereditarily slowed down

zymes that break down drugs. The transport proteins of the cell membrane affect the delivery of drug ingredients to their site of action. In the target tissue, the drug ingredient interacts with its target of effect.

“This results in a chain of events that brings about the desired drug effect. There are individual, partly hereditary differences in all these factors. It would be important to consider all these individual factors, including the genome, when selecting medication.”

In 2017, Mikko Niemi was granted substantial funding by the European Research Council for a project to develop an algorithm facilitating the selection of cholesterol medication. For this purpose, Niemi’s research group is building a so-called system pharmacological model.

“It is a kind of virtual patient that can be used to individually predict the effects of each alternative cholesterol drug.”

No similar algorithm has been attempted to date.

“If the algorithm works in the selection of cholesterol medication, a similar way of thinking could also be extended to other drug treatments.”

Of course, the algorithm cannot be built if there is not enough reliable research data available. Niemi’s research group has been compiling such data for years in their research projects. The biobanks established in Finland and the future genome centre will also speed up the collection of data needed for such research.

**“The algorithm uses data on the patient’s characteristics, illnesses, other medications and genome.”**

Better utilisation of genetic information is also desired by the Finnish state. Due to Finland’s exceptional settlement history, the genetic structure of the population provides special opportunities to combine genomic and health data. Pharmacogenetics is one of the four leading projects of the national genome strategy. The goal of the strategy is to have genetic data in efficient, health-promoting use already in 2020.

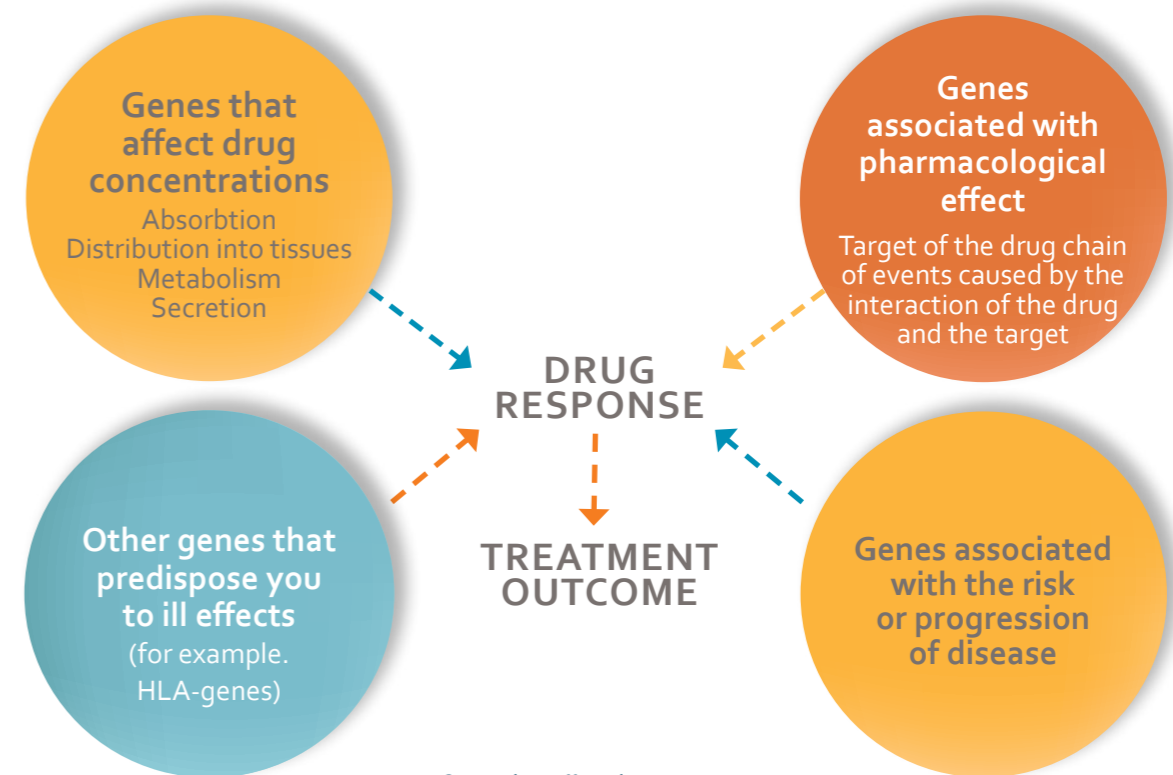
#### Pilot project: utilisation of genomic data in health care

At present, the number of genes with significant effects on the efficacy and safety of drug treatment is relatively low: less than 20 of the total of about 20,000 human genes.

Since the group of genes is so small, according to Mikko Niemi, it would be technically possible to test even large numbers of patients.

“The next step is to proactively test for all genetic variants affecting drug treatment.”

The National Institute for Health and Welfare (THL), HUSLAB’s Department of Clinical Pharmacology and CSC have launched a pilot project that will be implemented by combining the genetic data of THL Biobank and the patient document information of HUS (Helsinki University Hospital). The materials will be used to map the prevalence of the genetic variants affecting drug treatments in Finns. In addition, the project will look at how many patients in the sample receive drug treatment during or after the treatment period wherein genetic data could have affected the selection or dosage.



Genes that affect drug treatments.



# Striving for a national service to utilise genomic data in health care

The data obtained from the human genome will become part of health care decision-making. Combining a patient's genomic data with the information available on the current state of the patient's health enables the development of new algorithms, making it possible for a doctor to quickly select the best possible treatment and medication for the patient.

Medications have different effects due to the individual nature of a person's genome. For example, some antibiotics cause drug allergies. The body may break down the medication before it has time to take effect, or the patient may experience harmful side effects. That is why utilising genomic data in pharmacotherapy will reduce the number of incorrect

prescriptions. On the other hand, if a person knows that he/she has a digestion-related genetic trait that augments or weakens, for example, the breaking down of caffeine into energy and building materials, that knowledge may have a positive effect on his/her lifestyle. In the future, the algorithms of genetic databases linked to electronic patient record systems could automatically warn

against possible adverse drug reactions and provide advice on the most effective alternative.

In Finland, CSC - IT Center for Science, the National Institute for Health and Welfare (THL) and the Institute for Molecular Medicine of the University of Helsinki are creating a secure framework for storing the genomic data produced on Finns and interpreting the da-

ta for health care purposes. The aim of the Helsinki University Hospital (HUS), which is involved in the cooperation, is to investigate the benefits of digital health data on humans for research and care. The six-month pilot project was part of an assignment given to the Genome Center to be established in Finland, coordinated by the Ministry of Social Affairs and Health.

## Cardio Compass: a tool for assessing your health risks

Storing data becomes cheaper and capacity grows year after year. An exemplary file on the data collected on the health of Finnish people is the FINRISK cohort of THL. The analyses of the data collected for decades on Finns have been developed further in the GENRISK project studying the hereditary risk factors for cardiovascular diseases. An algorithm that calculates the risk points for an individual to suffer from cardiovascular diseases is tested at the same time. A tool called Cardio Com-



participants may also talk directly with experts on the interpretations made based on the data.

## Algorithms help with medication selection

In April 2016, the Finnish Government decided to establish a Genome Center in Finland with the aim of introducing genomic data as part of health care. In

data. Algorithms can be developed to select a suitable drug ingredient and to optimise the amount of medication with standardised software methods. This is called pharmacogenetics.

In 2016, Professor Mikko Niemi was granted substantial funding by the European Research Council for a project to develop an algorithm for finding a suitable cholesterol drug for a patient. The mathematical model takes into account the patient's genome, other medication, gender, age and weight.

However, effective utilisation of algorithms requires that there is enough different data available on patients. It is important to know the quality and purpose of the data. Sufficient meta-data describes the quality of the data, based on which decisions on the utilisation of the data can be made. The interpretation of the data will become easier once a functional technical distribution platform is provided for reference data, making it possible to design better interpretation algorithms for the data.

Creating interpretation algorithms for genetic data for clinical use is the long-term goal. In addition to algorithms helping doctors to, for example, determine the appropriate medication, they can even be suitable for predicting changes in the function of proteins. The goal is that once the interpretation algorithms are ready for clinical use, they would be available in patient information systems automatically instead of as a request for information to be ordered separately.



pass provides people with their current risk level and the development of the risk over the next few decades.

Cardio Compass is tested in practice by recruiting and testing 10,000 people from the Kotka region, the customer base of health care company Mehiläinen and blood donors in Helsinki. The people participating in the project receive important feedback on their own health, more accurately than ever before, through the combination of genomic data. The information is collected in Cardio Compass. The par-

## An algorithm that calculates the risk points is tested at the same time.

order to build the functions of the Genome Center, the data already collected and stored from the Finnish population will be utilised and combined in research which, if successful, will improve the accuracy of prescriptions. It would be possible to determine suitable medications or rule out the poor ones based on the patient's genomic



# BBMRI.fi: an IT infrastructure for shared biobanks

The BBMRI (Biobanking and Biomolecular Resources Research Infrastructure) is an infrastructure established by 17 European states. Its aim is to promote the high quality, research-based use of the sample collections and associated data of European biobanks. Use of such collections assists in the development of diagnostics and treatment, as well as health promotion and disease prevention. Finland has several biobanks in operation. A common IT infrastructure is being created for these based on cooperation between the BBMRI and ELIXIR.

The BBMRI operates through national centres that coordinate the biobanks of member states. Service centres serving the customers of biobanks are also being established in member countries and under the BBMRI. BBMRI.fi is a national cooperative body belonging to the BBMRI Network;

its membership is made up of Finnish biobanks.

## Finland's biobanks

Ten biobanks were operating in Finland in 2018. More than 100,000 Finnish sample collections were transferred to the National Institute for

Health (THL) in June 2015. Sample collections can be used to identify the causes of diseases and the related impact of heredity,

A total of 50 percent of samples from the Auria Biobank are cancer samples. The Auria Biobank focuses on cardiovascular, metabolic and cancer re-

search, as well as research into neurological diseases. The biobank was established by the University of Turku and the University Hospital Districts of Southwest Finland, Satakunta and Vaasa.

The FHRB – the Finnish Hematology Registry and Biobank – operates throughout the country and collects blood and bone marrow samples from patients with haematological diseases. Such samples are required for research into methods of treating serious haematological diseases, particularly leukaemia. The FHRB biobank is owned by the Finnish Association of Haematology, the Institute for Molecular Medicine Finland (FIMM) and the Finnish Red Cross Blood Service. The Association of Finnish Cancer Patients is also involved in its activities.

The mission of the Academic Medical Center Helsinki (AMCH) is to support research aimed at health promotion and the understanding of disease mechanisms, as well as the develop-

ment of products, diagnostic methods and treatment practices used in healthcare.

The HUB Biobank focuses on urological diseases and supports research in this field, based on biobank samples. The Biobank began sample collection at the beginning of 2015. Research based on samples and data is aimed at improving the prevention, diagnostics and treatment of urological diseases. The HUB Biobank was founded by FIMM and the Hospital District of Helsinki and Uusimaa (HUS).

## IT infrastructure of biobanks

Biobanks manage huge and important data sets. Tasks such as the alignment and management of genome data and imaging datasets are challenging. The aim is to create a national, web-based availability service for biobank data, based on which users can search for materials suitable for research and product development.

Juha Knuutila, Enterprise IT Architect at THL, coordinates IT cooperation between biobanks in Finland. Knuutila views the biobanks' IT infrastructure as central to national cooperation within the BBMRI.fi network.

fer practices. Consistent ethical principles and maintaining the confidence of the persons involved in research are another important area of national biobank activity.

In terms of IT, cooperation has been initiated through database pilots. For example, pathology archives form the key sample data of most Finnish hospital biobanks and biobank projects. A national digital pathology infrastructure has been created by digitalising pathology samples from university hospital archives, based on inter-biobank cooperation.

Digitalisation fosters the use of new applications such as DNA microchip technology and the development of tools for the analysis of large data sets, thereby promoting individualised health care. These services are part of the European BBMRI infrastructure.

"The goal is to create a unified Finnish interface connecting us to the European infrastructure."

## Harmonisation of data into a common database

However, much work remains to be done. With health care systems of several types in use, information is frag-

"ELIXIR provides a good cloud service while the BBMRI offers specialised IT systems in support of biobank activities."

"In Finland, IT infrastructure is highly developed in comparison to many other European countries. BBMRI.fi and the ELIXIR Finland are good examples of this. Both have a clear role to play. ELIXIR provides a good cloud service while the BBMRI offers specialised IT systems in support of biobank activities. The virtualised computing clusters of FIMM and CSC – IT Center for Science are available via a cloud service. Cooperation as smooth as this is still rare at European level," states Knuutila.

## Database pilots begun

The aim of the biobanks cooperation network is to agree on uniform practices relating to quality criteria, and to organise nationally consistent data trans-

mitted. Knuutila believes that the greatest task lies in the harmonisation of data.

"To facilitate research cooperation, clinical data, demographic data and sample data should be combined in one place and an easily searchable format. Biobanks should therefore agree on which variables can be combined in the databases in a realistic and useful manner."

Knuutila believes that this would force the biobanks to work together, which would also benefit researchers and pharmaceutical companies. Knuutila is leading the biobanks' joint IT group.

"Obtaining patient data in a structured format would be useful to both the biobanks and hospitals."

# Bank of million patient samples

More than a million tissue samples and tens of thousands of blood samples are stored in Auria, the first hospital biobank in Finland. The biobank is also able to combine donor-related data to the collections, providing significant assistance for research. The data can be requested from the donor of the sample, patient records or national registers.

An electronic health record accessed with your identity number has been in use in Finland for a long time. Registers requiring an identity number create good conditions for the efficient future utilisation of sample collections from people and the data linked to them. This is a great advantage over many countries.

The sample collections of Auria Biobank, which operates in connection with the Turku University Hospital and the University of Turku, are physically located in hospitals in southwest and west Finland. Samples are collected and combined with the necessary

metadata, indicating the clinical information on the sample donor, quantity, date and how the sample has been processed. The samples of Auria Biobank include tissue, blood and DNA isolated from cells.

## Providing consent once is enough

The Finnish legislation relating to biobanks is progressive. Consent provided once from the donor of the samples is sufficient for the stored samples to be utilised in various studies and in the future too. The law allows the biobank

to contact the sample donors who have given their consent, for example, to enquire about the willingness of the donor to participate in a study not covered by the consent or to donate additional samples.

"In most cases, the contact has to do with drug research. If the patient is interested, they will contact the author of the study directly, and then they will form a separate agreement with the research organisation, after which the matter no longer involves the biobank", says **Perttu Terho**, Vice Director of Auria Biobank.

**The samples of Auria Biobank include tissue, blood and DNA isolated from cells.**

The Personal Data Act and the Biobank Act are complied with in the transfer of data, safeguarding the privacy and confidentiality of patient information.

Consent for the donation of samples can be given in hospitals or online through an electronic form.

## Sample collection is growing and being digitised

High-grade prostatic intraepithelial neoplasia in the prostate issue. In addition to tissue samples, Auria collects fresh tissue left over from a diagnostic procedure. The biobank currently collects, for instance, prostate, intestinal, ovarian, pancreatic and liver tissue. Auria Biobank was established by the University of Turku and the hospital districts of Southwest Finland, Satakunta and Vaasa.

New samples are collected from consenting patients in connection with normal diagnostics and treatment. Tissue samples filed in hospitals are scanned, digitised and transferred to databases. Before transfer to the biobank, personal data is removed from the samples and replaced with a code. This ensures the efficient protection of personal data.

Auria collects tissue samples taken in connection with surgery that are left over from a diagnostic procedure.



ture, such as cancer tissue, and biobank blood samples taken in connection with laboratory visits.

“After surgery, the tissue sample is taken to a pathologist for examination. Typically, the sample is cast into paraffin and cut into slices with a thickness of a few micrometres which are stained with the colours needed for diagnostic purposes. The pathologist examines the stained tissue sections to establish whether there is a tumour present in the sample, for example. If some of the sample remains, it can be utilised in biobank studies. The sample must not run out, so there should be enough for the hospital to use. Once this has been confirmed, the tissue sample can be used for other research”, says Terho.

Auria Biobank digitises the samples that are needed for research projects.

“The purpose of digitisation is that we can, for example, ask a pathologist to assess the samples and mark the spots where cancerous tissue is found and where there is healthy tissue.

The pathologist can do this from anywhere on their own computer, and the samples themselves do not need to be transferred. The digitised images can also be analysed in an automated way using pattern recognition algorithms and methods based on artificial intelligence.”

Auria has previously isolated DNA only from those blood and tissue samples that were needed in projects. Now, DNA isolation is to be done from every blood sample stored.

“Isolating DNA from every sample enhances research. Samples are received and stored, but nothing is yet studied. The samples are left to wait for future research as it is not yet known what the samples may be needed for.”

DNA will be isolated from 16,000 blood samples this year. Going forward, more than 20,000 samples will be taken every year.

The blood sample is taken in conjunction with a normal diagnostic or clinical blood sample.

### Auria Biobank digitises the samples that are needed for research projects.

“We are talking about one extra 10 ml blood sample for the biobank. The blood plasma and white blood cells from the sample are placed in different tubes before being frozen.”

Perttu Terho emphasises that the donated sample is valuable when it can be combined with patient data.

“Researchers may want data on patients with a specific diagnosis, medication and blood count. In this case, it is possible to quickly check the biobank and see whether there are samples that meet these criteria and the associated data exist.”

Biobank material can be used to identify the special characteristics of diseases and drugs. For example, it is possible to learn more about why some patients have side effects from medication and others do not.

“It is important to collect a sensible amount of relevant patient data from as large a number of people as possible. This allows samples from patients who are of interest to research to be obtained for the biobank.”

Sample-related requests from researchers are received every week.

“Based on an analysis, we map the quantity of samples and data in the biobank that the researcher is interested in. If the researcher is satisfied with the outcome of the pre-analysis, they will submit a request for data and samples describing the study and defining the required samples and data.”

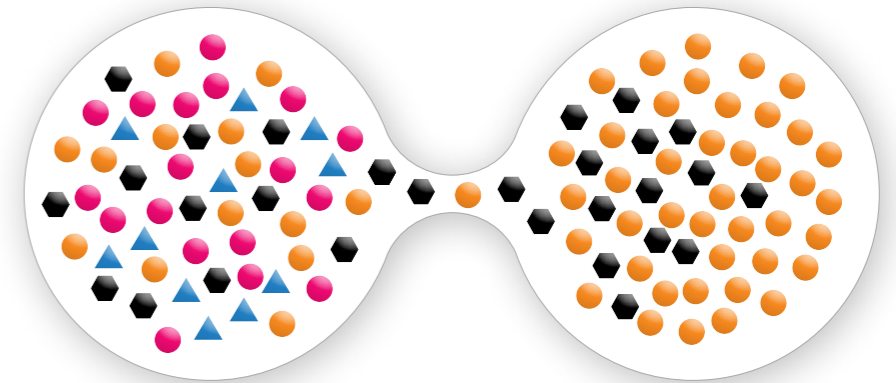
The requests for data and samples are processed by the biobank’s Scientific Steering Committee which convenes once a month. The steering committee evaluates the requests. If the steering committee decides in favour of the study, the applicant can proceed to the preparation of a Material Transfer Agreement.

### Availability service in the works

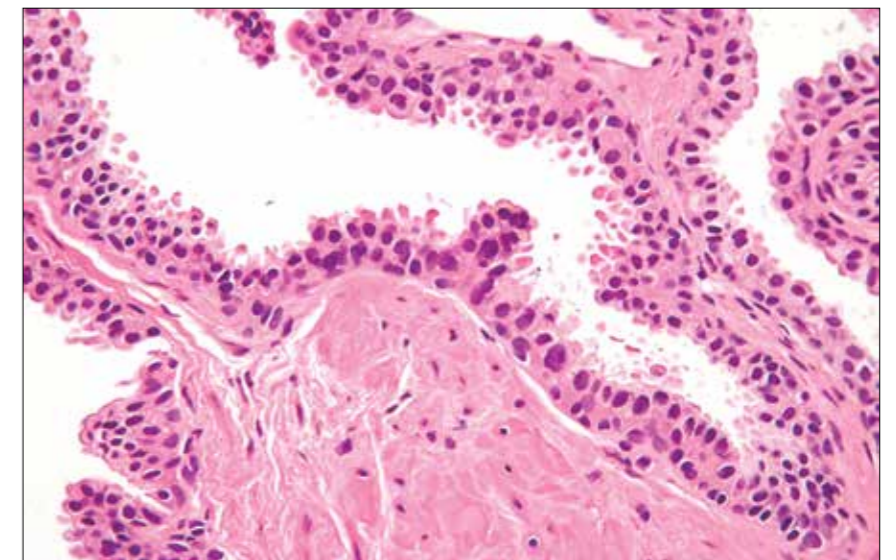
In principle, the operations of the biobanks operating in connection with Finnish hospitals are the same. They collect samples from their own hospital districts and store associated data. It would, of course, be a tempting idea to be able to search all the available sample collections in one go. The challenge is that, over the years, the different hospitals have stored and classified the samples in different ways. Different systems have different information registered, meaning that there is variation in the data provided on patient samples. Data should flow smoothly between the different biobanks.

“Hospital data is difficult to analyse. The expertise of a clinician is required to interpret what has been recorded. The data available is not directly commensurable. It would be important to create an availability service that is able to combine the data of the different biobanks so that at least the basic data would be available.”

The Finnish Biobank Cooperative was established in 2017. Its members include hospital districts and universities with faculties of medicine. The



About 9,000 years ago, a small number of settlers moved to the Finnish territory. The individuals of this new population represented small and narrow genetic material, resulting in the regional enrichment of certain disease genes. This is called a bottleneck phenomenon. The phenomenon is very useful for genetic research. Over-representation of genetic modifications is only observed in populations that have experienced a bottleneck. Auria Biobank is involved in the establishment of the future genome centre. According to Lila Kallio, Acting Director of Auria Biobank, the way in which the transfer and storage of research and diagnostic sequences will be organised is, so far, only at the consideration stage. “Genome legislation is being drafted and reform of the Biobank Act is underway. In addition, the new data protection regulation of the EU will also clarify the operations of biobanks.”



High-grade prostatic intraepithelial neoplasia in the prostate tissue. In addition to tissue samples, Auria collects fresh tissue left over from a diagnostic procedure. The biobank currently collects, for instance, prostate, intestinal, ovarian, pancreatic and liver tissue. Auria Biobank was established by the University of Turku and the hospital districts of Southwest Finland, Satakunta and Vaasa.

purpose of the biobank cooperative is to provide the material in the sample and the data collections of Finnish biobanks to be used by researchers under a one-stop principle. It would provide customers with a unified view and a centralised channel to the materials of Finnish biobanks. The biobank cooperative is responsible for the develop-

ment of information systems, among other things.

According to Terho, it is possible to combine the associated clinical data relevant to research to the samples. Biobanks will utilise the sensitive data platforms developed by CSC – IT Center for Science when designing the information services of their own.

# Storing the whole genome of the Finnish population? The data will benefit disease research

Extensive research projects are being conducted on Finnish genetic heritage and genomic data is being produced and analysed all the time. However, the national objective is to store the data produced on Finnish people in Finland, allowing analysts to combine the data with other health information. The utilisation of genomic data in health care is still in its early stages. Data analysis offers many opportunities for companies in the bio-industry.

Research-appropriate genetic data on the Finnish population exists fragmented all over the world in various databases and data storages with varying arrangements. Therefore, there is a need to create a domestic, secure service for the management of Finnish data that would cross organisational boundaries, be network-based and well-coordinated. By coordinating the data in different locations in just one place, the data could, with the permission of the owner, be released for legitimate purposes, such as research, product development and medication.

The human biology is very complicated, more complicated than previously thought. The expression, structure and function of genes and the building blocks of the body, or proteins, require advanced mathematical, computer science and statistical methods, i.e. bioinformatics.

New ways to study and prevent diseases are constantly being discovered through bioinformatics methods, such as gene sequencing. DNA sequencing is the starting point where the order of the four bases – adenine, guanine, cytosine and thymine (A, G, C, T) – within a DNA molecule is de-

termined when deciphering the genetic digital code. Each ACGT base is a similar nugget of information to a computer bit, zero or one, which, as a long chain, contains the instructions for a programme.

Sequencing methods have improved and become cheaper, and this has significantly increased the possibilities of

biology and medicine to produce this kind of data. The data is now being used to find out what digital messages have been written on the molecules of life for the survival of organisms.

However, data is only the first step towards interpretation. The interpretation of digital genomic data,

that is, how the information stored in the genome manifests itself in the body, is still under development. In the last ten years, for example, researchers in Sweden have been creating a map (HPA Human Protein Atlas) on how genes are expressed as proteins in different cells and then combining

this information with microscope images of cells. This allows you to see which gene is expressed in any given cell and is involved in the development of proteins and, hence, larger structures, such as neural fibres, hair follicles or light-sensing structures in the fundus of the eye. However,



a clear, deeper level map on how molecules that are operating on a nanometre scale produce these functional, microscopic structures does not exist yet. The structure of each cell requires millions of molecules in cooperation. The building instructions stored in genomes and the resulting molecules form a self-organising network that current research tries to understand.

Finland is fairly well positioned to be an international actor in the management of genomic data, but there are too few experts in individual organisations. The data masses required to understand genomic data are large and the analysis requires specialised expertise that does not sufficiently exist in Finland yet. There is a need for cooperation in genomic data management and for more interpreters specialising in data. Finland will gain more expertise once the creation of a framework for storing Finnish genomes is achieved. Initially, this would mean a national reference database created from the data of tens of thousands of people. It would be beneficial for diagnostics, such as in improving medical treatments, as it is already possible to determine, for example, suitable and safe medication based on the patient's genomic data.

When data is well organised and described, it can be combined. Combining supplementary information, such as a prescription, genome and long-term treatment results, is a prerequisite for developing a deeper understanding.

Data organised in the hands of skilled analysts will help achieve breakthroughs in research. The US company GRAIL, for example, seeks to understand the underlying mechanisms of cancer. The earlier the cancer is detected significantly improves the prognosis of the disease.

The GRAIL project has involved the collection of samples from 10,000 patients and their consent for the analysis of the diverse data created from the samples. The idea is to use the cancer tumours of this group of patients to create a database against which blood samples can be screened.

Cancer tumours are usually the result of a change in the genome of a cell of the person carrying the disease, making the cell abnormal. At the cellular level, each cancer is a rather unique disease that looks like its carrier; what they have in common is the reckless growth of abnormal cells. Cancer utilises the normal regeneration and healing mechanisms of the body to selfishly spread its own genetic instructions. The genomes and the digital information contained therein of two humans are, on average, 99.5% identical. That is why the progression process of many cancers is well-known despite the individual nature of cancers. Consequently, it is justified to study how changes in individual or multiple nucleotides (ACGT) in the genome affect the balance of the cell's molecular network so that the cell becomes a cancer cell.

In the GRAIL project, millions of unique changes in genomic data that may cause cancer are sequenced from the genomes and cancer tumours of patients. The project will create a database that allows health care professionals to detect early stages of cancer, even directly from the bloodstream. If the innovation is successful, cancer screening can be started earlier, meaning that the tumours are still microscopically small and easier to manage through, for example, medication.



Services provided by ELIXIR Finland.

Conducting similar research in Finland is possible by combining health and genomic data. The Finnish ELIXIR node, for example, has already started building the secure infrastructure necessary for the management and storage of genomic data.

#### Understanding the emergence of diseases at the molecular level

There are hundreds of times more data on the information contained in DNA available for science than ten years ago. Understanding of how the information stored in the genome is transmitted at the molecular level, for example, to proteins, and further to three-dimensional functional units of cells, is growing at a rapid pace. When human biology is understood from the cellular level to the

molecular level, it improves quality of life and the treatment of diseases.

One of the most important research subjects in bioinformatics is understanding the underlying mechanisms of diseases. The functional unit encoded by a gene is a protein. It is a chain of hundreds of units, or amino acids. There are 20 different amino acids. The protein chain guided by genes becomes a functional unit of the cell, such as an enzyme, only after it has folded into its three-dimensional state and can start interacting with other molecules in the cell. An incorrectly folded protein can lead to illness because it does not function as expected in the network formed by molecules important to life.

Sometimes, for example, there is a change in the genetic code at a critical point for the folding of this criti-

cal functional unit, or protein. Cells self-modify the composition of the proteins created, and thereby their structure and function. This may correct the error in the genetic code. On the other hand, what may also happen is that the protein breaks down in the cell's own process. Most diseases can be traced back to situations where a biochemical reading error has occurred in an important part of the dynamics of the cell's molecular network. On the other hand, this may just be a variation that only results in dietary recommendations to the person in question. The effect of molecular level changes on the data stored in the genome depends on many things, as DNA includes a "backup" of each gene from both parents. There are even several versions of some genes.

Even though the logic and knowledge on the main players in the network of biological processes are pretty much accounted for, the dynamic entity cannot yet be understood, let alone predicted or modified medically, as well as desired. Predicting the risks of contracting coronary artery disease, for example, has become more accurate thanks to the data obtained from the genome, but the understanding of molecular level events is at a stage where the components are known but there is a struggle to understand their interaction or defects occurring at the molecular level. However, molecular level understanding of diseases means more accurate and earlier diagnoses, that preventative measures can be initiated early and that those at risk, for example, may choose to change their lifestyle.

**The Finnish ELIXIR node has started building the secure infrastructure necessary for the management and storage of genomic data.**

#### Good organisation of data facilitates disease research

Analysing data from molecules, cells or whole organisms requires that the data is well organised. The data produced with sequencing, microscopes, mass spectrometry or computer simulations must have common file standards and sufficient machine-readable interfaces to be followed when the data is stored. A good indicator of the degree of data organisation is if another research group can utilise the data as well as its original producers.

# 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 dependencies 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.”

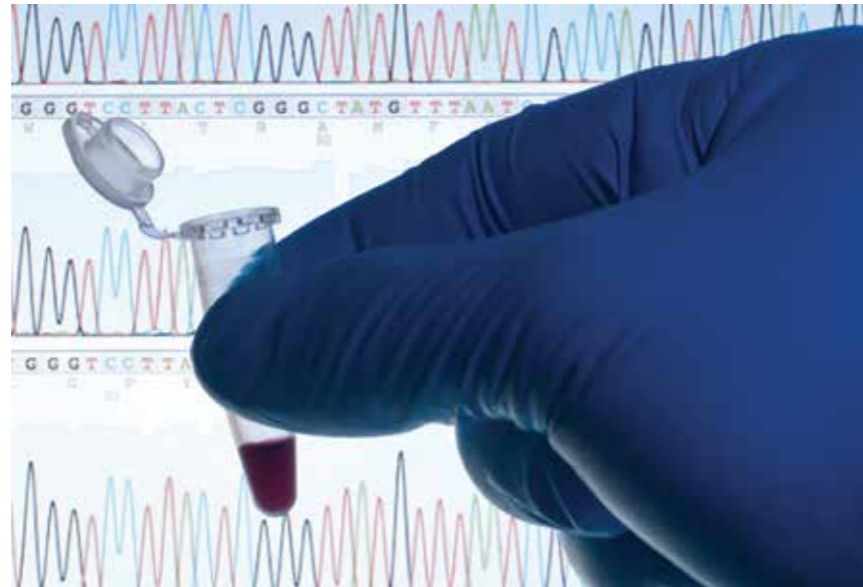
Analysis and interpretation of patient data is demanding work because it involves a lot of legislation and regulation. Blueprint Genetics provides medical doctors with processed information, but the doctors always make the actual decision.

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 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 geneti-



*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.*

cists 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 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.

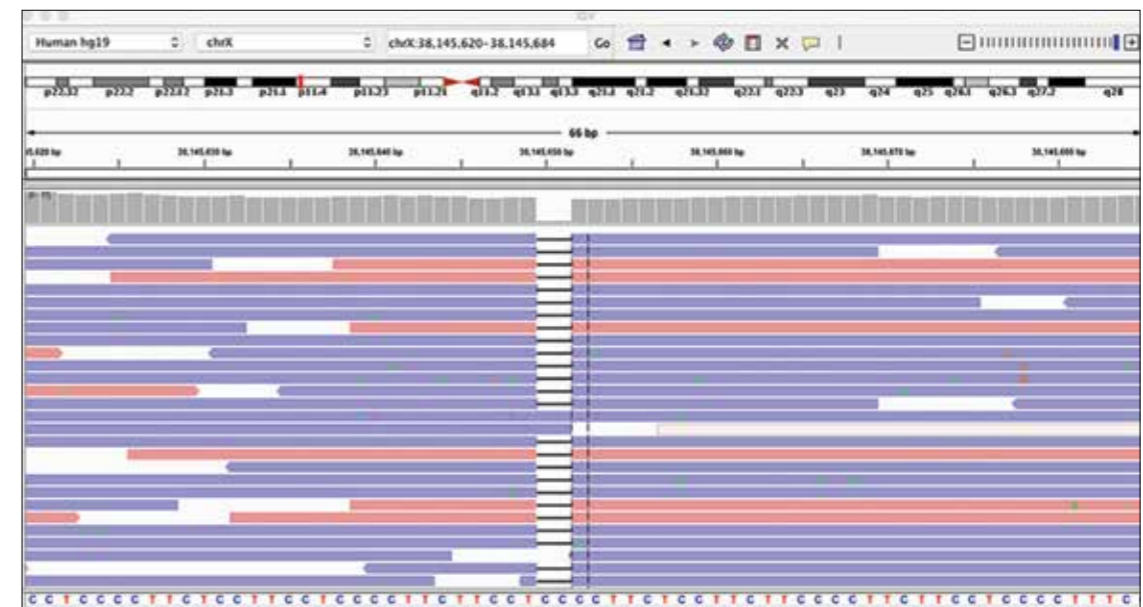
“We produce the DNA data ourselves and can decide what form it is in and which standards it conforms to. However, we utilise guidelines provided by others when interpreting the results.”

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.



*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.*

# Webmicroscope stores tissue samples in the cloud

An invention of the Finnish doctors Johan and Mikael Lundin provides an effective solution for the analysis and storage of tissue section images.

The volume of research data is increasing enormously year after year, requiring a continuously active approach from software developers. It must be possible to analyse large amounts of data with software that does not jam the workstation. Johan Lundin, Research Director at the Institute for Molecular Medicine Finland (FIMM), studies and develops image-based diagnostics using machine vision solutions. In future, it will become possible to produce individualised disease prognoses and treatments by combining various data sources, genetic data, tissue data and clinical patient data. This has been applied especially in the treatment of breast, prostate and colon cancer.

When he was working at the Helsinki University Hospital in the early 2000s, Lundin became frustrated with how difficult it was to process large tissue section images at the workstations. The size of tissue section images is 1–2 gigabytes, so storing them on your own hard drives does not make sense. Rotating the images is also slow. With his brother Mikael, Johan Lundin started thinking about a functional software solution for the problem.

The brothers developed a fully web-based software program, the essential components of which are an efficient image server and a web interface that

works with all browsers. With their compression algorithm, images take up less space and load quickly. A two gigabyte sample image can be compressed to the size of half a gigabyte. The tissue sample is stored in the cloud and large amounts of data can be processed quickly and easily from your own workstation.

The online microscope service can be used with all browsers and tablets, including smartphones. WebMicroscope® is also compatible with the image formats of different microscope manufacturers. WebMicroscope enables the study of very extensive materials and is also ideal for collaborative projects as a joint management and analysis space for digitised images.

“There has been a growing interest in the service. Doctors, researchers and teachers are shifting to digital microscopy. An online cloud-based service is a progressive solution for the users of digital microscopy all over the world”, says Kaisa Helminen, CEO of the service’s provider, Fimmic. Helminen is a trained biochemist and has previously worked for several companies in the bio-industry.

Fimmic was established in 2013 and the commercialisation of the service started the following year. Fimmic’s customers include universities, research institutes, pharmaceutical



companies and companies conducting external quality control. External quality control is enhanced when samples can be sent for analysis virtually instead of mailing samples on glass slides to laboratories.

Fimmic has tested ELIXIR Finland’s cloud service. It offers the WebMicroscope users their own server, a high-speed bandwidth and a massive amount of storage space. This ensures that the service works as efficiently as possible. WebMicroscope is also suitable for biobanks for tissue sample management. The service can be customised to suit a particular biobank.

## Samples stored directly in the customer’s account

Microscope scanners are expensive devices with the price typically varying between 150,000 and 300,000 euros. However, the number of scanners is increasing and, when images are scanned, the most convenient and least expensive solution for many users is to store them directly in the cloud.

“If a customer does not have the opportunity to use a scanner, the sam-

ples can be sent to us for scanning. We will store the digitised samples directly in the customer’s WebMicroscope account”, Helminen says.

Through the WebMicroscope portal, users can share their own microscope images with different research groups and partners around the world. This is an important feature because in drug design, for example, the rapid sharing of test results between research groups and pharmaceutical companies is a prerequisite for breakthroughs. Research related to drug development is one of Fimmic’s focuses.

With a traditional microscope, only a small portion of a sample can be examined at one time. A microscope scanner takes a picture of the sample with a large objective, digitising the entire sample in detail. With WebMicroscope, the resulting image can be viewed easily and quickly, regardless of location.

“You can select a section of the tissue sample to be examined, similarly to Google Maps, and only look at a part of it, quickly moving to another spot. The image is not saved on the workstations,

but is rather loaded over the network directly from the image server.”

All Finnish universities teaching medicine use WebMicroscope for educational purposes at anatomy and pathology courses. WebMicroscope allows digitised samples to be easily shared with students along with other documents and videos.

You can secure your own pages with a password and the software can also be used to complete exams. The virtual samples can be viewed using a tablet or a smartphone in distance education, or on a large screen in the classroom. The application is ideal for multi-touch screens that utilise multiple points of contact. Massive tissue section images can then be viewed easily and quickly on a large touch screen even with a larger group.

## Machine vision under development

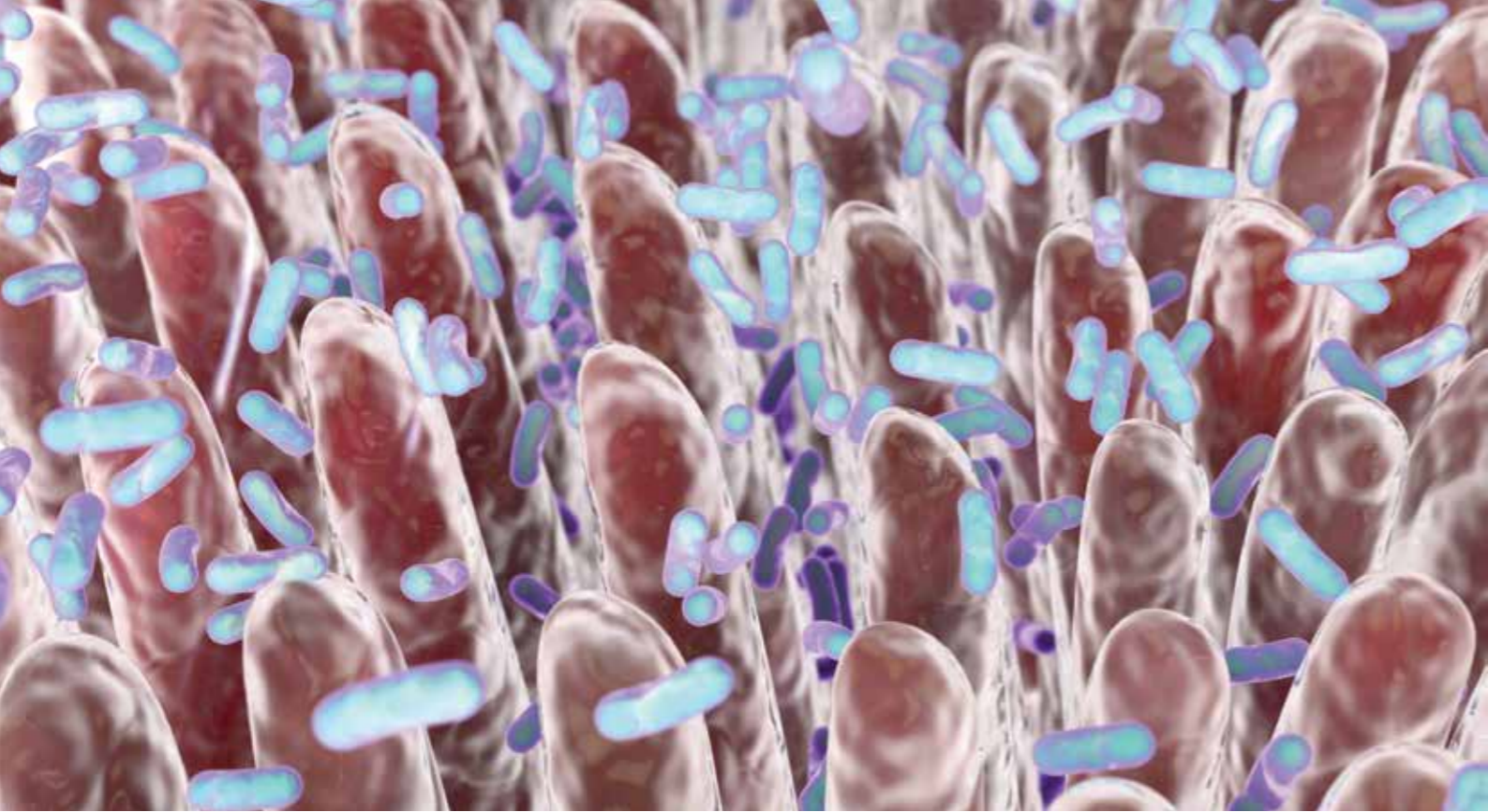
A microscope scanner produces a lot of data. There may be millions of observation points to examine, the processing of which requires computing power and good algorithms. Fimmic plans to

further develop the software and introduce quantitative image analysis tools, algorithms. According to Kaisa Helminen, the number of potential research subjects that algorithms can be used for is huge.

“Machine vision algorithms are based on signal processing. With dozens or even hundreds of images, the machine is taught to identify a particular signal from the background, for example stained cells from other tissue. Screening is case-specific and it varies how different samples have been processed. An algorithm is just as good as it has been taught to be.”

All of this requires computing power obtained, for example, from the supercomputers of CSC – IT Center for Science.

“A lot of computing power is required because the images being studied are so-called whole slide images. Smaller sections of these may, of course, be selected for analysis, but a lot of computing power is still required so that the analysis would not take too much time”, Kaisa Helminen notes.



# Secrets of the intestines

In recent years, intestinal bacteria have become more and more resistant to antibiotics. The bacteria that cause diarrhoea are also increasingly resistant. Antibiotic-resistant bacteria are a global threat. Professor Anu Kantele is interested in knowing what happens in the intestines as people travel to the tropics and back home again.

Regular Finns have been sent to Benin as subjects to test a new diarrhoea vaccine. The ETVAX vaccine is administered as an oral solution, similarly to heartburn medicine. Once the vaccine is on the market, it is planned to be sold to developing countries at a low price.

For tourists, diarrhoea is usually just an unpleasant experience, but it is life-threatening for the children in developing countries. Diarrhoeal disease is the second largest cause of death in children under five years old in the world. Every year, more than 1,7 billion children fall ill with diarrhoea. Of them, more than half a million under 5-year-olds die according to the World Health

Organization (WHO). Diarrhoea is also the main cause of malnutrition, short stature and impaired learning ability in small children.

Diarrhoea is a symptom brought on by disease-causing bacteria, viruses and parasites that have reached the intestines. They generally spread through water and food contaminated by faeces. Diarrhoea is transmitted especially when there is a lack of adequate hygiene and clean water for drinking and household consumption. According to UNICEF, in 2010, up to one fifth of the world's population had to relieve themselves outdoors. Almost 900 million people suffer from a lack of clean drinking water. On trips to the tropics,

the share of those who fall ill with diarrhoea may be up to 60%. Bacteria cause 80% of the cases of traveller's diarrhoea. Enterotoxigenic coliform bacteria (ETEC) are one of the most common causes of severe diarrhoeal disease. That is why there is demand for a vaccine that works against diarrhoea. The pathogens have been studied extensively both in developing countries and with tourists. The information has been used to develop a new vaccine designed to train the human immune system to identify and obliterate pathogens before they can cause symptoms.

The *Escherichia coli* bacteria, or coliform bacteria, normally live in the intestines of humans and animals. More than

700 types of *E. coli* have been identified. They are part of the normal microflora of everyone's intestines and are mainly useful. For instance, they protect us against many disease-causing microbes.

"*E. coli* is the most studied bacterium in the world. It is usually not dangerous, but some are disease-causing. There are several known so-called diarrhoea coliform organisms that cause diarrhoea.

ETEC is one of them. It causes severe watery diarrhoea", says Anu Kantele, Professor of Infectious Diseases from the University of Helsinki.

ETEC differs from other types of coliform bacteria in that it produces two toxins that cause significant fluid secretion from the small intestine, that is, watery diarrhoea.

"The tolerability of the vaccine and the immune defence it elicits are now being studied, while also investigating its efficacy against traveller's diarrhoea. Developing an ETEC vaccine is also one of the goals of the WHO. This so-called ETVAX vaccine generates a good immune response and is the most promising of the current ETEC vaccine candidates."

## Two-year protection

ETEC strains of bacteria were among the first pathogenic organisms for which molecular diagnostics were developed. Vaccine researchers are currently interested in several diarrhoea-causing microbes, such as ETEC bacteria, *Shigella* bacteria and norovirus. Vaccines already exist against other major causes of diarrhoea, such as cholera and typhoid bacteria and rotavirus.

The study is being conducted in collaboration between the universities of Helsinki, Gothenburg and Johns Hopkins and the vaccine manufacturer, Scandinavian Biopharma AB. United Medix Laboratories Ltd. is also involved. The safety of the vaccine has previously been tested on 140 Swedish adults and 450 Bangladeshi children. In both studies, the vaccine and placebo groups had an equal amount of side effects, so the vaccine is considered to be very safe.

The results of the Swedish research group have demonstrated that the ETVAX vaccine elicits a strong im-



900 million people suffer from a lack of clean drinking water. During the NEWAW WASH-project taps were installed in Puware Shikhar, Nepal. Photo: Jim Holmes/AusAID



Anu Kantele at the laboratory in Benin. Photo: Else Kyh  la.

mune response not seen in those who receive the placebo.

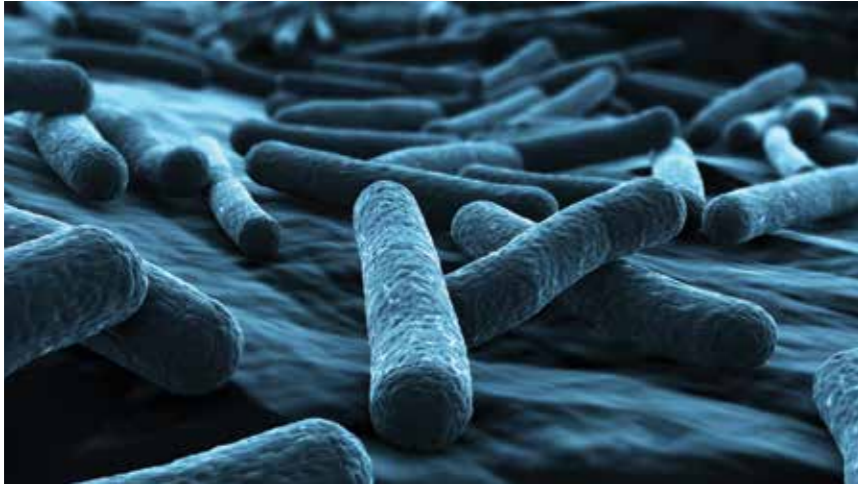
"Among the Bangladeshi children, the vaccine was well-tolerated and the response was good. Its safety was studied previously, and now we are studying the effectiveness of this oral vaccine for the first time", says Anu Kantele.

The vaccine has been estimated to provide protection against moderate to severe ETEC diarrhoea in 60-80% of those vaccinated. Studies on the cholera vaccine have been a great help in the development work for the ETEC vaccine. The pathogenic mechanisms of the ETEC bacteria and cholera bacteria (*Vibrio cholerae*) are very similar. They cause illness by attaching to the surface of the small intestine and producing en-

terotoxins, or intestinal poison, that are responsible for the symptoms of the disease. The toxin kills cells by preventing their protein synthesis. The toxin makes the mucous membrane of the intestine permeable, whereupon a lot of water passes from the tissues to the intestine. This causes very severe watery diarrhoea. The cholera toxin and ETEC bacteria toxins are structurally, functionally and immunologically similar.

ETEC bacteria produce both heat-labile (LT) and heat-stable (ST) toxin. Both have a protein structure and are toxic to humans. ETEC carries a plasmid, a circular DNA molecule that guides the production of the toxin.

"The ETVAX vaccine contains a number of components: killed ETEC



The most common causes of traveller's diarrhoea are the so-called diarrhoea coliform bacteria, of which there are five different types. One of the most important is the enterotoxigenic *Escherichia coli*, or ETEC. It is found in about 20–40% of disease cases. Of the other bacteria, the *Campylobacter*, *Salmonella* and *Shigella* species, for example, may result in very severe symptoms. Their combined share is about 15%.

bacteria, so-called colonisation factors that allow the bacteria to reproduce in the intestine, and detoxified LT toxin and an adjuvant produced from it", says Anu Kantele.

Once the samples of the 800 test subjects who travelled to Benin have been analysed, it will be possible to evaluate, in particular, how an immune system trained by the vaccine works against the ETEC bacteria contracted on the trip and its LT toxin. Half of the subjects have received the vaccine and the other half a placebo.

### What happens in the intestines?

Anu Kantele and her team study the microflora, pathogens and resistant bacteria in the human intestines from stool samples. The participants provide various samples for the study. The analysis of ETEC bacteria and other pathogens requires stool samples. Blood and saliva samples are used to study, for example, the immune response to the ETEC vaccine. Data is also collected on the possible adverse effects of the vaccine. Diarrhoea samples are collected and processed immediately on site in a laboratory in Benin. The number of samples obtained is essentially huge. One gram of human faeces can contain up to one million bacteria.

The researchers compare cultivation-based and molecular laborato-

ry methods used to identify ETEC and other causes of intestinal infections from the stool samples. They analyse the antibodies and genes involved in immune defence.

Anu Kantele has been studying diarrhoeal diseases for a long time.

"I am interested in knowing how new bacteria that arrive in the intestines manage to settle into the ecosystem formed by the native intestinal bacteria and how antibiotic treatment affects this", says Kantele.

Antibiotic-resistant strains of bacteria are most likely to develop in the poor countries of the world. The reason is the excessive use of antibiotics. If the sanitary conditions are inadequate, resistant bacteria spread easily and even from one country to another. According to the WHO, for example, the bacterial strain resistant to fluoroquinolone, which is commonly used to treat urinary tract infections caused by coliform bacteria, is widespread.

"The Benin test subjects will provide a lot of data that can be used to analyse the efficacy of the vaccine. In addition, the intention is to use genetic engineering techniques to analyse the microbes in the stool samples and to investigate the presence of antibiotic-resistant strains. The amount of data is enormous but, by combining data, it is pos-

**"The amount of data is enormous but, by combining data, it is possible to gain new insight into, for example, the spread of antibiotic resistance."**

sible to gain new insight into, for example, the spread of antibiotic resistance."

According to Kantele, the majority of the antibiotics used by tourists are taken for traveller's diarrhoea. The antibiotic shortens the duration of the disease, but it would almost always go away by itself also without antibiotics. Kantele emphasises that the symptoms can be alleviated with drugs that affect the functioning of the intestines without increasing the risk of contracting resistant bacteria.

"Antibiotics facilitate the settlement of resistant bacteria into the intestines, and so one of the best ways to avoid such colonisation is to not take antibiotics. Nowadays, antibiotic treatment is usually recommended only for severe diarrhoea; less severe cases are treated with fluid therapy and possibly drugs that affect the functioning of the intestines. The carriers of antibiotic-resistant bacteria may carry the bacteria to their home country and possibly spread them further there. To reduce the flow of resistant bacteria into the home country, antibiotics should be used with caution in the treatment of traveller's diarrhoea."

Antibiotic-resistant strains can now be quickly identified through genetic engineering techniques, particularly polymerase chain reaction (PCR). Sequencing can be used for even more accurate analysis. The rapid identification of the infection-causing bacteria prior to drug selection is one way to control the use of antibiotics.

One of the studies conducted under Kantele's leadership demonstrated that 80% of the tourists to high-risk areas who fell ill with diarrhoea and took antibiotics brought the ESBL super bacteria with them. ESBL (Extended Spectrum Beta-lactamases) is a special enzyme that breaks down antibiotics and makes the bacteria resistant to many common antibiotics.

The diarrhoea-causing ETEC coliform bacteria may also have the ESBL characteristic.

The premise of Anu Kantele is that we can learn from travel. What interests her is how the microbial activity of human intestines changes while travelling.

"I would like to find out what happens in the intestines of the 800 test subjects during the trip. We are talking about an ecosystem where the strongest bacteria win. It is exciting to combine the data on the changes in the microflora to how the body responds to them, what genes are activated, etc. We learn more about the intestines every day."

### Sequencing and intestinal metagenomics

Diarrhoea-causing bacteria can be detected in a stool sample through cultivation or PCR examination, or a combination thereof. PCR, or polymerase chain reaction, is one of the most important techniques used in molecular biology. It can be used, for example, to amplify a single gene or any segment of DNA multiple times. PCR is performed outside of living cells in a laboratory (*in vitro*) using a special PCR device. With PCR, a very small amount of DNA can be amplified to produce a billion times the amount of the same DNA in a few hours.

PCR technology is used for many purposes, including finding hereditary diseases, identifying individuals using genetic fingerprints, diagnosing infectious diseases and cloning genes. Microbial DNA is isolated from stool samples and amplified. By amplifying different gene areas, it is possible to quickly and efficiently identify pathogens from the stool sample. Microbes are identified by the base pair sequence. Cultivation is necessary alongside PCR because it allows the detection of antibiotic sensitivity.

The majority of the microbes in humans are located in the intestines. More than a thousand different species of bacteria live in the intestines of an adult human. The microflora in the intestines has up to a hundred times more genes than the human genome. 99% of intestinal bacteria are anaerobic, meaning that they grow in the ab-

sence of oxygen. Of the remainder, the most common are *E. coli* bacteria.

The intestinal microbes form their own ecosystem. Microbes have traditionally been studied and cultivated in laboratories. Now, with metagenomics, it is possible to also study them better in their natural habitat, be it soil or the intestines. DNA sequencing is used to try and ascertain the genome of an entire ecosystem. The human genome includes 20,000 genes. However, in addition to these genes, the intestinal bacteria of a single human being encode up to a million genes that affect the regulation of bodily functions. Almost 10 million genes originating from various bacteria have been identified in samples from human intestines. There is great genetic diversity and the amount of data is

enormous. Knowledge of the functions of the genes of the intestinal microflora is still in its infancy.

The metagenomics service of EMBL-EBI is an automated data transfer service (EBI Metagenomics Pipeline) for the analysis and archiving of metagenomic data. There are samples from the human digestive system, soil, water, animals and plants. The data to be studied can be submitted to the service for analysis and comparison.

The service can be used to gain additional information on the evolutionary history of different microbial species as well as the functioning and metabolism of microbes. The data archived by EMBL-EBI is publicly available. EMBL-EBI is part of the ELIXIR infrastructure.



Mono is a border river between Benin and Togo.



400 of the study participants will receive a vaccine containing inactivated ETEC bacteria. The remaining 400 will receive a placebo vaccine. The vaccination trip was arranged with the collaboration of the cultural center Villa Karo.



# Microbes and climate change

Genetic research has revealed that there are a lot more microbes with much more diverse communities than we even knew about. Studying the genetics of microbial communities gave rise to a new branch of life sciences, metagenomics. Jenni Hultman studies the significance of the microflora in Arctic regions on climate change.

**M**icrobes, or microorganisms, are what single-celled organisms or life forms consisting of a few cells are commonly referred to as. These include bacteria, protozoa, viruses and unicellular algae. Although microbes exist everywhere in our environment and even in extreme conditions, their genetic origin and function remain poorly understood. The vast majority of microbes are unknown.

The term metagenome references the idea that a collection of genes picked up and sequenced from the environment could be analysed in a way analogous to the study of the genome of a single species. With metagenomics, it is possible to investigate changes in microflora during the course of various diseases and, after treatment, find new pathogens and obtain information about their function during medication, for example. Metagenomics can also be used to study how microbes affect our environment.

## Arctic microbiology

**Jenni Hultman** is holding a sample that contains tens of thousands of different microbes.

In metagenomics, DNA is isolated from the microbial community. This has been relatively easy when microbes have been studied in the intestines and bodies of water, for example.

The examination of soil is considerably more challenging due to the large number of microbes in a single sample. One sample may include up to 10,000 different species. Since new technologies allow the DNA of different microbes to be isolated from the soil, microbial research is under constant change. New information about organisms as well as the origin of life on Earth is obtained all the time. However, microbial communities are challenging as research subjects. Microbial diversity is very high and microbes also affect each other in ways that are not well known yet.

“Microbes have traditionally been grown in petri dishes. But we are now talking about a huge number to be studied because the research subjects are microbial communities where the different microbes are dependent on other microbes or nutrients. Such communities cannot be grown in dishes. Now, the aim is to sequence the majority of the genes in the soil sample. Even if you find out what the species is, it is also important to know what the genes do. Since up to millions of genes are sequenced from a microbial community, this requires computing capacity”, says Academy Research Fellow Jenni Hultman.

Hultman is particularly interested in the microflora of Arctic regions. As microbes act as decomposers in nature, they may play a significant role in the formation of greenhouse gases, such as carbon dioxide and methane. In the short term, the impact of methane on the greenhouse effect is several dozen times that of carbon dioxide.

**“The microbes in the Arctic environment are not well known. They can have an effect on how the climate and conditions change.”**

“The microbes in the Arctic environment are not well known. They can have an effect on how the climate and conditions change. There are many questions. How is nature adapting to climate change? What do species do when the climate changes?”

The melting of peat bogs under the permafrost especially generates methane emissions. But what is the significance of microbes in this process? That is what Hultman wants to find out.

Hultman, who works at the Department of Food and Environmental Sciences of the University of Helsinki, collects research data on microbes in different parts of the Northern hemi-

sphere. In her research, Hultman analyses soil samples in Lapland, Alaska and Greenland. She is now looking for a survey site in Siberia, after which the samples she has collected would well represent the entire Northern hemisphere.

“20% of the Earth’s land surface is covered by permafrost. Within the permafrost are huge stores of carbon dioxide. The melting of permafrost may release the highest amounts of carbon dioxide ever measured into the atmosphere. This process is dependent on a microbial response but, at present, we know rather little about the activity of microbes under permafrost.”

**Data for climate models**

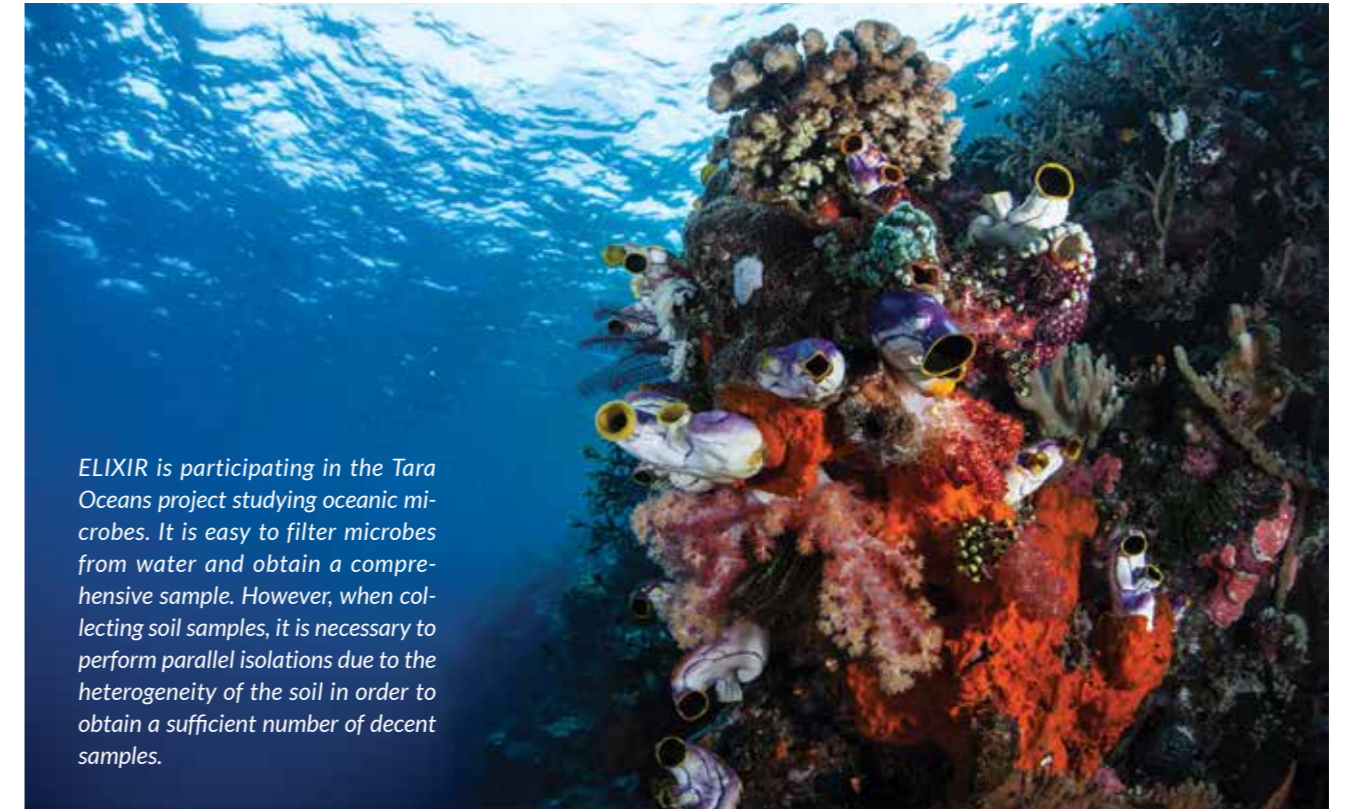
ELIXIR is participating in the Tara Oceans project studying oceanic microbes. It is easy to filter microbes from water and obtain a comprehensive sample. However, when collecting soil samples, it is necessary to perform parallel isolations due to the heterogeneity of the soil in order to obtain a sufficient number of decent samples.

Hultman is interested in the activity of microbial communities and especially in what the genes of the microbial communities do (metagenomics) and how active the genes of the communities are at a given time (metatranscriptomics).

Hultman isolates the total DNA and RNA from the soil samples of the field area in Kilpisjärvi, Lapland, divides them into smaller pieces and sequences them. She isolates the DNA and RNA from samples of 0.5 grams. The number of sampling points is over a hundred. The area has a microclimate, allowing Hultman to take into account various factors, such as humidity, pH and temperature. This makes it possible to study the significance of the activity of microbial communities on climate change on the scale of “mini climate change.”

“A high number of parallel samples weighing half a gram are needed because the microflora of the soil is diverse and because the soil itself varies greatly. Microbes can be present in stone, a dead worm, the root of a plant or just in a place that is more humid than another. So there is a lot to dig up and isolate.”

The essential thing is to know what the genes of the microbes are actively do-



*ELIXIR is participating in the Tara Oceans project studying oceanic microbes. It is easy to filter microbes from water and obtain a comprehensive sample. However, when collecting soil samples, it is necessary to perform parallel isolations due to the heterogeneity of the soil in order to obtain a sufficient number of decent samples.*

ing and how they affect climate change.

“I am studying what is happening in the soil sample at this moment. Which genes are active? Are some microbes accelerating climate change and some slowing it down? Do microbes just produce methane or do they utilise it?”

One important goal of Hultman’s research is to produce data obtained from metagenomics also for climate models. This may potentially improve the reliability of climate models.

**Only 1% can be made to grow in laboratories**

One gram of soil may contain up to ten billion different microbes. When microbial ecology research truly started in the late 1970s and microbial samples from the environment were compared with cultured microbial samples, it was found that the samples from the environment contained up to 99% more new and unknown microbes than the cultured samples.

Traditionally, the sequencing of genes is started by growing cells in a petri dish. When DNA from the cells is placed in a DNA sequencer, it determines the order of the DNA base pairs: adenine, guanine, cytosine and

thymine. However, early metagenomic studies revealed that there are large groups of micro-organisms that cannot be grown in laboratories and that, therefore, cannot be sequenced.

The early studies focused on the sequences produced by the 16S rRNA gene. The function of the 16S rRNA gene, which is found in all living creatures, is to produce the ribosomes in which protein synthesis occurs. In 1977, microbiologist **Carl Woese** started the sequencing of this gene when studying microbes. Because the gene is always slightly different in different microbes, Woese noticed that it can be used to study the development history of the microflora in the samples. However, Woese and his colleague **George E. Fox** were surprised when many of the isolated 16S rRNA sequences did not belong to any known species. The discoveries made with the 16S rRNA gene revolutionised microbial research.

Woese and Fox observed that the samples also contained unicellular, but anucleate microorganisms that externally resembled bacteria but were not. They called this group the archaea.

Archaea are involved in metabolism and affect the functioning of enzymes.

Archaea were initially observed only in extreme conditions, such as hot springs and salt lakes, but have since also been found in different soil types, marshlands, oceans and even human intestines, for example.

Organisms could thus be divided into three categories. Eukaryota, i.e. multicellular plants, fungi and animals, have nuclei. Bacteria and archaea, in turn, are anucleate microbes that make up most of the world’s biodiversity.

“As the sequencing of DNA is becoming cheaper all the time, metagenomics allows microbes to be studied on a much larger scale and in more detail than before”, Jenni Hultman says.

The mysterious archaea may play a greater role in the formation of methane than has previously been known. Some archaea break down organic carbon into methane. But how many of such archaea are there and how effective are they as decomposers?

The data on the secrets of the microbiome collected by Jenni Hultman and other researchers is stored in the public information resources maintained by ELIXIR, the European life sciences infrastructure for bioinformatics.



*Jenni Hultman is holding a sample that contains tens of thousands of different microbes.*





# Pups and pooches behind genetic discoveries

Would you have thought that the beloved tail-wagging pet resting on your couch could serve as a source for human genetic discoveries? Few people know or even come to think that the genome and diseases of dogs are 95% the same as those of humans. The genetic research conducted by Professor Hannes Lohi at the University of Helsinki brings forwards significant information regarding the eye, bone and neurological diseases of both dogs and humans.

The "Eureka!" moment occurred about ten years ago when research fellow **Hannes Lohi** pinpointed the epilepsy gene of miniature dachshunds with his research group in Toronto. At the same time elsewhere, the gene was also found in humans. This coincidence was the starting point of the cross-disciplinary canine genetic research led by the professor at the Faculty of Veterinary Medicine and Medicine of the University and Helsinki and the Folkhälsan Research Center. Since 2006, DNA samples from over 70,000 Finnish dogs have been collected in the DNA bank established by Lohi.



"Dog breeds provide a genetically excellent structure especially for behavioural studies and research into canine and human diseases in general. What animal species is socially gifted, shares the same environment and is exposed to the same pathogens other than man's best friend?"

Lohi noted that inbreeding within dog breeds, in particular, facilitates the identification of disease genes.

"It is easier to discover genes from bloodlines using smaller study cohorts. Compared with the mice and rats typically used in studies, dogs are closer to humans also in terms of vital functions due to their size", Lohi says.

If genes do not provide enough challenges, Professor Hannes Lohi also looks for them in his interest in epigenetic research on dog behaviour. His research group identified the **LGI2** gene depicting transient epilepsy in Lagotto dogs, providing a significant new perspective also for human childhood epilepsy research.

## Partially developed drugs for further development

The spectrum of the canine genetic research led by Lohi is extensive. The subjects include eye diseases, autoimmune diseases, neurological diseases as well as skeletal muscle diseases. The group has identified several new disease genes in dogs from factors causing, for example, epilepsy, dwarfism and anxiety disorders. With the genetic areas found, conditions such as anxiety disorders, from which about 5% of the human popula-

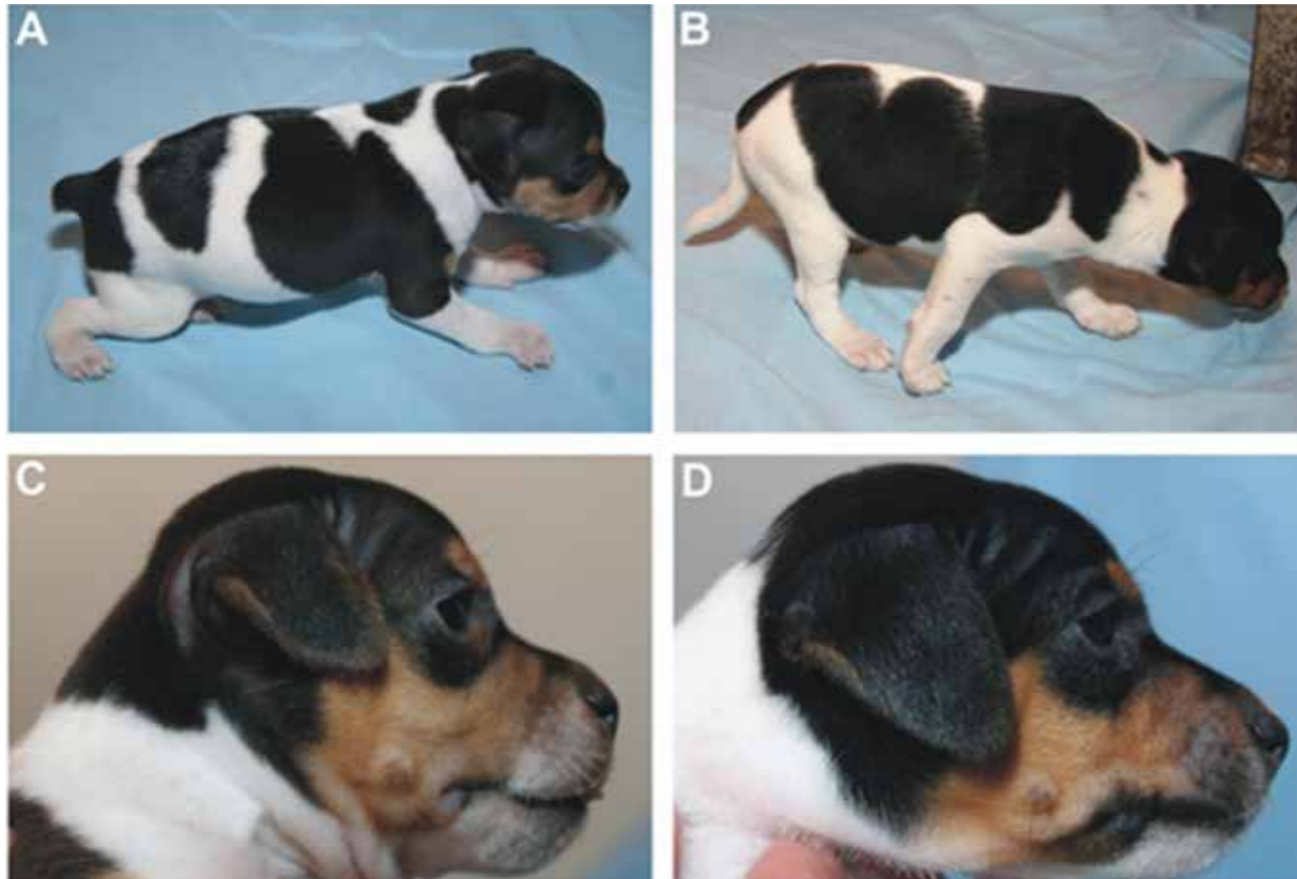
tion will suffer at some point during their life, gain a new research basis for the study of, for example, the genetic background and environmental factors in obsessive-compulsive behaviours.

"We look for the gene causing a disease in a dog breed and, at the same time, the breed provides a canine model for identifying the disease mechanism of human diseases", Lohi says, describing the benefits of the research.

The group identified the **CNGB1** gene that causes retinal degeneration and, at worst, blindness in Papillon dogs. The same gene has been found in human patients. One in ten people over the age of 65 suffers from this disease during retirement. The condition involves blind spots that limit the area of sharp vision, preventing the renewal of a driving licence, for example.

"With the further development of partially developed drugs, the degeneration of the human retina could be treated externally with gene therapy, for example, by applying to the retina a cream containing viruses carrying normal gene copies that would correct the functioning of the cells and may correct vision", Lohi describes the possibilities.

"After identifying the gene, it becomes possible to study the disease mechanism and make comparisons between humans and dogs. The gene may not always be the same in humans and the mutation can be located elsewhere, in another gene of the cell pathway. Understanding the gene function and disease mechanism are prerequisites for



The genetic research group led by Professor Hannes Lohi has studied, for example, the factors of a life-threatening bone disease in Brazilian Terrier puppies. In collaboration with the group of Docent Kirsi Sainio, the group figured out that the disease is caused by a genetic defect in the GUSB gene. Dysfunctional behaviour in the GUSB gene has previously been linked to an accumulation disease causing severe bone changes in humans (mucopolysaccharidosis type VII, MPS VII). A Brazilian Terrier puppy with MPS VII (A and C) has hypermobile joints, bone changes in the limbs and the typical round skull and short muzzle. A healthy litter sibling in images B and D. Sick puppies lag clearly behind in terms of growth, being approx. 35% smaller than their healthy litter siblings at the age of three weeks.

inventing treatments for the disease. On the other hand, when a mutation is found, it is possible to develop a genetic test for dogs and see which dogs are carriers of the disease. This allows dog breeders to quickly benefit from the research”, Lohi says.

He is involved in Genoscooper Laboratories Ltd., a company that, under his leadership, has built a unique and affordable genome-wide genetic test for dogs, MyDogDNA, which tests the dog’s carrier status for over 100 diseases and traits in one go, as well as genomic diversity and structure.

“The genetic diversity of dogs has been weakened by breeding. The number of dogs carrying disease genes has increased, and because many diseases arise in adulthood, sick dogs will have

already been used for breeding. To counter the negatives, breeding may lead to the gene causing the disease becoming more common in a particular dog breed. The candidate gene is more easily identified in dogs than in humans and with fewer samples.”

#### Goal: separate databases for dogs and cats

A large number of veterinarians and dog lovers around Finland have not been enthusiastic about participating in a DNA sampling effort for the benefit of a passing project. The aim of the research group is to build a separate, extensive sequence and variant database for Finnish dogs and cats, similar to the one that already exists for humans (1000 Genomes).

“Genetic research has always been the flagship of Finnish science. We have uniquely accurate health information on patients, including family trees. Equivalent pedigree databases and health data are available on dogs, and soon also on cats. Few countries have such a good, centralised system”, Lohi says.

“There are 400 breeds of dogs. At present, a total of 700 diseases have been depicted in dogs and more are found all the time. The aim is to have a database with the entire genome of each breed sequenced. This will speed up genetic discoveries”, Lohi says.

#### Dog families to be sequenced in the future

Lohi believes that the benefit of a large sequence database is a kind of consen-

sus. This is achieved once hundreds or thousands of genomes have been sequenced and the large number of variants can be accurately mapped. There may be many diseases in the same breed.

“For example, if the genomes of 1,000 dogs from 50 breeds have been sequenced into the database, it will include an estimated 25 million variants from the different breeds. The database will facilitate future projects in that a small family of dogs or cats can be studied with just a few of the individual animals sequenced to provide a sufficiently reliable result on the correct disease variant. The variants of a dog patient are compared with the variants of the thousand samples in the database and, if a particular variant is found in the patient but not in the reference samples in the database, it can be inferred to be disease-causing. After this, the matter is confirmed using a larger file.”

“An efficient and nationally significant database will help us catch disease genes faster. As things are now, you have to do a lot of work in research to obtain a sufficient picture of the location of a variant in the chromosomes. Going forward, a sample will be taken, the entire genome sequenced and compared directly with the variants in the database.”

#### Computing resources for sequence description methods and tools

It is estimated that new biotechnical methods will produce a million times the amount of data produced today by 2020. Lohi states that large amounts of computing resources are needed for both the methods and tools used in research.

“Before, short sections of the genome were sequenced. Now, genome lists are so long that managing them manually is completely impossible. If 200 dogs are studied and the entire genome, i.e. 39 pairs of chromosomes, is sequenced from each dog, the analysis would take several months with the traditional method. A single genome affords hundreds of gigabytes of raw data.”

“As we have shifted from the traditional Sanger method of sequencing to Next-Generation Sequencing (NGS) of the entire genome, huge quantities

of data are being analysed using new methods. The genome is first split into sections in the database, sequenced and assembled. The sequencing of a genome involves the processing of three billion pairs of genes for humans and 2,5 billion pairs for dogs as well as different variants and insertions that complicate the interpretation of the sequence”, Lohi says, describing the challenges of the research data.

“After the variants have been identified, it is examined whether the variant is pathogenic. Computing resources are required at this stage, too. Bioinformatics tools can be used to predict which amino acid change the variant causes in the genome. After that, the effects of the amino acid change are studied more closely, switching to use protein-level tools and various algorithms.”

The research group pinpointed the gene causing retinal degenera-

tion in Papillon dogs with six sick and 14 control animals. The genetic defect was identified using exome sequencing technology, analysing all of the protein-coding areas at once. Many disease-causing mutations are located in the exome, even though it only accounts for 1.5% of the genome. This technology, which is used especially to find disease forms present in the database, led to the identification of a mutation carried by almost one in five Papillon dogs in their genome.

Lohi’s research group participated as a pilot organisation in a project of CSC – IT Center for Science exploring what kinds of materials are created for researchers with extensive computing capacity and memory space. The aim of the project was to pilot models and solutions for the kinds of resources needed by researchers in the ELIXIR research infrastructure.



“It is easier to discover genes from bloodlines using smaller study cohorts. Compared with the mice and rats typically used in studies, dogs are closer to humans also in terms of vital functions due to their size.”

# Saimaa ringed seal aids the study of population genomes

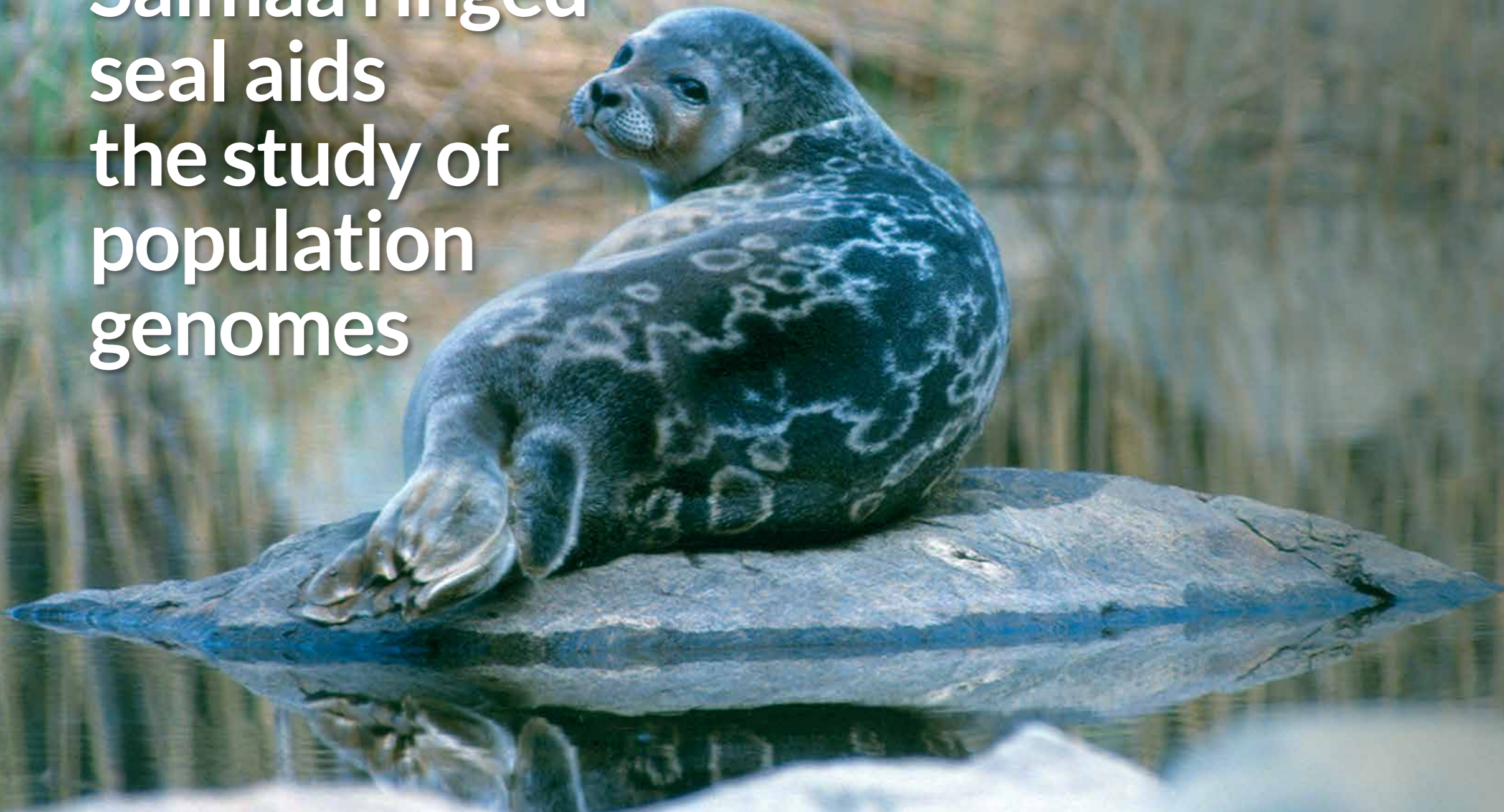


Foto: Suomen Luonnonsuojeluliitto/Juha Taskinen

WWF, Finland

The research groups of Jukka Jernvall and Petri Auvinen at the Institute of Biotechnology are investigating the genomes of different species and the structures of populations. The objective is to understand when the species arose and diverged from one another. The groups are particularly interested in the Saimaa ringed seal, whose full genome will be determined.

The Saimaa ringed seal is an excellent research subject for the study of genetic diversity, isolation and inbreeding. The Saimaa ringed seal has not been in contact with other seal species in more than ten thousand years. Its

eyes, brain and skull are different from those of other types of ringed seal. The Saimaa ringed seal developed from a seal population that probably came from the Baltic Sea to Lake Ladoga before moving to the Saimaa archipelago.

“If a Ladoga ringed seal was transferred to Lake Saimaa, it might not survive. The Saimaa ringed seal has adapted to the murky waters containing humus and the maze-like archipelago”, says **Petri Auvinen**, Laboratory

Director of the Institute of Biotechnology.

The DNA Sequencing and Genomics Laboratory of the Institute of Biotechnology specialises in gene sequencing, or determining the order of base pairs in DNA. The laboratory has sequenced the entire genome of several organisms, starting from the *Lactococcus piscium* bacteria that spoil cold food. Gene expression is also studied at the laboratory through sequencing. Key events in the evolution of organisms include cellu-

lar division and differentiation, which are highly temporally and spatially regulated.

Cellular differentiation takes place in stages. Sometimes a gene is switched on and sometimes it stops functioning. This active functioning is called gene expression. When gene expression can be measured, it is possible, for example, to monitor which genes start to function when a tree prepares for winter. The EST (Expressed Sequence Tag) technology provides information on the location and function of a gene. By identify-

ing the base pair sequence of genes, a tag can be provided for each expressed gene. Currently, the RNA-Seq method is mainly used to study gene function.

## Reference genome for population research

The aim of the researchers of the Institute of Biotechnology at the University of Helsinki is to obtain a reference genome of the highest quality possible from the Saimaa ringed seal. A reference genome is a digital sequence database on the full base pair sequence of a single species, compiled from one individual in the case of the Saimaa ringed seal and from numerous genomes in the case of humans. Collecting a good reference genome requires the use of various advanced technologies.

The reference genome and deviations in individual genomes enable the efficient study of the population. In the STR (Short Tandem Repeat) method, a specific locus on DNA where a few base pairs in a row are always repeated is compared with two or more DNA samples. The DNAs of individuals are clearly distinguished with STR. Mitochondrial DNA, in turn, can be used to trace the maternal lineage of individuals back thousands of years. The rapid development of DNA sequencing technologies has enabled the identification of single-nucleotide polymorphisms (SNP), providing a very accurate estimate on the differences between individuals. This method is also used in the Saimaa ringed seal genome project. The data collection requires a lot of storage space and computing power, provided by CSC – IT Center for Science via the ELIXIR infrastructure.

“Studying the genetic history of the Saimaa ringed seal is also helpful for human genome research.”

The genome of the Saimaa ringed seal is 2,5 billion base pairs in length, the same as the canine genome. In determining the genome of the Saimaa ringed seal, the group of Academy Professor **Jukka Jernvall** focuses on studying the teeth of seals while the group of Petri Auvinen focuses on population history and genome structure. Once

**“It is now possible to examine the impact of disease genes on population structure and the bottlenecks caused by nature and humans.”**



the genome has been determined, the genome of the Saimaa ringed seal will be compared with the genomes of the ringed seals in Lake Ladoga, the Baltic Sea and the Arctic Ocean.

The researchers are collecting data on the connections between the genotype (genetic factors) and phenotype (environmental factors) together with researchers from the universities of Oulu and Eastern Finland. A lot of data on developmental biology is obtained by analysing teeth. Once a tooth erupts, it will no longer develop or change due to the environment. However, there is huge variation in teeth. That is why it is studied as to which genes have affected unusual teeth. The teeth of the crabeater seal, for example, have become very polymorphic due to evolution and function like the baleen of whales because the seals eat krill.

“We have computer models of all ringed seal skulls. We can create accu-

rate phenotypes and look for the probable genes that caused a particular tooth. Gene function can be modelled on a computer and analyse which areas of the genome could affect the tooth.”

A different skull or teeth indicate adaptation or specification, adjustment to different conditions. Because the orbits of the Saimaa ringed seal are different from those of other, even closely related ringed seals, it can be concluded that, for example, it has adapted to murky and maze-like waters.

The groups of Auvinen and Jernvall have access to the DNA of the only known hybrid between a ringed seal and a grey seal in the world. In 1929, Skansen Zoo was the birthplace of a cub from whose tooth Auvinen managed to isolate DNA. The offspring of a huge grey seal and a small ringed seal only lived for a short time. The teeth and skull of the hybrid indicate an inter-

mediate form. According to Auvinen, it would probably be the equivalent of a hybrid between a chimpanzee and a human. It is now possible to compare why a specific kind of tooth or skull develops.

Auvinen considers this significant research also for human evolution because it is not known when modern humans differentiated into their own species. Hybrids have also occurred during human evolution. There have been findings of human skull fragments that are a cross between Cro-Magnon and Neanderthal. 2–5% of Europeans carry genes passed down from Neanderthals. Furthermore, the skeleton of a human subspecies named the Denisovan was found in the Denisova Cave in Siberia. It became extinct 40,000 years ago, earlier than its cousin, the Neanderthal. When DNA was isolated from the finger of the Denisovan’s skeleton, it was found that Tibetans have Denisovan

genes. One hereditary gene helps Tibetans survive in a high altitude climate.

### **Bottlenecks relate an interesting genetic history**

The researchers of the Institute of Biotechnology want to find out whether the Saimaa ringed seal is its own species or a subspecies. The researchers know exactly for how many generations the ringed seal has been isolated in Lake Saimaa. The population of the Saimaa ringed seal is small. There were only 140 individuals left in the 1980s, now the number is 320. By comparing the samples from Lake Saimaa, the Baltic Sea and Lake Ladoga to the reference genome of the Saimaa ringed seal, it is possible to study what kind of a population has passed through a so-called bottleneck.

Nowadays, there are also computational methods that make it possible

to determine reasonably accurately, even from a single genome, the kind of a population its ancestors have lived in. The bottleneck phenomenon faced by a population refers to an event where a large part of the population is destroyed or only a small number of individuals establish a new group, such as the people who once arrived in Finland. The reason behind the destruction may be changes in the environment or a transition to a new environment, which can prevent reproduction.

Studying the genetic history of the Saimaa ringed seal is also helpful for human genome research. Bottlenecks can increase inbreeding and thus also affect the disease heritage. In Finland, bottlenecks have given rise to about forty hereditary diseases in the population that are much common here than anywhere else. Finnish genetic bottlenecks have included the adoption of agriculture

4,000 years ago and the spread of settlements to the northern and eastern Finland in the 16th century.

“It is now possible to examine the impact of disease genes on population structure and the bottlenecks caused by nature and humans. The Finnish disease heritage is interesting in this respect. It can be determined what the disease heritage carried by Finns was like when they went through a bottleneck”, says Auvinen.

### **Data can be reused**

Creating a reference genome comes with many benefits. The reference genome data can always be reused. The better the reference genome is, the easier it is to analyse new data that can be compared to the data of the reference genome.

For example, analysing the reference genome of birch accelerates and enhances wood research for the needs of industry and medicine. New properties that affect the quality and quantity of wood can be looked for in the birch genome. This data can also be utilised in research on other wood species.

“Unlike birch, it will take 10 years to determine the properties of, for example, poplar and eucalyptus. Birch can be genetically modified. Since birch can be made to bloom up to three times a year, new properties can be introduced to birch in one to two years. These techniques can also be applied to other wood species. The genetic model of birch can be used, for example, in the study of eucalyptus”, says Petri Auvinen.

The birch reference genome project was also followed by industry representatives. Thanks to genetic data, birch properties can be refined and the forest industry can use the wood for purposes other than timber.

New applications include nanomaterials, wood processing industry side streams and, for example, hemicellulose. Auvinen also mentions the betulin in birch bark that has been reported to have anti-cancer and even antiviral effects. Betulin has already been used to create medicinal creams. Striving to produce birches with more betulinic acid using conventional breeding methods is also a possibility.

The background features a close-up of vibrant green leaves with prominent veins. Overlaid on this is a complex network of thin, light blue lines connecting various points, resembling a data network or a molecular structure. The overall aesthetic is clean, modern, and tech-oriented, with a focus on nature and data integration.

# Better harvests on the horizon?

**Data will also be collected in the future**

Plant growth and physiology are analysed with imaging methods, generating enormous amounts of data on the genomic and environmental response of plants. The aim of this is to improve the productivity of crops, allowing food and raw materials to be produced for the growing human race in an ecologically sustainable manner.

In NaPPI, a joint infrastructure of the universities of Helsinki and Eastern Finland, plants are measured and analysed automatically. The operation of the infrastructure and the data produced by it can be organised from the outset so that it is also compatible for the use of other European research organisations. This is a good goal because, until now, every laboratory around the world has collected data on the genome, phenotypes and environmental factors of plants in their own way.

The Viikki Plant Science Centre (ViPS) of the University of Helsinki is a research cluster with 36 groups studying plants. The research topics range from adaptation to a particular habitat and climate change to plant stress tolerance and plant breeding.

The activities of NaPPI (National Plant Phenotyping Infrastructure) focus on plant research and breeding. The aim is to produce comprehensive phenotypic data from a large number of plants. NaPPI provides the technical possibilities to combine the information on plant genomes to phenotypic data.

The phenotype of a plant is jointly produced by genes and the environment. The phenotype can take a very different shape due to the impact of the environment. Plants have a much wider capacity for non-hereditary variation than animals. Plant growth, for example, can be effectively influenced in various ways, including nutrients and light.

People have been cultivating plants for thousands of years due to a desire for better food. This has been done locally, and the information collected on plants has not been recorded systematically. A good example is the numerous varieties of grape, with more than a thousand in Europe alone. The origin of all the varieties is no longer known and that is why the origin is being investigated through genetic engineering.

“The data on plant phenotypes has not yet been standardised. Various research groups have been producing and categorising it in their own laboratories”, says **Kristiina Himanen**, Research Coordinator of the NaPPI infrastructure from the University of Helsinki.

The plants in front of Kristiina Himanen are about to enter a phytoscope. A phytoscope is an imaging device that analyses plant growth and physiology. The plants are measured and images are taken of them automatically, after which the computer calculates the height, width and, for example, the surface area and shape of the rosette based on the images.

### Studying plant architecture is important

The aim of the NaPPI infrastructure is to enhance and specify the collection and analysis of the data from plants with new imaging technologies. The infrastructure uses imaging devices that analyse plant growth and physiology. The plants are measured and images are taken of them automatically, after which the computer calculates the height, width and, for example, the surface area and shape of the rosette based on the images.

“The size, growth and form of a plant, i.e. plant architecture, are important in agricultural production,” Himanen emphasises.



Turnip rape plots in Viikki. Researchers are investigating whether the dwarf gene can increase the productivity of turnip rape by reducing the biomass of the stem in relation to the seed yield produced by the plant.



The plants in front of Kristiina Himanen are about to enter a phytoscope. A phytoscope is an imaging device that analyses plant growth and physiology. The plants are measured and images are taken of them automatically, after which the computer calculates the height, width and, for example, the surface area and shape of the rosette based on the images.

“Plant architecture can affect the yield or cultivation characteristics. As dwarf varieties of rice have been produced, they do not become lodged as easily anymore, and this affects the harvest. Genes can influence plant architecture and hence the quantity and quality of the harvest.”

What happens when a dwarf gene is fed into the genome of turnip rape is being studied in Viikki. **Tarja Niemelä**, PhD (Agriculture and Forestry), and partners are investigating whether the dwarf gene can increase the productivity of turnip rape by reducing the biomass of the stem in relation to the seed yield produced by the plant.

“There is a huge amount of genomic data available, but you have to be able to combine it with other data. We want to link the phenotypic data that we produce with imaging devices to genomic data. Ultimately, of course, we are interested in how the information obtained from genomes and phenotypes can be transferred to plant breeding.”

According to Himanen, the volume of plant research will increase thanks to new imaging methods.

### Spectral and fluorescence imaging produce a lot of data

Researchers are investigating whether the dwarf gene can increase the productivity of turnip rape by reducing the biomass of the stem in relation to the seed yield produced by the plant.

In addition to plant forms, the NaPPI infrastructure equipment is also used to analyse the physiological state of plants. The Spectromics Laboratory located at the Joensuu Campus of the University of Eastern Finland is the first research environment in Finland that focuses on the spectral imaging of plants and other biological samples. Spectral imaging consists of images taken at different wavelengths of light with their own colour channels. The Spectromics Laboratory is developing optical methods especially for the study of plant stress responses.

The human eye or a conventional camera sees colours as combinations of three wavelength bands (red, green and blue). With a spectral camera, however, it is possible to detect up to hundreds of different wavelength bands. It is also not limited to visible light, but is capable of taking images in the ultraviolet and infrared ranges. A separate image may be formed of each band and each pixel contains a complete spectrum.

“Spectral imaging enables very precise separation of colours, but it also multiplies the amount of data produced”, says Professor **Markku Keinänen** from the University of Eastern Finland.

“This, in turn, requires complex computational approaches in image analysis. So spectral imaging is, to a large extent, computation, and the images illustrating the results are not produced until the final stages of the analysis.”

When plants are analysed with thermal and fluorescence cameras, you can see things that are not visible in ordinary light. Fluorescence is visible light of a certain colour that is generated when

“Genes can influence plant architecture and hence the quantity and quality of the harvest.”

the atoms of a plant are excited due to, for example, invisible ultraviolet radiation. Thermal and fluorescence cameras can be used to calculate, one pixel at a time, the size of an area of a different colour in the plant and to study, for example, infections in the plant.

### Standardisation of data reduces redundant work

The Finnish ELIXIR node offers efficient capacity for the processing and storage of data. Since the data collection of phenotypes has been automated and digitalised, according to Kristiina Himanen, it is now possible to also start the standardisation of data.

“Data must have the same format. The Excelerate project is developing standards for phenotypic data and metadata. There are 22 countries involved. Although everyone has their own infrastructures, their operations are now being harmonised.”

In practice, researchers have access to information about the plant's genome and phenotypic data on growth conditions and other environmental factors. Once both data sources have been combined, it becomes possible to create comprehensive databases and the laboratories across Europe can avoid doing redundant work and divide data collection in a sensible way.

“The introduction of a single gene in plant breeding will become easier because the amount of work involved in the analysis of a single plant will become more reasonable.”

Going forward, the Viikki research groups will produce image-based data to which genomic data is linked. The Finnish ELIXIR node, in turn, is figuring out how to analyse and standardise the data and how to hand over the metadata to ELIXIR for a cloud database. The division of labour between the NaPPI infrastructure and the Finnish ELIXIR node CSC is a good example of how genotype and phenotype data on plants should be produced for research.

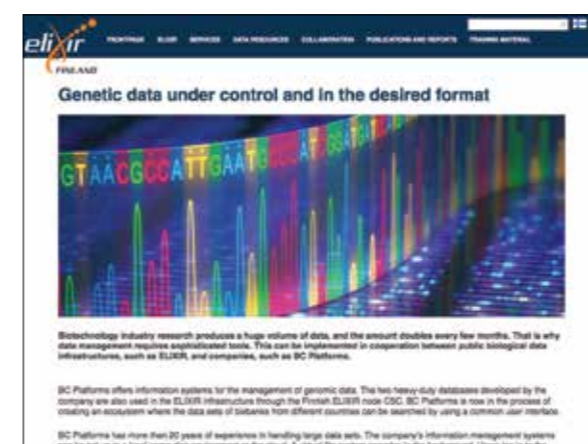


A phytoscope is an imaging device that analyses plant growth and physiology. The plants are measured and images are taken of them automatically, after which the computer calculates the height, width and, for example, the surface area and shape of the rosette based on the images.

Read more



ELIXIR Finland provides authentication and authorisation infrastructure (ELIXIR AAI), service to manage data access applications and access rights to sensitive datasets (AAI-REMS), analysis software for gene data (Chipster), cloud services (CSC Cloud) and training.



ELIXIR node in Finland is CSC – IT Center for Science



CSC – The Finnish 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>



[www.elixir-finland.org](http://www.elixir-finland.org)



Contact us at  
[info@elixir-europe.org](mailto:info@elixir-europe.org)



Ministry of  
Education  
and Culture



ELIXIR receives funding from the European Commission within the Research Infrastructures Programme of Horizon 2020.

[www.elixir-europe.org](http://www.elixir-europe.org)