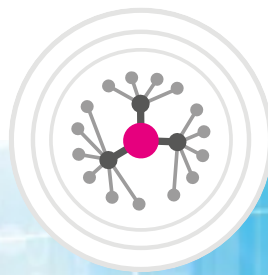


# DATA ANALYSIS FOR INSIGHTS INTO COMPLEX BIOLOGICAL SYSTEMS

Highlights from the German Network  
for Bioinformatics Infrastructure





# DEAR READERS

The generation of big data is one of the hallmarks of life sciences today. The German Network for Bioinformatics Infrastructure (de.NBI) was established five years ago with the goal to support researchers in the analysis of large amounts of data. The network provides services, training and computing capacities for the analysis of such vast quantities of data.



Prof. Dr. Andreas Tauch (left), Prof. Dr. Alfred Pühler (right)

The de.NBI network, funded by the German Federal Ministry of Education and Research (BMBF), consists of a large number of individual projects topically organised in eight service centres. Since March 2020, the network has been celebrating its fifth anniversary. To mark the occasion, this anniversary brochure was published to provide information on the network's activities. Particular emphasis was placed on application-oriented aspects from the areas of plants, microbes and medicine. The brochure is intended to help make the topics covered by the network accessible to a wider audience. You will be surprised at the diversity of our topics!

In addition to introducing the network, we also recorded an interview with the de.NBI coordinator and the head of the administration office. This interview deals with the structure and organisation of the network as well as the many activities that have been launched in the meantime. Finally, we report about the de.NBI network's various fields of activity - starting with the aspects of service and training followed by the de.NBI cloud and industrial forum.

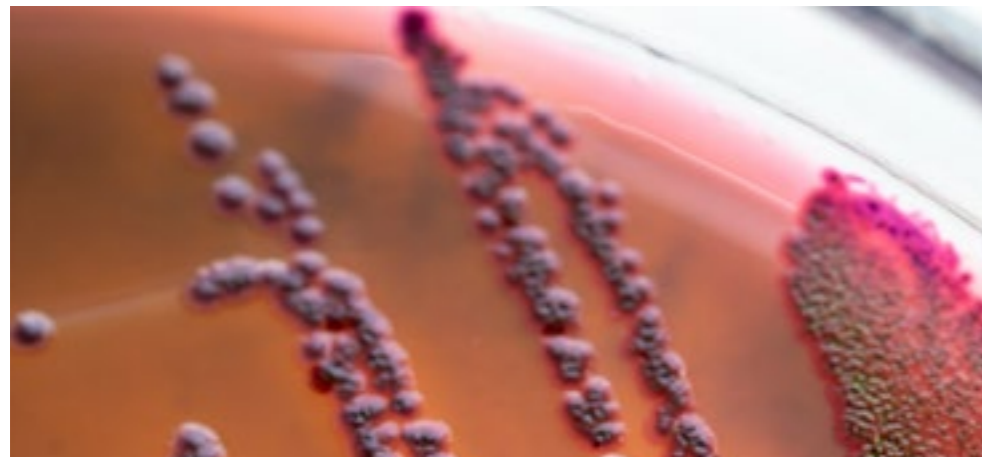
We wish you all an interesting and exciting read.

Alfred Pühler  
de.NBI Coordinator

Andreas Tauch  
Head of the de.NBI  
Administration Office

# CONTENT

PREFACE	3
CONTENT	4
<hr/>	
<b>PLANT BIOINFORMATICS – ADVANCING MODERN PLANT RESEARCH AND PLANT BREEDING</b>	<b>6</b>
<hr/>	
GREEN BIOINFORMATICS – DECODING THE ROOTS OF CIVILISATION	8
CHEMICAL DIVERSITY IN THE PLANT WORLD	14



<b>MICROBIAL BIOINFORMATICS – ANALYSING THE DIVERSITY ON OUR PLANET</b>	<b>20</b>
<hr/>	
MICROORGANISMS – THE INVISIBLE MAJORITY IN OUR OCEANS	22
EXPLORING THE DEEP SEA WITH BIOINFORMATIC IMAGE ANALYSIS	28
NON-CULTIVABLE BACTERIA – ACCESSING THE EARTH'S GREATEST GENETIC TREASURE	32
IDENTIFYING AND ANALYSING RESISTANT CLINICALLY-RELEVANT BACTERIA WITH THE HELP OF THE de.NBI CLOUD	36
PHYLOGENETIC ANALYSIS AS A TOOL FOR IDENTIFYING PATHOGENS	42
BRENDA – AN ESSENTIAL RESOURCE FOR THE DEVELOPMENT OF BIOTECHNOLOGICAL SUBSTANCE PRODUCTION ROUTES	48



<b>HUMAN BIOINFORMATICS – BENEFITS FOR MEDICINE</b>	<b>52</b>
<hr/>	
FROM PROTEIN STRUCTURES TO NEW DRUGS	54
LIPIDOMICS – HOW LIPIDS CONTROL BLOOD COAGULATION	60
MICROBIOME RESEARCH SHEDS LIGHT ON DISEASE DEVELOPMENT	64
WHAT THE PROPERTIES OF HUMAN CELLS TELL US ABOUT CANCER	70
PERSONALISED MEDICINE IMPROVING TREATMENT OF TUMOUR DISEASES	76
ANALYSING THE GENE REGULATION OF HUMAN CELLS WITH THE HELP OF MACHINE LEARNING	82
RNA IN MEDICAL DIAGNOSTICS	86
RESEARCH ON BIOMARKERS FOR THE EARLY DIAGNOSIS OF PARKINSON'S DISEASE	92
SYSTEMS MEDICINE OF THE LIVER – A CHALLENGE FOR DATA MANAGEMENT	96



<b>THE GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE (de.NBI)</b>	<b>102</b>
<hr/>	
THE GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE	104
INTERVIEW WITH THE de.NBI COORDINATION	106
de.NBI SERVICES	108
de.NBI TRAINING	109
de.NBI CLOUD	110
de.NBI INDUSTRIAL FORUM	111
ACTIVITIES IN THE de.NBI NETWORK	112
IMPRINT	114

# PLANT BIOINFORMATICS – ADVANCING MODERN PLANT RESEARCH AND PLANT BREEDING

The generation of large amounts of data has become an integral part of plant research and plant breeding. Yet data alone do not equate to scientific progress. Nonetheless, the bioinformatic analysis of sequence as well as transcriptome, proteome or metabolome data, can provide detailed information about important genetic and physiological processes in cultivated plants, thus helping us to better exploit their breeding potential. This is why the future of plant research and plant breeding is no longer conceivable without bioinformatics.

# GREEN BIOINFORMATICS – DECODING THE ROOTS OF CIVILISATION

Plants are our constant companions; as spices, as decoration, as the foundation of our nutrition and even as the basis of our civilisation. Today's varieties are the outcome of thousands of years of breeding. This process continues to this day, and new high-throughput methods provide data for the continuous improvement of our varieties. de.NBI contributes to making this available for research which, in turn, contributes to sustainable food production and supply.



## PLANT AND ANIMAL BREEDING ARE THE BASIS OF OUR CIVILISATION

Around 20,000 years ago, in the area known as the Fertile Crescent located between the Eastern Mediterranean and Mesopotamia (present-day Iraq), the transition to sedentary rural living began. One of the driving forces behind this conversion to the cultivation and breeding of crops was climate change. The interglacial period beginning at that time necessitated some way to compensate for dwindling food supplies of wild animals. In the course of this Neolithic revolution, useful plants and animals were domesticated for the first time. The first crops grown were cereals (Figure 1) and legumes, while the first domesticated animals were goats, sheep and cattle. This is considered to be the initial spark

and essential precursor of our current culture and took form in the first advanced civilisations in Mesopotamia and Egypt. The predictable and reliable availability of food laid the foundation for the culture and stability needed by a rapidly growing population. The breeding and selection of plants and animals beneficial to humans continues to shape our culture even today. Our landscape is dominated by organisms (plants) which did not evolve naturally, but rather are the product of systematic cultivation, selection and classical breeding by humans. Ancient motivating forces are more relevant today than in recent history due to new challenges. Rapid climate change, a dramatically growing global population and inferior soils present us with challenges comparable to those humanity faced 20,000 years ago.

### DECODING GENOMES IS HELPING TO OVERCOME CURRENT CHALLENGES IN PLANT BREEDING

Fortunately, millions of years of evolution to a constantly changing environment has been recorded in the blueprint of plants, the genome. In addition to common components or genes, each species or subspecies has developed its own,



sometimes unique genes, some of which encode favourable agronomical traits. By understanding these genes, which exist in every cell in the form of DNA molecules, we can draw on nature's repertoire of genetic solutions and attempt to introduce favourable traits into cultivated varieties – just as we did 20,000 years ago. This can be done either by classical cross-breeding and selection in the field, or by the targeted molecular analysis of genomes using plant gene banks. However, the genomes of many agricultural plants – maize and cereals such as wheat and barley, for example – are shockingly complex in size and structure, in some cases far surpassing the complexity of the human genome (Figure 2). However, the possibilities offered by modern biological and genomic

research are a much more promising starting point more than 20,000 years ago. Solutions for the identification of all genes, gene variants, genome structures and other trait-influencing properties have only been developed relatively recently. The resulting datasets enable us to ask completely new questions and investigate possible novel relationships. These techniques and technologies have largely

only been available to specialised laboratories or even entire consortia. However, we can observe a broad process of democratisation in (plant) biological genomic research and, linked to this, a massive digitalisation of areas once dominated by classic experimentation. To assist and support this process, specialised analytical software programs and prediction models are being made available by the expert groups in the plant service centre of the GCBN (German Crop BioGreenformatics Network) in a specially installed and customised analytical cloud. This is intended to support the widespread application of analytical processes formerly restricted to specialist groups and, ultimately, to achieve broad emancipation of *in silico*-based plant genomics research.

### The genomes of many agricultural plants far exceed the complexity of the human genome in size and structure.

The past two decades have already seen the emergence of broad interest and application of genomics not only in theoretical research, but also in applied breeding research. At the same time, integration and exchange between areas formerly considered basic research, and application-oriented research, as well as company-oriented development, has become very close. For example, breeding can inadvertently result in the selection of undesirable characteristics, i.e. in the accumulation of harmful substances; in soils containing cadmium, the heavy metal has been found to accumulate in modern durum wheat, but not in the original wild emmer. The associated gene has lost its function in durum wheat, thus allowing the accumulation of harmful cadmium. Breeding experiments are currently focusing on re-crossing the functional gene [1]. Similar aspects are being investigated in relation to common wheat varieties and gluten sensitivity, and will also be applied to breeding experiments [2, 3].

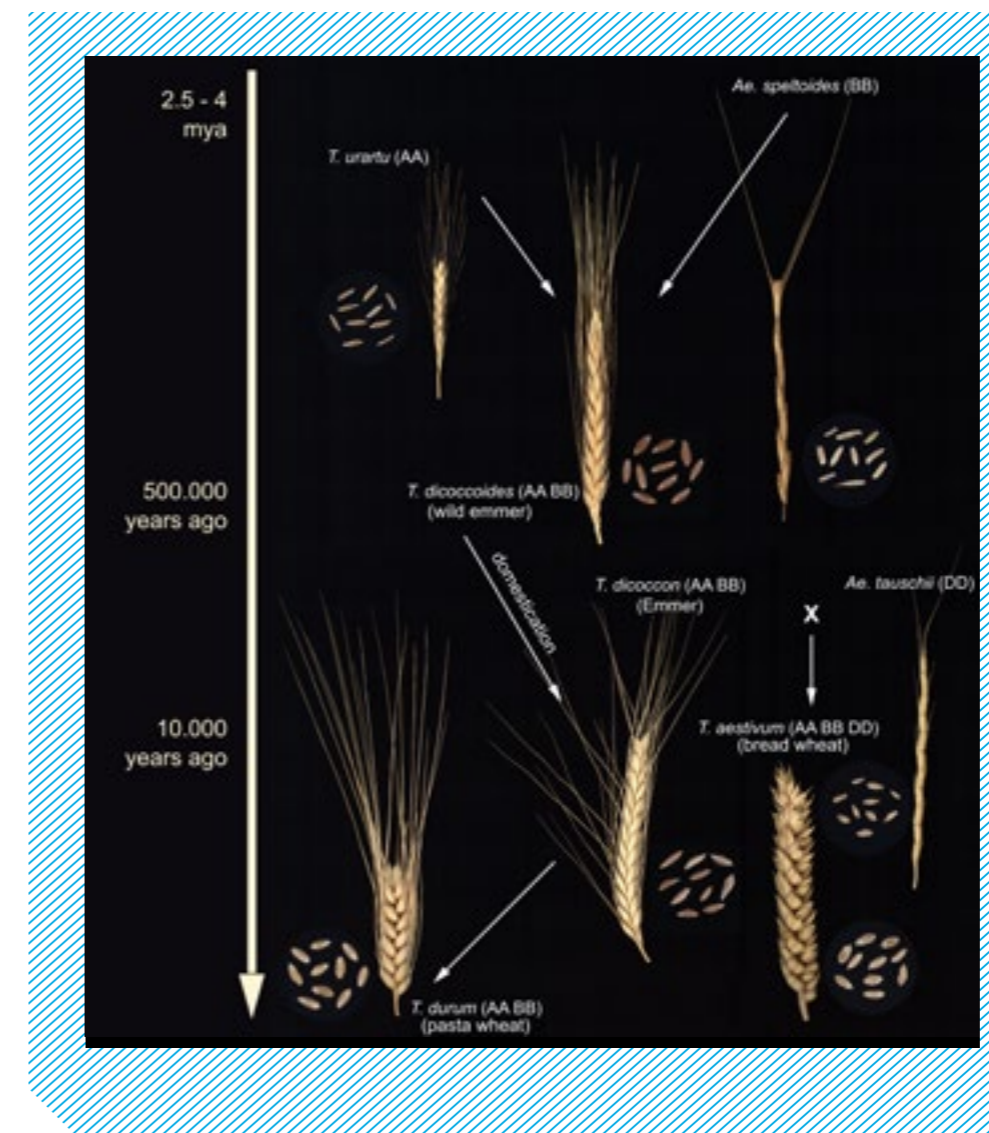
Other important steps include elucidating the interaction between the genotype, i.e. the genetic information, the phenotype, the traits of the plant, as well as interactions with the environment. Urgent environmental and climate problems, climate change, famine, civil unrest and migration flows are closely linked to those which are purely scientific questions at first sight. However, answers to such questions are crucial to solve some of our major worldwide challenges. The associated huge data amounts on phenotypes and genotypes require a more efficient data handling, for example, standardised

access and the structured provision of a wide range of 'omics' data using state-of-the-art methods of computer technology. Since this cannot be achieved in isolated laboratories, our goal is to give the broader user community – from plant molecular biologists to breeders – access to the accumulated bioinformatics expertise

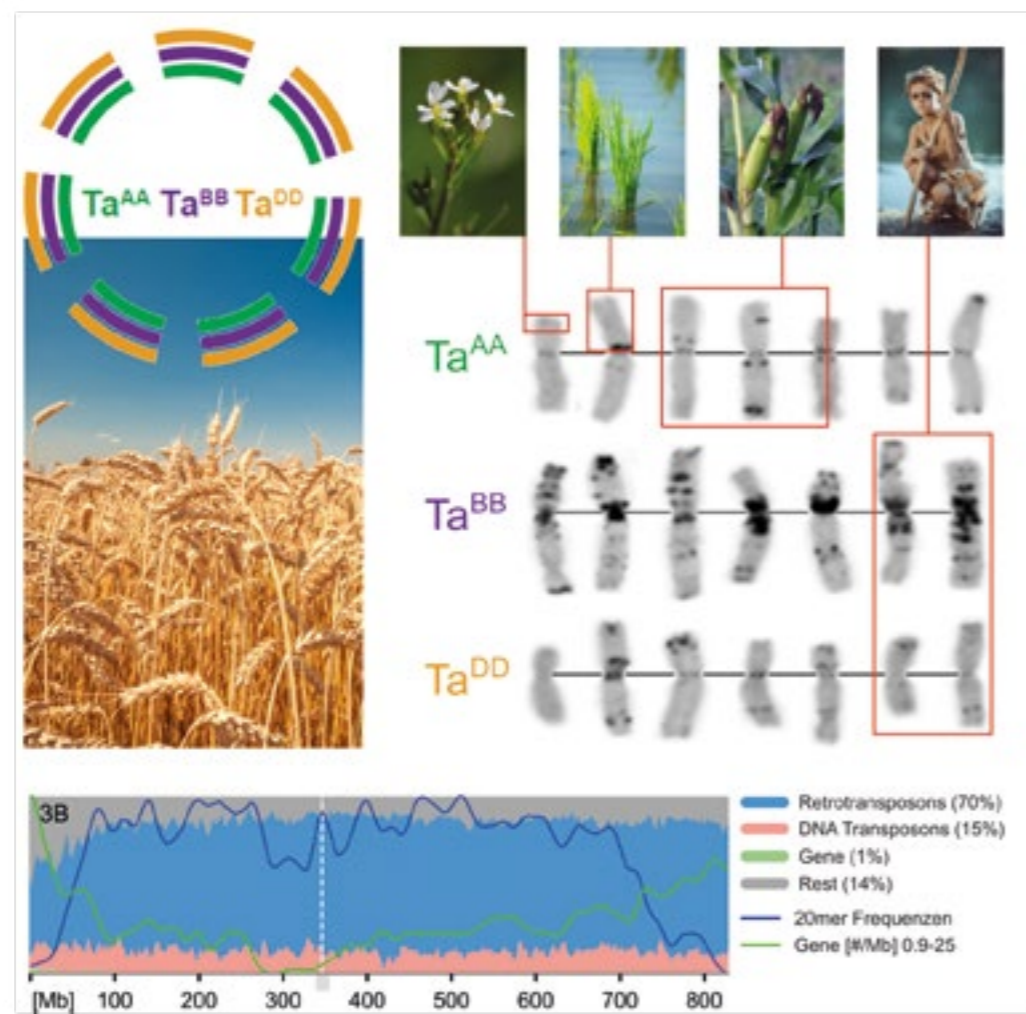
and the vast quantity of crop-based data available, in a structured and easily accessible way, and to provide appropriate software for inter-laboratory analysis and application [4].

In addition to reusable software, another main focus is the exploitation

of the generated data. To enable this, all data should be stored and managed in such a way that they are findable, accessible, interoperable with other data and reusable (Figure 3). These characteristics are summed-up under the acronym FAIR and form a key objective of the work carried out at the GCBN plant service centre.



**FIGURE 1:** The historical development of our present-day wheat. Around 500,000 years ago, the wild emmer (*Triticum dicoccoides*) was formed by a fusion of two diploid wild grasses, wild einkorn, *T. urartu* (AA) and a goatgrass, *Ae. speltoides* (BB), to form the tetraploid AABB genome. With the settlement of humans around 10,000 years ago, a process of selection began, giving rise to cultivated emmer (*Triticum dicoccon*), from which in turn pasta wheat (= durum wheat, *Triticum durum*) was developed. The hexaploid bread wheat (= common wheat, *Triticum aestivum*, AABBDD) originated at about the same time through the fusion of tetraploid emmer with another, rather inconspicuous goatgrass (*Aegilops tauschii*) and continued to be bred as a popular food source. (Image: Gudrun Schütze, IPK Gatersleben)



**FIGURE 2:** The complex structure of the bread wheat genome. With a size of 16 Gbp, the wheat genome is five times larger than the human genome. Bread wheat is hexaploid and consists of three very similar subgenomes, called A, B and D, each with seven chromosomes. The red boxes mark the genome sizes of *Arabidopsis thaliana* (0.13 Gbp) – the first plant genome sequence in 2,000 – rice (0.4 Gbp), maize (2.3

Gbp) and humans (3.2 Gbp) in relation to the wheat genome. The lower part shows the architecture of a typical grain chromosome as a stacked bar chart (0-100%) using wheat (3B) as an example. The genome landscape is dominated by transposons, predominantly LTR retrotransposons, whose high degree of repetitivity (blue line) greatly hinders the assembly of such genomes. As the principal agents of

traits, the genes are like needles in a haystack: they only account for 1% of the total DNA sequence and are highly enriched at the ends of the chromosomes (greenline). (Images) from left to right: photo of wheat © vovan/Adobe-Stock; photo of flower © lehic/Adobe-Stock; photo of rice © comzeal/Adobe-Stock; photo of maize © orestligetka/Adobe-Stock; photo of child Emotion-Photo/Adobe-Stock.



**FIGURE 3:** FAIR data for research and plant breeding. The figure on the left provides a typical overview of the essential characteristics of a (crop) plant genome, from the bioinformatician's point of view, using the tetraploid pasta wheat genome as an example. The data generated by the individual genome projects are currently being combined with phe-

notype data in pilot projects, with the aim of better understanding the biochemical basis of traits relevant to breeding. To ensure that the valuable data resources acquired can continue to be used in other contexts, the data is structured, indexed and archived in accordance with the FAIR principle. The map shows the world-wide data access to the eDAL

archive system from the IPK (Plant Genomics & Phenomics Research Data Repository, <http://edal-gppg.ipk-gatersleben.de>). Image top-left: © ktsdesign/Adobe-Stock; image bottom-left: © sdecoret/Adobe-Stock; image top-right: dppn.plant-phenotyping-network.de)

**REFERENCES:** [1] Nat Genet 2019;51(5):885–895. DOI: 10.1038/s41588-019-0381-3. [2] Science 2018;361(6403). DOI: 10.1126/science.aar7191. [3] Sci Adv 2018;4(8):eaar8602. DOI: 10.1126/sciadv.aar8602. [4] Genome Biology 2020. DOI: 10.1186/s13059-019-1899-5.

**AUTHORS:** Heidrun Gundlach<sup>1</sup>, Matthias Lange<sup>2</sup>, Marie Bolger<sup>3</sup>, Björn Usadel<sup>3</sup>, Uwe Scholz<sup>2</sup>, Klaus F. X. Mayer<sup>1</sup>  
<sup>1</sup> Plant Genome and Systems Biology, Helmholtz Zentrum München, Ingolstädter Landstrasse 1, 85764 Neuherberg,  
<sup>2</sup> Bioinformatics and Information Technology, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) Gatersleben, Corrensstrasse 3, 06466 Seeland  
<sup>3</sup> BG-2 Plant Sciences, Forschungszentrum Jülich, Wilhelm-Johnen-Strasse, 52428 Jülich

# CHEMICAL DIVERSITY IN THE PLANT WORLD

For many years, little attention was paid to the role of biodiversity on our planet. This has changed, however, both in science and in public perception. Today, research includes not only biodiversity, but also the investigation of the diversity of individual constituents, called chemodiversity.



The natural constituents of plants can be analysed for many purposes. For example, several chemical substances derived from plants have already been used as remedies in humans. In addition, secondary metabolic products control a multitude of interaction processes both within the plant and between different plants and the microorganisms in their environment. Chemical substances therefore provide insight into a variety of important biological processes. However, so far nothing is known about many of these natural substances – neither about their chemical structure nor their biological or ecological function. The research area of chemical ecology tackles such questions, as well as addressing the importance of chemical diversity.

The technical analysis of the natural ingredients of plants is often carried out with a mass spectrometer. First, samples of the plants are collected. Then their constituents are extracted in the laboratory, for example, by using water and methanol, and analysed by combining chromatography and mass spectrometry (Figure 1).

This generates a vast amount of complex raw data that provide information about the mass-to-charge ratio and the chromatographic retention time of the substances. These raw data can be interpreted as a plant's fingerprint and already enable researchers to examine the samples with statistical methods in order to address biological and ecological issues.

The illustrations in this article show some examples of research in the field of Eco-Metabolomics, in which the Center for Integrative Bioinformatics (CIBI) is actively involved.

## THE VALUE OF MOSS

Mosses are the oldest terrestrial plants on earth and can be found in almost all ecosystems. They are considered to be exceptionally good bioindicators, signalling changes in the environment such as pollutants in the air, which can lead to damage or impaired growth in mosses. Hitherto such changes have been considered mainly in terms of growth and morphological properties, however, not at the level of biochemical composition. To address this, the Leibniz Institute of Plant Biochemistry (IPB) used mass spectrometry to analyse the biochemical changes in various moss species over the different seasons, with regard to different living conditions and their relatedness to each other (phylogeny). They then evaluated the results using bioinformatics methods.

The study [1] analyses of the connections between the various lifestyles and selection strategies of mosses and their biochemical adaptation to changing living and environmental conditions. This untargeted Eco-Metabolomics approach thus provides valuable biochemical insights that can improve our understanding of key ecological strategies and serve as a basis for future research (hypothesis generation). Furthermore, we have created a representative data set and a bioinformatics workflow that can be reused in future metabolomics studies.

## MacBeSSt AT THE IDIV – USING THE PLANT FINGERPRINT AS A GUIDE

MacBeSSt is not about (classic) literature: it actually refers to the project "Metabolite Changes in Biodiversity Levels and Seasonal Shifts" at the German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, which also deals with (chemical) diversity in the plant world.

As opposed to medically relevant plants, such as sage or St. John's wort, little is known about the secondary constituents (metabolites) of grassland species. To investigate the metabolic fingerprint of these species, we studied plants that grew together with other plant species in the Jena experiment [2]. Since changing day lengths, warmer temperatures and water supply also play a major role in plant development, we took samples of 13 species at four different times between May and October in order to detect seasonal differences in the metabolic fingerprint.

The composition of these species communities is particularly important for the analysis of fingerprints, as a changed neighbourhood could also mean a



changed fingerprint. To examine these influences more exactly, we sampled species communities that consisted of a single species (monoculture) or two, four or eight species. The plant extracts are measured in a mass spectrometer connected to a liquid chromatograph. The data acquired can then be statistically evaluated and examined for correlations.

The examined external influences, species community and season are reflected in the altered quantities of the plant constituents – thus indicating the path the plant has taken so far. Yet, this does not change the dimension of the fingerprint, which makes it possible to identify all the species under investigation throughout the year on the basis of their unique pattern. The experimental design allows the project to investigate the relationships between plant species, species communities, seasons and the environment, simultaneously bridging the research areas of ecology, biochemistry and bioinformatics.

#### METABOLITE IDENTIFICATION

However, the tasks of bioinformaticians do not end with the analysis of fingerprints, since a biological (or ecological) interpretation requires the annotation of the chemical structure. There are two approaches to this, for which corresponding services are offered in the de.NBI network.

The spectra from the mass spectrometer can be compared with the entries of a reference database of known substances, for example. MassBank [3] contains more than 50,000 entries for over 13,000 substances. CIBI develops the software and helps to integrate new data provided from the user community. However, reference data are not always available, because the pure substances themselves

are often unavailable. In such cases, in silico predictions using bioinformatics methods (computational metabolomics) can help.

MetFrag [4], developed at the Leibniz Institute of Plant Biochemistry, can be used both online and in the de.NBI cloud. As part of our study of mosses (see above), we also analyse substance classes and have expanded MassBank with previously unknown spectra of mosses.

The need for automated data processing is increasing with the large number of samples and attributes in experimental results, especially in metabolomics. Workflow or pipeline tools are visual programming languages that enable biologists and biomedical researchers to apply state-of-the-art algorithms and data analyses to large data sets. These tools are already widely used in commercial data mining and in scientific fields such as pharmaceutical research or genomics. Time-consuming tasks can be outsourced to powerful cloud infrastructures. The establishment of the de.NBI cloud will thus make it easier to develop and operate metabolomics workflows. The cloud does it!

#### KNOWLEDGE IS THE ONLY THING THAT INCREASES WHEN SHARED.

Biological or ecological research also includes making data available to posterity. This is the predestined purpose of the MetaboLights metabolomics data repository at EMBL-EBI. The de.NBI network and the CIBI Service Centre provide support, particularly to the German user community, in publishing high-quality metabolomics data according to the FAIR principle. This means they are findable by means of meaningful metadata and corresponding search engines; there are regulations on how accessible they may be; they are interoperable, i.e. they can

be combined with other data, and they are reusable – in subsequent research projects, for instance.

The data pertaining to the examples described above can be found as studies MTBLS520, MTBLS709 and MTBLS679 in the MetaboLights research database.

There is a variety of educational and training opportunities to make these topics accessible to future generations of researchers and interested members of the public. To start as early as possible, interested high school students learn how to extract natural substances and evaluate the resulting data at the BioByte summer school at the Martin Luther University Halle-Wittenberg. More advanced de.NBI training opportunities are offered to scientists from various disciplines, from master's to the postdoc level. These include short workshops as well as longer offerings such as the one-week Metabolomics Winter School.



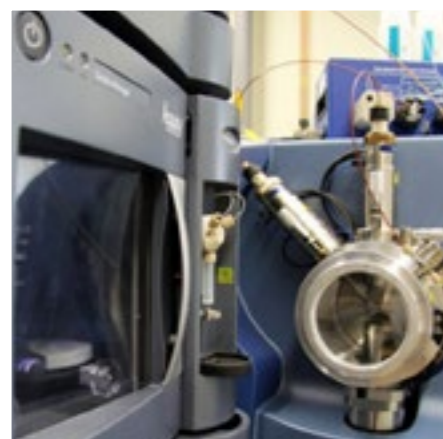


FIGURE 1: A modern mass spectrometer in the laboratory from [5].

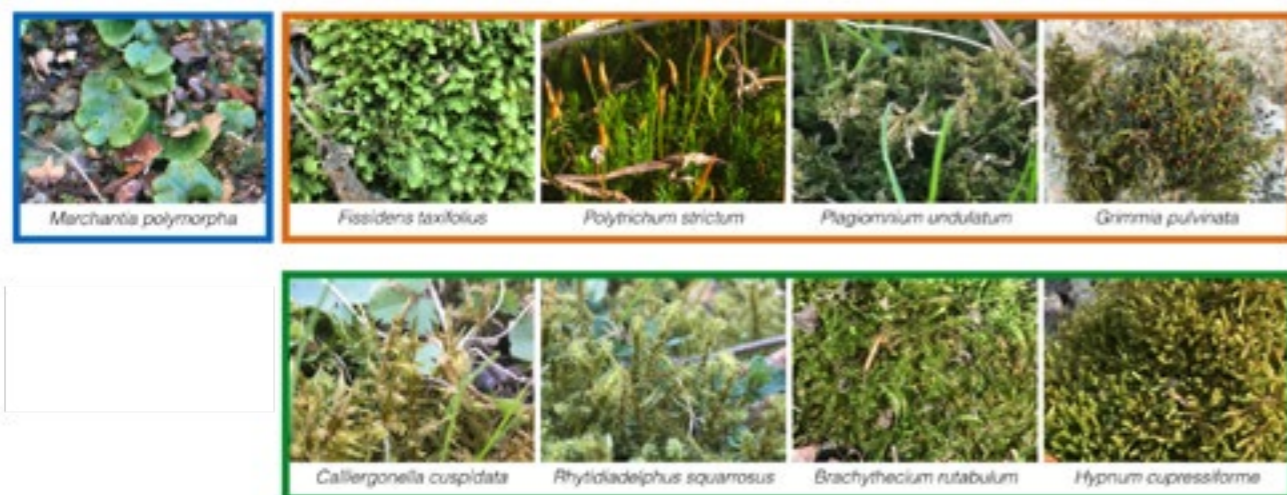


FIGURE 2: A variety of mosses in the botanical garden of the Martin Luther University Halle-Wittenberg from [6].



## CONCLUSION

Many of the challenges for (eco-) metabolomics described here also apply to other disciplines, which might not seem obvious at first glance. For example, one task in environmental research is the monitoring of water quality, which necessitates the comparison of samples across locations, over time or after water treatment. The biochemical composition of samples are also examined in food

control, a process that could profit from bioinformatics.

The de.NBI network covers different aspects of metabolomics in several of its service centres. This includes the Centre for Integrative Bioinformatics (CIBI). With the introduction of the de.NBI cloud, we can now handle the data management and processing even of large studies with

many samples. Bioinformaticians are thus an integral part of interdisciplinary teams working together with molecular biologists, biochemists and ecologists, to help clarify and conserve the diversity in the plant world of our planet.

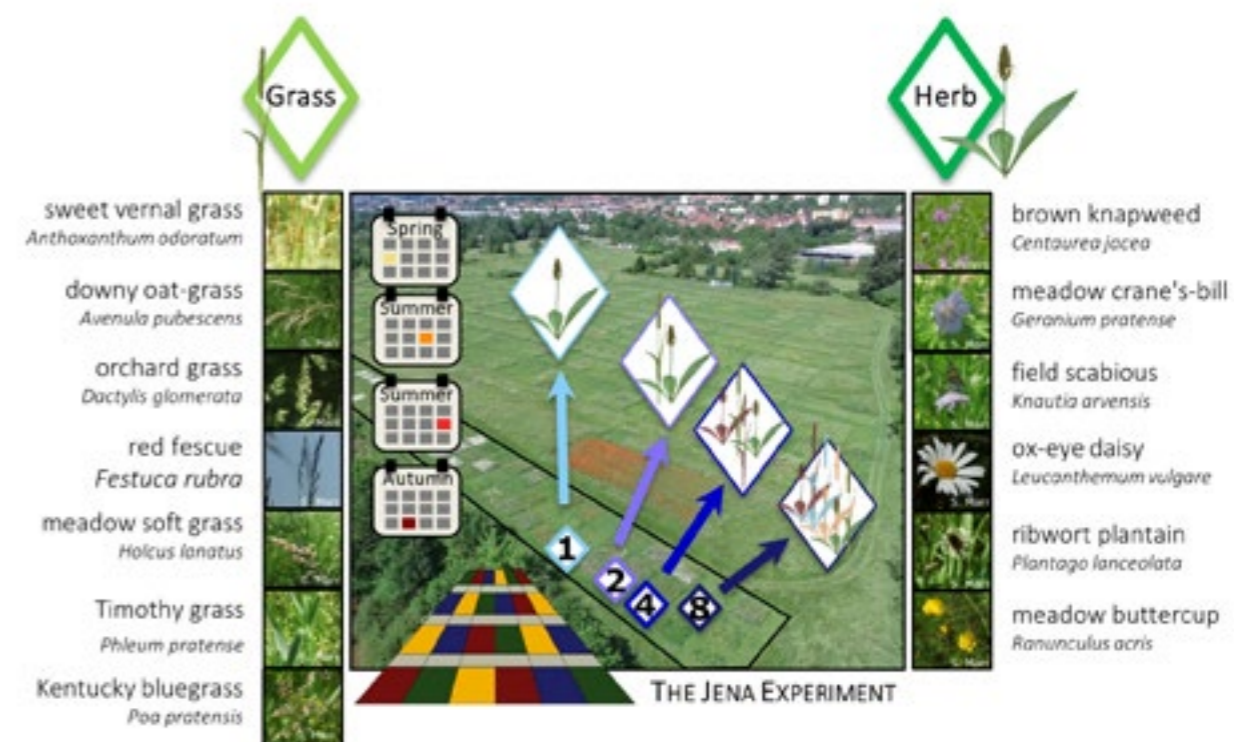


FIGURE 3: Metabolic changes in biodiversity level and seasonal shifts (Mac-BeSST) in the Jena experiment.




**REFERENCES:** [1] *Metabolites* 2019, 9(10), 222. DOI:org/10.3390/metabo9100222 [2] <http://www.the-jena-experiment.de/Video.html> [3] <https://massbank.eu/> [4] <https://msbi.ipb-halle.de/Metfrag> [5] <https://www.ipb-halle.de/forschung/technologie-plattformen/metabolomics/> [6] Presentation from K. Peters at <https://onlinelibrary.wiley.com/doi/abs/10.1002/ece3.4361>

**AUTHORS:** Kristian Peters<sup>1</sup>, Susanne Marr<sup>1,2,3</sup> and Steffen Neumann<sup>1,3</sup>

<sup>1</sup> Leibniz Institute of Plant Biochemistry (IPB), Weinberg 3, 06120 Halle (Saale)

<sup>2</sup> Martin Luther University Halle-Wittenberg, Universitätsplatz 10, 06108 Halle (Saale)

<sup>3</sup> German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Deutscher Platz 5e, 04103 Leipzig

A close-up photograph of a petri dish containing a bacterial culture. The medium is a light pinkish-orange color. Several dark, almost black, streaks of bacteria are visible, extending from the top left towards the bottom right. The streaks vary in thickness and density, with some appearing as thin lines and others as thicker, more textured bands. The background is slightly blurred, showing the edge of the petri dish and a white surface.

# MICROBIAL BIOINFORMATICS – ANALYSING THE DIVERSITY ON OUR PLANET

Life on our planet is profoundly affected in almost all respects by microscopically small creatures, the microorganisms. Nowadays, research of their life processes is being conducted in fascinating detail, with omics data and their bioinformatic analysis playing a key role.

# MICROORGANISMS – THE INVISIBLE MAJORITY IN OUR OCEANS

Man and the sea have always had a close connection. Oceans cover about 70% of the earth's surface, and about half of the world's population lives in coastal areas. Through fishing, the sea provides food for millions of people, and it has been one of the most vital trade routes for thousands of years. Over the past decades, tourism has increasingly marked coastal regions as an important economic factor. The oceans are also home to millions of animal and plant species and billions of microorganisms. Forming an invisible majority, they provide the foundation of the marine food web and are responsible for the recycling of virtually all nutrients across the globe. Exploring them is only possible through the skilful interaction of molecular techniques and their bioinformatic analysis on the basis of biodiversity, functional databases and environmental databases.

## THE IMPORTANCE OF MARINE MICROORGANISMS

Marine microorganisms are microscopically small, unicellular organisms that include bacteria, viruses, small algae and archaeae. They may be tiny, but they exist in great numbers everywhere in the oceans, from the deepest points on and in the seabed to the sun-drenched surface. One millilitre of seawater, or one thousandth of a litre, contains up to one million microorganisms (Figure 1). This means that there are more microorganisms in one litre of seawater than people on the entire planet. As they are responsible for global metabolism of nutrients and energy, they are indispensable for the proper functioning of the oceans [1].

Marine microorganisms affect our daily life and our well-being, no matter whether you live on the coast or inland. In addition to breaking down and converting nutrients, they also fulfil the important task of photosynthesis. Like plants, some marine microbes, such as cyanobacteria, can use the light energy of the sun to convert carbon dioxide (CO<sub>2</sub>) and water into sugar. During this process, oxygen (O<sub>2</sub>) is produced and released into the environment. Scientists estimate that about half of the world's oxygen production comes from the oceans, while the other half is supplied by other habitats such as forests or soils. This means that marine microbes produce the oxygen for every second breath we take.



**FIGURE 1:** The image shows microorganisms on an algae. The microbes were made visible by means of a fluorescent dye (photo: © Max Planck Institute for Marine Microbiology/ P. Gomez-Perreira/ B. Fuchs).

Another example of their importance is the ability of some microbes to break down oil. Some species feed on it, so they can help to clean up oil spills after tanker accidents. Recently microorganisms have been found that can even degrade certain types of plastic. Unfortunately, this takes decades and is therefore not an effective defence against the pollution of our oceans [2].

Researchers also have high hopes for the potential of marine microorganisms in the field of medical and biotechnological applications. Antibiotics are metabolic products of bacteria or fungi that have the property of harming other microorganisms by inhibiting their growth or killing them. As a result of the frequent use of antibiotics, many microorganisms no longer respond to them, i.e. they are resistant. Scientists are hoping to find hitherto unknown antibiotic substances in the sea. The feasibility of this strategy has been demonstrated by a recently completed research project in which an antibiotically active product originating from a previously unknown bacterium was discovered. However, the new antibiotic will initially only be used in aquacultures for fish farming with the aim of protecting the animals from pathogens. Its approval as a drug requires extensive series of tests, which usually take over ten years.

In biotechnology, biochemical reactions are needed to catalyze the conversion of organic substances, a task which is performed by enzymes. Enzymes are proteins that are formed by living cells and increase the reaction rate of biochemical processes. Cellulose, the main constituent of plant cell walls, is used as a raw material for paper production. Enzymes that break down cellulose are called cellulases. They help to make the material supple. One source of such enzymes is the bacteria that live in the deep sea or in Antarctic waters. The detergent industry has also placed its hopes in the cold waters of the oceans. In the past, it was common to wash white textiles at very high temperatures to remove impurities through the action of heat. Yet, high temperatures mean high energy requirements. With the increased use of enzymes that break down fat and protein in detergents, doing laundry has become

much more energy-efficient, despite relatively low temperatures.

#### HOW ARE MICROORGANISMS STUDIED?

Until recently, scientists required a pure microbial culture to answer seemingly simple questions such as "What species of microorganisms exist?", "What can they do?" and "How do they interact with their environment?" A pure culture means that individual microorganisms have to grow in the lab without their natural environment and without other microorganisms. Since these laboratory conditions differ greatly from those in the oceans, it is extremely difficult to cultivate marine microbes. It is estimated that only one to ten per cent of marine microorganisms can be cultivated in the laboratory. Fortunately, new molecular techniques have been developed in recent years, allowing

marine microbes to be researched without cultivating a pure culture in the laboratory (Figure 2).

#### Human DNA contains 25,000 to 35,000 genes.

The entire information of an organism exists in its genetic code, the so called DNA, which is why it is called the blueprint of life. It instructs the cell what to do and when to do it. DNA can be divided into small segments called genes. There are thousands of genes in the DNA of a living thing, and each gene has a specific function. For example, human DNA contains 25,000 to 35,000 genes, but very few are responsible for individual traits such as eye or hair colour. Next Generation Sequencing (NGS) technology enables scientists to read the DNA of an entire community of micro-

organisms with relatively little technical effort and at low costs [3]. This approach is also known as metagenomic sequencing and provides a list of the genes of all microorganisms that occur in a particular area.

#### BIOINFORMATIC ANALYSIS

Some genes are found in all organisms on earth and exhibit small but significant differences among organisms. Ribosomal RNA (rDNA) is an example of such a

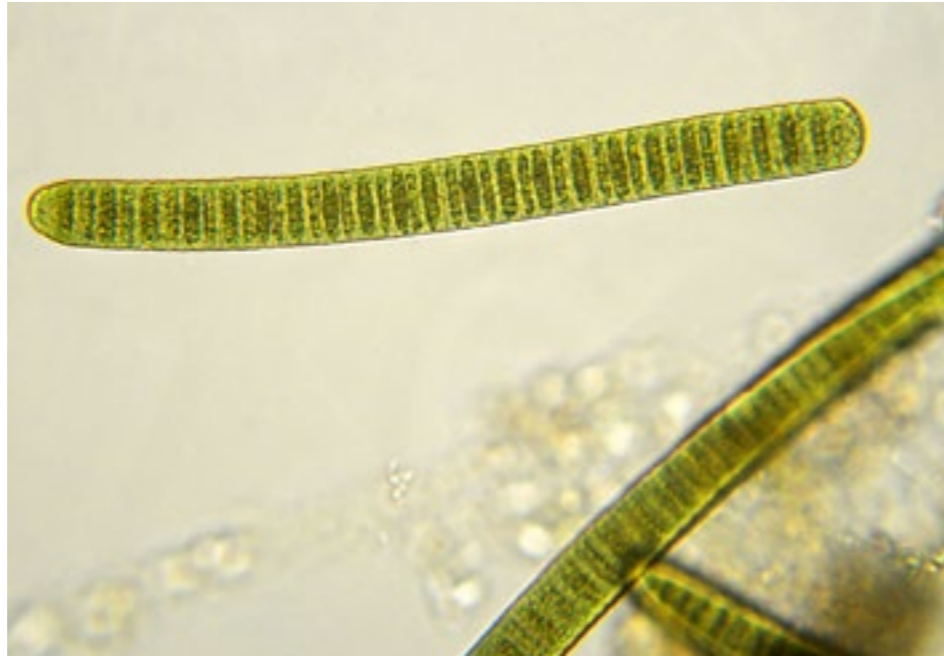
sequence of the rDNA is determined and then compared with the existing information stored in reference databases. The SILVA database [4], based at the BioData Service Centre, is one of the two leading rDNA reference databases worldwide. With almost ten million entries covering the entire tree of life, it is currently the most comprehensive repository of quality-tested rDNA sequences. Due to its systemic importance for the entire scientific community, SILVA was recently named an ELIXIR Core Data Resource.

functions. This makes it possible to create a model of the enzymatic functions and potential metabolic pathways present at the time of sampling. Not only does this improve our understanding of the ecosystem as a whole, gene sequences in general will be of great interest if medical or biotechnological applications can be found for them.

#### How do they interact with THEIR ENVIRONMENT?



temperature, salinity, water depth/pressure). In some cases, these factors can be identified at the same time as the microorganisms are sampled. However, an accurate characterisation of the environment often calls for complex analyses of the water and seabed samples in the laboratory.



Only when all information is combined, it will be possible to understand the complex interactions between the organisms and their respective environments, enabling us to make more accurate predictions of how global changes, such as the warming of the oceans as part of climate change, will affect them. For this purpose, individual measurements are often mere insufficient snapshots. Yet, technical developments over the past decades have now made it possible to measure a large number of environmental factors continuously and automatically. Both stationary and mobile measuring systems are used. The 3,800 Argo floats drifting all over the globe are an example of this. These systems automatically measure temperature and salinity at regular intervals in the upper 2,000 metres of the oceans. Using satellite links, these data are made available to the scientific community and the general public with just a short time delay [5].

The continuous provision of a large amount of data is essential for research

into global developments such as climate change and species extinction. This can only be ensured by storage in data archives. One of the world's leading systems for this is the Data Publisher for Earth & Environmental Science - PANGAEA. The certified World Data Center [6] is operated jointly by MARUM - Center for Marine Environmental Sciences at the University of Bremen and the Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research. With over 16 billion data points, the de.NBI database PANGAEA provides a vast collection of scientific data to a large user community [7]. This encompasses data from the earth and the environment as well as the occurrence and distribution of both living organisms and biochemical molecules. The data can be accessed on the website [8], but

experts also have the option of retrieving the data via machine interfaces to make them available for further analysis.

The mutual dependencies between microorganisms and larger life forms on our planet are held in a delicate balance and are endangered by environmental pollution and changing climatic conditions. To protect the environment, we need a sound knowledge of the microorganisms that inhabit the sea, their functions, and how they interact with each other and the environment. The BioData Service Centre provides internationally recognised databases for environmental and biodiversity research as well as medical and biotechnological applications.



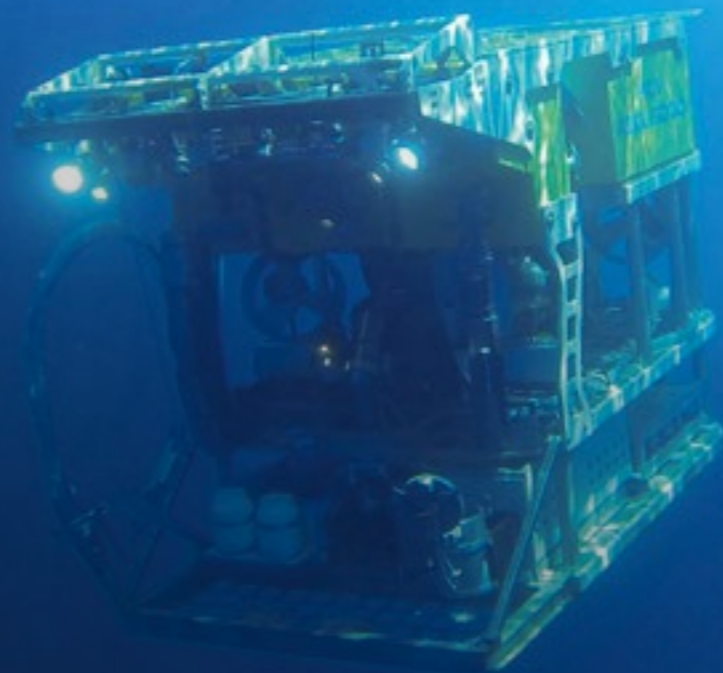
**REFERENCES:** [1] Nat Rev Microbiol 200;5(10):759-69. DOI: 10.1038/nrmicro1749. [2] Appl Microbiol Biotechnol 2018; 102:7669-7678. DOI: 10.1007/s00253-018-9195-y. [3] Nat Rev Genet 2016;17(6):333-51. DOI: 10.1038/nrg.2016.49. [4] Nucleic Acids Res 2013; 41 (Database issue): D590-D596. DOI: 10.1093/nar/gks1219. [5] <http://www.argo.ucsd.edu/> [6] <http://www.icsu-wds.org/> [7] J Biotechnol 2017;261:177-186. DOI: 10.1016/j.jbiotec.2017.07.016. [8] <https://www.pangaea.de/>

**AUTHORS:** Janine Felden<sup>1</sup>, and Frank Oliver Glöckner<sup>1,2</sup>

<sup>1</sup> MARUM - Center for Marine Environmental Sciences University of Bremen and Alfred Wegener Institute, Helmholtz Center for Polar and Marine Research, Bremerhaven

<sup>2</sup> Jacobs University Bremen, Bremen

# EXPLORING THE DEEP SEA WITH BIOINFORMATIC IMAGE ANALYSIS



Exploration and monitoring of the deep sea and the impact made by humans represent a major interdisciplinary scientific challenge. New and efficient bioinformatics approaches are needed to evaluate large quantities of underwater images. The new BIIGLE 2.0 system has rapidly developed into a valuable and internationally acclaimed tool for the management, visualisation, annotation and algorithmic analysis of underwater image data.

In marine research, image and video data are increasingly being recorded to capture the status and the development of ecosystems. The volume of data generated requires software-supported evaluation. For this research area, the Bio-Image Indexing and Graphical Labeling Environment (BIIGLE) was launched in 2009 as the first online annotation system for image data from marine research, and it has since gained continually increasing acceptance in the marine sciences.

Beyond humanity's habitual drive for discovery, the exploration and observation of the oceans has become even more essential in this millennium. On the one hand, scientists must evaluate the effects of climate change on marine ecosystems. On the other hand, other very direct impacts of humans on the world's oceans (e.g. overfishing, raw material extraction or tourism) must also be recorded, studied and assessed. Over the past ten years, technologies such as high-resolution digital photography have led to significant progress in the technical design of mobile or stationary underwater carrier systems. In this way, state-of-the-art systems such as the ROV (remotely

operated vehicle), AUV (autonomous underwater vehicle), OFOS (ocean floor observation system) and FUO (fixed underwater observatory) have made it possible to develop methods for surveying large expanses of the sea floor with high-quality photography or video recordings, or observing small areas over long periods of time in photo sequences [1]. The digital image data contain a wealth of information about the taxonomic composition and morphological properties of the megafauna. However, suitable algorithms and specialised software systems are urgently needed to help evaluate the rapidly growing amount of image data.

## METHODS OF IMAGE EVALUATION

In most cases, the evaluation of the image data aims to identify and mark a specific region in an image (step 1) and to provide a semantic annotation for this image region (step 2). Step 1 may consist, for example, of selecting a point, a circular or rectangular shape or a custom-drawn polygon at a defined location in the image. In step 2, a semantic category is either freely formulated or selected from a catalogue and attached to the image region. These may include predefined taxonomic cata-

logues from biology (for example, from the WoRMS database) or other catalogues describing various types of non-biological objects (for example, waste). Due to the relatively high level of diversity on the one hand, and the sometimes very low density per species on the other, the achievement of a complete automation of these two steps will not be a realistic prospect in the foreseeable future. Based solely on the circumstances mentioned above, there are generally not enough semantically annotated image sections available to apply modern machine learning algorithms (also called deep learning) to automatically detect and/or classify the objects in the image and video data.

In 2009, the Biodata Mining Group at the University of Bielefeld presented the first online annotation system for image data [2]. This system, called BIIGLE, gave marine biologists the unprecedented opportunity to retrieve, view and consistently evaluate their image data using an Internet connection. Furthermore, the system made it possible to mark objects of interest in the images with a very simple and efficient graphical tool and to link them to predefined semantic categories. Although the primary motivation behind BIIGLE was to collect training data for machine learning, the system quickly became popular in the research areas of marine biology and geology, where it was integrated into work processes for image data analysis.

### BIIGLE 2.0

In 2017, the BIIGLE system was completely reimplemented in order to add more features and to meet the increased demands that arose from a growing number of users with diverse research contexts [3]. Among the most important new features are new graphical annotation tools

(e.g. magic wand, polygons; Figure 1a), quality assurance tools for annotations (Figure 1b), a tool for video annotation, hierarchical catalogues of semantic categories that can be dynamically and interactively configured by the users (Figure 1c), as well as automatic laser-point detection, new geo-visualisations and an automatic tool for object detection based on machine-learning methods.

### TECHNICAL IMPLEMENTATION

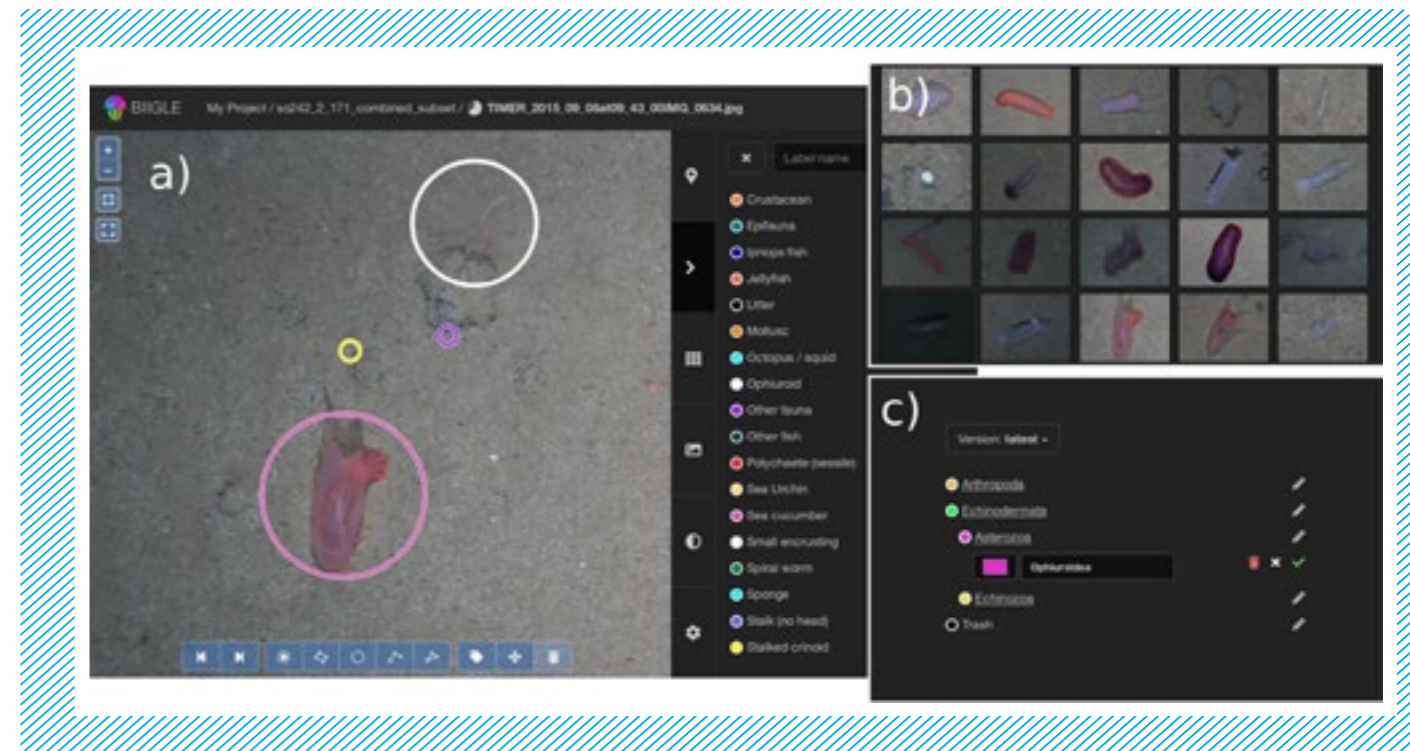
Since February 2018, BIIGLE has been operated entirely in the OpenStack cloud hosted by de.NBI in Bielefeld. The migration to OpenStack was a major step forward for the operation and further development of BIIGLE. By using more advanced hardware and software, the speed of the system has been more than doubled. Moreover, the utilisation of several separate virtual machines in OpenStack has improved the system's reliability and maintainability. The OpenStack service for storing large volumes of data was successively integrated into BIIGLE. In addition to image and video data, BIIGLE now uses this service to manage several million dynamically generated files. The availability of powerful special hardware in the form of graphics processors for scientific computing represented a further advance. This made it possible to implement state-of-the-art methods of machine learning in BIIGLE for the first time. One example is the method of machine learning-assisted image annotation [4], which has been available to all BIIGLE users since early 2019.

The use of the resources available in BIIGLE through the de.NBI cloud is planned to be further expanded in the future. One aim is to provide additional methods of machine learning operating with graphics processors. Another is to prepare the system for better scalability by using multiple virtual machines in

OpenStack to keep up with the system's growing popularity and number of users.

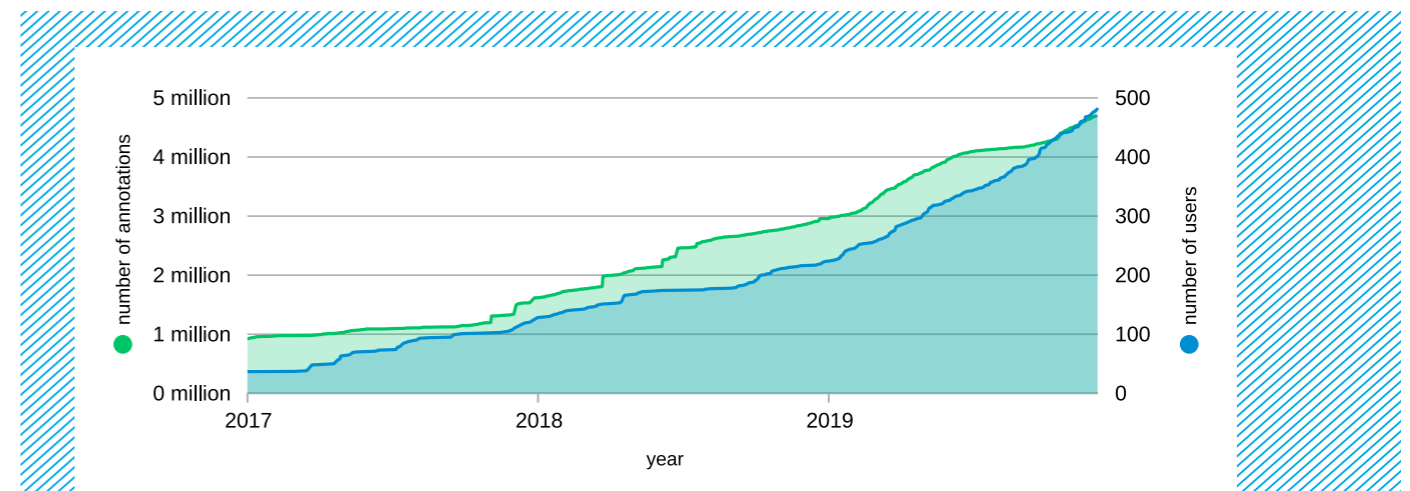
### HIGH ACCEPTANCE IN THE COMMUNITY

Since the release of BIIGLE 2.0 in 2017, the number of users and the number of annotations in BIIGLE has been steadily increasing (Figure 2). Users include marine research institutes such as the GEOMAR Helmholtz Centre for Ocean Research Kiel, the Senckenberg Research Institute in Wilhelmshaven, the French institute Ifremer, the British National Oceanography Centre and a number of universities and research groups from around the world. Marine research topics and image types are constantly increasing in number and diversity. Apart from images and videos from mobile or stationary carrier systems, the BIIGLE 2.0 system is now also used to analyse images from bright-field microscopy to classify plankton or diseased cell tissue as well as aerial photographs taken by drones.



**FIGURE 1 (above):** Elements of the BIIGLE user interface. a) The annotation tool with circle annotations in the main view and the available catalogue of semantic categories in the sidebar. b) Overview of existing annotations for quality assurance in the "label review grid overview" tool. c) View for editing a hierarchical catalogue of semantic categories.

**FIGURE 2 (below):** The number of annotations (green, left axis) and the number of users (blue, right axis) in BIIGLE 2.0 since its release in 2017. The initial values originate from the data transfer from the previous version of BIIGLE 2.0.

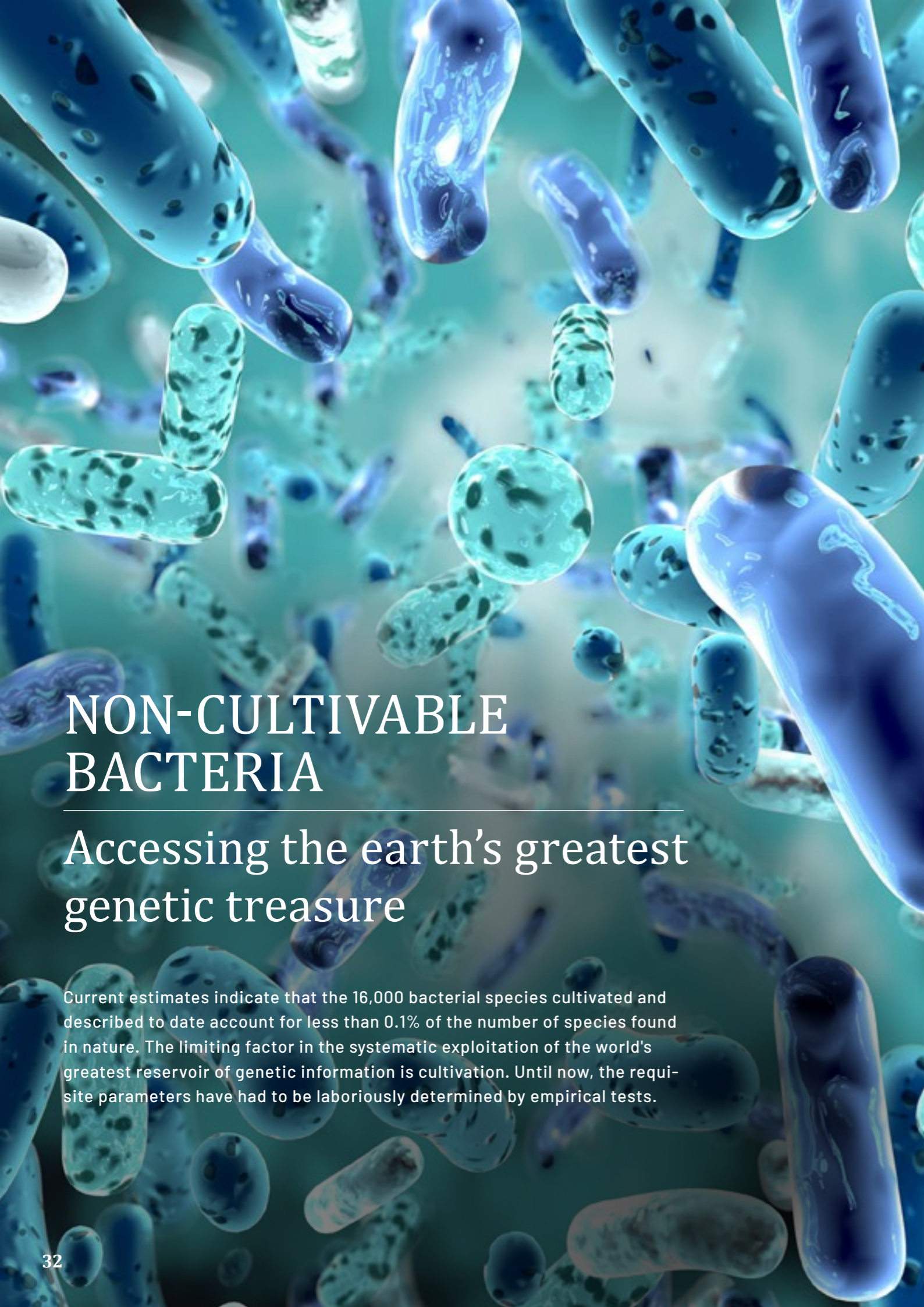


**REFERENCES:** [1] *Oceanography and Marine Biology*, 216, pp 9-80. DOI: 10.1201/9781315368597. [2] *OCEANS 2009-EUROPE*. DOI: 10.1109/OCEANSE.2009.5278332. [3] *Front. Mar. Sci.*, 28 March 2017 DOI: 10.3389/fmars.2017.00083. [4] *PLoS One*. 2018; 13(11): e0207498. DOI: 10.1371/journal.pone.0207498.

**AUTHORS:** Martin Zurowietz<sup>1</sup>, Tim W. Nattkemper<sup>1</sup>

<sup>1</sup>*Biodata Mining Group, Faculty of Technology, University of Bielefeld, Universitätsstrasse 25, 33615 Bielefeld*





# NON-CULTIVABLE BACTERIA

## Accessing the earth's greatest genetic treasure

Current estimates indicate that the 16,000 bacterial species cultivated and described to date account for less than 0.1% of the number of species found in nature. The limiting factor in the systematic exploitation of the world's greatest reservoir of genetic information is cultivation. Until now, the requisite parameters have had to be laboriously determined by empirical tests.

16,000

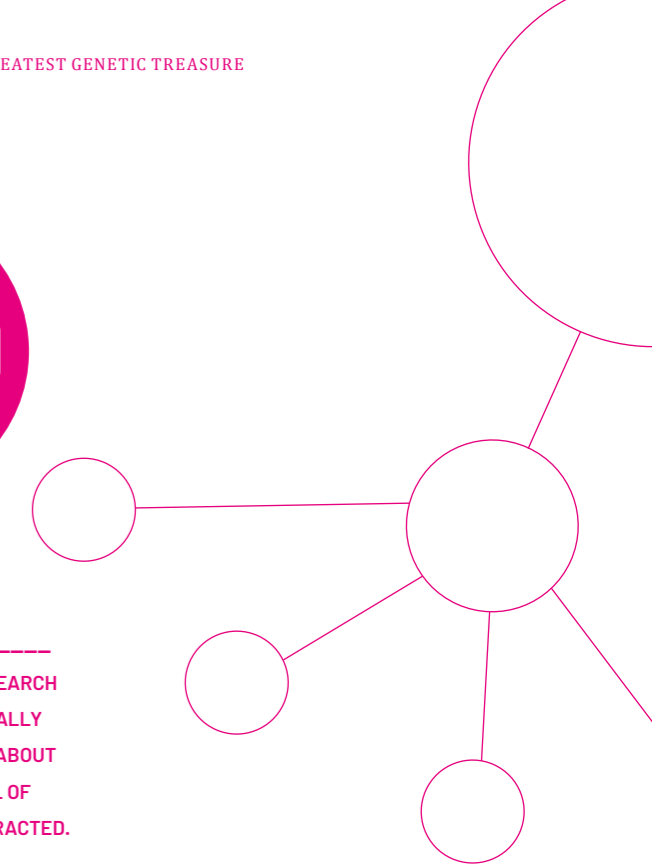
### Current estimates...

INDICATE THAT THE 16,000 BACTERIAL SPECIES CULTIVATED AND DESCRIBED TO DATE ACCOUNT FOR ONLY 0.001% TO 0.1% OF THE NUMBER OF SPECIES FOUND IN NATURE.

93,000

### So far...

150,000 REFERENCES FROM RESEARCH LITERATURE HAVE BEEN MANUALLY EVALUATED BY SCIENTISTS FOR ABOUT 93,000 ENZYMES AND A TOTAL OF 4.7 MILLION DATA HAVE BEEN EXTRACTED.



Antonie van Leeuwenhoek discovered the first bacteria along with the invention of the first microscope in 1676. For many years, the characterisation of bacteria was limited to the observation of their morphology. It was not until the end of the 19th and beginning of the 20th century that an increasing number of physiological tests were developed which showed differences in metabolism, the structure of the cell wall and resistance to antibiotics. Until today, new species of bacteria are described with up to 150 physiological characteristics with the aim of determining both the special abilities of newly discovered species, and differences compared to closely related species. Today, these phenotypic investigations are supported by sequence analyses. On the basis of sequences, scientists can elucidate the evolutionary relationships (phylogeny) to species already described. However, the sequencing of complete genomes is particularly useful in investigating the genetic potential of a new species. While new sequence data are safely stored in large repositories for ready access by scientists, phenotypic data are relatively hidden from view in laboratory books or publications. To improve the availability of phenotypic data in the

long term, the databases BRENDA [1] and BacDive [2] collect data manually extracted from publications, standardise them and make them systematically accessible.

### ENZYME DATA IN BRENDA

In the BRENDA database, enzymes have been characterised with all their properties for 30 years. BRENDA has become one of the world's most important and widely used information systems in the life sciences and is one of the ELIXIR Core Data Resources. In BRENDA, data from a wide array of sources are combined, researchable and processed for users. Manual text evaluation is by far the most time-consuming method, but it will remain an indispensable tool in the foreseeable future for providing scientists with structured information that is not otherwise accessible in the literature. So far, 150,000 references from research literature have been manually evaluated by scientists for about 93,000 enzymes, and a total of 4.7 million data have been extracted. However, to obtain a complete overview of the literature on the classified enzymes, additional text mining methods can be used. With their help, infor-

mation concerning the occurrence of enzymes in organisms has been quadrupled compared to the results of manual evaluation procedures. A total of 3.8 million citations from the literature could be collected this way. In addition, data from other databases are also automatically integrated, including protein sequences from the UniProt sequence database and 3D structures from the PDB protein structure data bank.

### METADATA ON BACTERIA IN BACDIVE

Since 2012, the Leibniz Institute DSMZ – German Collection of Microorganisms and Cell Cultures GmbH has been developing the Bacterial Diversity Metadata database (BacDive), which gives access to previously unavailable microbiological research data. The first version of the database contained basic data relating to taxonomy, cultivation conditions and place of origin for more than 23,000 BACTERIA and ARCHAEA. The potential uses of BacDive have been greatly extended over the past few years. New types of data were mobilised from the internal databases of the culture collections, which had previously not been accessible to the public. After having started in 2015, data

900,000

Currently...

BACDIVE IS THE WORLD'S MOST COMPREHENSIVE DATABASE FOR BACTERIAL METADATA, WITH OVER 900,000 DATA POINTS FOR 80,584 STRAINS.

31,826

To this end...

A DATA SET WAS GENERATED FROM THE TEMPERATURE DATA OF 31,826 ENZYMES AND GROWTH TEMPERATURE VALUES FROM 21,498 MICROORGANISMS.

6.5  
MILLION

In addition...

THE MODEL IS ABLE TO PREDICT OPTIMAL ACTIVITY TEMPERATURES FOR 6.5 MILLION ENZYMES.



FIGURE 1: Successfully cultured bacteria on agar plates. ©DSMZ.

pertaining to 152 data fields have been extracted from species descriptions in literature and integrated into BacDive. As a result, data from over 6,000 species descriptions are already available. With the goal to make all phenotypic information from species descriptions available and searchable in BacDive in the pure data-based form, this collection is continuously extended. Currently, BacDive is the world's most comprehensive database for bacterial metadata, containing over 900,000 data points for 80,584 strains.

#### DATA SYNTHESIS OPENS UP NEW POSSIBILITIES

The combination of data from different sources offers great potential and opens up completely new possibilities for analysis. The obstacles to be overcome include poor findability, limited access, technical incompatibility of formats and inadequate standardisation. This is why the publication of the FAIR principles (findable, accessible, interoperable, reusable) have initiated a cultural change in science, aimed at breaking down these barriers and improving the availability and reuse of scientific data. The following is an apt example of the

added value that can be achieved by recombining data. In his recent study [3], the Swedish researcher Martin Engqvist compared the cultivation temperatures of bacteria from BacDive with the optimal temperature data for the activity of enzymes obtained from BRENDA. To this end, he generated a data set from the temperature data of 31,826 enzymes and growth temperature values from 21,498 microorganisms. With these data, he was able to demonstrate a strong correlation between growth temperature and optimal enzyme temperature, indicating that there is a close relationship between these two parameters. Combining data this way offers a wealth of possibilities for systematically investigating enzyme functions as a function of growth temperature. At the same time, this data set is only the first step towards much more far-reaching studies for the prediction of hitherto unknown parameters.

#### THE PREDICTION OF CULTIVATION PARAMETERS FOR PREVIOUSLY NON-CULTURABLE BACTERIA

Widely available, standardised information is a precondition for making predictions for previously unknown param-

eters. In a follow-up study, researchers led by Martin Engqvist developed a model based on the previously generated data set that uses protein sequence data to precisely predict the optimal growth temperature for bacteria. In addition, the model is able to predict optimal activity temperatures for 6.5 million enzymes.

The optimal growth temperature is only one of many cultivation parameters required for the successful cultivation of a new isolate. However, other studies have already found a solution to this problem. For example, a research team led by Alice McHardy has developed the software Traitair which can predict up to 67 phenotypic parameters with varying degrees of certainty on the basis of the genome sequences of bacteria [4]. These parameters include the utilisation of nutrients such as sugars and amino acids, salt concentration of the medium, morphology and oxygen dependence. This shows that by combining data from different sources and by combining models and software from various developers, it is already possible to make many predictions about the growth conditions for bacteria that could not be cultured

before. Due to the systematic improvement of the data basis, these models will contribute to reducing the tedious and costly laboratory work in the future, thus significantly increasing efficiency and throughput rates in the investigation of new bacterial species.

#### IMPORTANCE OF PREDICTIONS BY ARTIFICIAL INTELLIGENCE FOR SCIENCE

Only recently could it be shown that an artificial intelligence trained with 100,000 images achieved significantly better results in the prediction of malignant

melanoma than experienced dermatologists [5]. In this study, the researchers used an artificial neural network (Convolutional Neural Network), which then correctly detected 95% of all melanomas from a test data set of 100 images. The support of artificial intelligence (AI) in data analysis and in the prediction of previously unknown parameters opens up new possibilities. Especially when it comes to recognising relationships within large amounts of data, a well-trained AI algorithm can be superior to humans and make predictions with a high degree of precision. These predictions in turn serve as a starting point for further

research. However, predictions alone are not enough. To confirm scientific hypotheses, the validation of predictions in the laboratory will always continue to be an essential part of the life sciences. We will need data sets of high quality and with a high degree of standardisation to better exploit the tremendous potential of AI-supported analyses in the future. To ensure this, databases such as BacDive and BRENDA have an essential role to play in compiling and standardising huge quantities of research data with great efficiency and making the results available to scientists.

**REFERENCES:** [1] BMC Microbiol 2018;18(1):177. DOI: 10.1186/s12866-018-1320-7. [2] Ann Oncol 2018;29(8):1836-1842. DOI: 10.1093/annonc/mdy166. [3] Nucleic Acids Res 2019;47(D1):D542-D549. DOI: 10.1093/nar/gky1048. [4] Nucleic Acids Res 2019;47(D1):D631-D636. DOI: 10.1093/nar/gky879. [5] MSystems 2016; 1(6): e00101-16. DOI: 10.1128/mSystems.00101-16.

**AUTHORS:** Lorenz C. Reimer<sup>1</sup>, Dietmar Schomburg<sup>2</sup>, Jörg Overmann<sup>1</sup>

<sup>1</sup> Leibniz Institute DSMZ - German Collection of Microorganisms and Cell Cultures GmbH, Inhoffenstr. 7B, 38124 Braunschweig

<sup>2</sup> Institute for Biochemistry, Biotechnology and Bioinformatics, Technical University of Braunschweig, Rebenring 56, 38106 Braunschweig

# IDENTIFYING AND ANALYSING resistant clinically-relevant bacteria with the help of the de.NBI cloud

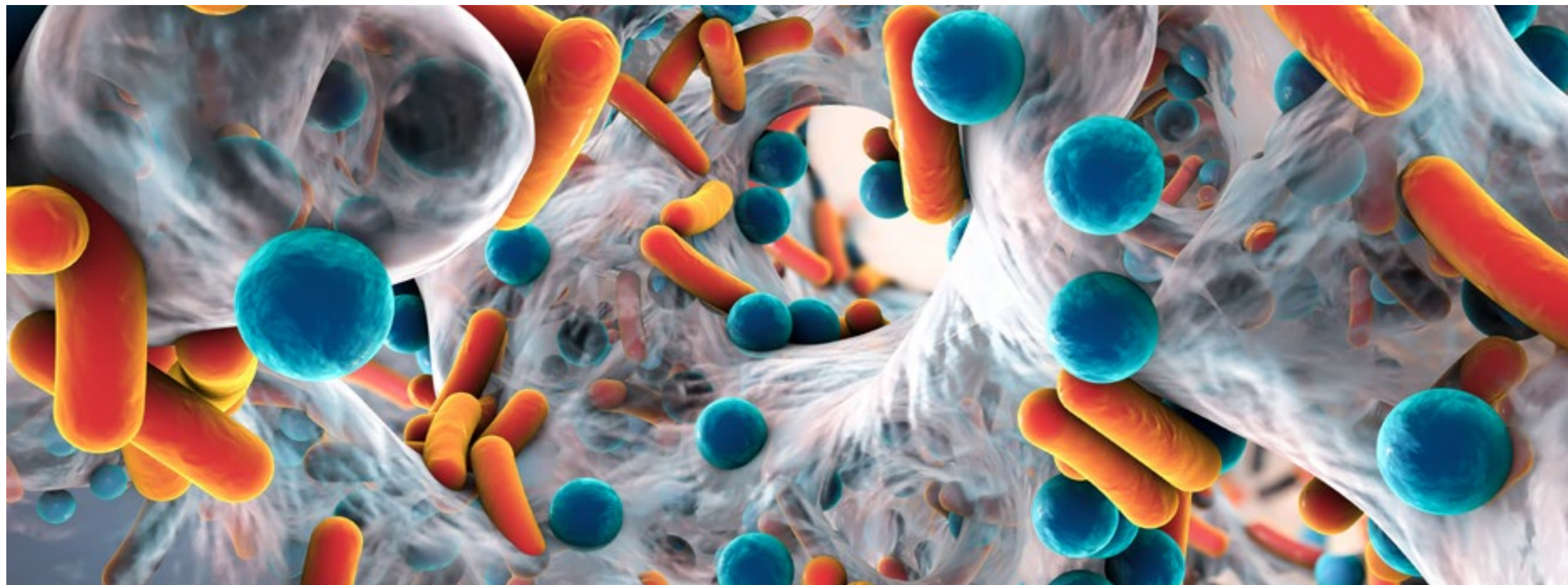
Antibiotic-resistant bacteria are becoming increasingly common in hospitals, farm animals, food and the environment all over the world. Owing to their increasing resistance – even to last-resort antibiotics – they are often difficult to confine and may even be untreatable. ASA<sup>3</sup>P software allows the comprehensive analysis of bacterial genomes, thus providing the basis for the development of new control strategies.

## THE GLOBAL THREAT POSED BY ANTIBIOTIC-RESISTANT BACTERIA

In 2015, about 670,000 infections and 33,110 deaths were attributed to antibiotic-resistant bacteria in the EU and the European Economic Area. By the year 2050, antibiotic-resistant bacteria may, on a global scale, lead to the death of up to ten million people at a cost of 94 trillion euros [1]. However, the increasing prevalence of antibiotic resistance is not

only a problem in the hospital setting. Antibiotic-resistant pathogenic bacteria have also been identified in many other areas, such as farm animals, food and the environment. In 2018, the World Health Organization (WHO) published a priority list for the development of new antibiotics against pathogenic bacteria. Carbapenem-resistant, Gram-negative bacteria (*Enterobacterales*, *Pseudomonas aeruginosa*, *Acinetobacter baumannii*, referred to as ESKAPE pathogens) were

of highest concern [2]. These multi-resistant bacteria in particular have been cropping up more and more frequently in recent years. There are growing concerns about reaching a post-antibiotic era, in which bacterial infections will become virtually impossible to treat with antibiotics. Counteracting this threat, by developing new antibiotics, for example, requires precise knowledge of the bacteria. For this purpose, their characteristics must be analysed as accurately as possible and for as many bacteria as feasible.



## THE USE OF GENOME SEQUENCING IN ANTIBIOTIC RESISTANCE RESEARCH

Bacterial characterisation methods have changed considerably over the last century. Significant progress has been made in the field of DNA sequencing over the last twenty years. Today, complete genomes of bacteria can be deciphered within a few hours. Prior to analysing the function of individual sequence segments via bioinformatics methods, the sequence of the individual nucleotides (letters) of the bacterial genome is identified. As costs are rapidly decreasing, these methods are now being used more often in combination with high-throughput methods to investigate antibiotic-resistant bacteria. This has led to a sharp increase of available bacterial genome data. For example, 219,763 strains of *Salmonella* and 106,458 of *Escherichia coli* have been sequenced until today [3]



**HIGHLY PARALLEL ANALYSIS OF BACTERIAL GENOMES THANKS TO ASA³P**

While the use of genome sequence data offers a number of advantages to characterise antibiotic-resistant bacteria, the generation and processing of such data in a high-throughput manner implies several challenges. On the one hand, a large amount of information related to these bacteria can be extracted from the genomic data – information that otherwise would not have been generated as easy and cost-efficient as with former methods. Meanwhile, sequenced genome data have become very accurate allowing researchers to generate a high-resolution genetic fingerprint of individual bacteria. This way, genes encoding for resistance to antibiotics or pathogenicity factors can be identified and relationships to other bacteria can be determined. These genetic fingerprints form the basis for the development of new strategies against antibiotic-resistant bacteria. They can also be reported back to hospitals or public health institutions in the form of simplified reports.

**FIGURE 1:** Automated analysis of bacterial genomes with ASA³P. Bioinformatics software ASA³P processes the raw data from state-of-the-art sequencing machines fully automatically and carries out comprehensive and highly specialised analyses. The diverse and complex results of the analysis are clearly visualised [2].

On the other hand, these methods quickly run into a general problem: genetic fingerprints must be extracted from a huge amount of raw sequencing data. This can still be done manually if only a few bacteria need to be analysed. But when analysing dozens, hundreds or even thousands of bacteria simultaneously, automated and highly parallel analysis software will be required, as the amount of output data generated is constantly increasing and is currently in the dimension of several terabytes already.

In order to achieve a focused and comprehensive analysis of genome sequence data, the analytical software ASA<sup>3</sup>P (Automatic Bacterial Isolate Assembly, Annotation and Analyses Pipeline) was developed in cooperation with the German Center for Infection Research (DZIF, led by Prof. Dr Trinad Chakraborty) and the working group headed by Prof. Dr Alexander Goesmann at the de.NBI site in Gießen [2]. ASA<sup>3</sup>P has been optimised to process sequence data obtained by applying leading sequencing technologies. In a first step, the analysis software subjects the genome sequence data to a quality control procedure and sorts out faulty data. The remaining data are then used to derive the genetic information of the individual bacteria (genetic fingerprint). At last, the genetic fingerprints of several bacteria can be compared. ASA<sup>3</sup>P

creates high-resolution genetic fingerprints of hundreds of bacteria within hours – a task that would have taken several weeks or even months to complete in the days of manual approaches. This was accomplished by special technical adjustments, allowing the optimal exploitation of the enormous capacities of de.NBI cloud computing infrastructure – if required. The de.NBI cloud provides scientists from various disciplines with extensive computing capacities to research both scientifically exciting issues and problems of urgent social concerns.

#### APPLICATION EXAMPLES OF ASA<sup>3</sup>P

The ASA<sup>3</sup>P software is used within national and international cooperations. As a result, more than 5,500 bacterial

pathogens from Germany, Europe and Africa have already been systematically analysed, leading to new findings on how to combat antibiotic resistance. Two application examples of ASA<sup>3</sup>P will be presented in the following.

#### HIGHLY RESISTANT BACTERIA DISCOVERED IN GERMANY

In collaboration with the DZIF, antibiotic-resistant clinically-relevant bacteria were collected, sequenced and analysed with ASA<sup>3</sup>P. Upon examination of the genetic fingerprints, it was discovered that there were extreme-drug-resistant bacteria among those analysed. These bacteria demonstrated resistance to antibiotics of many different classes, including the last-resort antibiotic agents colistin and carbapenems [4].

#### COMPARATIVE ANALYSIS OF WATER-BORNE BACTERIA

Another study with ASA<sup>3</sup>P was conducted in cooperation with journalists from NDR. The initial question was whether multiresistant bacteria could be found in water bodies and, if so, whether these bacteria had previously played a role in a

clinical context. Genome-based comparative analysis using ASA<sup>3</sup>P showed that water contains multiresistant bacteria that are highly similar to human-associated bacteria. This not only implies that water is a hitherto under-researched reservoir for multiresistant bacteria, but also that aquatic environments can pose a potential risk to humans [5].



#### OUTLOOK

The possible applications of ASA<sup>3</sup>P for the analysis of microbial genomes are almost unlimited. The genetic fingerprints generated can be combined with a wide range of clinical data to understand bacterial strategies of antibiotic resistance and develop new approaches to counteract them. The combined development of genome-based approaches and powerful software solutions is an emerging field in a systems biology approach aimed at gaining new insights into the antibiotic resistance of bacterial pathogens. In the medium term, these approaches will be transformed into diagnostic tools and used to predict future developments. The Microbial Genome Research Center (MGRC) was established as a new interdisciplinary platform to meet

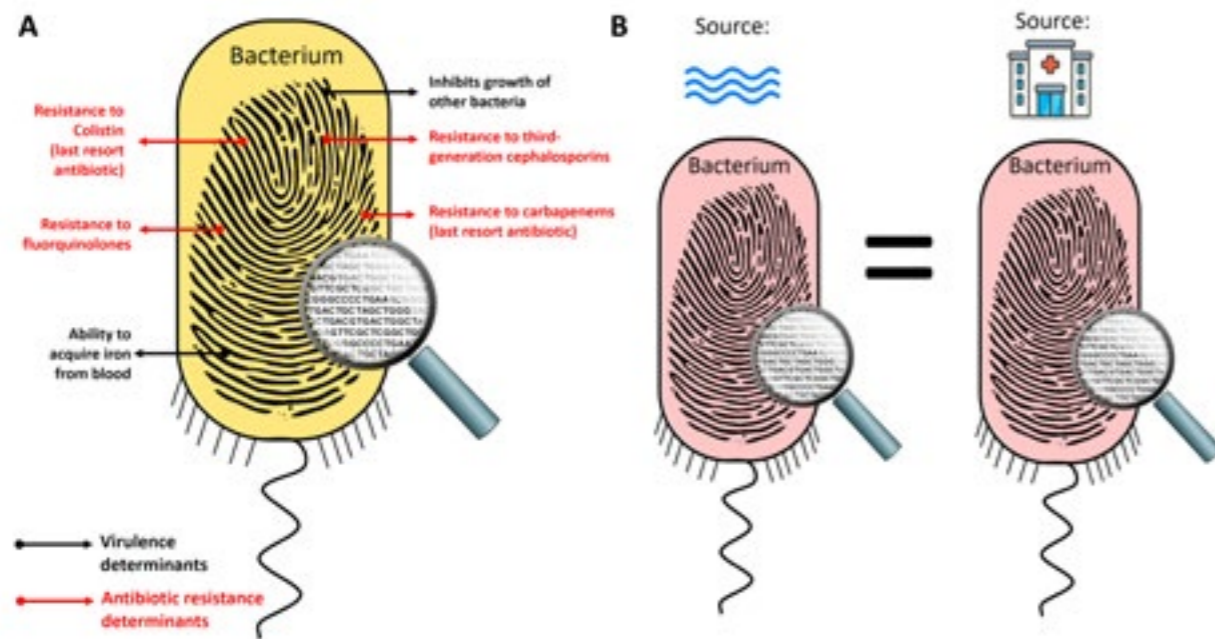
this demand. This platform includes a database component and a biobank component. The database component combines a variety of data (genetic fingerprints, data on antibiotic resistance, preclinical and clinical data sets, data from classical cohort and epidemiological studies). The biobank component gives scientists and stakeholders from industry access to well-characterised isolates, both current and historical, so that new approaches can be tested experimentally.

The MGRC thus closes the gap between basic bioinformatic analyses and medical informatics. Through the integrated evaluation of the various data available, the MGRC will contribute to assessing

the antibiotic resistance burden and to improving infection management and infection control. It aims to provide data for early warning systems to detect outbreaks and identify high-risk clones. Finally, it is intended to increase the effectiveness of measures against antibiotic-resistant bacteria and to reduce transmission in hospitals.

**REFERENCES:** [1] [https://www.ime.fraunhofer.de/de/presse/IMI\\_Project\\_GNA\\_NOW.html](https://www.ime.fraunhofer.de/de/presse/IMI_Project_GNA_NOW.html) [2] PLOS Computational Biology. DOI: 10.1371/journal.pcbi.1007134. [3] Lancet Infect. Dis. 18, 318–327. DOI:10.1016/S1473-3099(17)30753-3. [4] <https://www.dzif.de/de/wenn-antibiotika-versagen-neues-gen-fuer-antibiotika-resistenz-auch-deutschland-nachgewiesen> [5] [https://www.ndr.de/fernsehen/sendungen/panorama\\_die\\_reporter/Auf-der-Spur-der-Superkeime,panorama8258.html](https://www.ndr.de/fernsehen/sendungen/panorama_die_reporter/Auf-der-Spur-der-Superkeime,panorama8258.html)

**AUTHORS:** Oliver Schwengers<sup>1</sup>, Linda Falgenhauer<sup>2</sup>, Karina Brinkrolf<sup>1</sup>, Trinad Chakraborty<sup>2</sup>, Alexander Goesmann<sup>1</sup>  
<sup>1</sup> Bioinformatics & System Biology, University of Gießen, 35392 Gießen  
<sup>2</sup> Institute for Medical Microbiology, University of Gießen, 35392 Gießen und German Center for Infection Research, Gießen-Marburg-Langen site, University of Gießen, 35392 Gießen



**FIGURE 2:** Two possible applications for the ASA<sup>3</sup>P analytical software. a) creation of genetic fingerprints with selected examples of virulence and antibiotic resistance properties; b) comparison of bacteria from different sources.

# PHYLOGENETIC ANALYSES AS A TOOL FOR IDENTIFYING PATHOGENS

The continuous development of DNA sequencing over the last 15 years has made it possible to examine whole groups of bacteria for similarities and differences. One of the most established tools in this field is the EDGAR platform, which is provided in the de.NBI network and is used worldwide both in basic taxonomic research and to address practical clinical questions.

8,079

Currently...

PROJECTS FOR 322 GENERA WITH  
A TOTAL OF 8,079 GENOMES ARE  
AVAILABLE IN EDGAR.

4,400

In addition...

THERE ARE 226 PROJECTS ENCOMPASSING  
4,400 GENOMES IN WHICH TYPE STRAINS  
OF TAXONOMIC FAMILIES CAN BE ANALYSED.

The ongoing development of modern DNA sequencing methods has made it possible to examine whole groups of bacteria for similarities and differences, an approach known as comparative genomics. If the lineage relationships between the various bacterial species are the main focus, this is referred to as phylogenomics. One of the most established tools in comparative genomics and phylogenomics is the EDGAR platform, developed and provided by the Bielefeld-Gießen Resource Center for Microbial Bioinformatics (BiGi) at the University of Gießen as part of the German Network for Bioinformatics Infrastructure (de.NBI).

## THE EDGAR PLATFORM FOR PHYLOGENOMICS

Over the last ten years, the EDGAR platform [1] has become one of the standard tools in comparative genomics. EDGAR offers a wide range of analysis and visualisation functions such as calculating the divided and individual genetic configuration within genomic groups, Venn diagrams to represent the differential gene distribution, circular genome plots or multiple synteny plots. A particular focus of the software is set on phylogenomics. The web-based software gives users access to a wealth of tools to analyse lineage relationships and the taxonomic classification of bacterial species. In particular, it provides methods for calculating genealogical trees

and genome-to-genome distances; known methods include the average nucleotide identity (ANI) or the average amino acid identity (AAI).

## The EDGAR database contains 12,479 genomes.

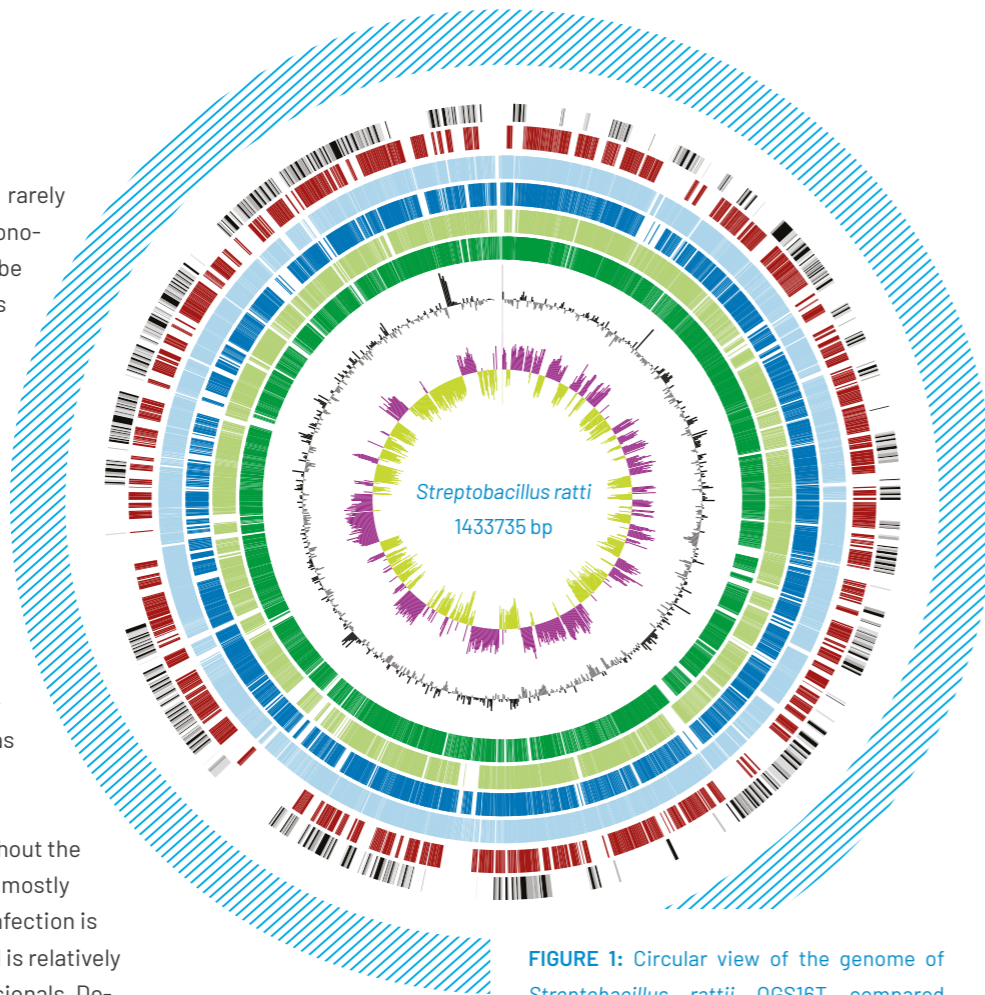
The methods implemented in EDGAR are freely available to scientists working in precomputed projects. This service encompasses a huge number of bacterial genomes contained in a public database. Currently, projects for 322 genera with a total of 8,079 genomes are available. In addition, there are another 226 projects with 4,400 genomes in which type strains of taxonomic families can be analysed. The EDGAR database, offered as a service of de.NBI, thus comprises a total of 12,479 genomes.

Besides the public EDGAR database, EDGAR also enables users to analyse unpublished data as part of scientific collaborations in password-protected projects. In recent years, a very successful cooperation has been developed with the Landesbetrieb Hessisches Landeslabor (LHL), the consumer protection agency of the State of Hesse, in the fields of veterinary medicine, food analysis and agriculture. The following section presents some of the scientific results achieved through cooperation between the LHL and de.NBI service EDGAR.

### RAT-BITE FEVER

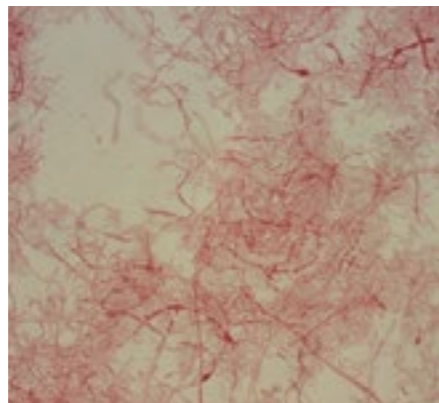
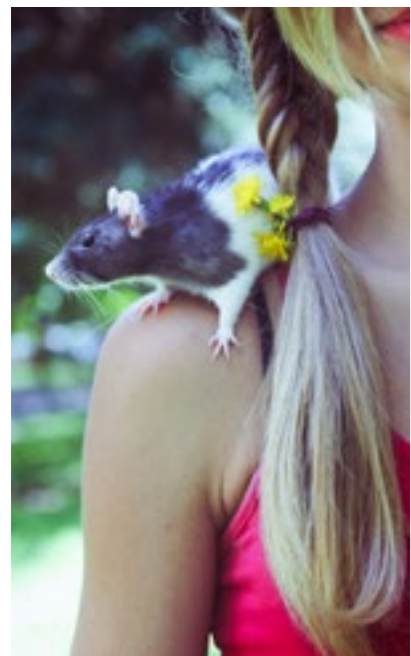
Rat-bite fever is a comparatively rarely diagnosed and largely unknown zoonosis – an infectious disease that can be transmitted from animals to humans (Figure 2). *Streptobacillus* (*S.*) *moniliformis* is the most important pathogen behind it. In humans, rat-bite fever is characterised by high fever, reddish skin rashes and inflammation of the joints; serious complications (brain abscesses, heart valve inflammation or bloodstream infections, for example) can be fatal. Occasionally, other animals also contract the disease, including turkeys, various rodent species as well as koalas and non-human primates.

Although rat-bite fever occurs throughout the world and the colonisation rate of the mostly unaffected rat can be over 90%, the infection is considered to be underdiagnosed and is relatively unknown even among medical professionals. Despite intensified research, especially the variability



**FIGURE 1:** Circular view of the genome of *Streptobacillus rattii* OGS16T compared to four other *Streptobacillus* genomes. The outer black ring shows the distribution of the genes in *S. rattii*. The red ring shows the genes conserved in the four reference strains. The green and blue rings each show the arrangement of matching genes in the selected *Streptobacillus* genome *S. moniliformis*, *S. hongkongensis*, *S. felis* and *S. notomytis*.

**FIGURE 2:** Rat-bite fever and a number of other zoonotic diseases can commonly be transmitted even by colour morphs of the brown rat breed, which are bred as pets. The often very careless handling of these pets frequently leads to illnesses, especially among children. (Photo left: <https://pixabay.com/de/photos/ratte-m%C3%A4dchen-park-457984/>, photo right: Tobias Eisenberg)



of the pathogen, its pathogenesis and its virulence factors remain unclear [2]. Our own investigations have shown that the genus *Streptobacillus*, which has consisted solely of *S. moniliformis* for almost 90 years, is actually more abundant in species. In the meantime, this genus has been extended by four species (*S. hongkongensis*, *S. felis*, *S. notomytis* and *S. rattii*, Figure 1), at least one of which has already been mentioned in connection with human rat-bite fever. Often, these new pathogens have only been described on the basis of single or few strains. Even for *S. moniliformis*, only about 24 isolates could be assembled in a strain collection at the LHL, despite worldwide acquisition efforts. The genome was subsequently sequenced from these strains. The range of the isolates over time and distance was enormous, extending over 90 years, almost all the continents, and various host species from which *S. moniliformis* had been previously isolated.

Since the similarity of the 16S rRNA gene in particular is very high within this lineage group, making differentiations on the level of species difficult, researchers have attempted to find more distinctive gene sequences in order to advance species-specific diagnostics [3]. Phylogenetic issues within the genus as well as closely related taxonomic groups were studied with EDGAR. The EDGAR platform was also used to identify virulence genes, resistance factors and phages in *Streptobacillus*. Thus, almost a century after the first description of the pathogen, scientists are shedding light on key aspects of this neglected zoonotic disease for the first time.

### EDGAR AND THE DIRTY DOZEN

In analogy to the terrifying hit list of toxins, the US health authority CDC has compiled a similar list for potential weapons-grade biological agents. *Brucella* is on the list of one dozen bioterrorism agents belonging to the second highest priority category, because it causes serious, sometimes fatal illnesses in humans that last for months. Beyond this, brucellosis is also a zoonosis that only occurs very rarely in this country, however, it is estimated to cause 500,000 new infections annually in endemic areas. These infections result from contact with infected animals or from the consumption of raw food of animal origin. Until now, *Brucella* has been

considered a pathogen solely affecting mammals. After the working group at the LHL succeeded in detecting *Brucella* in frogs for the first time in 2012 (Figure 3), and thus in an unexpected and relatively distantly related class of animals [4], *Brucella* was detected in other amphibians all over the world in subsequent years. A few years later, our cooperation again led to the initial detection of another class of animals, which was decoded in detail: the detection of *Brucella* in a tropical stingray [5] was followed by an extensive genomic characterisation with EDGAR, which included the participation of the Institute of Microbiology of the German Armed Forces. The strains infecting frogs and rays are very closely related to each other, currently holding an independent phylogenetic position within the *Brucella* genus. Little is presently known as to whether these bacteria cause the same serious diseases in humans as their relatives which are found to infect farm



**FIGURE 3:** At the LHL, researchers succeeded for the first time in proving that *Brucella* can also infect frogs. The photo above shows a glass frog (*Sachatamia ilex*) from Costa Rica. (Photo above: Tobias Eisenberg, photo below: iStock)



animals. However, similar strains have already been isolated from severely ill humans, without anyone having had contact with the poikilothermic host animals in question. This may be a still comparatively basal evolutionary form, and as such a transitional state between a soil dweller living on dead organic matter and an infectious agent highly adapted to mammals and humans. With EDGAR, genes from harmless soil bacteria as well as the same virulence genes of classical mammalian *Brucella* could be identified in the genomes of fish and frog strains. Further analyses will show whether these strains pose a similar threat.

#### EDGAR ON THE WAY TO THE FUTURE

These examples demonstrate the versatility of the EDGAR platform for the analysis of bacterial genomes both in basic taxonomic research and addressing specific clinical questions. Accordingly, EDGAR is one of the most widely used services offered by the de.NBI network, with users from over 200 universities and research institutes worldwide, plus an annual analytical volume of nearly 30,000 bacterial genomes.

30,000

#### The EDGAR platform...

IS ONE OF THE MOST WIDELY USED SERVICES OFFERED BY THE de.NBI NETWORK, WITH USERS FROM OVER 200 UNIVERSITIES AND RESEARCH INSTITUTES WORLDWIDE AND AN ANNUAL ANALYTICAL VOLUME OF NEARLY 30,000 BACTERIAL GENOMES.

Since the capacity of present-day sequencing systems continues to increase while costs are declining, EDGAR needs constant technical adjustments to keep up with the huge amount of data. For this reason, a complete replacement of the underlying data structure is planned, which will allow EDGAR analyses to be supplied with the required hardware resources in a way that is scalable according to the number of genomes examined. If necessary, researchers should also be able to use the extremely extensive resources of the de.NBI cloud. In conjunction with associated changes in data management, the aim is to ensure that EDGAR can be used in large-scale projects involving hundreds or even thousands of genomes. A further emphasis will be placed on the integration of new phylogenomic analyses. Various rapid alternatives to the established ANI/AAI methods are now available. They are currently being evaluated, and will be integrated into the EDGAR platform in the future.

By integrating state-of-the-art approaches based on marker genes, such as the use of the Universal Bacterial Core Genome (UBCG), EDGAR is well on the way to playing a key role in comparative genomics in general and phylogenomics in particular.

EDGAR  
WEB SERVER



**REFERENCES:** [1] In Bergey's Manual of Systematics of Archaea and Bacteria. DOI:10.1002/9781118960608.bm00038. [2] VVB Lauffersweiler Verlag 2018; URL: <http://geb.uni-giessen.de/geb/volltexte/2018/13567/>. [3] BMC Genomics 2016;17(1):864. DOI:org/10.1186/s12864-016-3206-0. [4] Appl Environ Microbiol. 2012;78(10):3753-5. DOI:10.1128/AEM.07509-11. [5] Antonie Van Leeuwenhoek. 2017;110(2):221-234. DOI:10.1007/s10482-016-0792-4.

**AUTHORS:** Jochen Blom<sup>1</sup>, Tobias Eisenberg<sup>2</sup>, Alexander Goesmann<sup>1</sup>  
<sup>1</sup> Bioinformatics & System Biology, University of Gießen, 35392 Gießen  
<sup>2</sup> Landesbetrieb Hessisches Landeslabor, Schubertstrasse 60, 35392 Gießen





# BRENDA – AN ESSENTIAL RESOURCE

## for the development of biotechno- logical substance production routes

This article describes the high importance of the BRENDA enzyme information system for the development of novel biotechnological processes leading to the production of complex drugs or valuable chemicals. Within the framework of such projects, BRENDA is used both in the design of new metabolic pathways, for the selection of suitable microorganisms and in the training of AI software for experimental design planning.

Biotechnological substance production is one of the fastest growing application fields in the bioeconomy. In addition to the use of naturally occurring metabolic pathways in known organisms, such as the production of alcohol in yeasts or the production of antibiotics in fungi, the production of novel products is now often conceived by combining the metabolic pathways of different organisms or by the targeted combination of enzymes leading to completely new metabolic pathways. For the selection of suitable enzymes, an exact knowledge of their properties is absolutely essential, including information on metabolism, stability, temperature, etc.

### BACKGROUND

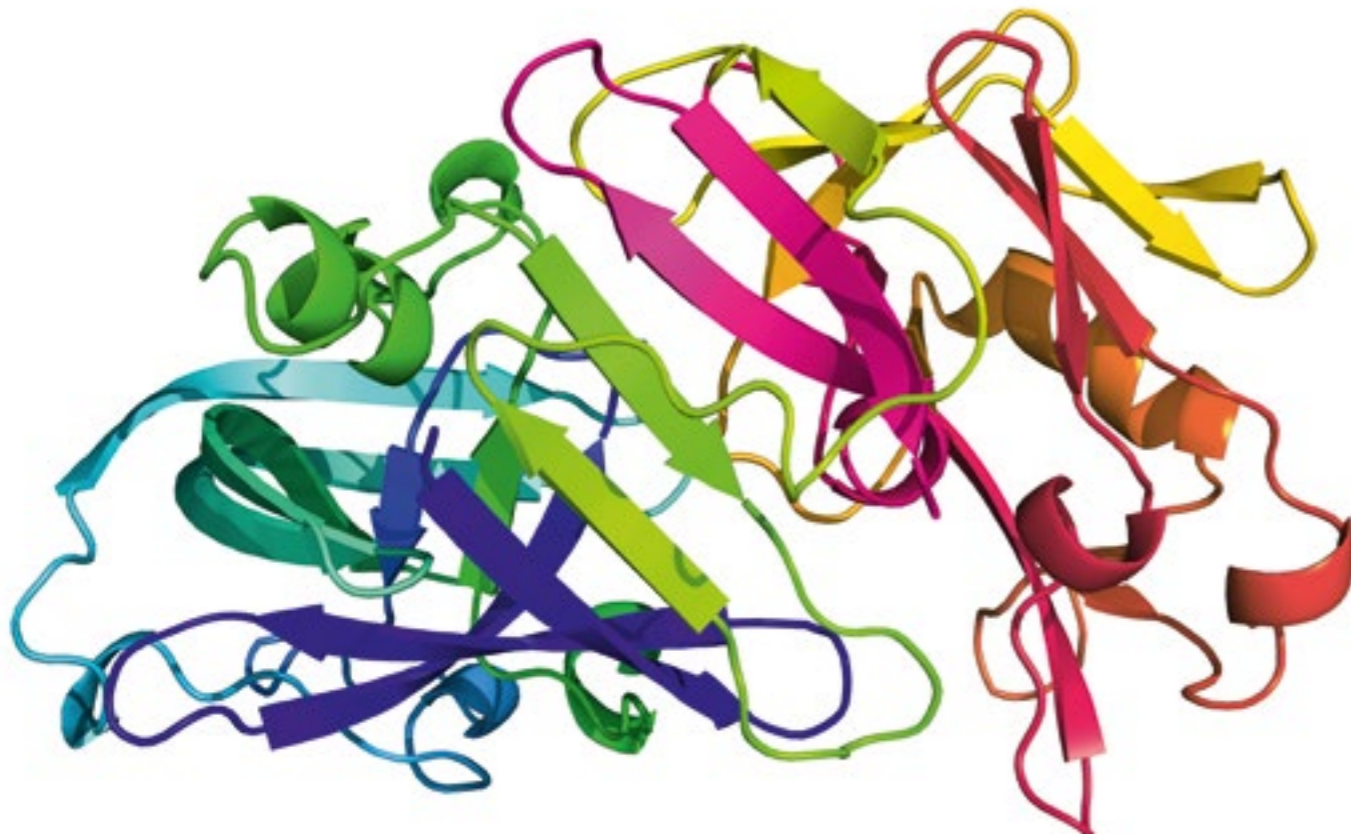
For many years now, advanced biological and biotechnological research has been unthinkable without the constant availability of facts databases. With the first sequencing of genes and proteins and the first protein 3D-structure determinations it became obvious that "big data" are created in the life sciences on a large scale. As they are absolutely essential to the efficient design of experiments, they cannot any longer be handled manually but require processing by clever algorithms.

However, whereas sequence and molecular structures are stored in repositories, most other data, such as the functions and properties of proteins, are now buried away in publications. To make them accessible in a structured form, these publications have to be evaluated, structured, standardised by means of manual work or – to a limited extent – by the use of text mining methods. Finally, they have to be made accessible to the scientific community as it is the case for enzymes in BRENDA[1].

### ENZYME DATA FROM BRENDA

For 30 years now, all functions and properties of enzymes are stored and presented in the BRENDA database. It has become one of the world's most important and most widely used information systems in the life sciences and has been selected as one of the ELIXIR Core Data Resources. We count over 80,000 users from all countries of the world every month. In BRENDA, data from a wide array of sources are combined, made researchable and processed for users. Manual text evaluation is by far the most time-consuming method, but it will remain an indispensable tool in the foreseeable future for providing scientists with structured information not otherwise accessible in the literature. So far, 150,000 references from research literature have been manually evaluated by scientists for about 93,000 enzymes and a total of 4.7 million data have been extracted. However, to obtain a complete overview of the literature on the classified enzymes, additional text mining methods are used. In particular, information on the occurrence of enzymes, the relationship between enzymes and disease, as well as certain kinetic data can be determined with high accuracy from the title or abstract of a publication by automatic text processing methods and subsequently integrated. As a consequence, information concerning the occurrence of enzymes in organisms has quadrupled compared to manual evaluation. A total of 3.8 million citations from the literature could thus be collected in this way.

Unlike pathway databases, BRENDA is not limited to naturally occurring reactions, but also includes reactions and substrates not occurring in organisms.



## BRENDA – one of the world's most important and widely used information systems in the life sciences.

Data from other databases are also automatically integrated, including protein sequences from the UniProt sequence database, 3D structures from the PDB protein structure data bank, sequenced genomes, taxonomic data, ontologies with reference to enzyme functions and much more. The addition of calculated data, e.g. the prediction of enzyme function (genome annotation), protein localisation, transmembrane regions or statistical distributions of kinetic parameters specific to particular classes of organisms, makes this information complete.

BRENDA is used for a variety of different projects from all fields of the life sciences, as can be seen in Figure 1, which shows the most common scientific terms

from the titles of about 1,600 publications citing BRENDA. The size of the letters represents the frequency of the respective keyword.

### BRENDA AND BIOTECHNOLOGICAL APPLICATIONS

A number of publications describe the use of BRENDA data for the design of enzymes with new properties as well as the design of entire metabolic pathways, which are either genetically engineered to be integrated in a specific organism or used for highly efficient *in vitro* production systems.

The authors often emphasise the fact that BRENDA data are created manually, making them of higher quality than automatically generated data. Of the approximately 50 data fields in BRENDA, researchers make use, in particular, of the broad description of the chemical conversions catalysed by each enzyme, including the reactions of synthetic

compounds, kinetic data, information on activators and inhibitors as well as information on the stability of enzymes at certain temperatures, pH values and with respect to oxygen and organic solvents. Information in BRENDA on the presence or absence in certain organisms and its cellular localisation also play a role, as does the influence the enzyme sequences and their substrate specificity or stability.

From the high number of applications, five different instructive examples from recent publications will be briefly described here. In one enzyme design project, the authors exploited the range of kinetic data for several 3,4-dihydroxyphenylacetaldehyde synthases listed in BRENDA to train an algorithm (M-path) that enabled them to construct a bifunctional enzyme that works alternatively as an aldehyde synthase and a decarboxylase and can be used to produce dopamine, for example [2].

As far as concrete applications in the construction of entire metabolic pathways are concerned, the construction of a synthetic biochemistry platform for the cell-free production of monoterpenes from glucose is particularly noteworthy [3]. The authors used kinetic values from BRENDA to construct a model for planning a system of 27 enzymes that produces, for example, limonene, pinene and sabinene stably, without any addition of ATP or NADH, with a yield of >95%, titres of >5g/l and a single addition of glucose. The product concentrations achieved with the system are an order of magnitude higher than the highest concentra-

tion being reachable by bacterial systems due to cytotoxicity of the product.

In a second project, the authors describe the production of glucaric acid from sucrose with a yield of 75% by means of metabolic engineering *in vitro* [4]. Glucaric acid is used in the food, cosmetics and pharmaceutical industries.

In a review article, the authors describe approaches for the use of “secondary activities” of enzymes integrated in BRENDA with respect to natural and artificial substrates in order to understand how new metabolic pathways evolve in

evolution and how they can be used to develop novel biotechnological processes [5]. These “secondary activities” often have a catalytic efficiency several orders of magnitude lower than their main activity and are not mentioned in typical pathway databases.

This information, which is stored exclusively in BRENDA, was used a few years ago to train an algorithm capable of detecting alternative metabolic pathways between two metabolites in organisms and to utilise this information [6].



FIGURE 1: The BRENDA Word Map – the most common keywords from the titles of 1,600 publications citing BRENDA.

**REFERENCES:** [1] Nucleic Acids Res 2019;47(D1):D542–D549. DOI: 10.1093/nar/gky1048. [2] Nat Commun 2019;10(1):2015. DOI: 10.1038/s41467-019-09610-2. [3] Nat Commun 2017;8:15526. DOI: 10.1038/ncomms15526. [4] ChemSusChem 2019;12(10):2278–2285. DOI: 10.1002/cssc.201900185. [5] Curr Opin Biotechnol 2018;49:108–114. DOI: 10.1016/j.cop-bio.2017.07.015. [6] Bioinformatics 2009;25(22):2975–82. DOI: 10.1093/bioinformatics/btp507.

**AUTHORS:** Dietmar Schomburg<sup>1</sup>, Ida Schomburg<sup>1</sup>, Lisa Jeske<sup>1</sup>, Antje Chang<sup>1</sup>, Sandra Placzek<sup>1</sup>

<sup>1</sup> Institute for Biochemistry, Biotechnology and Bioinformatics, Technische Universität of Braunschweig, Rebenring 56, 38106 Braunschweig

# HUMAN BIOINFORMATICS – BENEFITS FOR MEDICINE

In modern medicine, both individual sequence data and omics data will play an important role in the future. The use of these data offers new perspectives for research of diseases and their development – right up to early and individualised treatment.



# FROM PROTEIN STRUCTURES TO NEW DRUGS

Which proteins play a role in a particular disease and what do we know about them? What properties must an active substance have to affect these proteins? Are research data available and what about their quality? ProteinsPlus and BRENDA offer answers to questions that can already be asked as early as during rational drug design and prior to costly laboratory investigations.

Conventional drug development is an expensive and time-consuming process. Usually, the development cycle of a drug takes 14 years and costs over 800 million US dollars. Rational drug design is used to save time. It utilises computer models in advance to intensive laboratory investigations. In combination with the BRENDA enzyme information system, the web service ProteinsPlus provides important components for this process. We will demonstrate how exactly this is done in the case of the protein aldose reductase, which contributes to serious secondary diseases in cases of diabetes.

In rational drug design, the target of the drug to be developed is decided upon first. This is often an enzyme. An enzyme is a protein that facilitates or accelerates a specific chemical reaction. In the process, it comes into contact with other proteins or smaller molecules called ligands and interacts with them. Ligands may either be small molecules naturally occurring in the cell or the active substances of drugs. To develop new active substances for drugs, it is important to know not only how the protein functions, but also its spatial structure. Researchers focus on the region where the active substance is expected to bind, known as the "active site". On the basis of structural information about protein

and ligand, computer models can make predictions about the interactions between the two. This knowledge is helpful in the selection of small molecules that can serve as starting structures for the development of new active substances in drugs.

Diseases for which drugs have been successfully developed using rational design include HIV, tuberculosis, cancer, diabetes, rheumatism, and many others [1].

## APPLICATION EXAMPLE: INHIBITION OF ALDOSE REDUCTASE - REDUCTION OF DIABETES COMPLICATIONS

Aldose reductase is an enzyme (EC: 1.1.1.21); among other things, it converts glucose into sorbitol and reduces aldehydes, which are produced in various metabolic pathways. Diabetes often leads to transient high glucose levels in the blood. The conversion of glucose to sorbitol leads to an accumulation of sorbitol in the body because it can only be metabolised slowly. High sorbitol levels in turn are very harmful to the kidneys, nerves and eyes. To alleviate the secondary effects of diabetes, researchers pursue to develop drugs that inhibit aldose reductase, thus reducing the conversion of glucose to sorbitol.

2,000

### BRENDA offers...

RESEARCHERS COMPREHENSIVE  
INFORMATION ON ALDOSE REDUCTASE,  
WITH MORE THAN 2,000 RECORDS  
FROM 89 PUBLICATIONS.

#### DATA ON ALDOSE REDUCTASE FROM THE BRENDA ENZYME INFORMATION SYSTEM

The enzyme information system BRENDA [2] has become one of the world's most important and widely used information systems in life sciences and is one of ELIXIR's core data resources.

In BRENDA, data from a wide array of sources are combined and made searchable for users.

So far, 150,000 references from research literature have been manually evaluated by scientists for about 93,000 enzymes, and a total of 4.7 million data points have been extracted. In combination with text mining and data integration methods, data from a total of 3.8 million literature citations have been collected.

On the subject of the human aldose reductase discussed here, BRENDA offers researchers comprehensive information comprising more than 2,000 data entries

from 89 publications. The approximately 500 known inhibitors, which in BRENDA are linked with essential data such as inhibition constants, references to protein structure data and scientific publications, are of particular importance for the design of drug candidates. A total of 26 references lead to publications discussing the medical relevance of the enzyme for the development of drugs for the treatment of diabetes. This way, developers of new drug candidates can quickly and efficiently grasp the scientific background.

#### INVESTIGATION OF ALDOSE REDUCTASE USING THE PROTEINS- PLUS WEB SERVICE

ProteinsPlus [3] is a web service [4] that provides software tools for rational drug design developed at the Center for Bioinformatics in Hamburg. This enables scientists to select and analyse protein structures for their research online.

When working with protein structures their visualisation is of central importance, which is realised in the ProteinsPlus web server by means of the integrated NGL viewer [5].

If ligands are already present in the protein structure, they are visualised as structure diagrams and made available for use. With the help of the tools available in ProteinsPlus, important information about the properties of the active site of aldose reductase can be compiled and processed.

#### WHAT ALDOSE REDUCTASE STRUCTURES EXIST?

Protein structures form the basis of all calculations in the ProteinsPlus web ser-

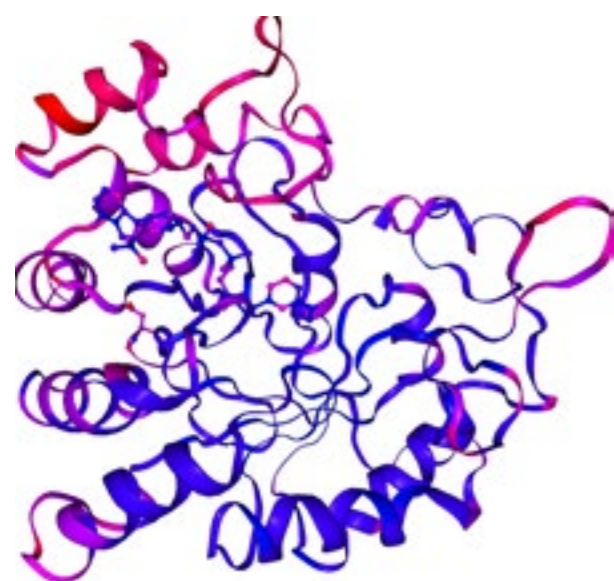


FIGURE 1: Aldose reductase structure with EDIA staining: red = poor quality, blue = good quality.

160,000

### The PDB...

PROVIDES APPROX. 160,000 BIOLOGICAL  
MACROMOLECULAR STRUCTURES.

The publicly accessible protein database PDB [6] provides approximately 160,000 3D structures of large biological molecules. Protein structures are archived using a four-digit alphanumeric code. Scientists are not required to know this code; services such as ProteinsPlus offer state-of-the-art text search functions like those we know from internet search engines. A text search for the sample protein "aldose reductase" yields 187 hits, which can be further filtered using a wide variety of criteria. We opted for

a holostructure with the code 1ah4, which, in addition to the protein, contains a co-factor that is important for its function.

#### QUALITATIVE ANALYSIS OF STRUCTURAL MODELS

When developing a drug, it is necessary to check the quality of the structural data. The ProteinsPlus web server provides two software tools for this purpose.

One of these tools is EDIA, a programme for checking a three-dimensional structural model with the underlying electron density. The electron density is the primary result obtained by structural elucidation. The 3D structure is then modelled on the basis of the electron density. Experimental data such as electron density maps contain variances and inaccuracies that are significant for further use of the structure. EDIA is used to

calculate and represent the accuracy of the model. EDIA calculations on aldose reductase structures show which parts of the protein are less well resolved (Figure 1). In our case, however, these areas are located outside the active site, which is relevant for further analyses and has a sufficiently high degree of accuracy.

#### IDENTIFYING THE ACTIVE CENTER

Following the qualitative analysis of the protein's structure, the numerical dimension of the active site is determined. Since the holostructure of aldose reductase does not yet contain a bound compound, DoGSiteScorer is used to determine the potential binding pocket. Based on topological and chemical properties, DoGSiteScorer examines the protein structure, lists possible binding pockets and calculates the probability of the binding pockets being able to interact with

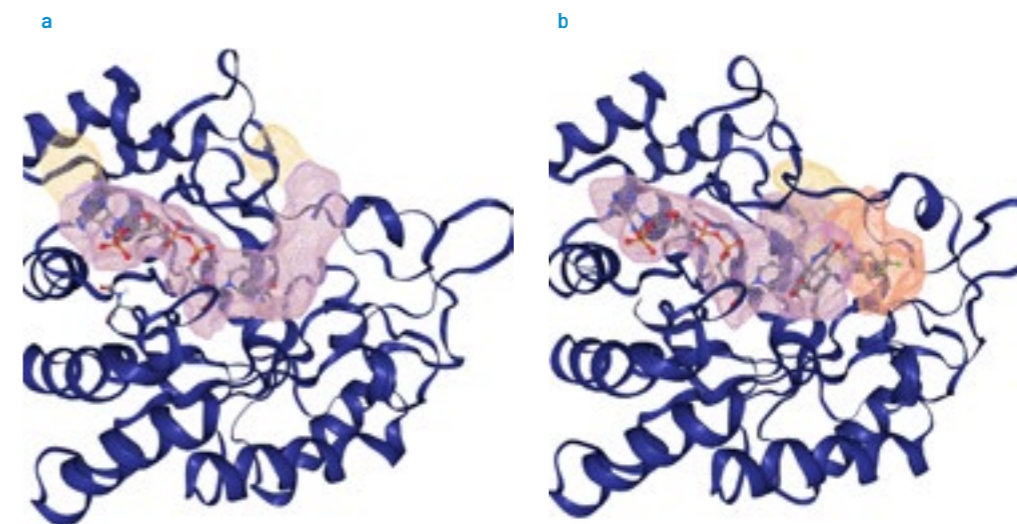


FIGURE 2: Aldose reductase structures with potential binding pockets determined by DoGSiteScorer. a) Holostructure 1ah4, purple = subpocket with cofactor; b) Aldose reductase structure with cofactor and zopolrestat 1frb, orange = opening of a further region of the binding pocket resulting from the drug zopolrestat.

an active substance. Eight pockets have been found to exist in aldose reductase, of which pocket P\_0 contains the cofactor and also has enough space for an additional small molecule, which might act as a drug (Figure 2a).

#### KNOWN LIGANDS AND THEIR STRUCTURAL EFFECTS

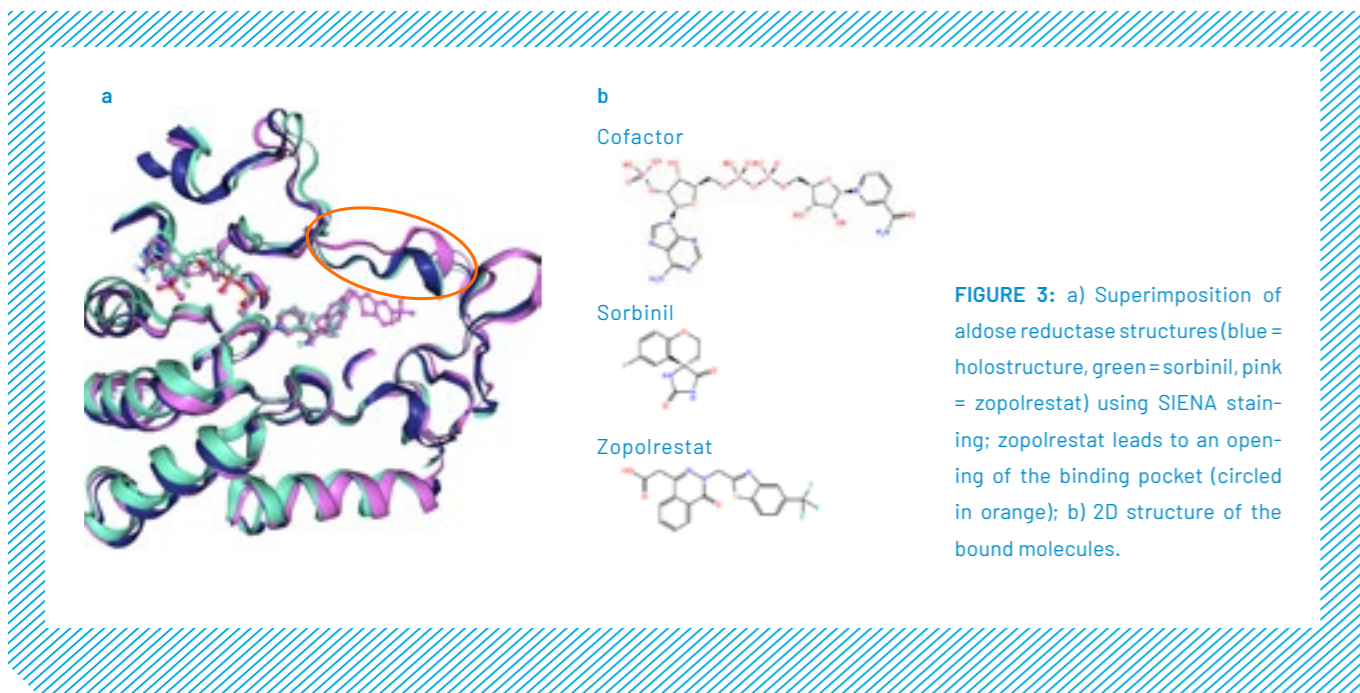
After defining the binding pocket, the next question is how flexible it may be and whether there are already known ligands, both natural substrates and drugs. To answer these questions, SIENA is used to search for highly similar binding pockets in the PDB. On the basis of the binding pocket defined by DoGSiteScorer, SIENA searches for binding pockets possessing an almost identical amino acid sequence, but which may differ in their spatial structure. The search with SIENA resulted

in 164 hits, which can now be further examined visually. Two structures – with the PDB codes 1ah0 and 1frb – contain inhibitors of aldose reductase: 1ah0 contains sorbinil, a relatively small molecule, whereas 1frb contains the much larger compound zopolrestat, which lies outstretched in the binding pocket (Figure 3). In comparison with the holoprotein, we notice that the 3D arrangement of the structure has changed and the large molecule zopolrestat protrudes into an area that was not accessible in the holoprotein. Thus, the binding pocket of aldose reductase has opened further due to the larger ligand zopolrestat (Figure 2b).

#### CONCLUSIONS FROM THE STRUCTURAL ANALYSIS OF ALDOSE REDUCTASE

The structural analysis of aldose reductase has shown that the binding

pocket, which forms the active site of the protein, is very flexible and that a ligand can lead to structural adaptations. This phenomenon, technically referred to as induced fit, illustrates the relevance of precise structural analysis in drug design. For the development of new drugs, this means that as many as possible different structures should be used to integrate a broad spectrum of structural variations. ProteinsPlus plays an important role in this process, enabling researchers to make effective use of structural data in a computer-assisted search for new active substances.



**FIGURE 3:** a) Superimposition of aldose reductase structures (blue = holoprotein, green = sorbinil, pink = zopolrestat) using SIENA staining; zopolrestat leads to an opening of the binding pocket (circled in orange); b) 2D structure of the bound molecules.

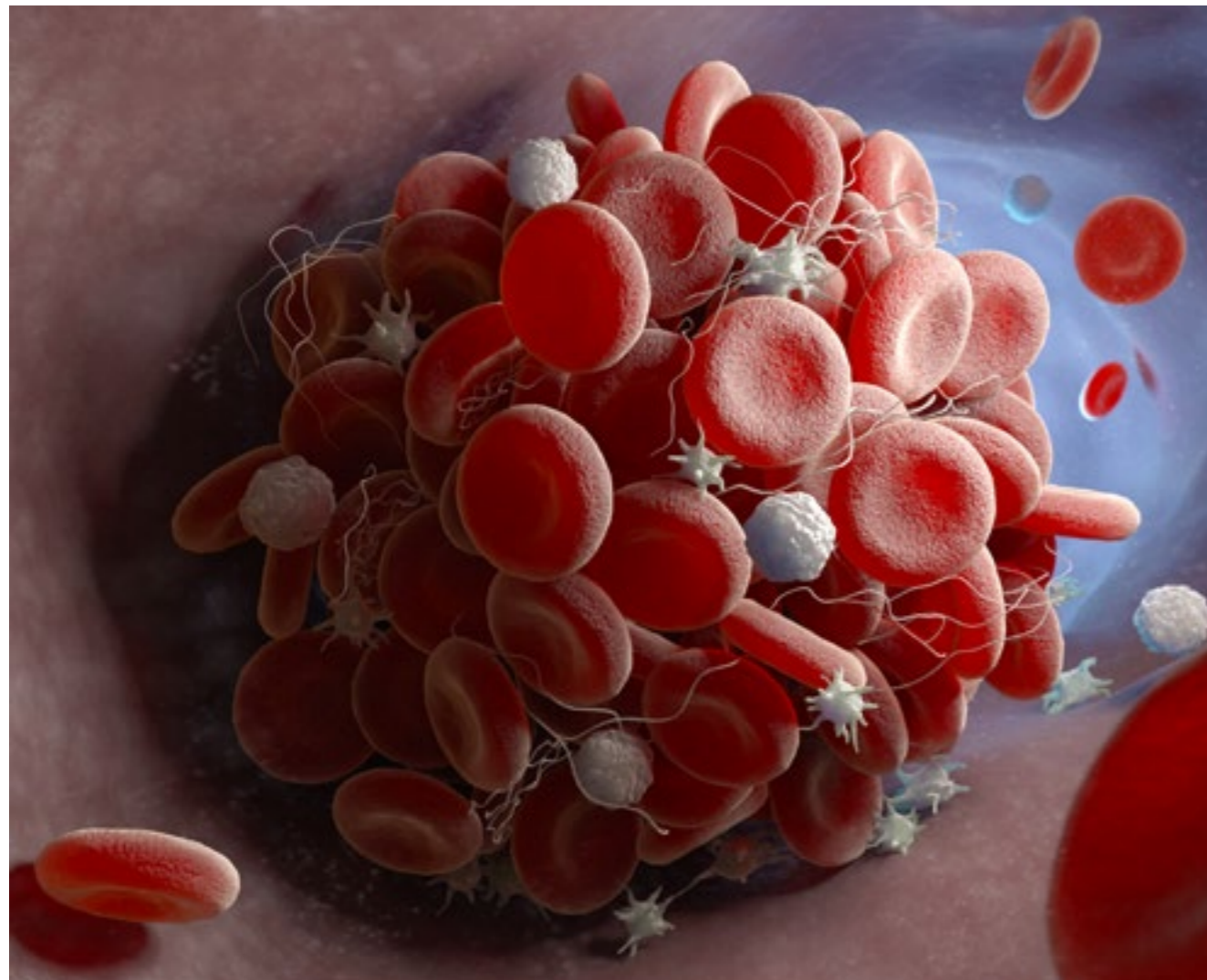


**REFERENCES:** [1] Int. J. Mol. Sci. 2019, 20 (11). DOI: 10.3390/ijms20112783. [2] Nucleic Acids Res. 2019, 47: D542–D549. DOI: 10.1093/nar/gky1048. [3] Nucleic Acids Res. 2017, 45 (W1), W337–W343. DOI: 10.1093/nar/gkx333. [4] <https://proteins.plus/> [5] Bioinformatics 2018, 34 (21), 3755–3758. DOI: 10.1093/bioinformatics/bty419. [6] <https://www.rcsb.org/>

**AUTHORS:** Katrin Schöning-Stierand<sup>1</sup>, Eva Nittinger<sup>1</sup>, Dietmar Schomburg<sup>2</sup>, Ida Schomburg<sup>2</sup>, Matthias Rarey<sup>1</sup>  
<sup>1</sup> University of Hamburg, ZBH – Center for Bioinformatics, Bundesstrasse 43, 20146 Hamburg, <http://uhh.de/zbh>  
<sup>2</sup> Technical University of Braunschweig, BRICS, Rebenring 56, 38106 Braunschweig

# LIPIDOMICS – HOW LIPIDS CONTROL BLOOD COAGULATION

Lipids – derived from the Greek word for fat – along with proteins and carbohydrates, are the most common biomolecules in every cell, responsible for various functions such as protection, energy storage and signal transduction. Taking blood platelets as an example, we present here how bioinformatics techniques can be used to analyse the lipidome, the total-ity of all lipids, and to gain important insights into blood coagulation that have medical implications.



## WHAT IS LIPIDOMICS?

Lipidomics is a relatively new field of research that uses modern mass spectrometric and other high-throughput chemical-analytical methods to determine the structure, composition and exact amount of lipids in biological samples. This makes it possible to compare lipid concentration across different patient groups, identify biomarkers or monitor the treatment progress in different diseases. However, an isolated consideration of lipids alone is not enough. This is why lipidomics also carries out interdisciplinary research to cross-link information on genes, proteins, regulation and exposure (e.g. to environmental toxins), pursuing the objective of obtaining a better systemic overview, thus to provide the basis for personalised medicine.

## BIOINFORMATIC APPLICATIONS FOR LIPIDOMICS

In Germany, de.NBI is contributing to the bioinformatics side of lipidomics within the scope of the subproject "Lipidomics Informatics in the Life Sciences" (LIFS) [1] with research partners from the Leibniz-Institut für Analytische Wissenschaften – ISAS – e.V. in Dortmund (Lipidomics Research Group, Robert Ahrends, now at the University of Vienna), Forschungszentrum Borstel, Leibniz-Lungenzentrum (Research Group Bioanalytical Chemistry, Dominik Schwudke) and the Max Planck Institute of Molecular Cell Biology and Genetics in Dresden (Biological Mass Spectrometry Research Group, Andrej Shevchenko). For this purpose, the partners are developing and maintaining programmes to perform and evaluate mass spectrometric measurements such as LipidXplorer [2] and LipidCreator & Skyline [3] to determine the identity and concentration of lipids and to compare the lipid profiles derived from different measurements with LUX Score [4]

and Clover (Figure 1). The LipidCompass reference database for lipid concentrations in various tissues is currently being developed at the Dortmund site. To this end, data samples of the model systems platelets and plasma are first archived as a reference with quantified lipids obtained from various national and international cooperation partners.

## APPLICATION: CONTROLLING BLOOD COAGULATION WITH LIPIDS IN BLOOD PLATELETS

Blood platelets (thrombocytes) play an important role in blood clotting after injuries to blood vessels. When activated as a result of such an injury, they change their shape and cross-link with their neighbours, a reaction mediated by fibrin. This leads to the formation of a blood clot (thrombus), which clogs the injured area and prevents further blood loss. Unfortunately, blood platelets are also activated by other factors, which leads to the formation of thrombi in blood vessels that are otherwise uninjured but may have been affected by previous illnesses. This has undesirable side effects, as the blockage of important blood vessels, partially or completely, interrupts the supply of nutrients and oxygen to vital organs and other parts of the body. Typical acute consequences include heart attacks and embolisms, which lead to numerous deaths worldwide year after year, and very often to severe health impairments for the affected patients. However, there are also (mostly hereditary) diseases that disrupt blood coagulation. As a result, internal as well as external injuries can lead to massive blood loss in the persons affected, since the formation of a stable thrombus to close the wound does not occur. Our application example [5] provides a preliminary inventory of lipids extracted from murine platelets and their concentrations at rest as well as after activation, and validates

them against human platelets. We focused, in particular, on gaining a better understanding of the metabolic mechanism of Niemann-Pick disease type A/B in blood platelets. Among other things, this hereditary lipid storage disease leads to a considerably reduced life expectancy of persons affected and, as a consequence of the impaired lipid metabolism, to a greatly reduced blood coagulation capacity.

We found that a particular lipid (species PI 18:0-20:4, Figure 2) serves mainly as a precursor to other lipids important to the coagulation mechanism during platelet activation. Patients with Niemann-Pick disease type A/B lack a specific protein, which can no longer convert the precursor of lyso-sphingomyelin (SPC) into ceramides, instead, it leads to an accumulation of SPC during platelet activation. SPC in turn interferes with the formation of blood clots, which has been validated using healthy human platelets after activation. However, the study also provided indications of other mechanisms that will be investigated in more detail in future work.



LIFS offers a variety of events, training courses and workshops on lipid bioinformatics

LIFS.ISAS.DE



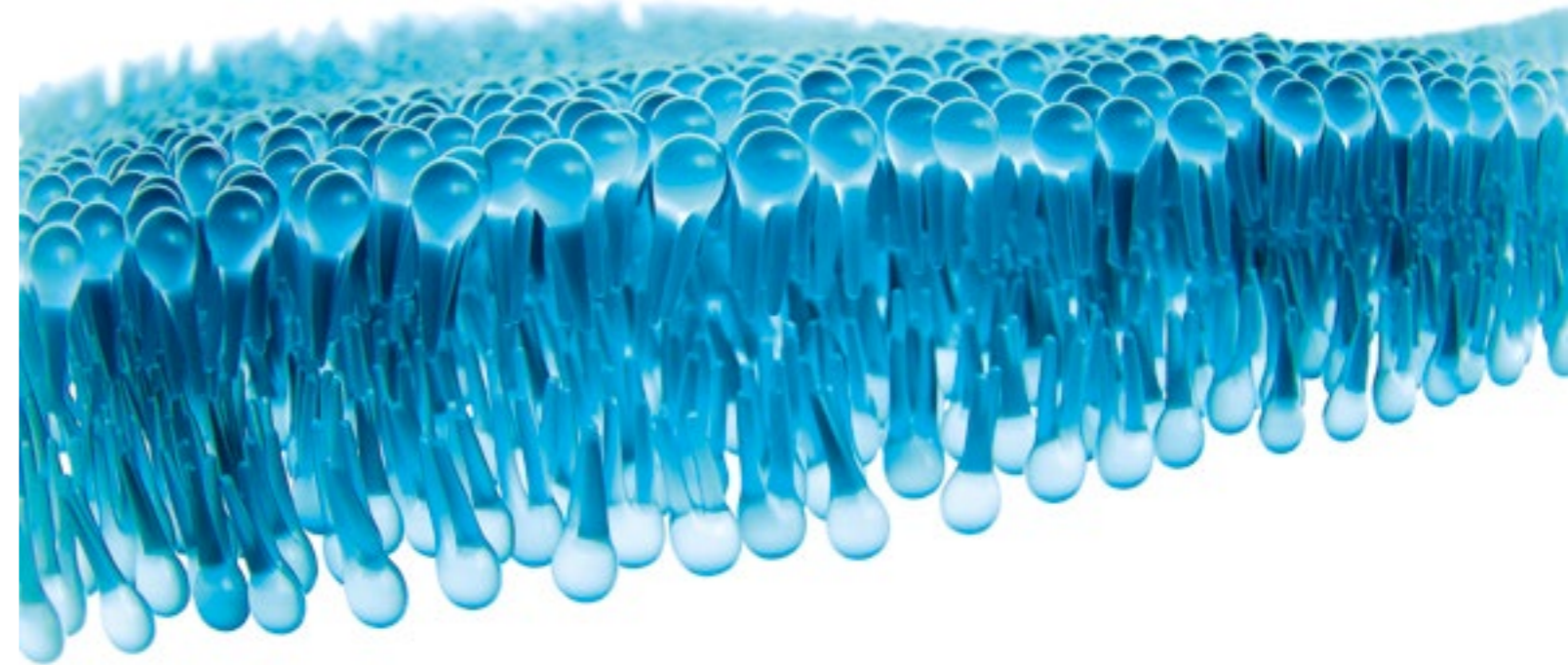
**FUTURE APPLICATION  
POSSIBILITIES**

In the future, a thorough understanding of the biochemical mechanisms behind the formation of thrombi will help physicians and pharmacologists to develop targeted drugs and treatments that can help prevent infarctions, thromboses and embolisms and better control blood coagulation. Furthermore, diagnostic biomarkers derived from the lipid profiles enable the early detection of thromboses and the monitoring of treatment progress. This research will therefore help to provide faster and more precise

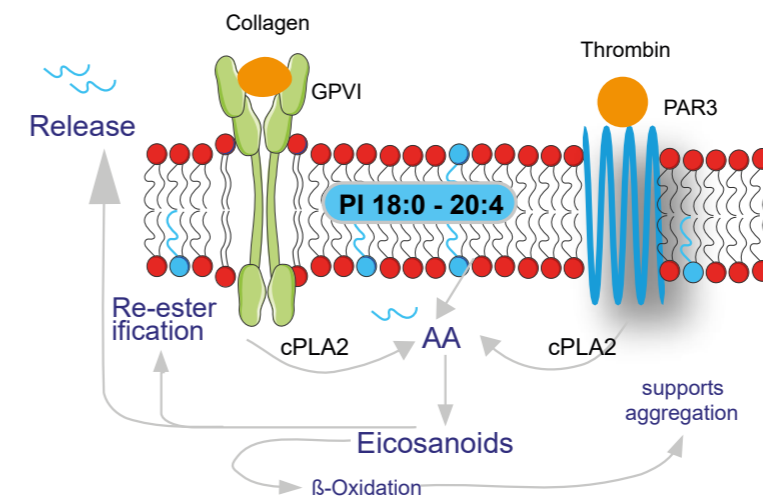
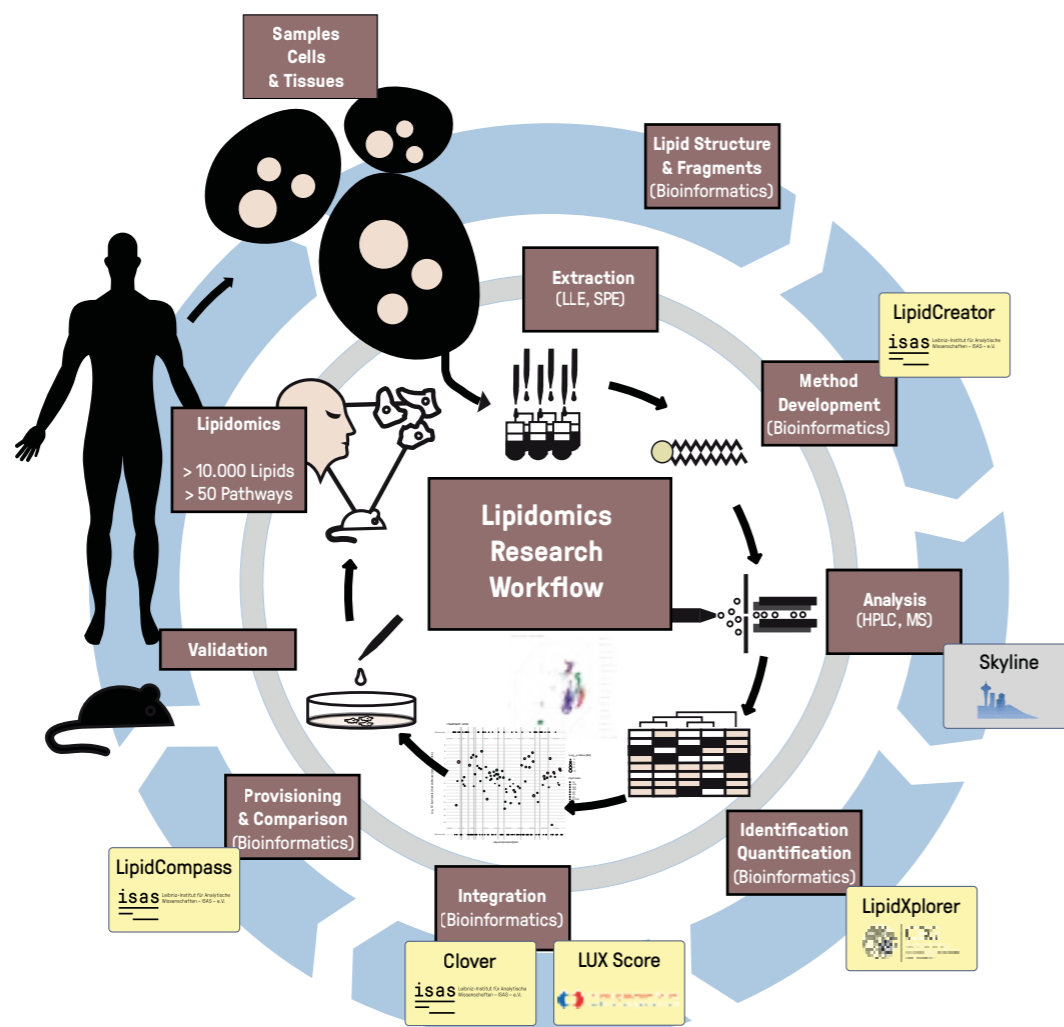
help to many patients with acute and chronic blood-clotting disorders, so that deaths and negative long-term health consequences caused by infarction and thrombosis can be more effectively prevented.

The programmes LipidXplorer and LipidCreator in combination with Skyline were used for the automated analysis and evaluation of the mass spectrometric measurements. Data integration and statistical comparisons were performed with Python and R and mapping onto metabolic networks was realised with Cytoscape [6]. Quantitative visualisa-

tions of lipid concentrations between humans and mice were implemented on the basis of an R/Shiny application. These tools significantly helped to combine, compare and interpret the large number of measurements and lipid concentrations. The LIFS project makes these self-developed applications available free of charge, so that other researchers in the field of lipidomics will also be able to use them for their own work.



**FIGURE 1:** Work steps of lipidomics research. The LIFS partners are developing programmes customised for the individual steps, such as LipidCreator, LipidXplorer, LUX Score, Clover and LipidCompass (highlighted in yellow), enabling a smooth flow of information from sampling and analysis to data integration, data visualisation and data provision. For this purpose, 5 programmes that have already been established, such as Skyline (highlighted in grey), are also being specifically expanded, in this case for instance by the Skyline plugin LipidCreator, in order to integrate them into the work process.



**FIGURE 2:** Collagen- and thrombin-induced generation of arachidonic acid during platelet activation. The figure shows two paths in which lipid mediators are formed from phospholipid PI 18:0 - 20:4, a precursor molecule, and subsequently converted or metabolised to lipid mediators. Figure adapted from [4].

**REFERENCES:** [1] Journal of Biotechnology 261, 131-136 (2017). DOI: 10.1016/j.jbiotec.2017.08.010. [2] PLoS ONE 7, e29851 (2012). DOI: 10.1371/journal.pone.0029851. [3] Nature Communications 11, 2057 (2020). DOI: 10.1038/s41467-020-15960-z. [4] PLoS Computational Biology 11, e1004511 (2015). DOI: 10.1371/journal.pcbi.1004511. [5] Blood, blood-2017-12-822890 (2018). DOI: 10.1182/blood-2017-12-822890. [6] Genome Res. 13, 2498-2504 (2003). DOI: 10.1101/gr.1239303.

**AUTHORS:** Nils Hoffmann<sup>1,5</sup>, Dominik Kopczynski<sup>2,5</sup>, Fadi Al Machot<sup>3</sup>, Dominik Schwudke<sup>3</sup>, Jacobo Miranda Ackerman<sup>4</sup>, Andrej Shevchenko<sup>4</sup>, Robert Ahrends<sup>1,6</sup>

(additional colleagues from outside de.NBI: Bing Peng<sup>5</sup>, Cristina Coman<sup>5</sup>, Canan Has<sup>5</sup>)

<sup>1</sup>LIFS 1, <sup>2</sup>BioInfra.Prot 2, <sup>3</sup>LIFS 2, Research Center Borstel, Leibniz Lung Center, Borstel, <sup>4</sup>LIFS 3, Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, <sup>5</sup>Leibniz-Institut für Analytische Wissenschaften – ISAS – e.V., Dortmund,

<sup>6</sup>Department of Analytical Chemistry, University of Vienna, Vienna, Austria



# MICROBIOME RESEARCH SHEDS LIGHT ON DISEASE DEVELOPMENT

## and opens up new treatment approaches

Microbiome research explores our microbial co-inhabitants and their influence on our health. Bioinformatic data analysis plays a critical role to reach a better understanding of the interactions between the human host and its microbiome and for us to be able to use this information to derive biomarkers – for the early detection of colon cancer, for example. In the future, it will thus contribute to the prevention of diseases and the development of new therapies.

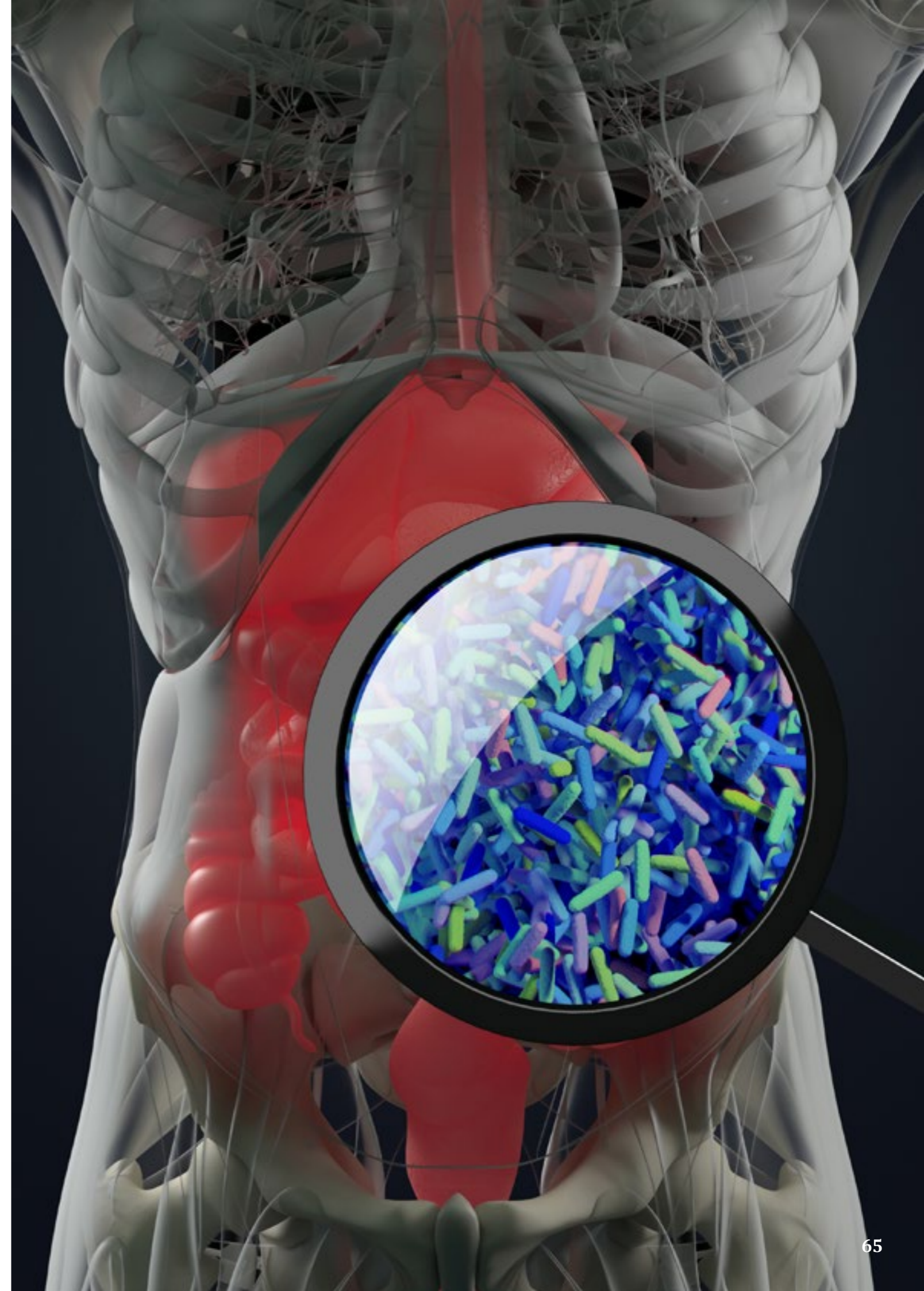
Recent research is increasingly revealing the extent to which human microbial colonisation affects our health; characteristic changes in the microbiome can be detected for a wide range of diseases. For example, microbial biomarkers applicable to the early detection of colorectal cancer have been identified and are currently undergoing clinical trials. In addition, researchers, including members of the de.NBI network, have begun to systematically investigate interactions between drugs and intestinal bacteria. The necessary tools for this have been provided not only by progress made in high-throughput sequencing of microbial genomes (DNA), but also by developments in com-

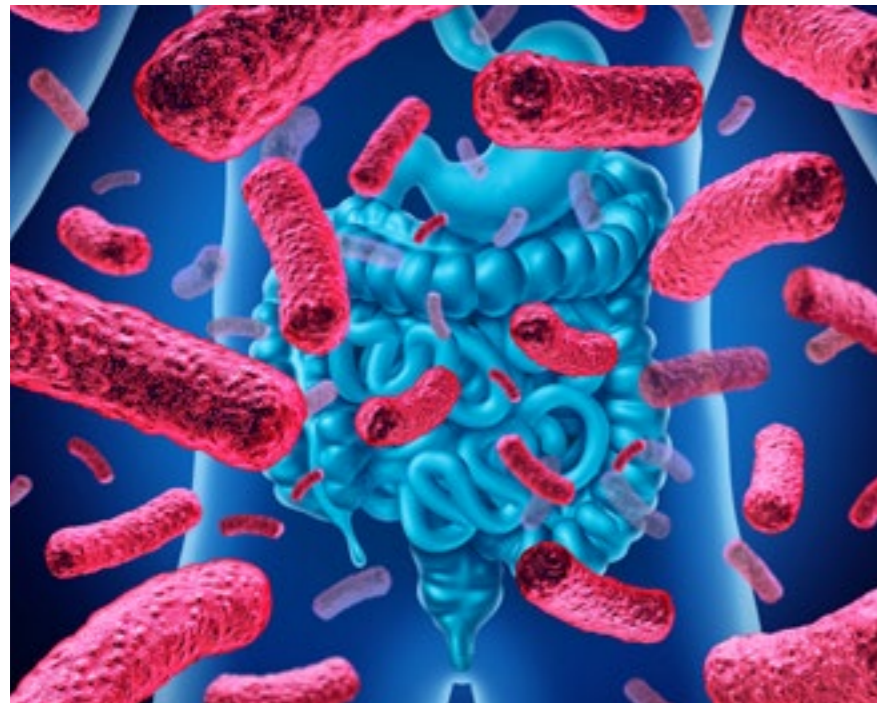
putational biology directed to evaluate the sequenced data. We would like to illustrate the key contribution of computational research in particular, using recently published studies as examples.

### THE HUMAN MICROBIOME IS SPECIES-RICH AND AS INDIVIDUAL AS A FINGERPRINT

The microbial co-inhabitants of the human body, their genes and – last but not least – their metabolic products, which have a decisive influence on the environmental milieu, are collectively referred to as the microbiome. In addition to other microorganisms such as yeasts and vi-

ruses, more than 1,000 different species of *BACTERIA* and *ARCHAEA* can colonise our intestines. The composition of this highly diverse microbial community – formerly known as intestinal flora – varies from person to person; even identical twins have different intestinal microbiomes [1]. The genetic diversity of the microbiome is even greater – the gut metagenome, defined as the total complement of all intestinal microbial genes, comprises about 100 times more genes than the human genome.





**MICROBIOME RESEARCH INVESTIGATES INDIVIDUAL BACTERIAL GENES AND THE GENETIC MATERIAL OF ENTIRE MICROBIAL ECOSYSTEMS – THE METAGENOME**

We largely owe these insights to the revolutionizing development of new sequencing technologies, which today permit the decoding of genetic information at enormous throughput rates. With “shotgun metagenomics”, it is even possible to sequence all the genes in all organisms present in a given sample simultaneously. The fact that the microorganisms do not have to be cultivated in the process is a crucial advantage, since many microorganisms do not grow under laboratory conditions.

However, analysing metagenomic sequence data poses an enormous challenge to bioinformatics. For example, categorising the bacterial diversity of a sample (Figure 1) and determining the frequencies of individual bacterial species in it (taxonomic identification and quantification) is a key step in bioinformatic analysis. To achieve maximum

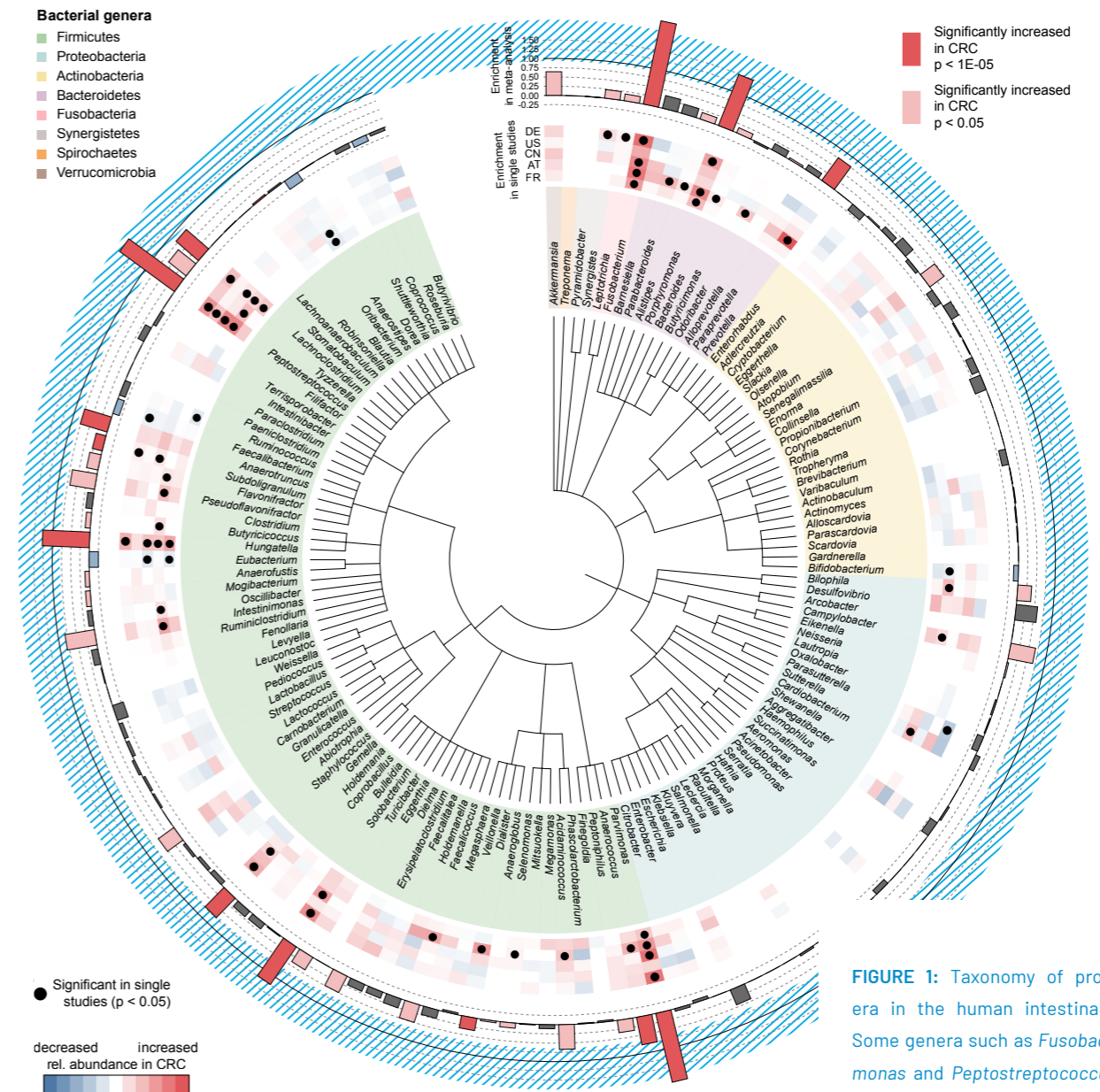
accuracy, researchers at EMBL have developed the software tool mOTUs. Its centrepiece is a comprehensive database containing genes of all the bacteria that have been cultured to date and whose genome has been decoded as well as genes that have been obtained directly from metagenomic data. Their exact classification in the bacterial phylogenetic tree allows the mOTUs software to determine the frequency of previously uncultivated bacteria in metagenomes, which significantly improves the accuracy of bacterial biodiversity analyses compared to all other analytical tools that are currently available.

In addition to such biodiversity analyses, researchers can also examine a metagenome to find out which metabolic pathways are available to the microbes, which biochemical products result from them and what significance these might have for the health status of the human organism. The fact that our understanding of microbial metabolism in its enormous diversity is still very incomplete makes such analyses difficult and often requires statistical inferences and ex-

trapolations. Comprehensive databases that map the evolutionary and functional diversity of microbial genes and metabolic pathways known to date also serve as a basis. For over ten years, researchers at EMBL have been maintaining and expanding just such a database, called eggNOG. In particular, they thoroughly investigated the accuracy and completeness of the information in these databases and compared them with other databases. On this basis, the quality of the database is constantly being improved by means of manual curation, which requires a great deal of time and money.

**THE HUMAN MICROBIOME PLAYS A DECISIVE ROLE IN HEALTH**

In-depth analyses of microbial metabolism in the human intestine have helped to revise the view that bacteria generally cause disease. On the contrary, countless studies show that a healthy intestinal microbiome contributes to our well-being. These health-promoting bacteria train the immune system, provide highly effective protection against an uncontrolled growth of pathogens, and their metabolism supplies us with many important – sometimes essential – vitamins and nutrients. Microbial metabolism is so closely interwoven with host metabolism that it even affects neural control processes and cellular regeneration [2] (Figure 2).



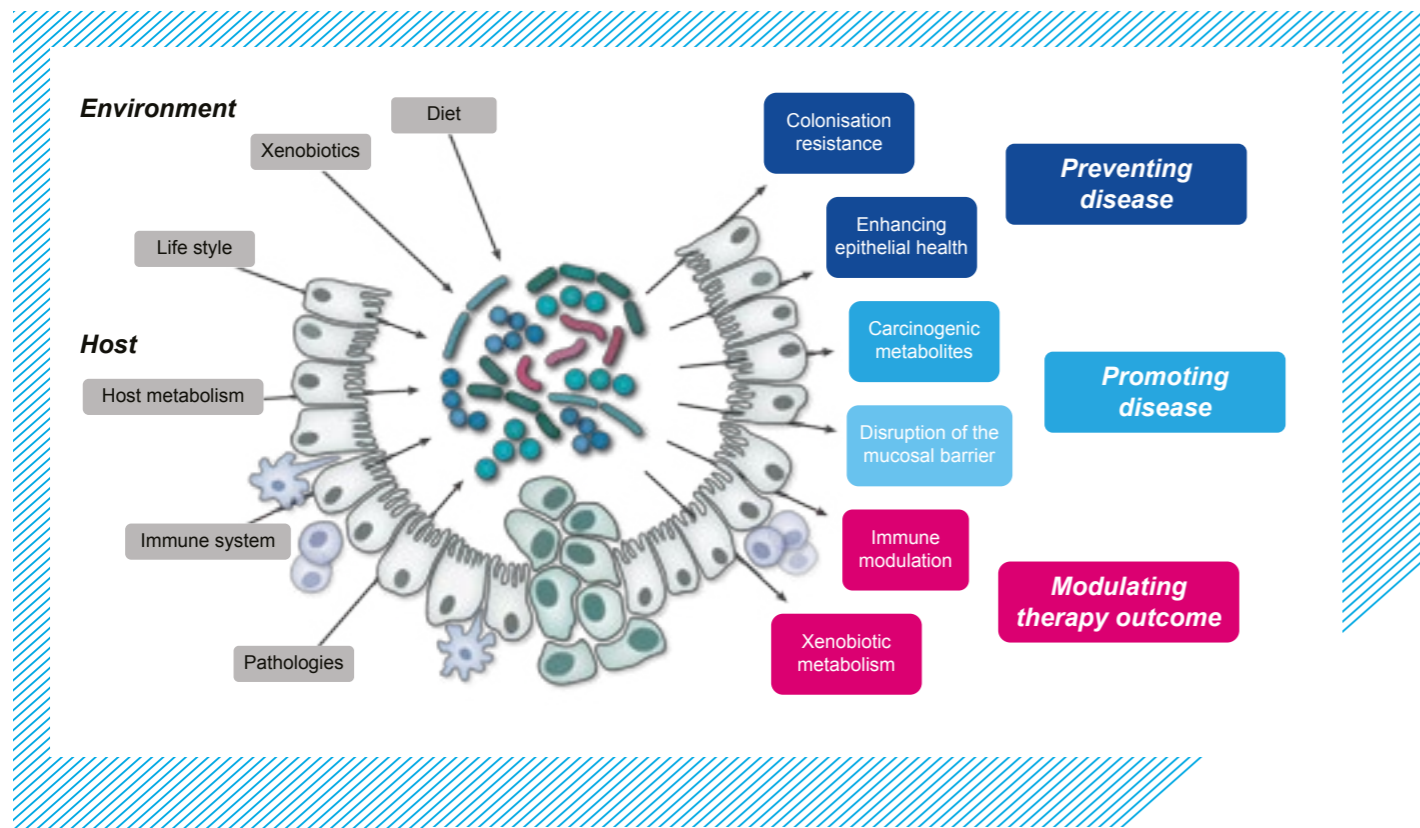
**FIGURE 1:** Taxonomy of prokaryotic genera in the human intestinal microbiome. Some genera such as *Fusobacterium*, *Parvimonas* and *Peptostreptococcus* were found to be enriched in colorectal cancer (CRC) patients relatively consistently in data from several studies.

**CHANGES IN THE INTESTINAL MICROBIOME ARE ASSOCIATED WITH MANY DISEASES**

Despite the fact that the positive effects of intestinal microbiomes on human health have been proven many times over, scientists have not yet succeeded in defining what exactly constitutes a healthy intestinal microbiome. For this purpose, comparative analyses known as association studies have also been carried out, in which the microbi-

ome of patient groups is compared with healthy subjects. This way, researchers can systematically identify changes in the microbiome associated with the disease under investigation. In fact, countless microbiome association studies have been published in recent years on a variety of diseases. Since these are often based on a small number of patients, it cannot always be guaranteed that the results are reproducible. The statistical methodology for such studies therefore plays a key role. Researchers at EMBL have provided

a whole range of such statistical tools specifically for microbiome analysis as a package on the R/Bioconductor platform. This software package, called SIAMCAT, permits the precise statistical evaluation of microbiome association studies, taking into account other potential influences (of technical nature, but also due to differences in ethnicity, diet, etc. of the host individuals) that could otherwise lead to erroneous disease associations.



**FIGURE 2:** The intestinal microbiome is influenced by a variety of environmental and host factors. These influences can lead to changes in the microbiome, which can have disease-promoting effects or affect the success of

drug treatments. Because of its individuality, the intestinal microbiome thus represents an individual-specific risk factor in the development of disease and for therapeutic complications.

So far, the statistical evaluation of many microbiome association studies and further investigations on animal models have made it very clear that the exact composition of the microbiome is decisive. While high diversity of microbial species is usually positive, individual microbes can accelerate the course of certain diseases and influence the efficacy of medications and occurrence of side effects.

#### MICROBIAL BIOMARKER RESEARCH

In the case of colorectal cancer, researchers at EMBL have shown in several publications that the composition of the intestinal microbiome can be used to distinguish tumour patients from

cancer-free subjects. A recent article published in the renowned journal "Nature Medicine" illustrates how the concepts and bioinformatics tools described above can be combined to clarify in detail changes in the intestinal microbiome in colorectal cancer patients. In a cross-study comparison (meta-analysis), EMBL scientists led by Georg Zeller and their international research partners describe the significantly increased abundance of 29 bacterial species in colorectal cancer patients in the eight studies investigated [3] (Figure 1). Their results show that the variability in the composition of the human intestinal microbiome does not exclusively depend on external factors such as nutrition and lifestyle, but that

certain types of bacteria are generally found in larger numbers in colorectal cancer patients than in the healthy population. In principle, these are therefore globally applicable as microbial cancer biomarkers. A corresponding diagnostic procedure for the early detection of cancer (non-invasive colorectal cancer screening) is currently undergoing clinical trials.

Furthermore, a detailed analysis of microbial gene functions in colorectal cancer metagenomes is shedding light on which metabolites are enriched in cancer patients. The researchers at EMBL found that the metabolic pathways for the decomposition of foods containing fat and

meat and for the synthesis of carcinogenic secondary bile salts were found in the patient samples at significantly higher levels, while those needed to break down plant carbohydrates from dietary fibre were found in smaller quantities than in samples from healthy people.

These findings regarding the intestinal microbiome are consistent with epidemiological studies on nutritional risks for the development of colorectal cancer and could be further developed into improved approaches to personalised cancer prevention in the future.

**Development of non-invasive and accurate methods for the early detection of colorectal cancer.**



#### OUTLOOK

Although microbiome research is only in its infancy, it holds great promise for improving our health and well-being.

The human microbiome is a hot topic currently being addressed by many researchers worldwide. Over the past few years, the diverse influences our microbial co-inhabitants have on our bodies have increasingly been appreciated. They regulate the immune system, chemically transform drugs, and control our sense of satiety. With the aim of enabling non-scientists to participate in their research work, researchers at EMBL led by Peer Bork initiated the study my.microbes, which is intended to contribute to a better understanding of the interaction between humans and their microbiomes with the help of a large number of test participants from the general population [4].

In the long term, it is hoped that the findings obtained about the intestinal microbiome can be used systematically for disease prevention and personalised therapy. For example, the composition of the intestinal microbiome has already been recognised as an important factor determining the outcome of immune therapies for cancer patients. Although little is known about the molecular mechanisms by which the microbiota activates the immune system, clinical studies are underway which seek to modify the intestinal microbiome to make immune therapies more effective [5].

Another milestone in microbiome research was the discovery that not only antibiotics can disturb the balance of the beneficial microbial community in our gut: other drugs have a similar effect, as a study by scientists at EMBL led by Peer Bork shows [6]. According to this study, one in four of the over 1,000 medications

investigated from all non-antibiotic pharmacological classes inhibit the growth of our intestinal bacteria – from anti-inflammatory to antipsychotic drugs.

Microbiome research is an emerging – and highly interdisciplinary – field of research in which computational biology plays a key role. The quickly and continually expanding volume and complexity of research data require ever more powerful bioinformatics algorithms and software tools, which increasingly incorporate developments in the field of artificial intelligence and machine learning. Exploiting this potential for intelligent analyses promises further rapid progress in deciphering the complex interactions between the human organism and its microbial inhabitants.

**REFERENCES:** [1] Dtsch. Med. Wochenschr. 142:267-274. DOI: 10.1055/s-0043-124940. [2] N Engl J Med. 2016 Dec 15;375(24):2369-2379. DOI: 10.1056/NEJMra1600266. [3] Nat Med. 2019 Apr;25(4):679-689. DOI: 10.1038/s41591-019-0406-6. [4] <http://my.microbes.eu> [5] Nat Med. 2019 Mar;25(3):377-388. DOI: 10.1038/s41591-019-0377-7. [6] Nature. 2018 Mar 29;555(7698):623-628. DOI: 10.1038/nature25979.

**AUTHORS:** Ulrike Trojahn<sup>1</sup>, Jakob Wirbel<sup>1</sup>, Peer Bork<sup>1</sup>, Georg Zeller<sup>1</sup>  
<sup>1</sup> European Molecular Biology Laboratory (EMBL), Meyerhofstrasse 1, 69117 Heidelberg

# WHAT THE PROPERTIES OF HUMAN CELLS TELL US ABOUT CANCER

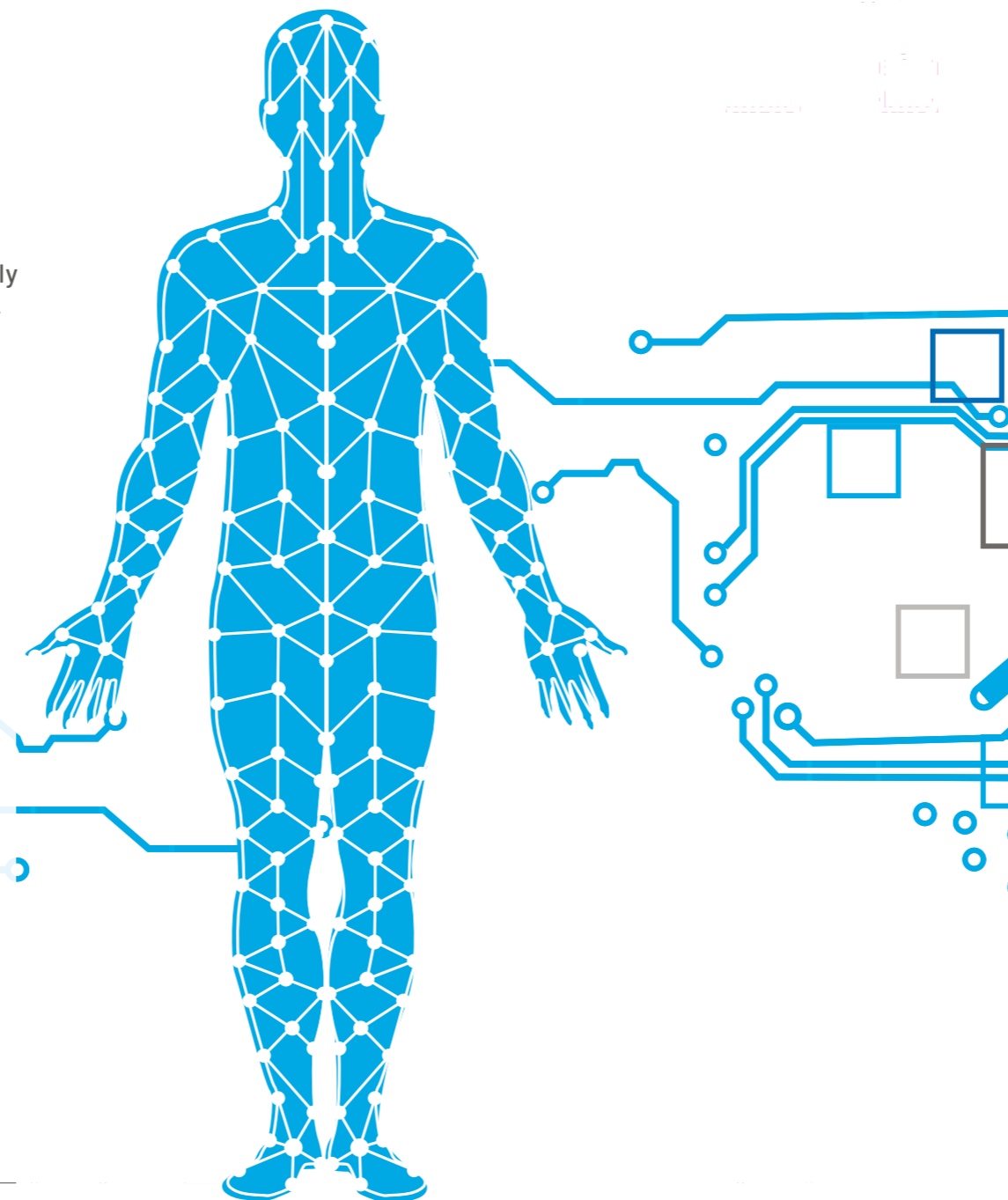
The observable characteristics – also known as the phenotype – of human cells provide information about their function and the development of diseases. A well-established IT infrastructure alongside advanced bioinformatics tools are necessary to reliably analyse the enormous quantities of digital data from cells. High-throughput methods, microscopy imaging, computer-based image analysis and biological databases play a particularly important role.

The human body is composed of trillions of cells that form many different tissues in the body. If cells deviate from their predetermined function, diseases can develop. One goal of research in the life sciences is to understand how cells function, so that treatment approaches can be found for diseases caused by malfunctioning cells or genes. For example, scientists have been examining cells in tissues with microscopical methods. Other approaches use molecular biological technologies such as RNA interference or the CRISPR/Cas9 gene scissors on a large scale to decipher the function of genes by specifically inhibiting them. Current high-throughput screening methods allow researchers to examine a large number of cells in a very short time. This creates enormous amounts of digi-

tal data (big data), which must be stored, analysed, put into a biological context and made accessible on a long-term basis. This poses high demands on the IT infrastructure and bioinformatics tools (Figure 1).

## FROM AUTOMATED MICROSCOPY TO THE IDENTIFICATION OF PHENOTYPES

A number of bioinformatics tools are now available for the analysis of large amounts of image data, such as those generated in high-throughput human cell imaging using automated microscopy, such as CellProfiler, or libraries for various programming languages such as R, Python or Matlab. Special bioinformatics knowledge is required to use many of these programmes



## The KNIME and Galaxy software platforms allow

the integration of different data sources and the execution of complex workflows.

efficiently. In contrast, the “Konstanz Information Miner” KNIME ([www.knime.org](http://www.knime.org)) software offers a simple and intuitive approach and is currently being used by Dr. Holger Erfle’s working group at Heidelberg University. With it, individual processing and analytical steps can be graphically integrated into complete workflows.

First, the image data are processed and then the higher-level data (metadata) belonging to the respective experiment – such as coordinates and the experiment-specific treatment of the cells – are assigned to them. The images are then used to identify the individual cells and record individual properties such as their brightness, shape or structure. Based on these values, the cells are categorised according to their observable properties, called phenotypes. The occurrence of certain phenotypes or their changes allow us to draw conclusions as to how cells react to the effects of various treatments, for example, the inhibition or up-regulation of individual genes.

The advantage of these workflows is that they can be reused, shared and applied to different image data by adjustment of the parameters. It is also possible to group individual parts of the workflow and link them together in a modular way. In addition, a wide range of further possibilities will open up if automatic image analysis is integrated into microscopic image acquisition, so that representative cells or rare phenotypes can be specifically captured or the resolution of selected areas can be increased in a feedback mechanism. This also reduces the time required and the data volume compared to standard high-throughput methods, which record all the data first and then evaluate them.

This technique of image acquisition with several resolution levels has been used, for example, to examine telomeres – the ends of the chromosomes – in prostate

cancer tissue. Telomeres shorten with each cell division; tumour cells must therefore be able to actively lengthen them again, so that they can continue to divide and multiply unhindered. In an approach for examining tissue samples from several patients, the microscope first acquired an overview of the samples on a slide (tissue microarray). Cell nuclei were automatically identified in these images, and the telomeres were then recorded and analysed using high-resolution 3D microscopy (Figure 2). This enabled the researchers to obtain specific information on the distribution and size of the telomeres and thus on the mechanisms of telomere elongation [1].

## CLOUD TECHNOLOGIES FOR THE WEB-BASED ANALYSIS OF MICROSCOPY IMAGES

With the constantly growing quantities of data in biomedical research, automatic analysis is becoming more and more important. Particularly large quantities of data are generated by microscopy imaging, which can be very large in number and encompass several gigapixels in size. This poses increasing challenges to the computing capacities and methods used for the computer-based analysis of the image data. Cloud-based solutions facilitate the use of a centralised, high-speed computing infrastructure. With cloud technologies, complex computing infrastructure can be made available in a transparent manner and scientists no longer need to copy image data on individual computers. Efficient and reliable automatic analysis of microscopy image data has the potential to improve the identification of disease-relevant biomarkers.

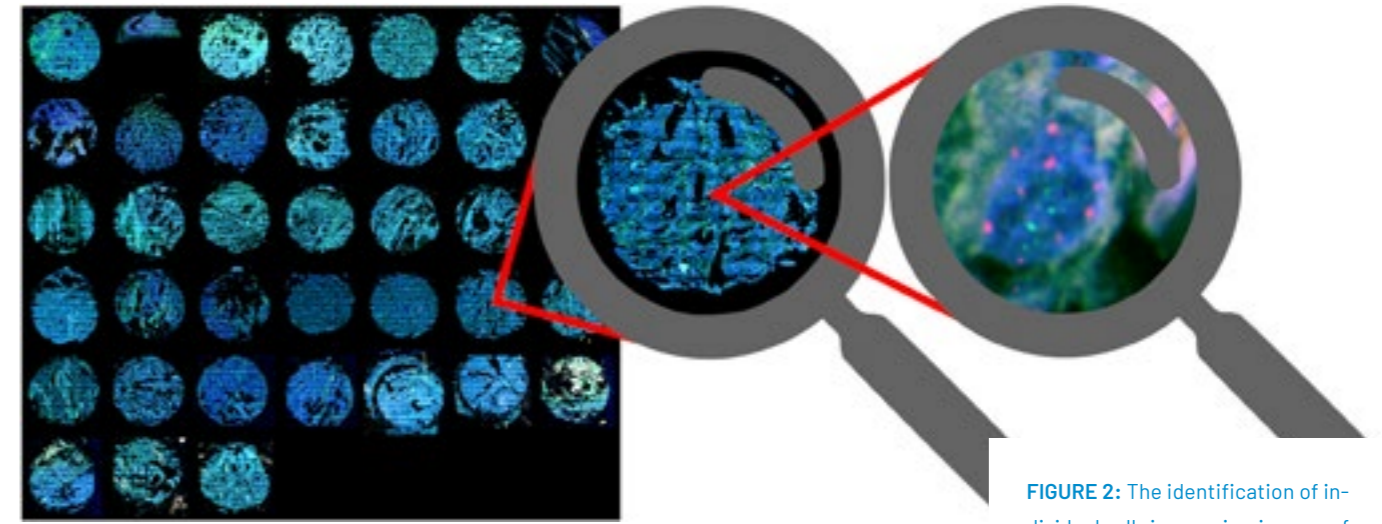
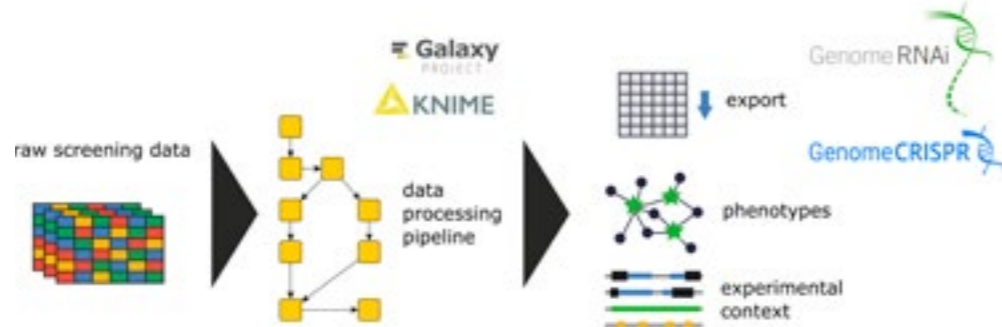
To evaluate large data sets of microscopy images automatically in the cloud, the working group headed by PD Dr Karl Rohr at Heidelberg University has extended

the web-based platform Galaxy and developed the system Galaxy Image Analysis [2]. The use of a web-based interface for the cloud allows users to perform automatic analysis in the cloud with the aid of a standard web browser. The advantage is that they no longer have to install any software on their own computers. In addition, computer scientists can use the platform to efficiently provide biologists and physicians with new image analysis methods from a central place. For example, this includes methods for image segmentation and image registration. Image segmentation is important to identify the outline of important objects such as cells or tissue. Image registration is required to bring objects taken from different viewing angles or using different imaging modalities into relation (Figure 3). Particularly good results are achieved with machine learning approaches or methods of artificial intelligence, such as deep learning. In deep learning, deep neural networks, i.e. networks of artificial neurons with a large number of network layers, are trained using examples. The training requires the IT infrastructure to have a high computing capacity, which is provided by the cloud-based system used. Galaxy was

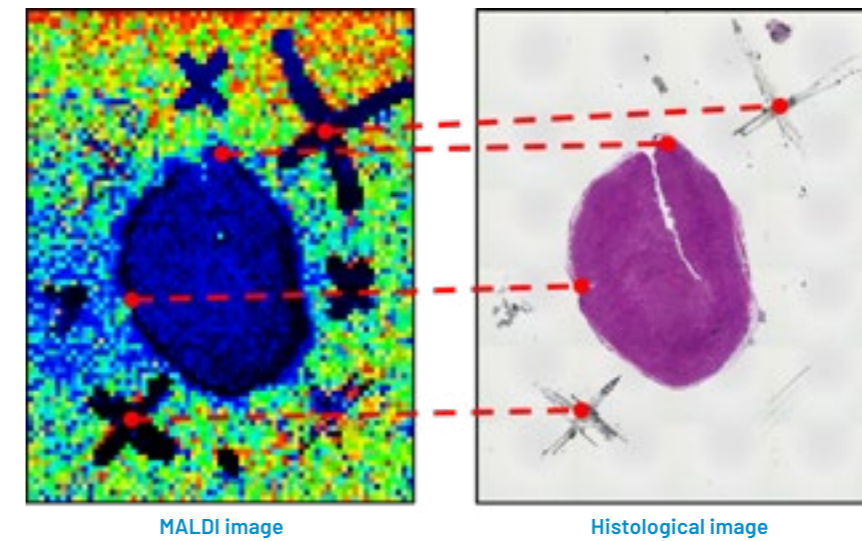
originally developed to analyse genome data. With our extension, image data can now be analysed as well as a combination of genome and image data.

In the scope of an interdisciplinary cooperation, Galaxy Image Analysis is currently being used, for example, for the combined analysis of histological microscopy images and mass spectrometry data (MALDI) (working groups of O. Schilling, University Medical Center Freiburg; B. Grüning, University of Freiburg; K. Rohr/T. Wollmann, Heidelberg University). MALDI permits the acquisition of a spatially resolved mass spectrogram of tissue relatively efficiently (Figure 3). This allows physicians to make more precise cancer diagnoses on a routine basis. A computer-based method (workflow) was developed for the automated analysis of the images [3]. In this workflow, new image segmentation and image registration methods are combined. Biologists and physicians can integrate them into their own workflows via the Galaxy platform. Galaxy Image Analysis is provided for various applications, in particular, via the Galaxy Europe Platform (ELIXIR) and the de.NBI cloud.

**FIGURE 1:** Example of processing steps in high-throughput procedures from the extraction of the original data to automatic evaluation for the identification of phenotypes (cell changes) and long-term storage all the way to classification in the biological context.



**FIGURE 2:** The identification of individual cells in overview images of a specimen can be used to achieve different magnification levels thus, enabling multi-scale imaging.



**FIGURE 3:** MALDI and histological example images with marked corresponding landmarks, which are used for registration to relate complementary image information.

The two databases GenomeRNAi and GenomeCRISPR provide structured access to data from large-scale, high-throughput experiments involving millions of measurements.

CREATING GENE CIRCUIT DIAGRAMS WITH THE HELP OF DATABASES

With the aid of high-throughput experiments and data analysis workflows, measurements can be performed systematically on billions of cells. Analysing and interpreting these data volumes efficiently requires a powerful data infrastructure. To provide structured access to data from these large-scale, high-throughput experiments that carry out millions of measurements simultaneously, the data must be stored systematically in specially designed databases. For this purpose, the working group of Prof Michael Boutros of the German Cancer Research Center (DKFZ) and Heidelberg University operates the two databases GenomeRNAi and GenomeCRISPR [4]. These databases contain results from hundreds of high-throughput experiments in which the function of genes was specifically influenced by molecular biological methods such as RNA interference (RNAi) or CRISPR/Cas9. Researchers from Germany and around

the world can access these data and use them to address biomedical questions. For example, the GenomeCRISPR database contains data from experiments in which the CRISPR/Cas9 gene scissors were used systematically to switch off individual genes in many different forms of cancer, after which the effect of gene loss on tumour growth was measured. For their growth, cancer cells depend on mutated genes that are not found in healthy body cells. The altered genes enable the cancer to grow and spread. Since these changes are vital for the disease, but not for healthy cells, the mutated genes represent interesting target for new therapies. However, it often occurs that particularly these genes cannot be targeted for technical reasons. The GenomeCRISPR database helps scientists to circumvent the problem by enabling them to create gene circuit diagrams from the comprehensive data sets and to identify further targets. For example, cancer cells often react sensitively to the loss of genes that are located close to the genes altered by the cancer in these circuit diagrams.

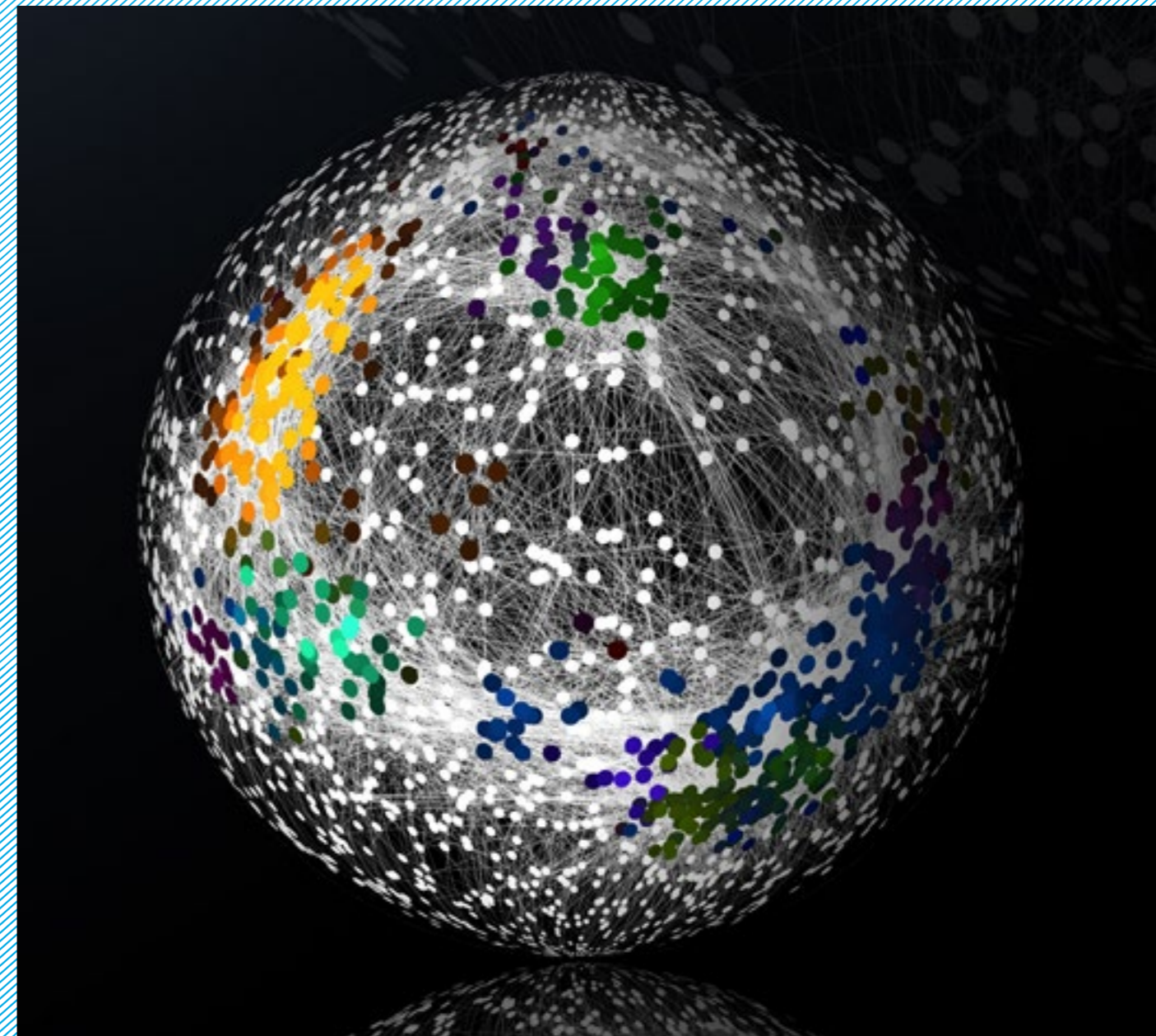
In a recent study, scientists in the working group led by Prof M. Boutros used the data in GenomeCRISPR to create a comprehensive map of the genetic circuits of cancer cells (Figure 4). They discovered that the two genes GANAB and PRKCSH control the release of Wnt ligands [5]. Because of these signalling molecules, neighbouring cancer cells can stimulate each other to grow – a process that plays a particularly important role in pancreatic, colorectal and liver cancer. This work shows how a large number of genetic screens can be integrated, thereby providing new insights through bioinformatic analyses. As more data become available, larger gene circuit diagrams can be created and yield further information about the function of (cancer) cells.

**REFERENCES:** [1] *Methods*. 2017 Feb 1;114:60-73. DOI: 10.1016/j.ymeth.2016.09.014. [2] *J Biotechnol*. 2017 Nov 10;261:70-75. DOI: 10.1016/j.jbiotec.2017.07.019. [3] *GigaScience* 2019, 8:12, giz143, DOI: 10.1093/gigascience/giz143 [4] *Nucleic Acids Res*. 2017 Jan 4;45(D1):D679-D686. DOI: 10.1093/nar/gkw997. [5] *Mol Syst Biol*. 2018 Feb 21;14(2):e7656. DOI: 10.15252/msb.20177656.

**AUTHORS:** Manuel Gunkel<sup>1</sup>, Thomas Wollmann<sup>1</sup>, Benedikt Rauscher<sup>1,2</sup>, Holger Erfle<sup>1</sup>, Michael Boutros<sup>1,2</sup>, Karl Rohr<sup>1</sup>

<sup>1</sup> Heidelberg University, Im Neuenheimer Feld 267, 69120 Heidelberg

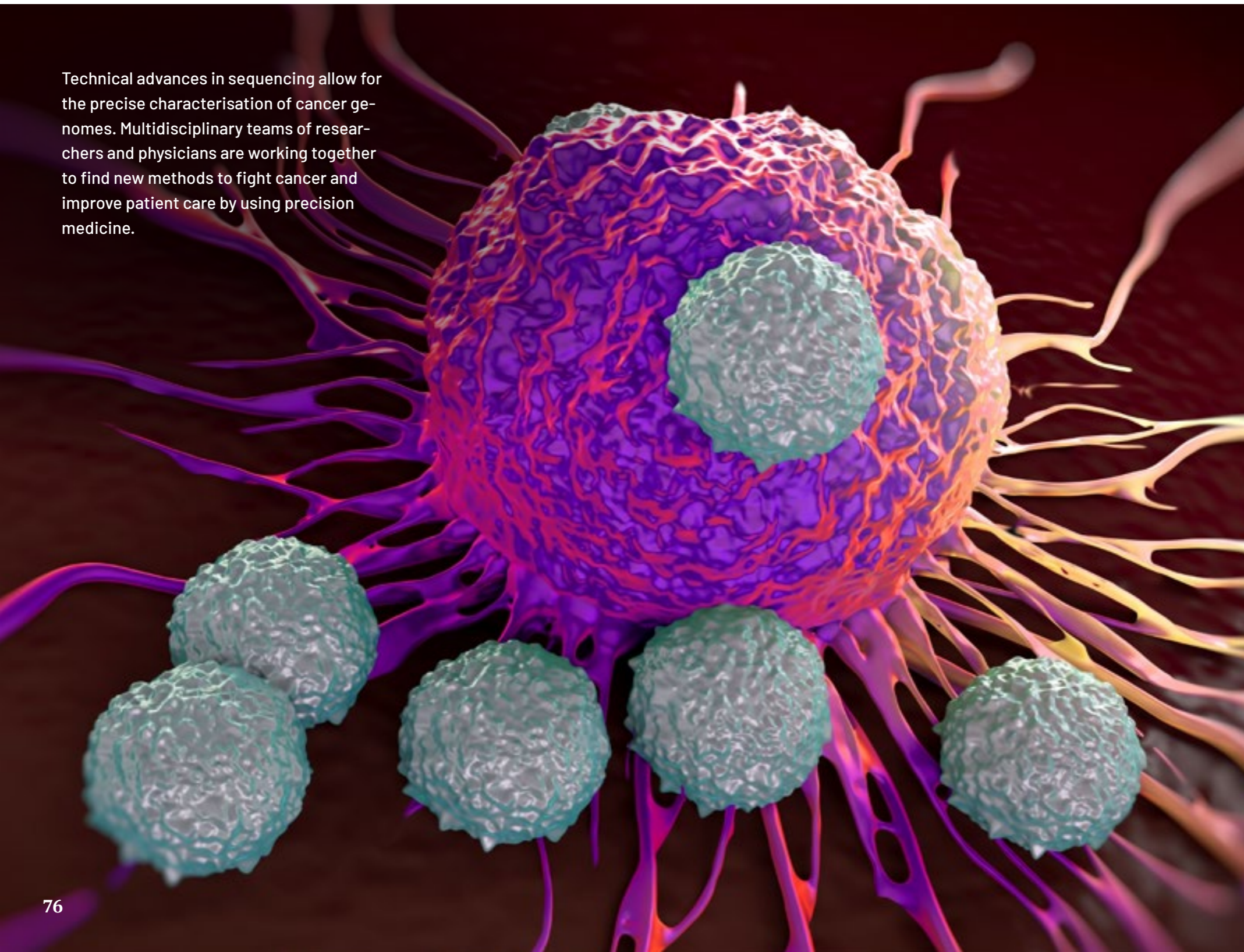
<sup>2</sup> German Cancer Research Center, Im Neuenheimer Feld 280, 69120 Heidelberg



**FIGURE 4:** A genetic circuit diagram based on the GenomeCRISPR database. Each dot represents one gene. In the circuit diagram, these will be connected to each other if they are functionally related – that is, if they are responsible for regulating the same biological processes. Specific processes that play an especially important role in cancer are highlighted in colour. This helps researchers to determine whether and to what extent hitherto unexamined genes might play a role in cancer.

# PERSONALISED MEDICINE IMPROVING TREATMENT OF TUMOUR DISEASES

Technical advances in sequencing allow for the precise characterisation of cancer genomes. Multidisciplinary teams of researchers and physicians are working together to find new methods to fight cancer and improve patient care by using precision medicine.



## DNA IS THE BLUEPRINT OF HUMAN LIFE

A C G T - these are the four DNA bases that form the building blocks of life as we know it. For over four billion years, these DNA sequences have been encoding the genetic information passed on to our descendants; they fulfil several basic functions of life: growth and reproduction. We humans have 3.2 billion DNA bases distributed across 23 pairs of chromosomes, together making up our genome. Our genome provides the template for over 20,000 genes. The product of each of these genes has a precise function, working in a complex network with other gene products, and together they govern every biological process in our body. Differences in our DNA lead to the variety of phenotypes we see all around us. However, if parts of our DNA are damaged or mutated, this can have negative consequences and lead to illnesses.

When mankind realised the importance of decoding the human DNA sequence, the Human Genome Project was launched: this was an international collaboration between 20 different institutes which sequenced and assembled the human genome over a period of 13 years (from 1990 to 2003). The cost of this endeavour was immense - 2.7 billion US dollars. Decoding the human genome sequence has helped us to better understand not only human biology, but also the origin of many diseases, including cancer.

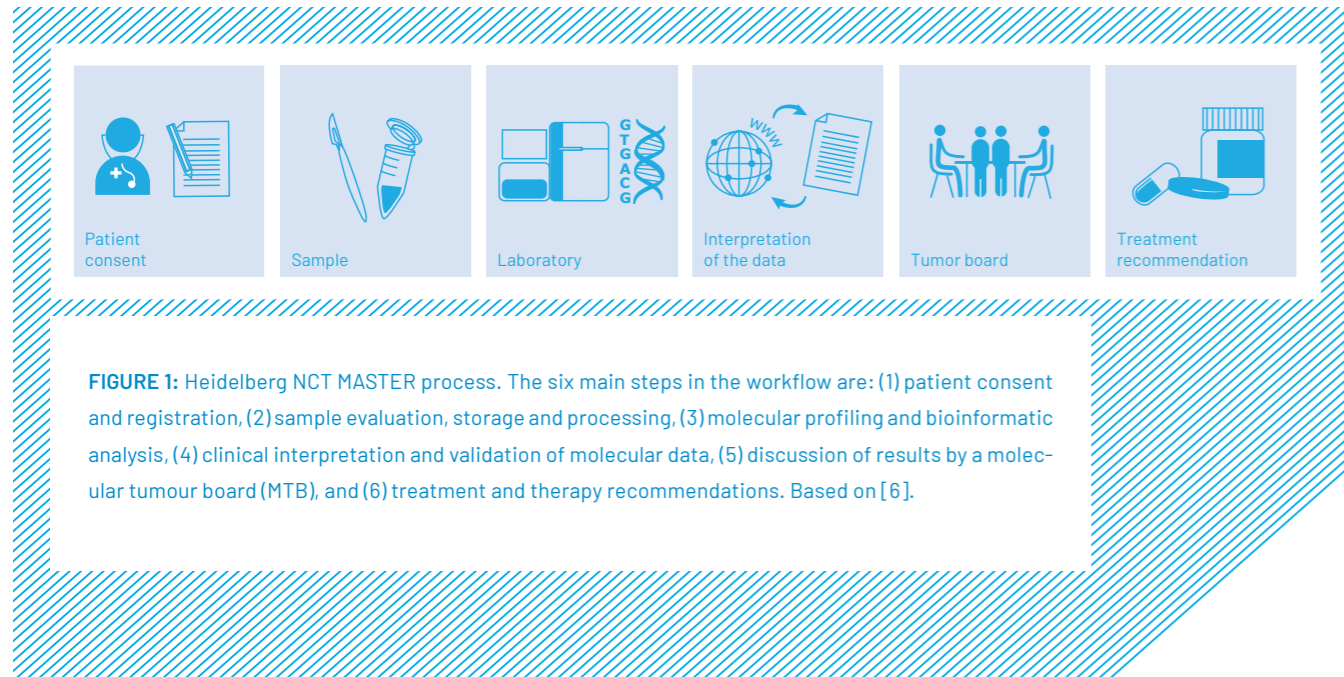
## THE BEGINNING: BIG DATA FROM GENOMICS SHEDS LIGHT ON CANCER DRIVERS

Recently, new generations of DNA sequencing technologies have been developed that have become faster, more affordable and more accessible. Using the latest technologies, we can sequence a human genome for less than 1,000 euros and in less than a week. This remarkable reduction in cost and time required to sequence a human genome has made it possible for researchers to investigate the causes of a large number of diseases, with cancer genomics becoming a focus area in recent years. This has led to international efforts to understand how DNA mutations influence the development of cancer in different forms of cancer.

The largest consortia in this field are the International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (TCGA), which together have sequenced over 23,000 patients for more than 30 different cancer types.

**That's big data!  
Over 23,000 patients have  
been sequenced for more  
than 30 different cancer  
types.**

In addition, we saw the establishment of the Heidelberg Institute of Personalised Oncology (HIPO) as a joint effort between the DKFZ, NCT and Heidelberg University.



**FIGURE 1:** Heidelberg NCT MASTER process. The six main steps in the workflow are: (1) patient consent and registration, (2) sample evaluation, storage and processing, (3) molecular profiling and bioinformatic analysis, (4) clinical interpretation and validation of molecular data, (5) discussion of results by a molecular tumour board (MTB), and (6) treatment and therapy recommendations. Based on [6].

HIPO has initiated nearly 100 projects and analysed over 3,000 patient samples to date. These consortia have brought together clinicians and researchers to address the medical and technical challenges involved in analysing these data. Overall, the most important advances have come from big data analysis methods and from multidisciplinary teams working to understand this data.

**THE ACTION: EVERY TUMOUR IS DIFFERENT AND MUST BE TREATED ACCORDINGLY**

Ever since the inception of cancer genomics, great importance has been attached to ensuring that the knowledge gained can be quickly put to use in translational research projects for patients suffering from different types of tumours. As a result, genomics has made a decisive contribution to the development of what is known as precision medicine and precision oncology. Besides extensive sequencing programmes, molecular tumour boards have been established, where physicians specialising in various fields consult with bioinformaticians

and other scientists on individual cases [2, 3]. Patients who meet certain inclusion criteria (extremely young patients; extremely rare tumour; patients having undergone all established therapies without being cured) can be provided with this state-of-the-art, but very comprehensive diagnostic tool. The recognition that drivers are not only specific for certain tumour types has led to the establishment of personalised sequencing programmes such as NCT MASTER and INFORM in Heidelberg [2, 3] (Figure 1). These programmes combine logistics, sample processing, sequencing, analysis and clinical evaluation with the objective of obtaining a therapy recommendation within four to six weeks after the biopsy. Biologists, pathologists, bioinformaticians and physicians jointly coordinate, analyse and interpret the genome sequences of cancer patients who have not responded to standard therapy. The findings are discussed by a molecular tumour board with experts from various disciplines, and then a therapy recommendation is made (Figure 2). With this personalised approach, at least one mutation is identified in 75% of cases, which can be used to guide further therapy. Two-thirds

of these are supported by clinical evidence, and the recommended therapy is implemented in over 35% of cases.

The success of these programmes is reflected in a large number of individual treatments, the use of medications approved for other tumours and the implementation of new cancer therapies such as immunotherapy. In summary, it can be said that the use of high-throughput procedures in combination with a team of specialists brings substantial additional diagnostic, therapeutic and prognostic benefits for patients.

**THE FINDING: MOLECULAR DISCOVERIES LEAD TO NEW THERAPEUTIC OPTIONS – “BIOMARKERS”**

This type of diagnosis involves the search for certain treatable constellations irrespective of the original tissue of the sequenced tumour. Treatability can result either from the presence of specific mutations in certain genes (targetable lesions) or from more general combined features. A diagnostic feature that leads to a therapy-relevant consequence is called a “biomarker”.

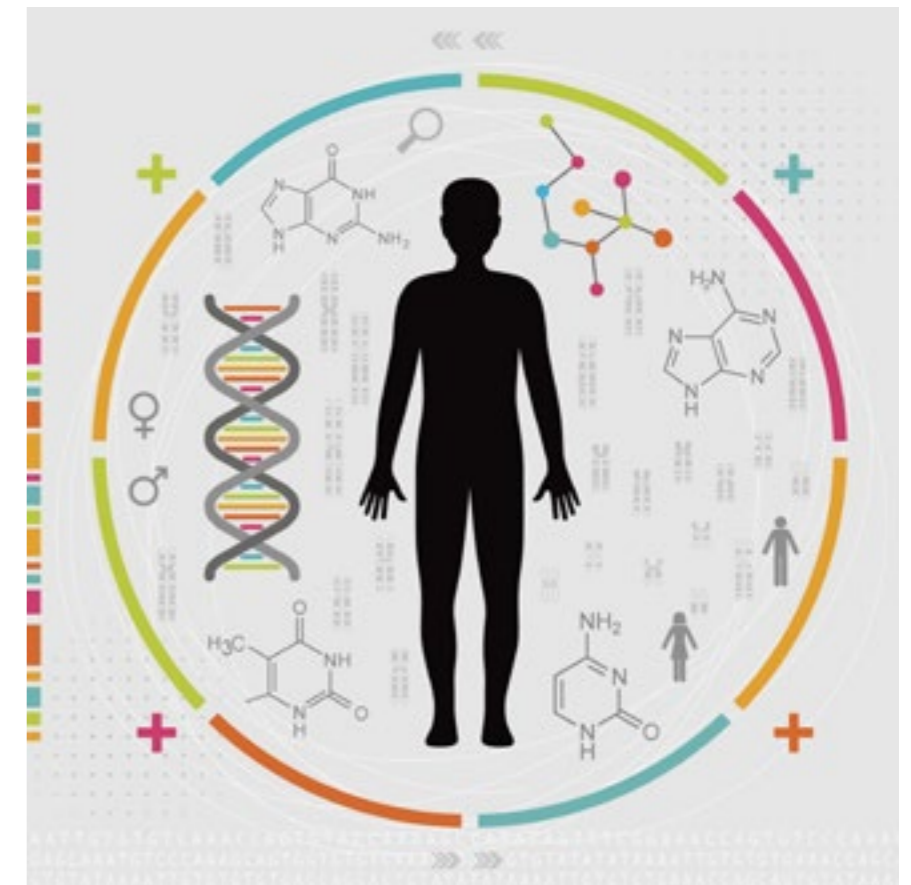
Activating mutations in tyrosine kinase receptors or signalling cascades downstream of them are a very common group of targetable lesions in tumours of different entities. In healthy tissue, these receptors regulate the communication of a cell with its environment and the cell's response to external stimuli. A constitutive overactivation of a signalling pathway that stimulates cell proliferation, for example, might lead to the development of a tumour.

The detection of certain DNA repair defects is an example of a combined biomarker. Every cell has several molecular mechanisms to detect and repair mutations. One highly efficient mechanism for the error-free correction of mutations is homologous recombination. If this malfunctions in a tumour cell because, for example, one of the genes involved in this repair pathway is itself mutated, the affected cell will accumulate more and more mutations over time. These can be corrected at least partially by other, still intact DNA repair mechanisms. However, if a patient is given a drug that inhibits another repair mechanism, the overall repair capacity of the cancer cell may be exhausted, whereas that of the healthy cells in the same patient is not, because homologous recombination still works in these cells. Such a constellation, consisting of a mutation or biomarker and the efficacy of a medication, is known as synthetic lethality. To exploit this, the underlying feature, e.g. the defect in the homologous recombination, must be precisely defined. Yet, constellations exist in which the mutation causing the defect is not found. In such cases it may be important to use pattern recognition to identify the imprint of this DNA repair defect on the genome. In the case of homologous recombination, var-

ious methods and measures have been developed (mutation signatures, HRD score and combined measures such as the score on which the TOP-ART study is based [4]).

Another example of a combined biomarker is measuring the total number of mutations in a sample, especially the total number in coding regions of the genome (only 2% of the human genome

immune system can recognise it as a tumour cell and, under certain circumstances, it can be killed and removed by T cells. This is the body's natural defence against tumours, which functions very efficiently and eliminates about 6,000 newly malignant degenerate cells in each person every day. However, some tumours have the ability to make T cells in their environment weak and sluggish by means of certain signals. A new class of drugs



directly codes genes). Any non-synonymous mutation can cause a change not only in the corresponding protein, but also in fragmented pieces of this protein (peptides), which cells present on their surface for recognition by the immune system (neoepitopes). If a cell presents a peptide containing a mutation, the

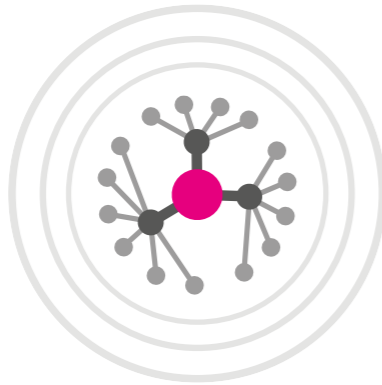
called immune checkpoint inhibitors (ICI) suppress this weakening signalling cascade, thereby leading to a reactivation of the cytotoxic T cells. The aforementioned total number of mutations in the coding region that determines the number of neoepitopes is a predictive value for the efficacy of therapy with ICI.



### THE REQUIREMENT: LARGE AMOUNTS OF DATA REQUIRE A LARGE INFRASTRUCTURE

Analysing these data requires not only great expertise, but also a large, specialised IT infrastructure. To store the data involved for sequencing the tumour and blood of one cancer patient alone, 500 gigabytes of memory is needed. With several thousand patients, this can quickly amount to several petabytes (1 petabyte = 1,000,000 gigabytes). These data must also be analysed in complex and CPU-intensive steps. To conduct research on the data, they often have to be compared with large data sets such as

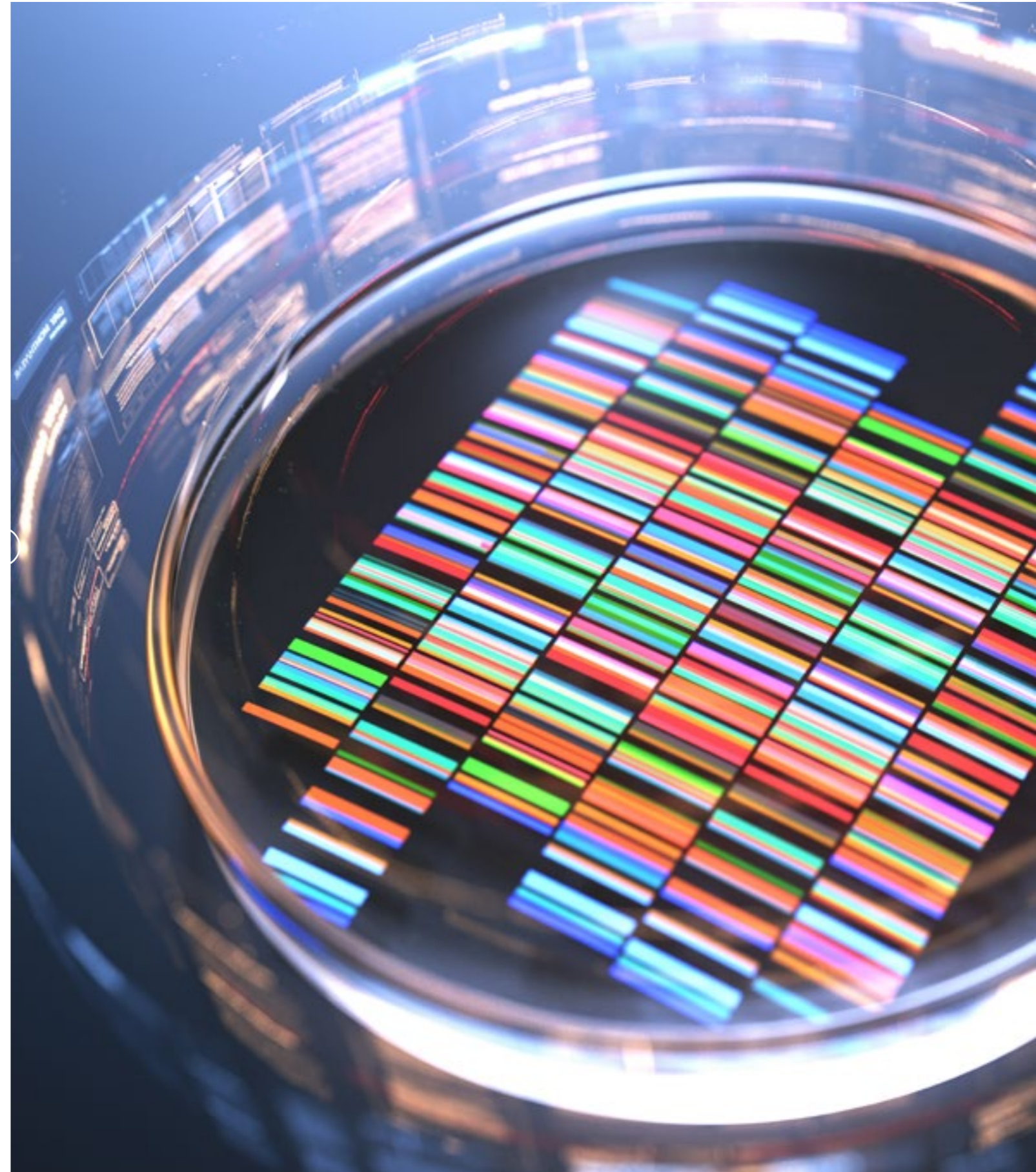
those of the ICGC or TCGA. Unfortunately, few institutes have the capacity to download, store and analyse several petabytes of data. To get around this, increasing attempts are being made to store the data records in clouds. After demonstrating that they are authorised to work on the data, scientists can analyse the data in these clouds. For example, a mirror of the ICGC data is currently being established in the de.NBI Cloud. Alongside increased efficiency, this form of data analysis also makes it possible for all scientists at all institutes to work on the large data sets. The sharing of resources releases capacities urgently needed in this area of science.



### CONCLUSION: THE FUTURE OF PERSONALISED ONCOLOGY

Driven by the success of the first precision oncology programmes, their number continues to grow. More and more university hospitals and centres are starting their own programmes, and existing programmes are being expanded, for example, by establishing a second NCT in Dresden. To ensure the same quality at all locations, standardised, easily divisible analysis procedures are required. Experience has shown that installing common software in clouds is the most efficient way to do this.

Although precision oncology has already made great strides and gained many new insights, even for rare cancers, it is still making rapid progress. Countless doctors and scientists around the globe are working to make the motto of the German Cancer Research Center a reality: *research for a life without cancer.*



**REFERENCES:** [1] Nat Com 2019;10(1):368. DOI:10.1038/s41467-018-08069-x. [2] Int J Cancer 2017;141(5):877-886. DOI: 10.1002/ijc.30828. [3] Eur J Cancer 2016;65:91-101. DOI: 10.1016/j.ejca.2016.06.009. [4] <https://www.nct-heidelberg.de/das-nct/newsroom/aktuelles/details/top-art-studie-den-krebszellen-gezielt-das-reparaturwerkzeug-wegnehmen.html> [5] Nature 2018;555: 469-474. DOI: 10.1038/nature26000. [6] [https://www.nct-heidelberg.de/fileadmin/media/nct-heidelberg/forschung/nct%20master/nct\\_HD\\_master\\_k6.pdf](https://www.nct-heidelberg.de/fileadmin/media/nct-heidelberg/forschung/nct%20master/nct_HD_master_k6.pdf)

**AUTHORS:** Naveed Ishaque<sup>1</sup>, Ivo Buchhalter<sup>2</sup>, Daniel Hübschmann<sup>2,3,5</sup>, Barbara Hutter<sup>2</sup>, Franziska Müller<sup>1</sup>, Matthias Bieg<sup>1</sup>, Nina Haberman<sup>4</sup>, Jan Korbelt<sup>4</sup>, Benedikt Brors<sup>2</sup>, Stefan Fröhling<sup>2,3</sup>, Roland Eils<sup>1,6</sup>

<sup>1</sup>Berlin Institute of Health (BIH) and Charité – Universitätsmedizin Berlin

<sup>2</sup>German Cancer Research Center (DKFZ), Heidelberg

<sup>3</sup>National Center for Tumor Diseases (NCT), Heidelberg

<sup>4</sup>The European Molecular Biology Laboratory (EMBL), Heidelberg

<sup>5</sup>Heidelberg Institute for Stem Cell Technology and Experimental Medicine (HI-STEM), Heidelberg

<sup>6</sup>Faculty of Medicine and University Hospital Heidelberg, Heidelberg

# ANALYSING THE GENE REGULATION OF HUMAN CELLS WITH THE HELP OF MACHINE LEARNING

Machine learning methods, especially deep learning, have proven to be of enormous importance in recent years in the pursuit of new insights into gene regulatory mechanisms. We have provided a new software package called Janggu, which supports the establishment of deep learning applications with genomic data. Janggu reduces the time and effort required for software development and makes it possible to answer biological questions more efficiently.

## FROM GENES TO CELLS

Each cell in the body contains our entire genetic material in the form of the DNA sequence, which is divided into various sections. Perhaps the best known of these sections are referred to as genes. Most genes contain building instructions for proteins, which in turn are essential as molecular tools for the execution of biochemical processes in our body.

Although each cell contains the entire DNA sequence, liver cells have other tasks to perform compared to muscle or nerve cells, for example. This is because specific genes are active, or expressed, in different cell types. For example, there are genes specific to the liver or muscles which are actively read and translated into the corresponding proteins in the respective cells only. This process requires a high degree of coordination, referred to as gene regulation. Although

the human DNA sequence has become known in its entirety since the beginning of the millennium, the regulation of many genes and associated processes are still far from being understood in detail. One reason for this is that gene regulation is a coordinated and highly complex process governed by DNA packaging, epigenetic modifications and the binding of proteins to the DNA sequence.

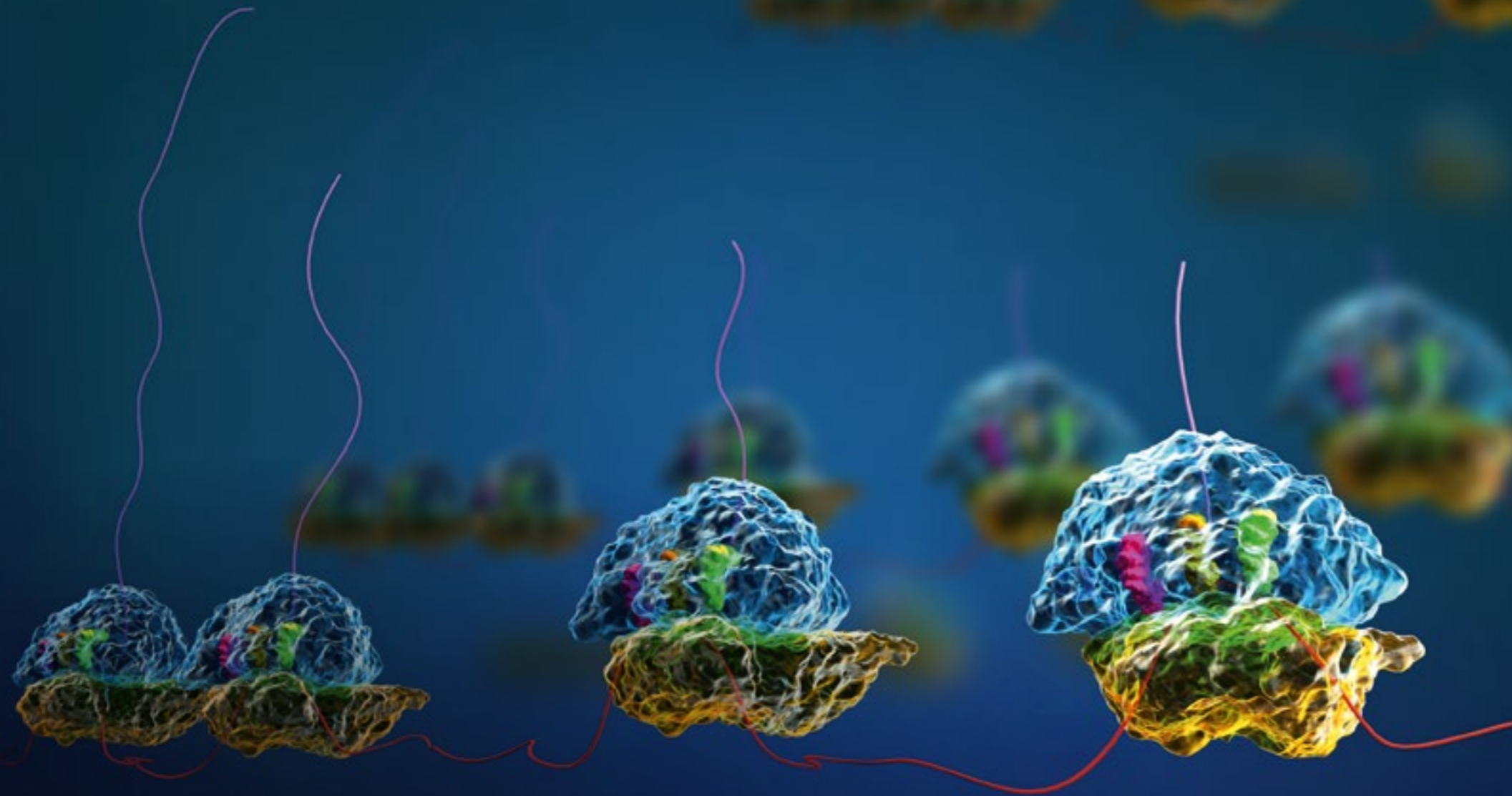
## BIOTECHNOLOGICAL METHODS HELP TO UNDERSTAND GENE REGULATION

In recent years, biotechnological advances have contributed to new insights into gene regulation - most notably high-throughput sequencing. These methods allow researchers to detect millions of short DNA or RNA sequence sections that, directly or indirectly, result from gene regulatory activities, thus allowing conclusions to be drawn about gene regulation. Examples of

such high-throughput protocols include: ChIP-seq, which identifies protein-bound regions in DNA or detects epigenetic modifications; RNA-seq, which quantifies gene expression; and ATAC-seq, which distinguishes openly accessible DNA regions from tightly packed ones. Yet, measuring these processes has led to an explosion in the volume of data: For a single experiment, data volumes totalling hundreds of gigabytes are no longer a rarity. Moreover, conducting such experiments under different conditions, e.g. other cell types, species or diseases, multiplicatively increases data growth in genomics.

In recent years, such measurements have even become possible in individual cells. They facilitate a hitherto unsurpassed resolution of processes in cell biology and developmental biology. In single-cell RNA sequencing, for example, gene expression profiles for over two million cells were reported in a single study [1].





# RNA

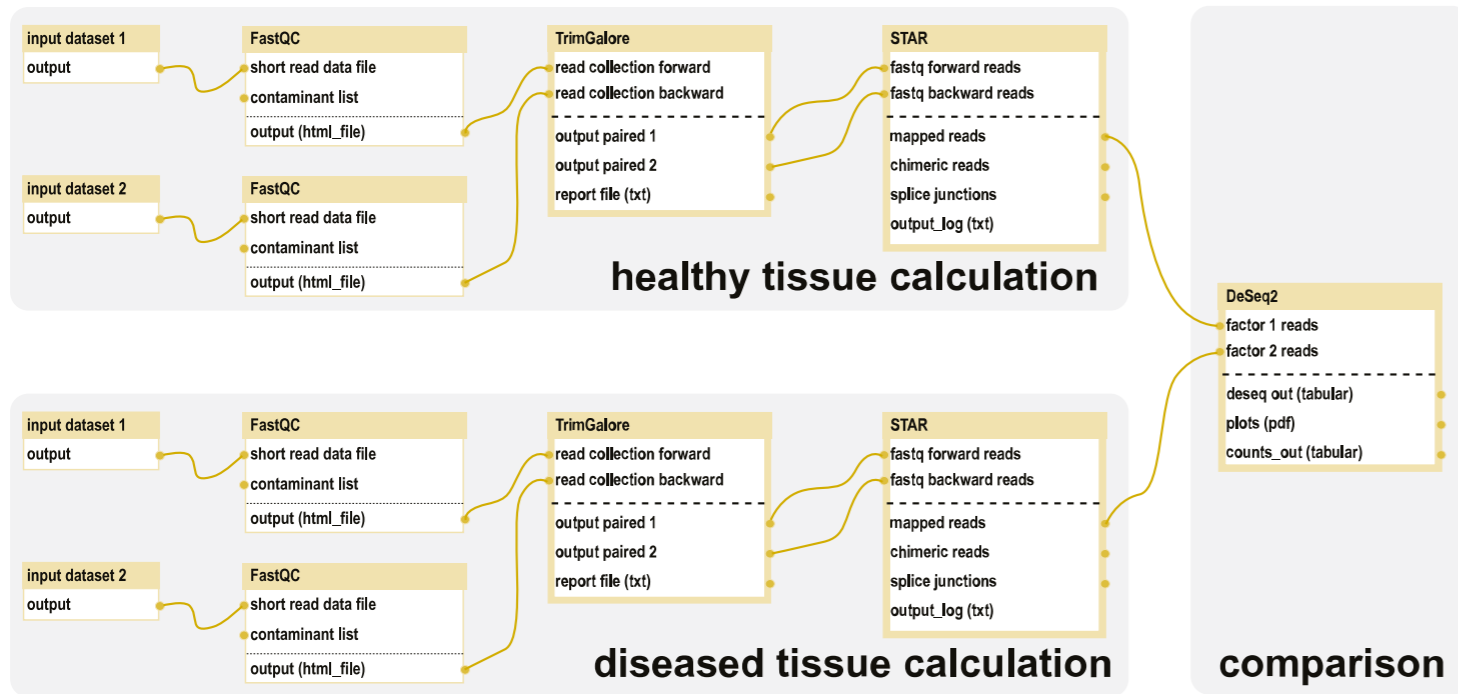
## in medical diagnostics

The role of RNA molecules in cell function is substantially more important than previously thought: Non-coding RNA have a function in cell regulation and offer completely new therapeutic options. RNA can even be used to measure the expression of genes in individual cells. This way, the few cells that can form metastases can be identified in a tumour. One of the main tasks involved is the high-quality analysis of the data.

As a functional unit, the cell is dominated by three groups of molecules: DNA, RNA and proteins. DNA is the carrier of the genetic information which is encoded in the genes by means of four bases: A, C, G and T. These gene sequences are translated into proteins, a process that additionally involves an intermediate mRNA copy consisting of A, C, G and U. Proteins then carry out certain functions. Not all genes are read in every cell type. Instead, the type and number of genes read and expressed determine whether a cell is a liver cell, a heart cell, or one of the approximately 300 other cell types found in the body. In a manner of speaking, DNA can be compared to legislature which determines the possible functions, while the proteins represent the executive that carries out the function. In that case, what are the RNA molecules? For a long time, it was believed that their only property was to act as templates for proteins. However, it has become clear in recent years that RNA has a substantially more important role than previously assumed. For example, there is a large number of non-coding RNAs, i.e. RNAs that are not translated into proteins that play an essential role in the regulation of cell function. For this reason, RNA can perhaps best be compared to the judiciary, since it controls the function of proteins (i.e. the executive) via regulatory mechanisms. However, the proteins also control the function of the RNA, resulting in a complicated control loop. If the control loop is disturbed diseased cells will be the result.

### RNA MOLECULES AS MEDICINE

The abovementioned control loop offers a completely new therapeutic option that has hitherto only been used to a limited extent. Existing drugs usually specifically target proteins involved in a disease. However, the success of intervention may be limited, especially in cases involving genetic diseases. One example is spinal muscular atrophy, one of the most common genetic causes of death in infants [1]. It is caused by a mutation in a gene (SMN1), which results in a lack of sufficient proteins from this gene to ensure the correct function of muscle cells. However, there is a slightly modified copy (SMN2) of the gene in our DNA, which often exhibits no genetic defect. A new drug with the active ingredient nusinersen, which was approved in the EU in 2017, uses an RNA and its regulating action to produce the SMN2



**FIGURE 1:** Workflow for the comparison of healthy and diseased tissues by means of RNA sequencing. The sequence data, which are available in two files for the forward and backward strands of the DNA, are first quality checked with the FastQC programme. This allows us to detect errors in sequencing. The next step, as described, is to remove the adapters (TrimGalore) and assign the sequences (called reads) to the known genes. This is done by mapping to the reference genome. This means that the STAR tool assigns each read to its exact position on the genome and then determines the number of reads per gene. This then becomes the expression profile of the healthy

tissue. The same process is carried out with the diseased tissue, and the DeSeq2 programme executes the comparison. It identifies which genes are very different in the two tissues. These are then candidate genes for a disease.

remove the adapters (TrimGalore) and assign the sequences (called reads) to the known genes. This is done by mapping to the reference genome. This means that the STAR tool assigns each read to its exact position on the genome and then determines the number of reads per gene. This then becomes the expression profile of the healthy

gene in higher numbers of copies, which in turn alleviates the consequences of the disease.

### RNA BIOINFORMATICS AS DETECTIVE: WHICH IS THE MALIGNANT CELL?

The cell type is thus not defined by the genome, but by the genes used or, more precisely, by the number of RNA copies per gene read. This is called the expression of a cell, which means that the type of a cell is defined by its expression profile. RNA sequencing makes it possible to determine the expression profile of a group of cells (or the cells of a tissue). With it, diseased tissue can be identified by comparing abnormal expression profiles with those of healthy tissues.

However, this definition applies not only to healthy, but also to diseased cells. Originally, cancer cells are also nothing more than altered body cells that share most of their hereditary information with normal body cells. However, a cancer does not consist solely of tumour cells: they need the support of neighbouring cells (stroma) to maintain the cancer [2]. To put it very simply, someone has to do the shopping (blood vessels and the corresponding cells) or keep the household together (connective tissue cells), so that the cancer can continue to grow through the uncontrolled division of tumour cells. There are also major differences among tumour cells. Many are simply couch potatoes that may divide uncontrollably, but do not break out of the tumour. Far worse, however, are cancer cells that do break out of the

tumour and search for new regions in the body, thus spreading the cancer. However, these cells are often very few in number, making them easy to overlook, especially in early stages.

This is where single-cell sequencing comes in. Typically, several thousand cells of a tumour are individually sequenced and their RNA profiles identified. Cell types are determined by comparing their expression profiles. This method can then be used to detect particularly malignant tumour cells [3].

As simple as this may sound in theory, the technical execution and the demands on digital data processing are extremely complex. Clarifying this requires taking a closer look at the sequencing process. For technical reasons, sequencing machines attach a specific, characteristic RNA sequence to each RNA molecule that is read and digitised. This is known as the adapter. The simple, but ingenious technique in single-cell sequencing is to extend these adapters by a short piece and use them to identify the individual cells. This is the only way to collect enough RNA from all of the cells so that they can be sequenced. To illustrate an example of this, let us assume, for the sake of simplicity, that the adapter required by the sequencer consists of a sequence of five Gs, i.e. exactly GGGGG. Now, in each cell, this adapter and a sequence of three additional nucleotides are attached to each RNA molecule (a sequence of A, C, G and U), with which the cell is identified. These sequences of three are then interpreted as numbers, i.e. AAA = 1, AAC = 2, AAG = 3, AAU = 4, ACA = 5 and so on. In this way, the sequence GGGGAAA is attached to all RNAs of cell 1, the sequence GGGGAAC is attached to all RNAs of cell 2, etc. This trick can

then be used to uniquely identify 4 to the power of 3 or 64 cells. In reality, these sequences are longer, enabling us to identify several thousand cells.

### AND HOW ARE WE TO INTERPRET ALL THESE DATA?

This procedure, already complicated in itself, is further complicated by the fact that not only five or ten RNA molecules from the cells have to be sequenced, but many more. An actual data set consists of 100 million sequences, for example, in form of GGGGAAUUUUUJAGACCCCAUCAAA, and a hundred other bases. How can we possibly interpret this and confirm that it is an RNA molecule of a gene with the DNA sequence TTTAGACCCATCAAAC...in the fifth cell? How can this information then be manually correlated to thousands of cells to discover the malignant tumour cells?

The answer is simple: it is not humanly possible. This must be done by computer programmes. A large number of programmes have been designed for this purpose, which are managed and adapted by the RNA Bioinformatics Center and provided free of charge to a broad research community. These include programmes that remove aptamers, assign them to the cells, assign the attached RNA sequence to specific genes (called mappers) and use them to create expression profiles for the individual cells (Figure 1). Typically, around 2,000 to 3,000 genes and their expressions are determined. But even then, comparing whether the expressions of the 2,000-3,000 genes in cell X are similar to the profile of cell Y would exceed the capacity of any human being. Consequently, there are pro-

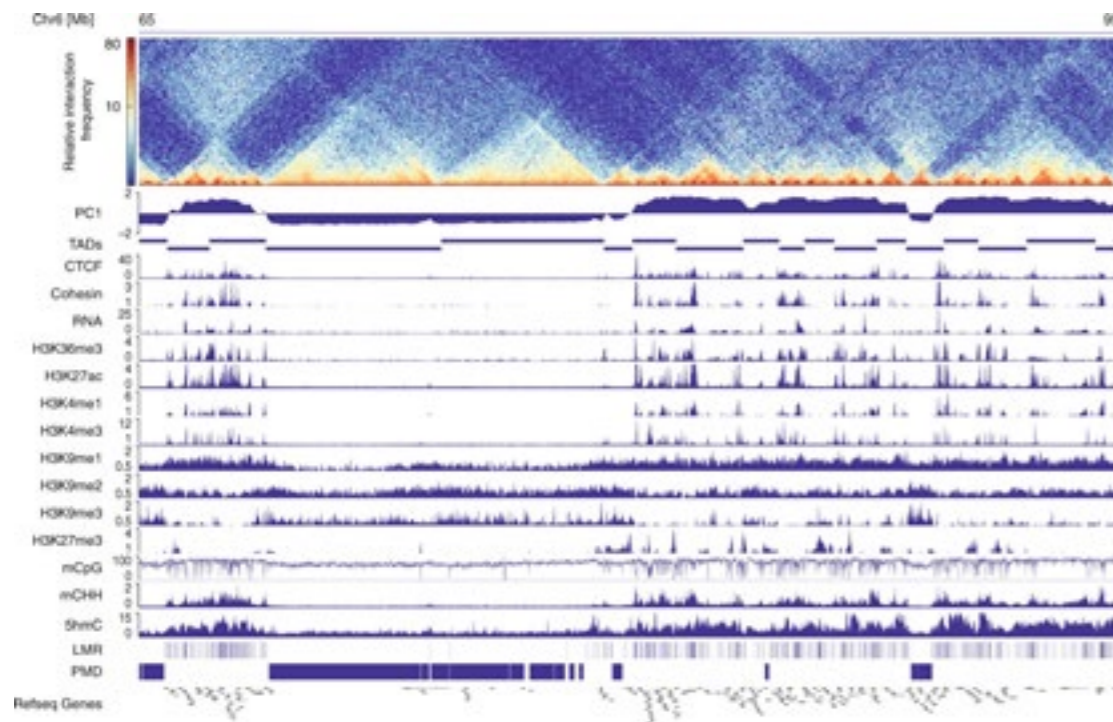
100  
MILLION

sequences...

COMPRISES AN ACTUAL DATA SET.

grammes that perform this comparison, ultimately showing the groups of cells and their frequency to the physician. A physician or life scientist can then use this visualisation of the digital data to draw conclusions. For an analysis problem, workflows are created that typically combine one to several dozen programmes into a meaningful sequence to ensure that this visualisation is successful and new insights can be gained (Figure 2).

The RNA Bioinformatics Center of the German Network for Bioinformatics Infrastructure consists of seven partners from all over Germany who have committed themselves to developing the necessary tools, workflows and visualisations and making them accessible to everyone. On our Galaxy server, for example, we provide access to over 2,000 different tools that can be linked as required to analyse highly complex data.



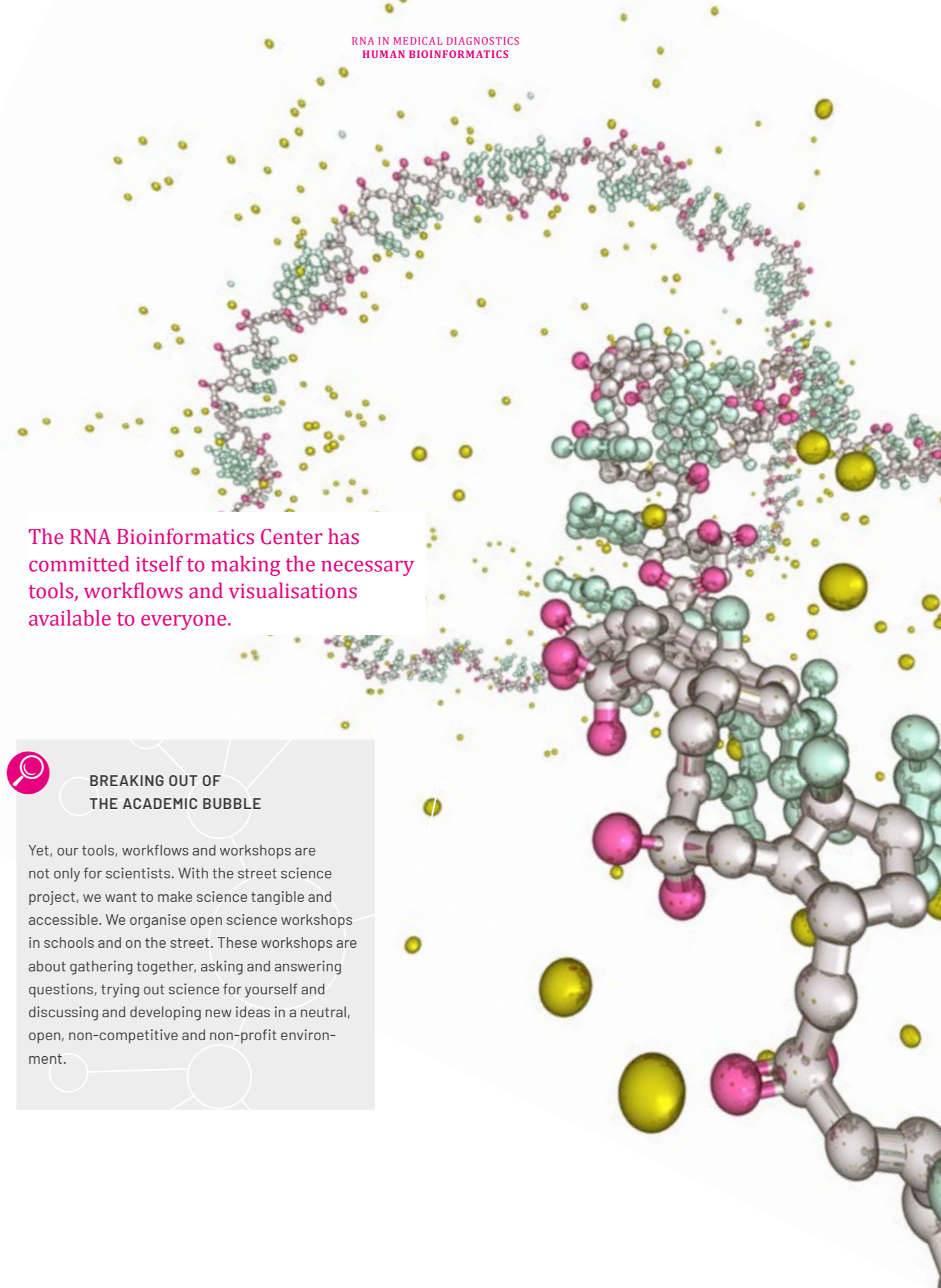
**FIGURE 2:** Visualisation of several sequencing operations that investigate different properties of the genome. Each series (i.e. PC1, TADs, etc.) corresponds to one sequencing experiment determining, for example, the structure of DNA (TADs) or its epigenetic changes (H3K36me3, etc.). The last row represents the reads of a normal RNA sequencing. This com-

pact chart enables the life scientist to interpret the results of the various experiments correctly. (Image from: Nothjunge, S., Nührenberg, T.G., Grüning, B.A. et al. DNA methylation signatures follow preformed chromatin compartments in cardiac myocytes. *Nat Commun* 8, 1667 (2017) doi: 10.1038/s41467-017-01724-9)

**REFERENCES:** [1] <https://www.presseportal.de/pm/102449/3651447> [2] <https://www.uniklinikum-leipzig.de/einrichtungen/dermatologie/Seiten/forschung-prof-simon-tumor-stroma-interaktionen-.aspx> [3] <https://science.sciencemag.org/content/352/6282/189>

**AUTHORS:** Rolf Backofen<sup>1</sup> and Björn Grüning<sup>1</sup>

<sup>1</sup> University of Freiburg, Department of Computer Science, Georges-Köhler-Allee 106, 79110 Freiburg



The RNA Bioinformatics Center has committed itself to making the necessary tools, workflows and visualisations available to everyone.



#### BREAKING OUT OF THE ACADEMIC BUBBLE

Yet, our tools, workflows and workshops are not only for scientists. With the street science project, we want to make science tangible and accessible. We organise open science workshops in schools and on the street. These workshops are about gathering together, asking and answering questions, trying out science for yourself and discussing and developing new ideas in a neutral, open, non-competitive and non-profit environment.



A PROJECT FROM THE PERSPECTIVE OF AN MD STUDENT  
RESEARCH ON BIOMARKERS

for the early diagnosis of  
Parkinson's disease

At the Medizinisches Proteom-Center (MPC) in Bochum, MD student Petra Weingarten is researching biomarkers for the diagnosis of Parkinson's disease. In doing so, she is supported by bioinformaticians and statisticians from the de.NBI service centre BioInfra.Prot: From the pre-processing and analysis of the data, to the publication of the results. This example illustrates the importance of cooperation between different disciplines in a research project.

and bioinformatician Michael Turewicz, who provide consulting and analyses in the field of bioinformatics and statistics of proteomics data within the de.NBI service centre Biolnfra.Prot. First, the goals of the project and the preparatory work done so far were discussed in a preliminary meeting. According to Karin Schork, an initial meeting should take place as early in the project as possible: "People think that statistical analysis comes at the very end of projects like this. The study design at the very beginning is also a crucial element. It's important to contact a statistician or bioinformatician before measuring the data and talk about

the planned project. This way, possible challenges can be identified early on and some problems in the analysis can be prevented later on."

#### THE CHALLENGE: METABOLITES

During this first meeting, it became clear that this project would involve a great challenge: analysing and processing metabolite data. As opposed to the analysis of protein data, there has been hardly any pertinent experience with metabolite data regarding both the laboratory part and the data analysis part. The metabolite data were measured using a

commercial kit, which necessitated the establishment of this technology in the laboratory. In the data analysis, it was mainly the pre-processing of the data that raised questions: Which pre-processing steps still have to be carried out and which have already been included in the supplied software? Can metabolite data actually be treated in the same way as protein data when it comes to statistical analysis? After many discussions, reviews of different methods and a conference call with the kit manufacturer, a strategy for pre-processing the data was found and the data were prepared for statistical analysis.

#### de.NBI TRAINING SUPPORTS CONSULTING

For the statistical analysis of metabolite and protein data, scripts were prepared in the R programming language, which allow researchers to compare the data at different points in time and between patients and control persons and create appropriate graphics. As part of de.NBI training, Michael Turewicz and Karin Schork offer an introductory course to R once a year so that participants can practice the basic use of this programming language. Petra Weingarten also attended this introductory course, enabling her to adapt the scripts provided to new data and make minor changes herself: "The R course has helped me a great deal in my work. During the course, I was introduced to the relevant functions in a slow and easily understandable way, so that I was able to read the R scripts that were provided to me and better understand some of the analysis steps using the scripts."

#### PROMISING RESULTS

The results of the statistical and bioinformatic analyses were presented during one of many consultations. Several promising biomarker candidates

were found, which will be validated in subsequent experiments. Biomarker panels were also examined next to the analysis of individual proteins and metabolites. "Against the background of natural biological variability, individually varying disease progressions and diverse disease subtypes, it is assumed that a small number of combined biomolecules are better suited as diagnostic biomarkers than individual proteins or metabolites. These are sought using machine learning methods and can better reflect the complex molecular patterns that enable us to identify Parkinson's disease. Such a set of biomolecules is called a biomarker panel," explains Michael Turewicz. The final analysis is currently still pending.

#### ON THE PATH TO PUBLICATION

The results of the study will also be summarised and published in a scientific paper. In proteomics, it is com-

mon practice and a requirement of many scientific journals for a publication to also include the raw data of the measurements belonging to the study. For this, we have the PRIDE-Archive [3], into which these data can be uploaded and will be publicly available for re-analysis after a successful publication. Since uploading large amounts of data can often be problematic, Biolnfra.Prot offers an upload service that assists users with questions and problems. In addition, a tool has been developed that converts the data into standard formats that are accepted by PRIDE.

#### SUCCESSFUL PROGRESS IN THE PROJECT

In general, both sides are quite satisfied with the progress the project has made so far; some further analyses and the compilation of a report for

publication will follow in the coming weeks. Petra Weingarten is currently writing the final chapters of her dissertation, for which her collaboration with Biolnfra.Prot was highly beneficial: "The close cooperation with bioinformatics and statistics has given me a secure feeling when it came to handling the data. The opportunity to clarify questions directly with experts in a way that was easy to understand was indispensable – I don't know how I would have managed the job effectively without their help."

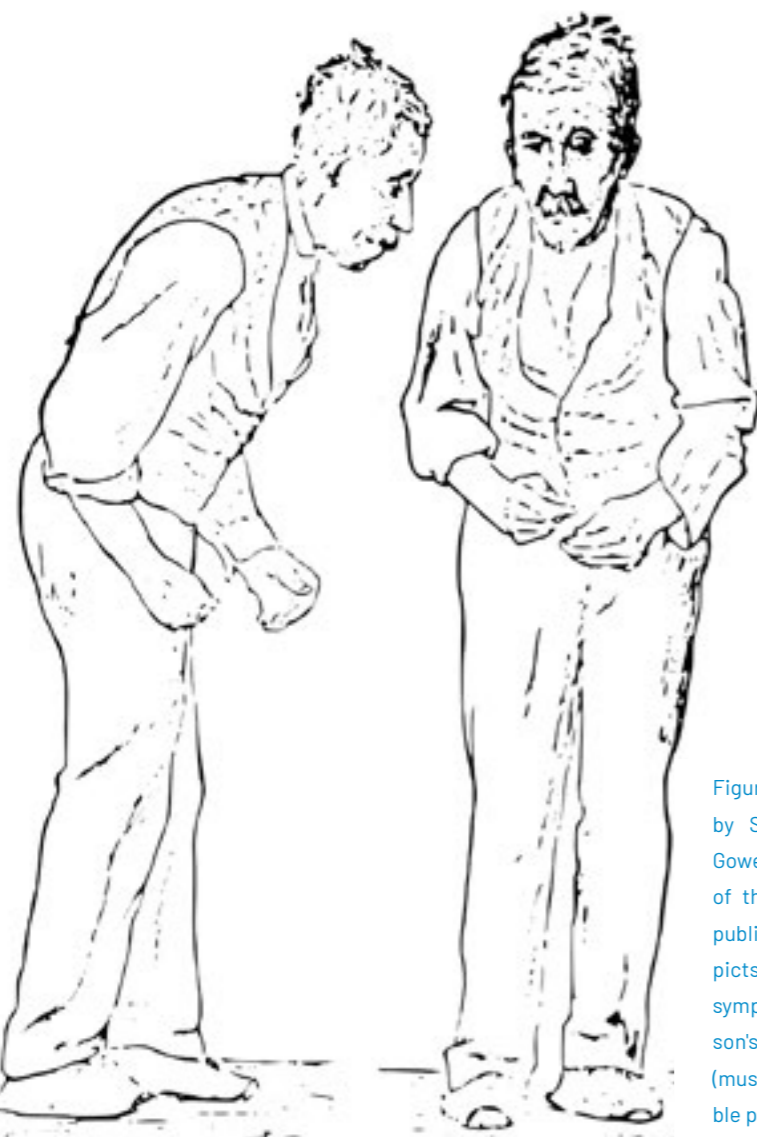


Figure 1: This illustration by Sir William Richard Gowers from "A Manual of the Nervous System" published in 1886 depicts some of the typical symptoms of a Parkinson's disease patient (muscle stiffness, unstable posture) [4].

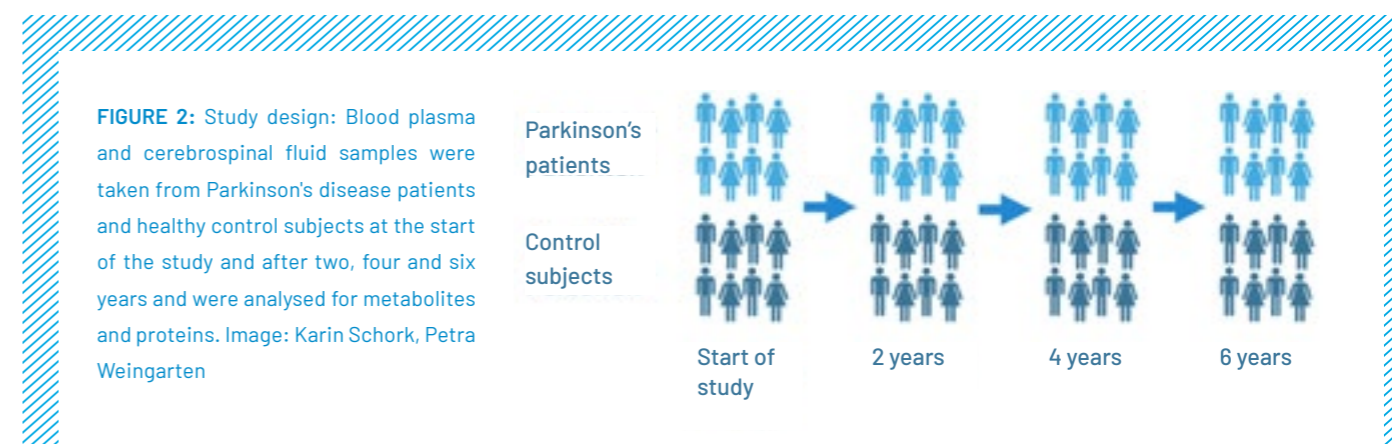


FIGURE 2: Study design: Blood plasma and cerebrospinal fluid samples were taken from Parkinson's disease patients and healthy control subjects at the start of the study and after two, four and six years and were analysed for metabolites and proteins. Image: Karin Schork, Petra Weingarten

REFERENCES: [1] <https://www.parkinson-gesellschaft.de/aktuelles/36-von-der-forschung-in-die-klinik-die-deutsche-parkinson-gesellschaft-mit-neuer-praesenz-im-web.html> [2] <https://www.denopa.de/> [3] <https://www.ebi.ac.uk/pride/archive/> [4] [https://de.wikipedia.org/wiki/Parkinson-Krankheit#/media/Datei:Sir\\_William\\_Richard\\_Gowers\\_Parkinson\\_Disease\\_sketch\\_1886.svg](https://de.wikipedia.org/wiki/Parkinson-Krankheit#/media/Datei:Sir_William_Richard_Gowers_Parkinson_Disease_sketch_1886.svg)

AUTHORS: Karin Schork<sup>1</sup>, Petra Weingarten<sup>1</sup>, Martin Eisenacher<sup>1</sup> and Michael Turewicz<sup>1</sup>  
<sup>1</sup> Ruhr-University Bochum, Faculty of Medicine, MPC, Gesundheitscampus 4, 44801 Bochum



# SYSTEMS MEDICINE OF THE LIVER – A CHALLENGE FOR DATA MANAGEMENT

One of the special features of the liver is its healing ability. The aim of the LiSyM project is to gain clinically relevant insights into how long-term stresses can nevertheless damage the liver and lead to progressive liver disease. In this project, scientists are contributing their expertise from laboratory research, theoretical studies and clinical practice. This article describes how this complex project is represented in data management.

The liver is an extraordinary organ with many different functions in our body and amazing capabilities. Even the ancient Greeks knew that parts of the liver can be regenerated again and again and remain functional. Legend has it that the Titan Prometheus was cruelly punished by the gods for having taught humanity the art of making fire. He was chained to a rock in the Caucasus mountains and an eagle ate parts of his liver every day, which then continually regenerated until the eagle came back the next day.

The stresses on a modern liver are more mundane, but much more diverse. Despite the ability to regenerate itself, a liver can gradually deteriorate in the long term. This long-term liver deterioration usually begins with the appearance of a fatty liver. About 20% of the western population suffer from something called a non-alcoholic fatty liver,

i.e. a liver that has stored fat droplets and is considered damaged, but this damage is probably not caused by alcohol. Some non-alcoholic fatty livers become inflamed and develop non-alcoholic steatohepatitis, a form of hepatitis. Other patients remain healthy except for the fatty deposits. What is this difference based on? What leads to a progression of the disease? What protects the liver from it? These and other questions are the main focus of the research network Systems Medicine of the Liver (LiSyM), funded by the German Federal Ministry of Education and Research. LiSyM follows a systems medicine approach: using various approaches, the researchers are trying to understand biological systems with the aid of computer models that can be simulated, so that they can then apply this knowledge in clinics or make it more applicable to clinical settings by developing new therapeutic approaches.

In LiSyM, 37 research groups from 23 different research centres and organisations are collaborating. Of course, this does not work spontaneously, but requires a meaningful structure and organisation supported by a central data management system. Another key reason for data management is to make the data traceable and reusable. This is referred

to as FAIR data: findable, accessible, interoperable and reusable. FAIR is not a precise guideline for structuring and formatting data, but rather a complete spectrum of very simple and basic rules on how data can be made “fair”. The objective is a useful compromise that achieves data FAIRness with minimum effort.

Within the framework of the LiSyM network, personnel for data management experts are also supported for the further development of the software platform used, as well as for the collection of requirements which different users place on data management and for community management (including user training). These project-funded experts



work closely with personnel from the de.NBI-SysBio team and, in addition to their own developments, also draw on the results of development work from de.NBI-SysBio. Joint, cross-project conceptual and development work for data management is also carried out, creating synergies from which all projects involved benefit.

### LiSyM is organised into four thematic pillars.

LiSyM is organised into four thematic pillars, each of which investigates one question (from animal experiments to clinical practice). Each pillar has partners from experimental research, modelling and clinical practice.

► **Pillar 1: Early Metabolic Injury** deals with how fatty liver evolves into hepatitis.

► **Pillar 2: Chronic Liver Disease Progression** deals with the transition from inflammation to cirrhosis.

► **Pillar 3: Regeneration and Repair in Acute-on-Chronic Liver Failure** deals with how to promote liver healing in cases of acute failure of a chronically diseased liver.

► **Pillar 4: Liver Function Diagnostics.** This pillar's goal is the non-invasive diagnosis of liver damage.

This large-scale project is being completed by the coordinating programme directorate under the leadership of Prof Peter Janssen. Data management is also established here.

One of the challenges for data management is the existence of very diverse data, which must be combined or, in other words, integrated, in order to generate and simulate the computer mod-

els developed in LiSyM. The data differ in terms of their modality (image data, measurement data, gene or protein sequence data, clinically collected data, etc.), the formats used for data and meta-data (data that describe and correlate the actual data) and requirements for privacy and data security. Mice may not have any special protection in terms of privacy, but humans do. As a result, data concerning humans must be treated differently from data obtained from animals or cell lines in the laboratory.

The diversity of computer applications being used simultaneously and the geographic distribution of data from the network pose further challenges. Patient-related data are usually not allowed to leave the organisation where they were obtained. This means that the project partners cannot store all the data together in one central location, but nevertheless need to be able to correlate them, even beyond the borders of these organisations. Furthermore, some users have local data storage for other reasons or other tools that they use. The central data management system of such a research network must be able to communicate with these as well.

### THIS LEADS US TO THE LISYM DATA MANAGEMENT ARCHITECTURE SHOWN BELOW:

At the heart of the architecture is LiSyM SEEK, an installation of the SEEK software. It has been jointly developed over the last ten years by the University of Manchester, the Heidelberg Institute for Theoretical Studies (HITS) and other partners in the FAIRDOM initiative [1], and is used in various European and national research consortia. SEEK has been developed in the knowledge that data management in interdisciplinary research projects must be able to catalogue data from a wide variety of sources. It is ca-

pable of storing data centrally, but is also able to refer to distributed data and link them together – which is exactly what LiSyM needs.

We use a separate instance of SEEK for the LiSyM network for reasons of flexibility and increased data security: not every change to the system that is necessary or useful for LiSyM is also useful or necessary for other instances of SEEK, such as FAIRDOMHub, which is used in parallel by several different research projects and consortia. This SEEK installation, which – like LiSyM-SEEK – is operated on a separate server at HITS, is the most frequently used instance, managing data for over 100 projects of various sizes. Users of LiSyM can transfer their data here as they wish. This makes it much easier to share individual and networked data sets with other projects (i.e. outside LiSyM) in the shared platform FAIRDOMHub. Access rights that can be fine-tuned for each data set allow users to exchange confidential data with other individual users or user groups within a project and across project boundaries.

With the SEEK software, including FAIRDOMHub and LiSyM SEEK, users can store, catalogue, add metadata, link to other data and finally share data with other users. For every single data set, users can determine exactly who is allowed to see and download the data, who can only see some basic metadata and receive the actual data only upon request to the owner of the data and who does not have any access at all. This may also be changed throughout the life cycle of the data. For example, data from new laboratory experiments may be stored by the experimenter, but initially only shared with a few close cooperation partners. At a later stage, this data will be linked with other data from the project – in order to develop a computer model, for example. Most SEEK users initially only share their data

within their project until publication. In addition, data can also be kept essentially hidden but shared with reviewers using secret links, which can also be given an expiration date: users with the secret link get access to the file – even without a user account. Finally, data can be shared with the world or, in other words, be published, and be provided with stable, permanently citable *digital object identifiers*, for example, to refer to supporting material from a publication.

### HOW DO I GET MY DATA AND METADATA INTO LISYM SEEK?

In addition to the classic manual upload via web interface by the user, the follow-

ing options are available (Figure 1). API for data transfer: There is a web-based programming interface that allows users to upload data directly into SEEK under programme guidance. This can be done, for example, by using Python programmes. For this purpose, we offer specific example codes and help you to adapt this example code.

Link to openBIS [2]: The Laboratory Information Management System (LIMS) openBIS [2], developed by our cooperation partners at ETH Zurich in Switzerland, is used in laboratories for data management. More recent versions of this system offer a configurable interface to SEEK, which was developed in collab-

oration with partners from Manchester and Edinburgh. This makes it possible to show specific experiments and groups of experiments within SEEK. They are then visible in both systems and can be shared. Our partners at the DKFZ in Heidelberg, for example, make use of this option.

Upload via data media: Finally, for large data, there is the option of uploading data by exchanging data media. This is particularly useful for large amounts of data such as stacks of related, high-resolution microscopic images.

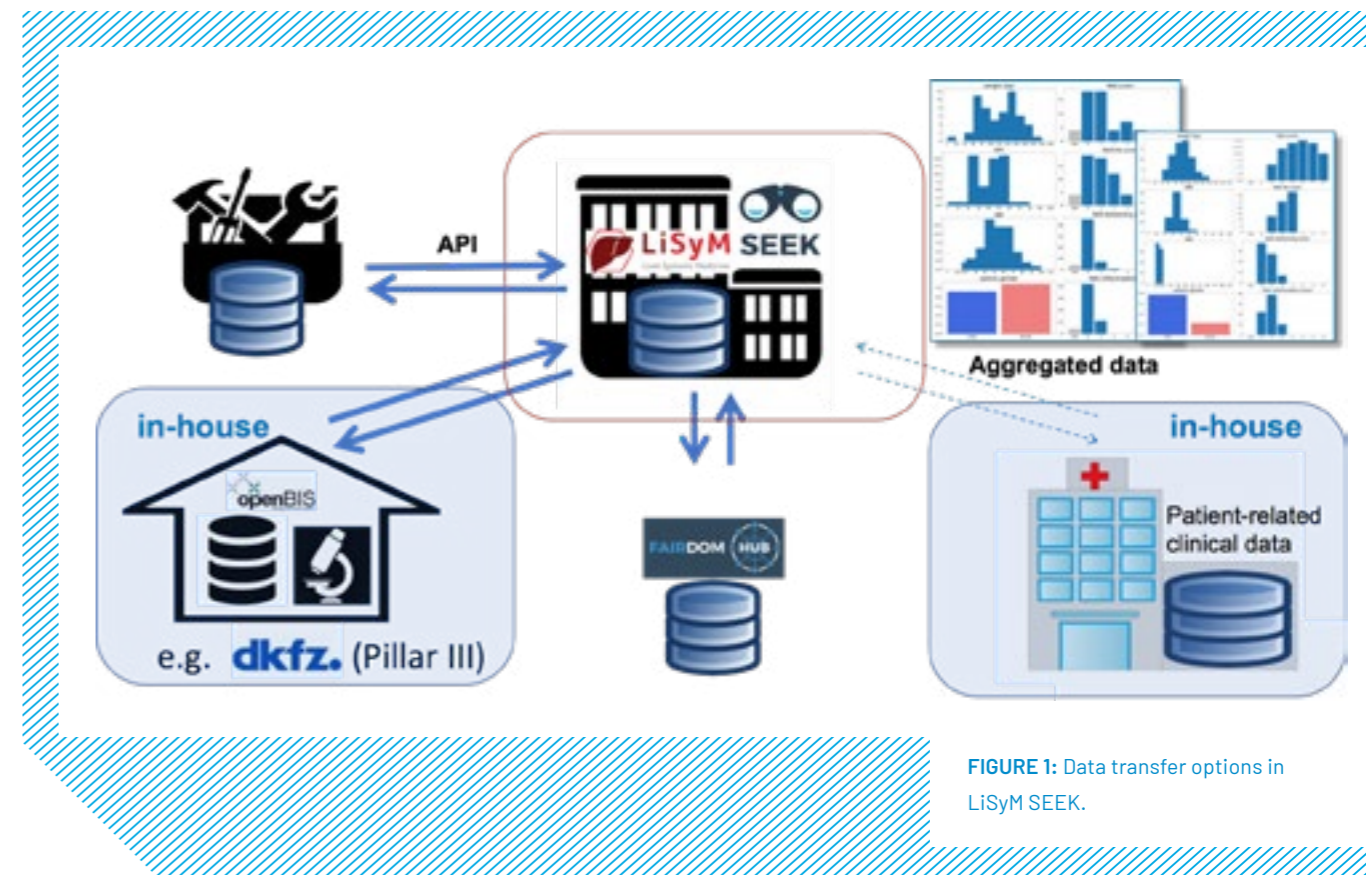


FIGURE 1: Data transfer options in LiSyM SEEK.

### EXCHANGING CLINICAL DATA (AS AGGREGATED ANONYMOUS DATA)

Clinical data that are destined for use in research projects are a major challenge for scientific data management in cooperative research networks, as they cannot be readily sent across organisational boundaries. Normally, they cannot be stored in FAIRDOMHub or LiSyM SEEK, as this is generally not in accordance with data protection regulations. However, it is possible to share these data with cooperation partners outside the clinic by grouping (aggregating) the data of individual patients locally in the clinic and sharing only the data concerning patient groups, not individual patients. What characteristics do the participants in a study have? Might one of the partners have the longed-for data on young liver patients that could complement your data on older patients? What is the distribution of certain liver values among the clinics participating in a study?

**A major challenge for research is that for reasons of data protection, patient data cannot simply be passed on. Only a summary can be stored legally in LiSyM SEEK and exchanged between the partners.**

A good way to do this is to provide a mobile code. We have demonstrated how this can be done using Anaconda and Jupyter: The clinical partners agreed on table templates, i.e. example structures for clinical data available in Excel. We then implemented code in Python that reads, directly analyses and aggregates these data on site at the clinic without the need to transport the data. The summaries of the analyses did not identify patients;

this is why they could be legally stored in LiSyM SEEK and exchanged between the partners. The advantage of this solution is that it is easily manageable for all parties concerned. On the part of the clinical partners, there is relatively little software to install and it is easy to administer. It does not require administrator rights on the computers on which it runs. On the other hand, the degree of automation is small and the partners start the software themselves and also merge the data themselves. This requires more manual work, but is easier to secure.

In SEEK, associated data can be grouped into assays, studies and investigations and correlated with standard operating procedure (SOP) protocols, descriptions of the biological samples used, resulting computer models and publications. As long as they are based on, or referred to, each other, they can all be described and networked in SEEK. This structuring is based on corresponding metadata, which describe the relationships between the data. The result is a FAIR representation of the data and metadata of the LiSyM research network (Figure 2).

In this way, SEEK and our data management service based on it, which also covers key aspects of user and expectation management, are ideally suited to meet the requirements of a data management concept for such a distributed, cooperative and interdisciplinary research consortium as LiSyM. This is complemented by offers such as user participation in planning the further development of the SEEK software and training for the various user communities from laboratory, theory and clinical practice. We also offer this complete package to other users within the framework of de.NBI.

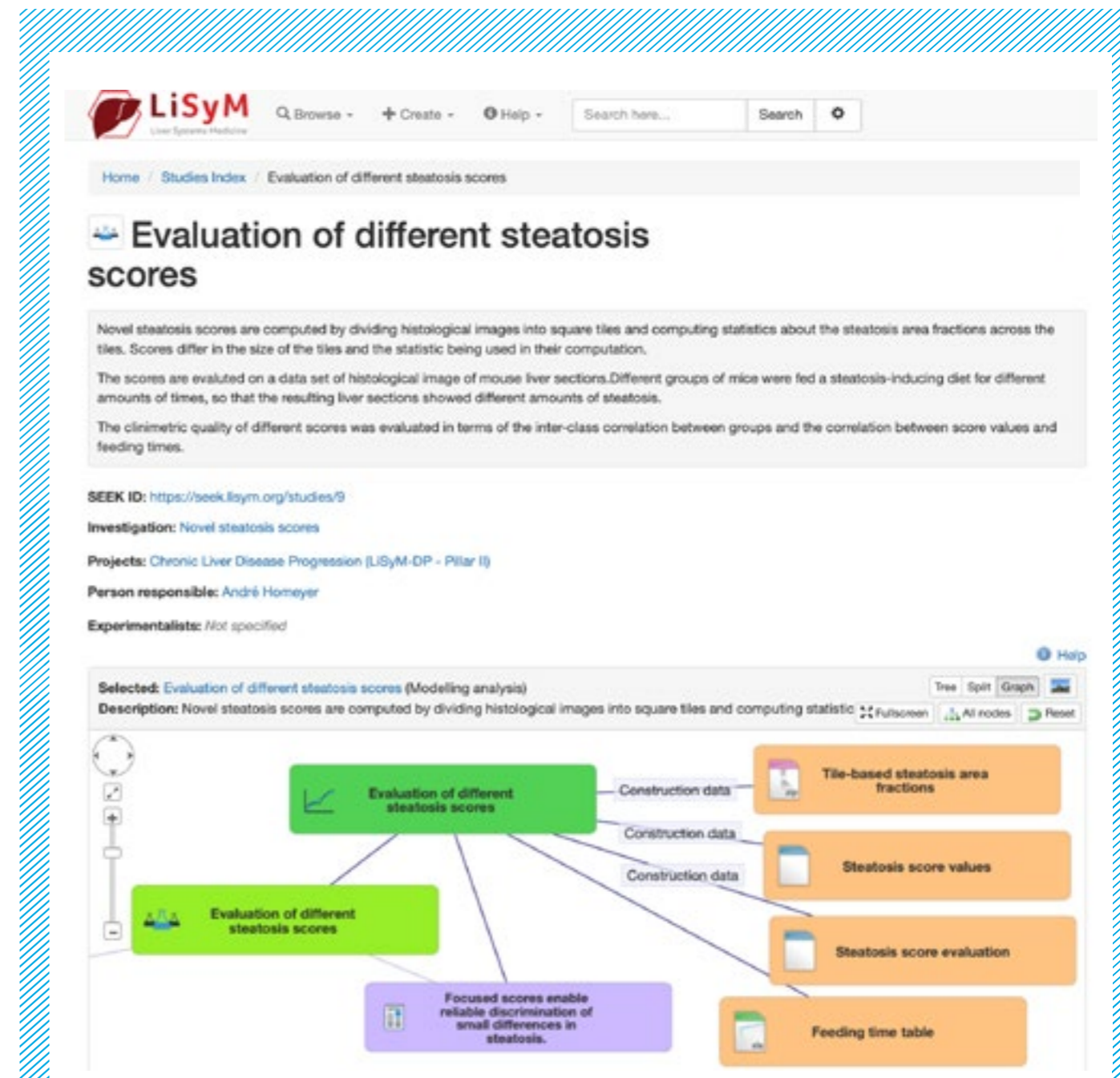


FIGURE 2: Screenshot from LiSyM SEEK to demonstrate the structuring of related data sets (orange) and resulting publications (purple) for grouping in assays and studies (green).

REFERENCES: [1] <https://fair-dom.org> [2] <https://sis.id.ethz.ch/software/openbis.html>

AUTHORS: Martin Golebiewski<sup>1</sup> and Wolfgang Müller<sup>1</sup>  
<sup>1</sup>Heidelberg Institute for Theoretical Studies (HITS), Heidelberg

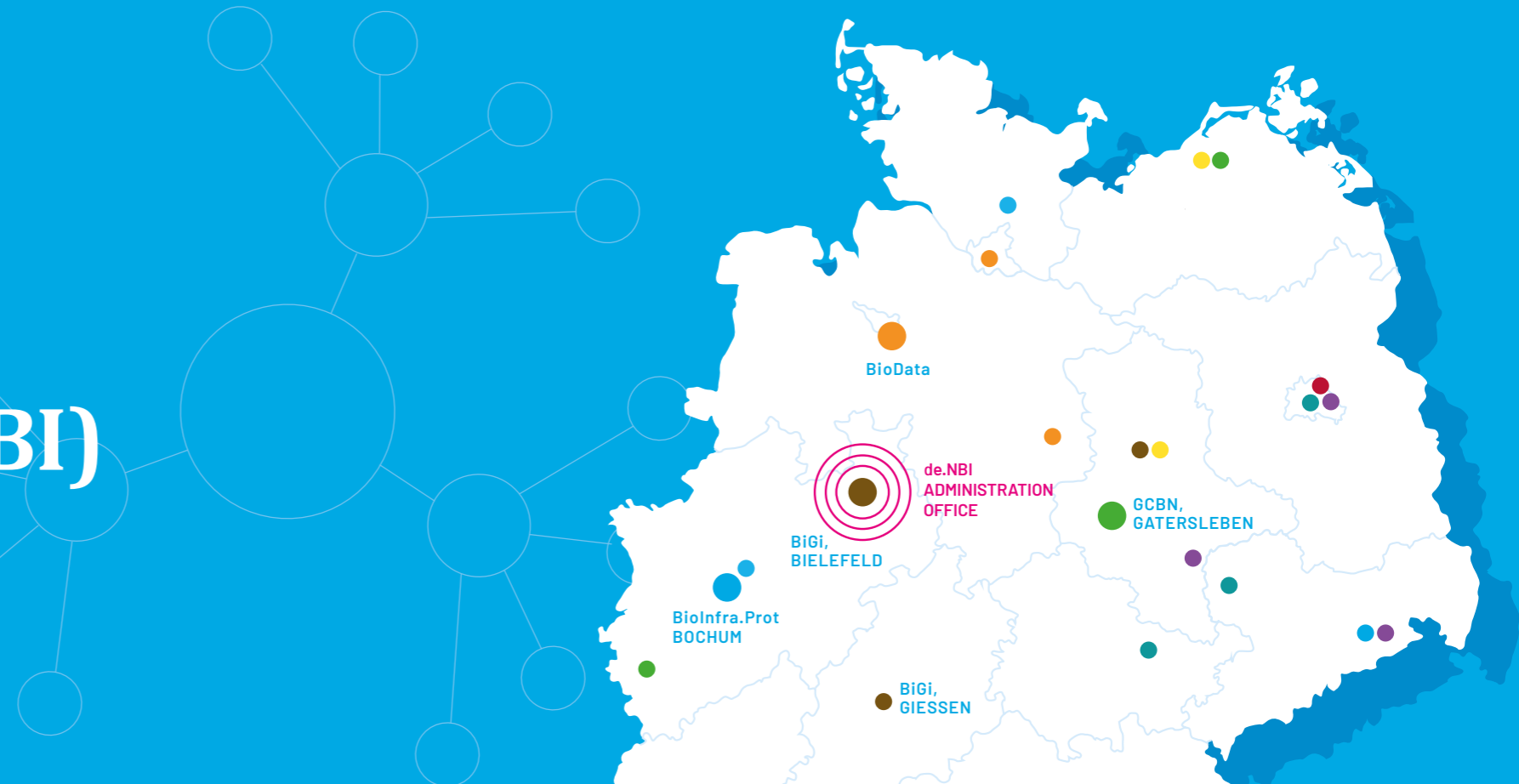


# THE GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE (de.NBI)

The German Network for Bioinformatics Infrastructure (de.NBI) is a national, academic infrastructure financed by the Federal Ministry of Education and Research (BMBF) since 2015 to provide bioinformatics solutions to researchers in life sciences and medicine for the analysis of large amount of data. With its wide range of bioinformatics expertise and reputable partner institutions, the de.NBI network guarantees the delivery of high standards bioinformatics services, comprehensive training, as well as powerful computing capacity that contributes to the advancement of bioinformatics research in Germany and elsewhere in Europe.

# THE GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE (de.NBI)

The German Network for Bioinformatics Infrastructure - (de.NBI) consists of eight interconnected service units that serve life science research communities by offering tools, training, compute resources, as well as connections to major industrial companies within Germany and Europe. de.NBI also offers large computing power and storage capacity through a free cloud environment that allows researchers to process and analyse their own data. The network is managed by a Coordination and Administration Unit consisting of the de.NBI Coordinator and the team of the Administration Office (AO).



32

Institutions...  
BELONG TO THE NETWORK.

42

Projects...  
ARE INTEGRATED INTO THE NETWORK.

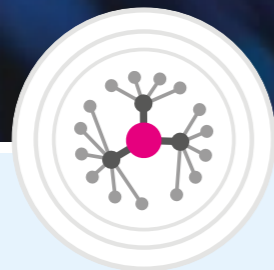
250

Scientists...  
ARE WORKING IN THE NETWORK.

## THEMATIC FOCUSES & SERVICE CENTRES:

- **HUMAN BIOINFORMATICS**  
HEIDELBERG CENTER FOR HUMAN BIOINFORMATICS (HD-HuB)
- **MICROBIAL BIOINFORMATICS**  
BIELEFELD-GIESSEN RESOURCE CENTER FOR MICROBIAL BIOINFORMATICS (BiGi)
- **PLANT BIOINFORMATICS**  
GERMAN CROP BIOGREENFORMATICS NETWORK (GCBN)
- **RNA BIOINFORMATICS**  
RNA BIOINFORMATICS CENTER (RBC)
- **PROTEOME BIOINFORMATICS**  
BIOINFORMATICS FOR PROTEOMICS (BioInfra.Prot)
- **INTEGRATIVE BIOINFORMATICS**  
CENTER FOR INTEGRATIVE BIOINFORMATICS (CIBI)
- **BIODATABASES**  
CENTER FOR BIOLOGICAL DATA (BioData)
- **DATA MANAGEMENT/SYSTEMS BIOLOGY**  
de.NBI SYSTEMS BIOLOGY SERVICE CENTER (de.NBI-SysBio)

- LOCATIONS OF SERVICE CENTRES
- LOCATIONS OF PARTNERS



## CONTRIBUTION OF THE de.NBI NETWORK

# to solving the big data problem in the life sciences

The de.NBI network has existed for five years now. In order to take a closer look at the tasks of the network and examine the results achieved in the meantime, Irena Maus, who is responsible for public relations at the de.NBI administration office, interviewed de.NBI Coordinator Alfred Pühler and de.NBI Head of Administration Office Andreas Tauch.

**IRENA MAUS:** The de.NBI network is celebrating its fifth anniversary. Why was it established?

**ALFRED PÜHLER:** The de.NBI network was established in 2015 to provide all researchers in the life sciences with an infrastructure that enables them to analyse large amounts of data. This infrastructure initially included the areas of **service** and **training**. The service area offers a wide range of analysis programmes that can be used to evaluate life science data. In addition to this service area, the training area of the de.NBI network is of crucial importance. In the training area, researchers are taught how to deal with bioinformatic tools and the results achieved. These two areas have been consistently expanded over the past five years. In the meantime, over 100 modern analysis programmes have been made available, with over 280 courses on how to use them. To date, over 5,000 participants have been trained.

**IRENA MAUS:** How is this network structured and how are decisions made?

**ANDREAS TAUCH:** The de.NBI network

is structured in themed research units. It consists of **eight service centres** that cover different sub-disciplines in the life sciences such as human bioinformatics, RNA bioinformatics or biodatabases. The network is managed by a **central coordination unit**, which includes the de.NBI coordinator and the heads of the eight service centres. This body meets quarterly to make pending decisions. For this purpose, it draws on the advice of **seven expert groups** from the network. This approach has worked extremely well over the years.

**IRENA MAUS:** How has the network developed over the last five years?

**ALFRED PÜHLER:** The de.NBI network has become involved in other tasks in addition to the areas services and training additional. One of the tasks is the establishment of a compute facility that allows de.NBI users to analyse large amounts of data. Right from the start, the de.NBI network has relied on future-oriented technology and has set up a **de.NBI cloud** at several locations in Germany. The network also had the task of establishing a European coop-

erations, a task facilitated by Germany's entry into the **ELIXIR** organisation. Finally, we worked successfully on establishing an industrial branch of the de.NBI network. In recent months, a **de.NBI Industrial Forum** has been established, which currently has 26 companies as members.

**IRENA MAUS:** How successful was the establishment of the federated de.NBI cloud?

**ANDREAS TAUCH:** The establishment of our own cloud was made possible in 2016 by additional funding from the BMBF. We opted to set up a federated cloud at six locations in Germany. This project is managed centrally at the de.NBI administration office. What makes the **de.NBI cloud** special? It is a fully academic cloud federation, that provides storage and computing is free to charge for academic use. The scientific success of the de.NBI cloud can be judged by its numbers: over **700 registered researchers** with over **200 ongoing large-scale projects!**\*

**IRENA MAUS:** How does de.NBI participate in the European ELIXIR organisation?

**ALFRED PÜHLER:** ELIXIR is a European infrastructure network with the mission of supporting all aspects of handling life science data in its member countries. ELIXIR thus pursues analogous goals in Europe as the de.NBI network in Germany. After Germany joined ELIXIR in July 2016, the de.NBI network was commissioned to develop the **German ELIXIR node**. This was achieved through participation in ELIXIR activities. In the service area, for example, de.NBI bioinformatics programmes were provided to users throughout Europe by the ELIXIR organisation. A cooperation with ELIXIR partners was also made in the area of training. Furthermore, several ELIXIR member states have integrated the de.NBI cloud into cooperation projects. Finally, the de.NBI Industrial Forum is also attracting considerable attention at the ELIXIR level, as it too promotes the integration of European industry into a bioinformatics infrastructure.

**IRENA MAUS:** What is the role of the de.NBI Industrial Forum?

**ANDREAS TAUCH:** The **de.NBI Industrial Forum** represents the latest development of the de.NBI network. This is a **loose association of currently 26 companies** that was organised over the course of 2019. In November, members of the forum met for the first time for a one-day information event in Berlin. The forum is intended to facilitate scientific cooperation between de.NBI and industrial partners at the project level, with the aim of transferring de.NBI's expertise in the analysis of large amounts of data to the industrial sector. The member companies in turn have access to de.NBI training activities and to scientific de.NBI events, and they can themselves contribute to shaping the forum.

**IRENA MAUS:** What efforts have been made to make the offers and services of the de.NBI network available in the long term?

**ALFRED PÜHLER:** Over the past five years, the de.NBI project has helped to

establish a future-oriented infrastructure, the continued existence of which is to be secured by means of a **stabilisation step**. This is one of my main tasks as de.NBI coordinator. Intensive research has shown that an incorporation of the de.NBI network into the Leibniz Association is a possible option. However, there are still a number of negotiations and talks to be held before admission to the Leibniz society. Thus, the envisaged stabilisation will not be a seamless continuation of the de.NBI project drawing to a close. Fortunately, the BMBF has agreed to support the de.NBI network with **bridging funding** until the end of 2021. The members of the de.NBI network are very grateful for this solution and hope that the planned stabilisation will ensure the long-term existence of the de.NBI network in the future.

**Prof. Dr Alfred Pühler**  
de.NBI Coordinator (right)

-----  
puehler@cebitec.uni-bielefeld.de

**Prof. Dr Andreas Tauch**  
Head of the de.NBI  
Administration Office (left)

-----  
tauch@cebitec.uni-bielefeld.de



\* At the Galaxy site in Freiburg, there are roughly another 10,000 users, who mainly perform micro jobs in the de.NBI cloud there.



# de.NBI SERVICES

## Tools, Workflows, Databases, Consulting

One of the main tasks of the de.NBI network is the service area. de.NBI offers a diverse portfolio of web tools, workflows and databases that are available to life science researchers for the analysis of large amounts of data. Besides statistical consulting, advice on the tools offered is also available. All de.NBI tools are open source.

100

### de.NBI SERVICES...

PROVIDE MORE THAN 100 TOOLS FOR THE ANALYSIS OF LARGE QUANTITIES OF DATA IN THE LIFE SCIENCES.

Protein List Comparator  
 EDGAR RNA-seq end-to-end workflow  
 Excmplify Quality-standards Freiburg RNA tools  
 webserver PIPmiR IPK-Blast-Server Github-repository-galaxytools BiBiServ tools INFO-RNA TPP Pan-Cancer-alignment-workflow microMUMMIE rightfield data-standardisation-and-conversion-service circBase iPATHKNIME  
 Cellular phenotyping of microscope image data MORRE ReadXplorer SABIO RK services blockbuster GotohScan roddy CopraRNA tRNadb PlabiPD  
 TargetThermo specl SIACAT OTP Conveyor-workflows eggNOG NGS Pipelines CRISPR iTOL  
 PIA Unique-peptide-finder GBIS galaxy rna workbench Patient-Searchtool SEEK SDA Hardware-Sharing PAA Vienna RNA package SABIO-RK PicTar MOTUs motifSearch pSILAC PLEXY RSVP SILVA IntaRNA MARNA DARIO MGX CARNA DEXSeq DELLY Bioinformatical consulting and statistical analysis of proteomics data IceLogo PIA memeris BRENDA e!DAL RNAsnoop ProMeTra EURISCO PlantsDB SNV-calling-pipeline Docker-images:-galaxy-stable PeptideShaker Enterotyping EBI-image-&-RBioFormats AntaRNA COMBINE-Archive-Toolkit CrossPlatformCommander GenomeRNAi doRNA S-Peaker SILVAngs SpliceMap MeltDB segemehl ExpaRNA MITOS LocARNA BiVes Freiburg-Galaxy-Server RNApIex ProCon OpenMS BacDive snoStrip WaRSwap KNIME EMMA2 PANGAEA workflows-and-recipes Cloud/HPC IONiser Pan-Cancer-alignment-workflow

450,000

Users...

PER MONTH.



"For the evaluation of biological data sets, de.NBI provides researchers with over 100 bioinformatics tools, including consultation with experts."

**Rabeaa Alkhateeb**  
 de.NBI Service Coordinator  
 contact@denbi.de  
 www.denbi.de/services



# de.NBI TRAINING

## Workshops, Hackathons, Summer Schools

Next to service, the training area of the de.NBI network also plays a major role. In a variety of training courses, de.NBI users are trained in the use of bioinformatics tools, thus enhancing their understanding of the results achieved. Current developments in the field of bioinformatics are also addressed in de.NBI symposia, special workshops and annual summer schools.

5,000

Participants...

HAVE BEEN TRAINED IN de.NBI COURSES SO FAR.

Advanced modeling with Copasi  
 Analyzing metabolic networks with CellNetAnalyzer Applied Metaproteomics Workshop  
 Bioimage Analysis Course Data Management For Plant Genomics & Phenomics Differential analysis of proteomic data using R Galaxy for linking bisulfite sequencing with RNA sequencing Galaxy workshop on HTS data analysis Genomics and Metagenomics training course Genomics training course Introduction to BRENDA and ProteinPlus Introduction to Python Programming Linux Command Line & Basic Scripting course Machine Learning in R Microscopy Image Analysis Course Nanopore Best Practice Workshop  
 de.NBI Cloud User Meeting  
 Proteomics and Metabolomics with OpenMS SILVA/BacDive Workshop: From Primer to Paper and Back Single-Cell Omics workshop Software Carpentry workshop Spring School "Computational Biology Starter" Statistical analysis & qualitative and quantitative comparison of lipidomics data Tool-Training for Proteomics Tools for Systems biology modeling and data exchange Training on microbial phylogeny and diversity analysis Metabolomics Data Clinic Data Interpretation of Whole-Genome and Exome Data in Cancer Research Statistics and Computing in Genome Data Science The Linux Command Line: From Basic Commands to Shell Scripting Phylogenetic reconstruction course  
 DNA Methylation: Design to Discovery

Eukaryote genome annotation workshop

Big Data Training Course in Plant Genomics

280

Training courses...

HAVE BEEN HELD.

"To ensure that our tools are used optimally for data analysis, we offer a wide range of training courses, workshops, hackathons and summer schools."

**Daniel Wibberg**  
 de.NBI Training Coordinator  
 contact@denbi.de  
 www.denbi.de/training

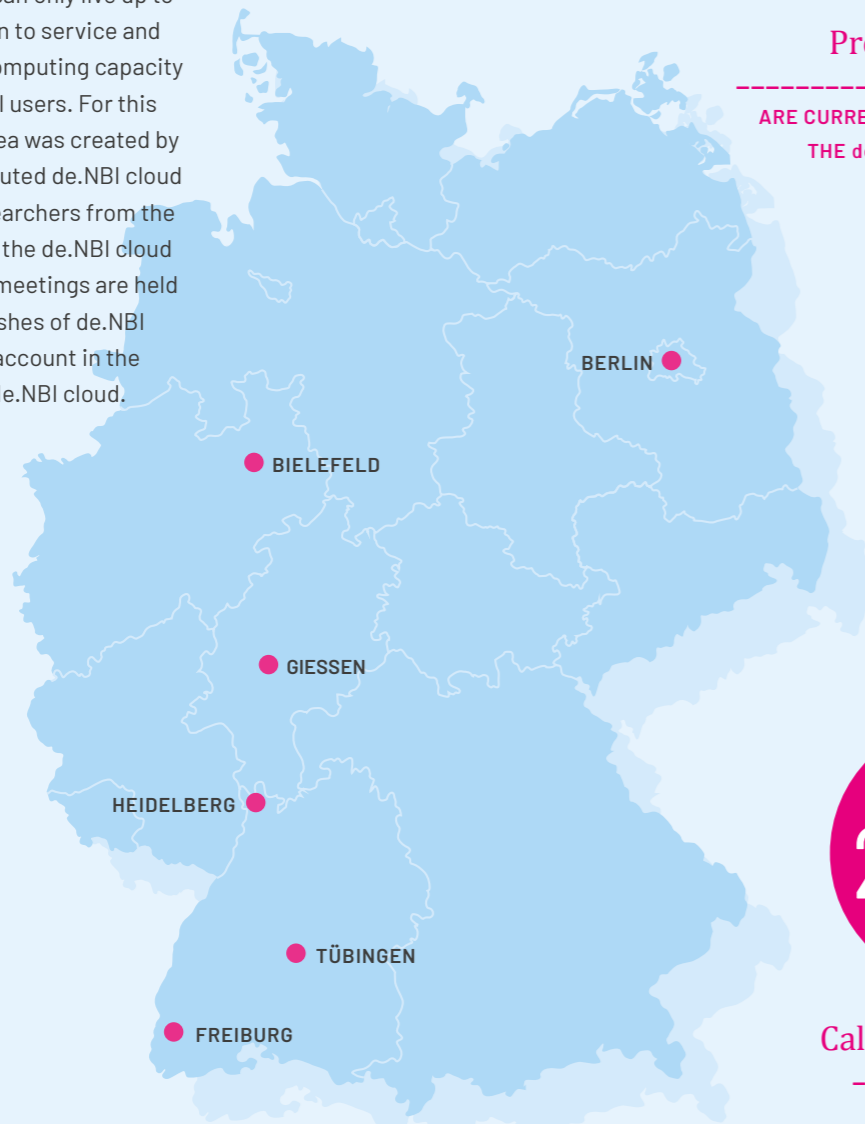




# de.NBI CLOUD

## Infrastructure, Platform and Software as a Service

The de.NBI network can only live up to its tasks if, in addition to service and training, adequate computing capacity is available for de.NBI users. For this reason a compute area was created by establishing a distributed de.NBI cloud at six locations. Researchers from the life sciences can use the de.NBI cloud free of charge. User meetings are held to ensure that the wishes of de.NBI users are taken into account in the development of the de.NBI cloud.



250

Projects...

ARE CURRENTLY RUNNING IN THE de.NBI CLOUD.

38

Petabytes of storage space

20,000

Calculation engines



“With the establishment of the de.NBI cloud, we are responding to the international trend in bioinformatics to develop scalable approaches for analysing large amounts of data.”

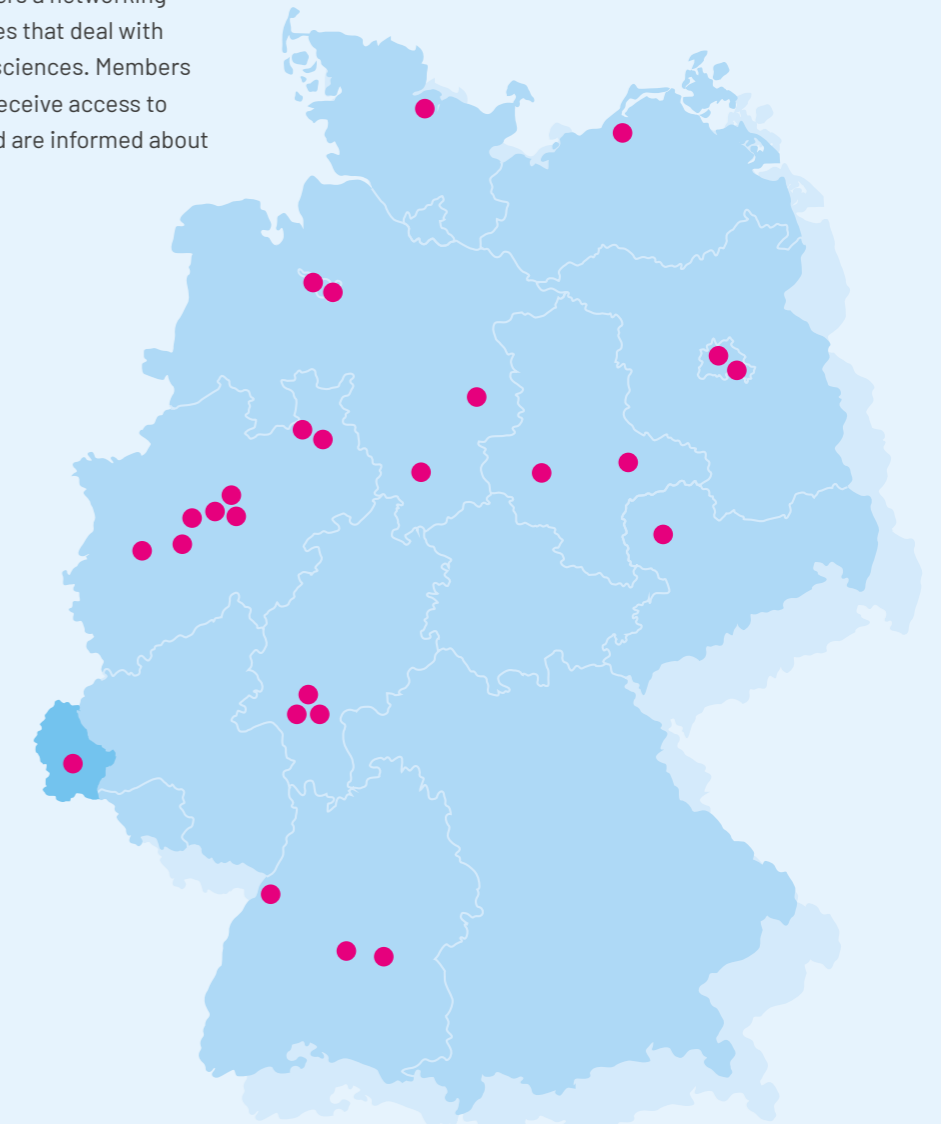
**Peter Belmann**  
de.NBI Cloud Governance  
cloud@denbi.de  
www.denbi.de/cloud



# de.NBI INDUSTRIAL FORUM

## Software Solutions, Consulting, Networking

The de.NBI Industrial Forum offers a networking platform for industrial companies that deal with huge amount of data in the life sciences. Members of the de.NBI Industrial Forum receive access to de.NBI services and training and are informed about developments in the network.



26

Members ...

IN GERMANY AND LUXEMBOURG.

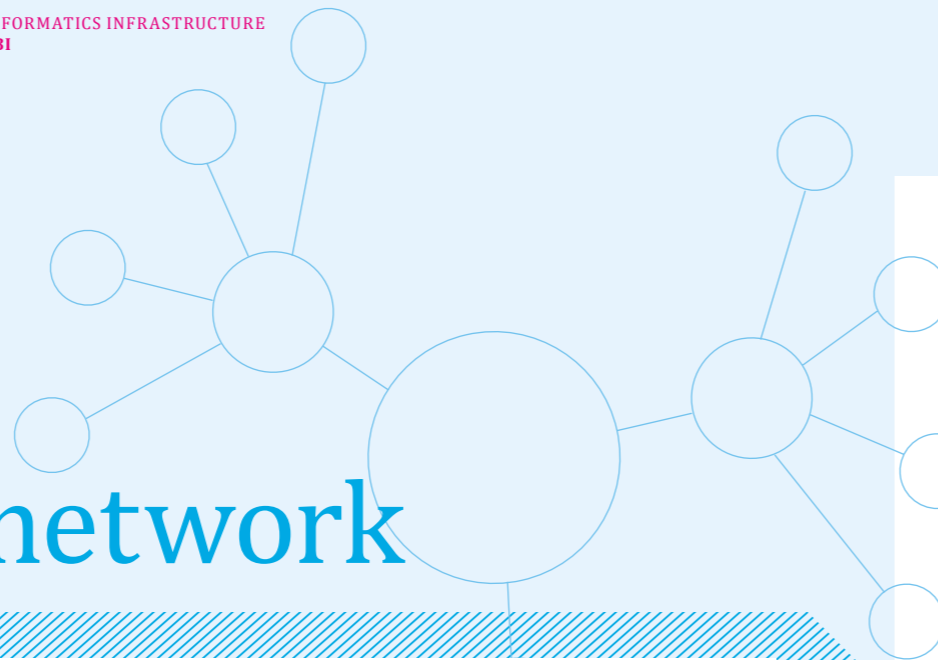
“Nowadays the analysis of large amounts of data in the life sciences is also extremely relevant for industrial companies. With the de.NBI Industrial Forum, we provide a transfer of expertise between academy and industry.”

**Manuel Wittchen**  
de.NBI Industrial Forum Manager  
contact@denbi.de  
www.denbi.de/industrial-forum





# Activities in the de.NBI network



**de.NBI Training Course 2018:  
Introduction into Targeted and  
Untargeted Metagenome Analysis**

-----  
Gießen



**de.NBI Plenary Meeting 2018**

-----  
Berlin



**Editorial team of the de.NBI Administration Office**

From top to bottom: Peter Belmann, Manuel Wittchen,  
Doris Jording, Daniel Wibberg, Andreas Tauch, Irena  
Maus, Tanja Dammann-Kalinowski, Alfred Pühler

-----  
Bielefeld



**de.NBI Summer School 2018:  
Riding the Data Life Cycle**

-----  
Braunschweig

**Spring School 2019:  
Computational Biology Starter**

-----  
Gatersleben





**de.NBI Training Course 2017:  
High-Throughput Genome Analysis  
and Comparative Genomics**

-----  
Bielefeld

# IMPRINT

Prof. Dr Alfred Pühler  
German Network for Bioinformatics Infrastructure (de.NBI)  
de.NBI Administration Office  
Center for Biotechnology  
Universitätsstraße 27  
33615 Bielefeld

Tel: +49 (0)521 106 8750  
Fax: +49 (0)521 106 89046  
E-Mail: [contact@denbi.de](mailto:contact@denbi.de)

[www.denbi.de](http://www.denbi.de)  
 [@denbiOffice](https://twitter.com/denbiOffice)  
 [linkedin.com/company/de-nbi](https://www.linkedin.com/company/de-nbi)

Date: August 2020

Photo credits:  
iStockphoto, Pixabay, ROV-Team/GEOMAR (CC BY 4.0)

Design and Layout:  
MEDIUM Werbeagentur GmbH, Bielefeld

Translation:  
Sprachenfabrik GmbH, Bielefeld

Lektorat:  
Kern AG, Bielefeld

Printing:  
Bruns Druckwelt GmbH & Co. KG, Minden

SPONSORED BY



Fkz 031A532B  
(de.NBI administration office)



