

# systembiologie.de

THE MAGAZINE FOR SYSTEMS BIOLOGY RESEARCH IN GERMANY

ISSUE 10 JUNE 2016

**special:**  
bioinformatics,  
a key technology

an introduction to the german  
network for bioinformatics  
infrastructure – de.NBI

from page 8

ethical and legal  
issues in genome  
research

page 40

something to remember

page 72

interviews with  
Niklas Blomberg,  
Heyo K. Kroemer,  
Peter Jansen and  
Olaf Wolkenhauer

page 28, 60, 69 and 66

SPONSORED BY THE



Federal Ministry  
of Education  
and Research

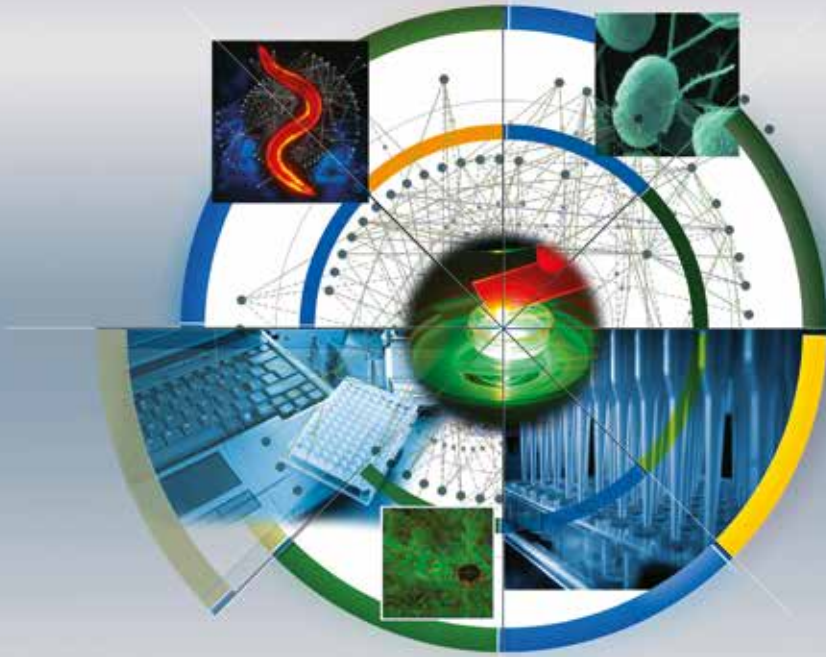


Photo: Detrichs Kommunikation GmbH, Jülich, using pictures from Prof. Dr. Ralf Baumeister

## systembiologie.de

Systems biology is a young and dynamic discipline that sees the whole picture. As part of the life sciences it builds a bridge between sophisticated laboratory experiments and mathematical modelling, between high-tech data measurements and computer-aided data evaluation. Its research subjects are the network-like entangled activities of signal transduction and metabolism in cells, tissues, organs and organisms. Systems biology research deals with this complexity by organising itself into interdisciplinary networks. Experience this fascinating, upcoming branch of science and what answers it provides to previously unresolved questions about human life.



Cover photo: Sergey Nivens – Fotolia.com



# welcome note

Esteemed Reader,



Modern information technologies have changed the world of biosciences dynamically in recent years. Entire genomes can now be sequenced in just a few days. Large numbers of proteins can be analysed using the latest technologies. Bioinformatics has established itself as the key technology for evaluating the wealth of data thus produced. It provides the basis for using data efficiently and for constructing systems-biological and systems-medical models.

The Federal Ministry of Education and Research is therefore supporting the establishment of a “German Network for Bioinformatics Infrastructure” with funding of approximately 22 million euros over a five-year period. The aim is to provide bioinformatics analysis tools and the appropriate expert advisory services nation-wide – so that researchers working in the life sciences and biomedicine can profit from the latest bioinformatics procedures.

The interaction between experiments, bioinformatic analyses, mathematical modelling and computer simulation is producing an increasingly detailed and comprehensive depiction of biological processes. This could bring us closer to being able to model and simulate entire cells, making it possible to predict the effects of medical drugs in the overall context of systems biology.

This edition of [systembiologie.de](http://systembiologie.de) provides a fascinating insight into the opportunities which the fusion of informatics and the life sciences offers systems biology and biomedical research in the 21st century.

A handwritten signature in blue ink that reads "Johanna Wanka". The signature is fluid and cursive, written in a professional style.

Prof. Dr. Johanna Wanka

Federal Minister of Education and Research



# greetings

Dear Readers,

Life without computers has become inconceivable. Just as the daily use of PCs, smartphones and digital cameras is now the norm for most people, scientific research cannot be imagined without the help that computers provide.

While the use of supercomputers and computer models during the last millennium was confined mainly to disciplines such as particle physics or climate research, over the past fifteen years methods such as these have also become an integral component of life sciences. A prime example of this is systems biology, in which an entire discipline has been built around the use of mathematical computer models to simulate biological systems. Comparison between computer-generated model predictions and experimental results provides the basis for the further adjustment of theoretical models to fit biological reality. A state-of-the-art IT infrastructure is also essential for bioinformatics. This magazine will introduce you to the broad range of activities in which the German Network for Bioinformatics Infrastructure engages. The Jülich Research Centre with its Jülich Supercomputing Centre operates the Biology Simulation Laboratory in order to meet these requirements (see p. 56).

But it is not just within the context of basic research that it has become impossible to imagine day-to-day life without computers. Work within the medical sector is becoming increasingly digitised, also. These days, the daily volumes of data created during routine care – by imaging techniques, laboratory tests or long-term studies, for example – are enormous. In addition, we have seen the emergence of new technology such as genome sequencing, which increasingly is becoming a proven method of obtaining molecular diagnoses. However, when it comes to using this wealth of data productively, the greatest challenge is no longer to collect data, but rather its correlation and integral analysis. In this respect, there is still significant potential for development in both basic medical research as well as patient care. The researchers involved in the international project *Pan-Cancer Analysis of Whole Genomes (PCAWG)* (see p. 34) have designed an excellent approach to handle data integration. The amalgamation of over 2,800 tumour sequences from various projects enables the development of answers to highly relevant questions relating to the link between mutations and the emergence of cancer.

In order to facilitate research projects such as this more often in Germany, the Federal Ministry of Education and Research has launched a medical informatics initiative (see interview with Prof. Kroemer, p. 60), which will lay the foundation for further digitization of medicine and of biomedical research over the coming years.

This current issue of *systembiologie.de* offers interesting insights into the huge potential of information technology within the field of life sciences. I hope you enjoy reading it!

Prof. Dr. Otmar D. Wiestler

President of the Helmholtz Association



# foreword

## Who would have thought it...



Dear readers, in your hands you have a special edition on bioinformatics of a magazine that is usually dedicated to systems biology. What has happened? Life sciences and health research have been overrun by developments, which just a decade ago would have been almost impossible to predict. The production of the first blueprint of the human genome at the start of the 2000s still required the combined efforts of all the world's sequencing laboratories, at the immense cost of approximately three billion euros. Only a decade later, a modern genome centre is capable of sequencing dozens of these human genomes each day for a price of around one thousand euros. And the prices continue to fall. Obtaining genome data is no longer a limiting factor, but the analysis of this data is proving to be a stumbling block when it comes to making advancements in the field of life sciences.

So where will we be in another ten years? Some would say the answer to this question requires prophetic capabilities. "The best way to predict the future is to invent it," according to Alan Curtis Kay, who studied mathematics and molecular biology in the 1960s and actively shaped the future of the information society in remarkable ways, in his role as a pioneer of object-oriented programming and an architect of window-based graphical user interfaces.

There is great opportunity for bioinformatics to shape the future of life sciences. Anything which is achievable using current information technology and which is desired by society appears feasible. In ten years we will be able to sequence a human genome at a fraction of today's costs. Genome sequencing will therefore be possible for anyone, at any time. This also applies to the recording of the human epigenome, which acts as a mirror image of the genome's interactions with its environment. In ten years' time we will probably collect and sequence biological samples such as blood or saliva on a regular basis throughout our lifetime, beginning at birth, which will allow us to predict the emergence of specific diseases later in life. A kind of (epi)genetic early warning system for health. In addition, when a disease emerges, we will be able to describe in detail the point at which an (epi)genetic dysregulation occurred, hence enabling an even more precise treatment of that specific disease.

Is this the brave new world of (epi)genetics? Besides all the legal and ethical challenges associated with comprehensive genetic monitoring, it is important to remember that continuous recording of our (epi)genetic condition in this form will yield vast quantities of data. Today, the storage of large quantities of sequence data is already a tremendous challenge. Anyone hoping that the problem will resolve itself as the result of the rapidly falling prices in the IT sector must remember that the price of sequencing is dropping at an even faster pace. In other words, time is not on our side: the life sciences are producing more data than we are able to store, in increasingly shorter cycles. If bioinformatics is to keep this flood of information under control, it must uncover the most important patterns within this data. And it needs to do so ideally in real-time. This is already a reality within other disciplines such as particle physics, where tremendous quantities of data are produced with hitherto unrivalled precision, so as to allow the laws of physics to be investigated. In future, those of us working in life sciences will follow the example set by the LHC, the Large Hadron Collider at CERN, and retain only a fraction of the data, discarding the vast majority in real-time. This is where bioinformatics will set the standards, push the boundaries of what is possible and help provide as yet unknown insights into the worlds of health and illness.

The articles featured throughout the following pages of this magazine will shed light on bioinformatics from a hugely varied range of perspectives. We will hear from bioinformaticians and systems biologists as well as users and promoters of this discipline. What they all have in common is their immense enthusiasm for bioinformatics.

I hope that this enthusiasm proves infectious as you enjoy the entertaining reading material that you are sure to find in this issue of *systembiologie.de*.


A blue ink handwritten signature of Roland Eils, consisting of a stylized 'R' and 'E'.

Yours, Roland Eils

Editor in Chief

# index

welcome note	3	
Prof. Dr. Johanna Wanka, Federal Minister of Education and Research		
greetings	4	
Prof. Dr. Otmar D. Wiestler, President of the Helmholtz Association		
foreword	5	
Prof. Dr. Roland Eils, Editor in Chief		
german network for bioinformatics infrastructure – de.NBI	8	
A BMBF infrastructure measure to solve the Big Data problem in life sciences by Alfred Pühler		
rna bioinformatics under one roof – the rna bioinformatics service center	14	
Challenges and solutions for a readily accessible research infrastructure by Björn Grüning		
BRENDA	18	
From a database to a centre of excellence by Dietmar Schomburg and Ida Schomburg		
NBI-SysBio – the de.NBI data management hub	22	
From specialised data silo to interconnected data by Wolfgang Müller		
the development of software solutions for microbial bioinformatics	25	
Institution portrait: Justus-Liebig University Giessen Bioinformatics Center by Alexander Goesmann		
ELIXIR – keeping data flowing	28	
Interview with Niklas Blomberg by Marcus Garzón and Vera Grimm		
omics infrastructures for research and teaching	30	
A concept by Leopoldina for a change in the field of life sciences by Alfred Pühler		
a global initiative for cancer research	34	
The Pan-Cancer Analysis of Whole Genomes (PCAWG) project by Jan O. Korbel, Sergei Yakneen, Sebastian M. Waszak, Matthias Schlesner, Roland Eils and Fruzsina Molnár-Gábor		
sensitive genome data	40	
EURAT is addressing ethical and legal issues in genome research by Sebastian Schuol and Eva C. Winkler		
big data: perspectives in cancer therapy	44	
Impressions from an industry-based point-of-view by Ajay Kumar		

<i>fachgruppe bioinformatik – representing interests with one voice</i>	49	
FaBI – the alliance of the bioinformatics special interest groups of five German scientific societies from the field of life sciences and informatics by Matthias Rarey		
news from the BMBF	52	
news from the helmholtz association An introduction of the Simulation Laboratories at the Jülich Supercomputing Centre (JSC) by Olav Zimmermann	56	
new opportunities for medicine Interview with Heyo K. Kroemer by Katja Nellissen, Marco Leuer and Bettina Koblenz	60	
i:DSem – integrative data semantics in systems medicine A new initiative by the German Federal Ministry of Education and Research promotes innovative data management in biomedical research by Christian Rückert	63	
“projects can fail in a good way” Interview with Olaf Wolkenhauer by Melanie Bergs and Gesa Terstiege	66	
“the liver is the first choice of both scientists and our network” Interview with Peter Jansen by Melanie Bergs and Gesa Terstiege	69	
something to remember A systems biological view of the epigenetic basis of memory by Tonatiuh Pena Centeno, Ramon Vidal, Magali Hennion and Stefan Bonn	72	
AptaBodies DNA aptamers as an alternative to antibodies in Western blotting by Jasmin Dehnen and Frieda Anna Sorgenfrei	76	
events	80	
news	86	
imprint	89	
about us	90	
contact data	91	



# german network for bioinformatics infrastructure – de.NBI

## A BMBF infrastructure measure to solve the Big Data problem in life sciences

by Alfred Pühler

The Big Data problem in life sciences is the consequence of a paradigm shift. “Omics” technologies – which include genomics, transcriptomics, proteomics and metabolomics – now enable the complete mapping of cellular components in any organism. This produces enormous quantities of data, which can only be saved and analysed via an extensive bioinformatics infrastructure. However, most experimental working groups do not have tools of this kind at their disposal. This is where the German Network for Bioinformatics Infrastructure (de.NBI) comes into play, assisting the experimental working groups in the field of life sciences to analyse large quantities of data produced with high throughput technologies.

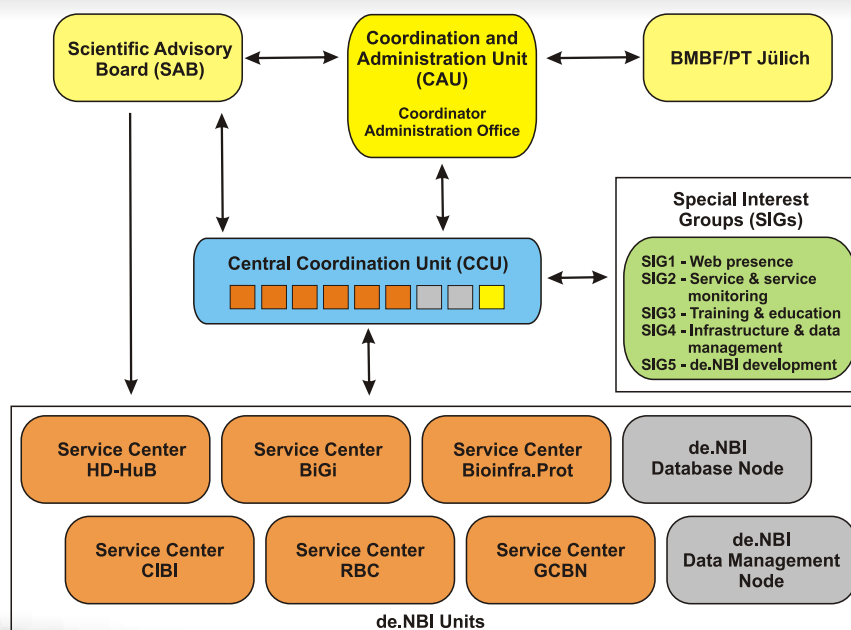
### Establishment and functions of the de.NBI network

The setup of a Network for Bioinformatics Infrastructure is based on a recommendation by the German Bioeconomy Council, which in 2012 published a statement proposing that a network of local, well-equipped and specialised centres should be created, to be led by a coordination body vested with wide-ranging powers [1].



This proposal was received by the German Federal Ministry of Education and Research (BMBF) and was implemented in May 2013 by means of a tendering process for the creation of a Ger-

Figure 1: de.NBI organizational diagram



Source: de.NBI



Figure 2: Participants in the first meeting of de.NBI's central coordination unit (Photo: de.NBI).

man Network for Bioinformatics Infrastructure (de.NBI). The de.NBI network was established in a phased process. Firstly, from among the large number of applications submitted, eight service centres were selected by an international evaluation panel, charged with the preparation of a complete proposal and responsible for ensuring cooperation between these centres as part of a network. Preparation of the proposal as a whole was placed in the hands of a coordinator and an administration office. The complete de.NBI proposal was submitted to the established evaluation panel in July 2014. The de.NBI network was officially founded in March 2015 following approval of the complete application.

The de.NBI network comprises eight service centres, covering a wide spectrum of offered services. An overview of these service centres can be found in **Table 1**. Firstly, three de.NBI centres have an organism-oriented focus. These are the Heidelberg Center for Human Bioinformatics (HD-Hub), the Bielefeld-Gießen Resource Center for Microbial Bioinformatics (BiGi) and the GCBN Center in Gatersleben, devoted to plant bioinformatics. Then there are centres with a methodical approach. The Freiburg RBC Center focuses on RNA bioinformatics, while the Bochum-based BioInfra.Prot Center specialises in proteome bioinformatics. The Center for Integrative Bioinformatics in Tübingen provides software libraries for mass spectrometry and sequence analysis. Finally, the remaining service centres are data-oriented. The first of these is the Database Service Center in Bremen, followed by the NBI-SysBio Data Management Center in Heidelberg. These eight service centres are the workhorses of the de.NBI network and are ultimately responsible

for ensuring that de.NBI achieves the goals for which it was established.

The key words “service, training and education” ideally describe the tasks entrusted to the de.NBI network. “Service” stands for the willingness to assist experimental working groups in the field of life sciences with the analysis of large data packets. Courses explaining the use of available bioinformatics programmes to experimental researchers help to fulfil the aspect of “training”. de.NBI-specific workshops and summer schools also contribute here. Additionally, the de.NBI consortium is charged with promoting integration of de.NBI within the Europe-wide ELIXIR bioinformatics network, facilitating the involvement of industrial companies in the de.NBI network and developing strategies for the consolidation of de.NBI beyond the five-year period for which it will receive funding.

### Organisation and procedures within the de.NBI network

A clearly ordered organisational structure provides for the integration of the eight selected service centres within a single network (Fig. 1). The most important element in this organisational chart is the central coordination unit, which makes all decisions relating to issues that affect the network. Within the coordination unit, each of the eight service centres is represented by a delegate, with one further seat held by the de.NBI coordinator. The coordination unit has set up special interest groups to perform the groundwork for the coordination unit and most notably to prepare any upcoming decision-making processes. The de.NBI specialist groups work on subject areas such as web

**Table 1: List of de.NBI Partners**

<b>CENTRES</b>	<b>PARTICIPATING PARTNERS</b>
<b>Heidelberg Center for Human Bioinformatics – HD-HuB</b> Centre Coordinator: Roland Eils, Heidelberg	<ul style="list-style-type: none"> <li>• Heidelberg University</li> <li>• DKFZ Heidelberg</li> <li>• EMBL Heidelberg</li> </ul>
<b>Bielefeld-Gießen Resource Center for Microbial Bioinformatics – BiGi</b> Centre Coordinator: Jens Stoye, Bielefeld	<ul style="list-style-type: none"> <li>• Bielefeld University</li> <li>• Gießen University</li> </ul>
<b>Bioinformatics for Proteomics – BioInfra.Prot</b> Centre Coordinator: Martin Eisenacher, Bochum	<ul style="list-style-type: none"> <li>• Bochum University</li> <li>• Leibniz-Institut für Analytische Wissenschaften - ISAS - e.V., Dortmund</li> </ul>
<b>Center for Integrative Bioinformatics – CIBI</b> Centre Coordinator: Oliver Kohlbacher, Tübingen	<ul style="list-style-type: none"> <li>• Tübingen University</li> <li>• Free University of Berlin</li> <li>• Konstanz University</li> </ul>
<b>RNA-Bioinformatics Center – RBC</b> Centre Coordinator: Rolf Backofen, Freiburg	<ul style="list-style-type: none"> <li>• Freiburg University</li> <li>• Leipzig University</li> <li>• Max Delbrück Center for Molecular Medicine, Berlin</li> </ul>
<b>German Crop BioGreenformatics Network – GCBN</b> Centre Coordinator: Uwe Scholz, Gatersleben	<ul style="list-style-type: none"> <li>• Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben</li> <li>• Helmholtz Centre Munich</li> <li>• Jülich Research Centre</li> </ul>
<b>Databases</b> Centre Coordinator: Frank-Oliver Glöckner, Bremen	<ul style="list-style-type: none"> <li>• Jacobs University Bremen gGmbH</li> <li>• Bremen University</li> <li>• Braunschweig Technical University</li> <li>• Leibniz-Institut DSMZ GmbH, Braunschweig</li> </ul>
<b>Data Management Node – NBI-SysBio</b> Centre Coordinator: Wolfgang Müller, Heidelberg	<ul style="list-style-type: none"> <li>• Heidelberg Institute for Theoretical Studies</li> <li>• Rostock University</li> </ul>

presence, service and service monitoring, training and education, infrastructure and data management as well as the de.NBI development. The network is controlled by the aforementioned coordinator together with a branch office. The branch office itself consists of a branch office leader, a research assistant, a web and service expert and an expert for training and education.

In the first year of its existence, the most important function of the de.NBI network was to introduce coordinated structures to the work of the network as a whole. This task was discussed in multiple meetings of the central coordination unit, during which the de.NBI specialist groups contributed important ideas. As such, the specialist **web presence** group made preparations



for de.NBI's online presence, which must be considered a key element of the network as a whole. The de.NBI webpage, accessible via the link [www.denbi.de](http://www.denbi.de), provides information on the current state of development. The specialist **service and service monitoring** group created a list of all the bioinformatics services offered by the individual service centres. The list is composed of around 80 offered services and in future will also be listed on the de.NBI homepage. The specialist **training and education** group, together with the training expert, coordinate the range of training courses which are developed within the individual service centres. In 2015, 16 training courses attended by 329 participants in total have already been conducted. The specialist **infrastructure and data management** group focuses primarily on the computing capacity available within the de.NBI network, which is currently regarded as somewhat below par. This group developed initial procedures for the handling of data from an ethical perspective, especially in the medical sector. The specialist de.NBI development group usually handles overarching issues such as the ongoing development of the de.NBI network with regard to the addition of specialisations, international cooperation, involvement of industrial companies and consolidation within the network. So far, this area has seen vital work relating to the addition of specialisations within the de.NBI network by so-called partner projects, and the integration of de.NBI into the pan-European bioinformatics network ELIXIR as a national hub.

The central coordination unit meets every three months at the various locations of the individual service centres. The first meeting, which took place in Berlin on 10 March 2015 (Fig. 2), defined the administrative regulations to enable the de.NBI network to commence proper operations. The kick-off meeting of the de.NBI network was then held on 26 March 2015 in Bielefeld at the same time as the inaugural general assembly (Fig. 3). Since that time, the individual service centres have cooperated intensively, hence ensuring that the de.NBI network's objectives are achieved.

Another function of the de.NBI network was to establish a scientific advisory committee. To this end, a list of potential candidates was submitted to the BMBF by the de.NBI coordinator and the central coordination unit. Six appointments were made to the advisory committee after confirmation of this list. The six members are seasoned bioinformaticians with outstanding knowledge of the bioinformatics infrastructures field. An initial meeting between the scientific advisory committee and the de.NBI network took place in November 2015. The de.NBI network submitted a written report giving an account of the establishment of de.NBI and the work accomplished. The meeting between the advisory committee and the network was held as part of a de.NBI workshop, which focused on the results achieved by the de.NBI representatives so far. The advisory committee compiled a detailed report examining the individual elements of de.NBI. This report confirmed that the de.NBI network carried out excellent work during the first months of its existence. In addition, it made recommendations on the direction in which the de.NBI network should develop. The opinion of the scientific advisory committee is of great importance to de.NBI. It will form the basis for a de.NBI development plan, which will be agreed following detailed discussions.

### A development plan for the expansion of the de.NBI network

Following the end of the first year of funding, it is clear that the de.NBI network was set up promptly and that it has worked emphatically towards implementing its objectives since then. Discussions with the scientific advisory committee have shown that the network's aims should be extended. Consequently, a development plan comprising the items outlined below (Table 2) is currently under discussion within de.NBI.

The initial task will be to close some topical gaps in the de.NBI network. Subsequently, a research programme with postgraduate students is intended to be put in place in order to establish a

**Table 2: The de.NBI Development Plan 2016**

- Integration of partner projects
- Establishment of a research programme with postgraduate students
- Integration of de.NBI into ELIXIR as a national hub
- Development of an industrial branch of the de.NBI network
- Development of a de.NBI cloud
- Consolidation of the de.NBI network

research component within the de.NBI network. Its purpose will be to train bioinformaticians in the field of bioinformatics infrastructure. Furthermore, de.NBI is to be integrated into the Europe-wide ELIXIR bioinformatics infrastructure network in the form of a national hub. To this end, legal and personnel-related prerequisites must be defined in order to successfully achieve close cooperation. The involvement of industrial companies in the de.NBI network is to be promoted through the creation of an industrial branch of de.NBI. Here it is particularly important to clarify how this industrial branch can interact with the academically oriented de.NBI network. A de.NBI cloud will be set up to resolve the issue of

insufficient computing capacity within the de.NBI network. Of course, this de.NBI cloud will be coordinated with other cloud activities in Germany and Europe. Finally, de.NBI must still develop a plan for consolidating the network following the end of the five-year period during which it will receive funding from the BMBF. In order to do this, legacy plans in other European countries participating as national hubs in ELIXIR will be evaluated and analysed in order to assess their transferability.

## German Network for Bioinformatics Infrastructure (de.NBI) profile

The de.NBI network, a project of the German Federal Ministry of Education and Research, was founded on March 1st, 2015. It consists of eight service centres with a total of 23 partners and is led by a coordinator with an administration office. A central coordination unit was set up to manage the de.NBI network in which all eight service centres are represented by coordinators.

The de.NBI network was established to assist experimental working groups operating in the field of life sciences with the analysis of large quantities of data. The main de.NBI functions are divided into service, training and education. In addition, de.NBI is to pave the way for cooperation with other bioinformatics infrastructures in Europe, to involve relevant industrial companies and to develop legacy models for its continued existence.



Figure 3: Participants in the kick-off meeting of the de.NBI network in Bielefeld (Photo: de.NBI).

### Partners involved:

#### Coordinator:

Prof. Dr. Alfred Pühler, Center for Biotechnology (CeBiTec),  
University of Bielefeld, Germany

#### Head of Administration Office:

Prof. Dr. Andreas Tauch, Center for Biotechnology (CeBiTec),  
University of Bielefeld, Germany

#### Heads of Service centres involved:

Prof. Dr. Rolf Backofen, Freiburg, Germany

Prof. Dr. Roland Eils, Heidelberg, Germany

PD Dr. Martin Eisenacher, Bochum, Germany

Prof. Dr. Frank Oliver Glöckner, Bremen, Germany

Prof. Dr. Oliver Kohlbacher, Tübingen, Germany

PD Dr. Wolfgang Müller, Heidelberg, Germany

Dr. Uwe Scholz, Gatersleben, Germany

Prof. Dr. Jens Stoye, Bielefeld, Germany

### Contact:



#### Prof. Dr. Alfred Pühler

Coordinator

puehler@cebitec.uni-bielefeld.de



#### Prof. Dr. Andreas Tauch

Co-Coordinator and

Head of Administration Office

tauch@cebitec.uni-bielefeld.de

Center for Biotechnology

Bielefeld University

Bielefeld, Germany

contact@denbi.de

### References:

[1] Requirements for a Bioinformatics Infrastructure in Germany for future Research with bio-economic Relevance, Bioeconomy Council Recommendations 06 (2012) [http://bioekonomierat.de/en/publications/?tx\\_rsmpublications\\_pi1\[publication\]=20&tx\\_rsmpublications\\_pi1\[action\]=show&tx\\_rsmpublications\\_pi1\[controller\]=Publication&cHash=4dd14c6e2719375142ac9c79e5961151](http://bioekonomierat.de/en/publications/?tx_rsmpublications_pi1[publication]=20&tx_rsmpublications_pi1[action]=show&tx_rsmpublications_pi1[controller]=Publication&cHash=4dd14c6e2719375142ac9c79e5961151)

[www.denbi.de](http://www.denbi.de)





# rna bioinformatics under one roof – the rna bioinformatics service center

## Challenges and solutions for a readily accessible research infrastructure

by Björn Grüning

The RNA Bioinformatics Center (RBC) is a service centre within the German Network for Bioinformatics Infrastructure (de.NBI), which deals with the RNA-based mechanisms of gene regulation. The role of the RBC is to develop a comprehensive, integrative platform for RNA analysis and in doing so, to explain the significance of RNA with respect to gene regulation. The services of the RBC range from advice on experimental study design, to the preparation of protocols for data analysis and its associated infrastructure, to the development of specific solutions to individual scientific issues.

### The significance of non-coding RNAs in medical research

For a long time, non-coding ribonucleic acids (ncRNAs) and RNA-protein interactions were neglected by the scientific community. After all, the focus of research until just a few years ago was placed mainly on the protein-coding regions of DNA. However, genome-wide sequencing showed that the majority of DNA does not code for proteins, but for ncRNAs. In order to investigate regulatory RNAs, e. g. micro-RNAs (miRNA), and RNA-protein interactions, new technologies were developed which showed that the complexity of gene regulation at the post-translational level is comparable to transcriptional gene regulation. The human genome contains thousands of miRNAs and at least 800 RNA-binding proteins. With this new knowledge, scientists have already demonstrated that many diseases are triggered not only by mutations in specific genes, but that their cause can lie specifically in post-transcriptional gene regulation<sup>1</sup>.

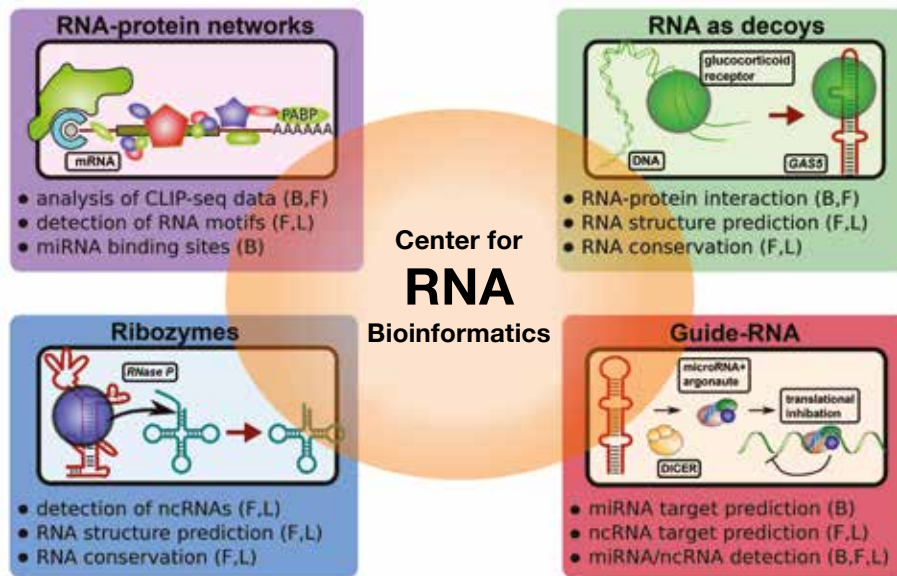
### Objectives and tasks of the RBC

The RNA Bioinformatics Center (RBC) brings together the internationally renowned German RNA bioinformatics work-

ing groups from Freiburg (Rolf Backofen, coordinator), Berlin (Nikolaus Rajewsky and Uwe Ohler, MDC) and Leipzig (Peter F. Stadler). The RBC pursues three objectives:

- 1.) The establishment of an easily accessible RNA workbench. This analytical environment can be used on any PC or alternatively it can be made available in a HPC environment (university computing center or cloud computing).
- 2.) Creation of a sustainable infrastructure. The RBC works together with numerous other centres, firstly to promote interoperability between the centres, and secondly to provide the scientific community with a broad foundation and to ensure that the developed solutions are sustainable.
- 3.) In addition to productive use, the RNA workbench also serves as an experimental platform for courses and other forms of training. Comprehensive training sessions and workshops on the topic will be held additionally to allow junior researchers and scientists without knowledge of bioinformatics to access the RNA-based analysis tools, and to pass on knowledge relating to the significance of RNA-dependent regulation.

RBC applies the globally utilised Galaxy<sup>2</sup> workflow management system in order to achieve the objectives described above. This platform allows life scientists to fully analyse independent sequence data in a transparent and reproducible way and to share the information with colleagues. Until now, over 50 different bioinformatics applications have been integrated within the RNA workbench, which was specially programmed to analyse RNA-based data. Besides tools for analysing transcriptome data, the RBC also includes within the workbench a variety of tools for RNA structure analysis (LocARNA, GraphClust, ExpaRNA), for ncRNA target structure prediction (ViennaRNA Suite) and for RNA transcript definition and classification. The purpose here is to fully analyse and understand the different levels of regulation (transcription, splicing, translation and degradation) of RNA transcripts and their interdependencies.



**Figure 1: The RNA Bioinformatics Center (RBC)**, composed of AG Prof. Dr. Rolf Backofen in Freiburg (F), AG Prof. Dr. Peter F. Stadler in Leipzig (L) and AG Prof. Dr. Uwe Ohler and Prof. Dr. Nikolaus Rajewsky (B), is the central point of contact for all questions relating to RNA bioinformatics (Graph: Prof. Dr. Rolf Backofen).

Moreover, the tools already developed by RBC and used worldwide for the characterisation of ncRNAs are implemented in Galaxy. They include applications for detecting miRNAs (MirDeep, PipMir, BlockClust, RNaz) and new transcripts (LncRNAs), for mapping HTS data (Segemehl Suite<sup>3</sup>) and for predicting RNA-RNA and protein-RNA interactions (IntaRNA, RNAup, CopraRNA, PARalyzer<sup>4</sup>, Dorina<sup>5</sup>, GraphProt<sup>6</sup>). Many of these tools are already available as individual web services<sup>a</sup>.

### Galaxy – a web-based, open-source platform for data-intensive biomedical research

There are numerous commercial and freely available RNA analysis tools. Nevertheless, easy access to a unified, integrative system offering standardised examination of RNA-based gene regulation has been missing so far. Furthermore, most software tools used by life scientists are almost impossible to operate without advanced IT skills, as their installation can sometimes be anything but simple and access often requires programming expertise. With this in mind, Galaxy offers a unique solution due to the fact that freely available and self-developed tools as well as visualisations and databases can be integrated, so that these components can be easily accessed and used in a transparent manner by any life scientist. There are currently more than 2,000 different tools available in Galaxy, from simple text manipulation to HTS analysis (e.g. mapping, differential gene expression analysis). This platform allows any user to combine the full set of tools within workflows and in doing so to operate

the system in a transparent and reproducible way. The Galaxy server in Freiburg is one of the largest Galaxy instances in Germany. Here, RBC makes an immense contribution to the further development of the Galaxy platform. Many renowned universities use the RBC server as a template to create local Galaxy instances and to tap into RBC's docker-based virtualisation solution. Besides the provision of data analysis, a key area of the RBC service centre is to train operators in the use of the tools and infrastructure. All RBC partners offer various de.NBI workshops and practical tutorials on the use of Galaxy and the analysis of RNA data several times each year. The workshops cover a number of different topics, e.g. genome annotation or high-throughput sequencing (HTS) data analysis, and run over a period of 1-5 days. RBC also takes part in the summer schools organised by the de.NBI network in order to teach knowledge relating to RNA bioinformatics. Moreover, the RBC organises six-monthly hackathons during which developers from Germany and abroad meet to integrate new tools into the RNA workbench and to run tests. What's more, RBC develops additional continually updated analysis workflows for standardised HTS data analysis, which it makes freely available to the scientific community.

### Access to the RNA workbench

The Galaxy-based RNA workbench is an easily operable browser system that does not require installation, so users do not have to download any large data sets and analysis will not occupy local computer memory. RBC has released a Galaxy server with the RNA

<sup>a</sup> <http://rna.informatik.uni-freiburg.de/>  
<http://www.bioinf.uni-leipzig.de/webServices.html>  
<https://ohlerlab.mdc-berlin.de/software/>

workbench tools and others for testing. The system places high demands in regard to computing power, storage capacities for HTS data and individual requirements such as security and data protection, which means that not all analyses can be carried out centrally on one server. With this in mind, RBC has developed an alternative, which allows each user to access the RNA workbench independent from the public servers. This alternative is called Galaxy Docker. Galaxy Docker is a preconfigured Galaxy platform with a set of tools that can be individually compiled and used on any system. The remote architecture of the Galaxy Docker project means the analysis environment can also be operated within closed networks, making it ideal for the analysis of highly sensitive data. Indeed, the Galaxy Docker project that we developed has been successfully implemented in this way at the University of Oslo and others, where it is used in a context with sensitive data.

### Collaboration with the de.NBI centres and other projects

The RBC works in close collaboration with the various de.NBI service centres on a number of different projects. For example, RBC works closely together with the Center for Integrative Bioinformatics (CiBi) to establish a common description language for command line tools that would permit automatic integration of these tools within Galaxy. The use of Docker, which is also being tested as a potential solution for the virtualisation of all de.NBI tools, is coordinated in collaboration with the Heidelberg Center for Human Bioinformatics (HD-HuB) and the de.NBI database division. There is also an ongoing cooperation with database division to investigate integration of the SILVA database and BacDive, which prepares rRNA data, within Galaxy.

Additionally, there are collaborations with the ELIXIR project, which is responsible for uniting Europe's leading life sciences organisations. The aim here is to develop common management strategies and strategies to protect the huge volumes of data pro-

duced by the publicly funded projects each day. RBC is involved in this process and promotes collaboration through participation in ELIXIR project events and by inviting ELIXIR partners to RBC events. Collaborations with the BioJS community and the Jupyter project allow them to be used directly through Galaxy and promote the exploratory character of the RNA workbench.

The long-term scientific aim of the RBC is to achieve a comprehensive examination of gene regulation with respect to RNA. This knowledge will help to improve understanding of the influence of RNA-based regulation on the subsequent regulatory processes of gene regulation, e.g. DNA methylation (epigenetics), transcription and translation. At the moment, the restricted availability of bioinformatics tools continues to impede findings relating to the role of the structure of non-coding RNAs. A great deal of research is still required with particular regard to the structure of mRNAs, which is influenced by the interplay between various interactions with RNA-binding proteins. In the services that it offers and in the development of the RNA workbench, RBC is making an important contribution towards the comprehensive, integrative analysis of RNA-based gene regulation.

---

### Research project profile:

#### Name of the project:

RBC – The RNA Bioinformatics Center:  
Prof. Dr. Rolf Backofen (RBC Coordinator)  
Prof. Dr. Peter F. Stadler  
Prof. Dr. Uwe Ohler  
Prof. Dr. Nikolaus Rajewsky

#### Members of the RNA Bioinformatics Center:

Dr. Björn Grüning, Dr. Anika Erxleben, Dr. Torsten Houwaart, Dr. Altuna Akalin, Dr. Bora Uya, Dr. Dilmurat Yusuf, Dr. Sebastian Will, Prof. Dr. Nikolaus Rajewsky, Prof. Dr. Uwe Ohler, Prof. Dr. Peter F. Stadler, Prof. Dr. Rolf Backofen

## RBC workshop program in 2016:

### Freiburg, Germany:

RBC Hackathon – January 2016  
Galaxy HTS - Data Analysis Workshop I – February 2016  
Galaxy HTS - Data Analysis Workshop II – September 2016  
Genome-Annotation Workshop – June 2016

### Berlin, Germany:

Computational Genomics Workshop – February 2016  
Galaxy Workshop – May 2016  
Computational Genomics in Precision Medicine Workshop  
– September 2016

### Leipzig, Germany:

Summer School with Jan Gorodkin (ELIXIR)





Figure 2: The RBC Team (from left to right: Dilmurat Yusuf, Sebastian Will, Björn Grüning, Anika Erxleben, Torsten Houwaart, Rolf Backofen, Peter F. Stadler, Uwe Ohler, Altuna Akalin, Bora Uya) (Photo: Lukas Jelonek).

## References:

- <sup>1</sup> S. Gerstberger, M. Hafner, and T. Tuschl. A census of human RNA-binding proteins. *Nature Reviews Genetics* 2014; 15: 829–845. doi:10.1038/nrg3813
- <sup>2</sup> Goecks, J, Nekrutenko, A, Taylor, J and The Galaxy Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* 2010 Aug 25;11(8):R86. doi:10.1186/gb-2010-11-8-r86
- <sup>3</sup> Christian Otto, Peter F. Stadler and Steve Hoffmann: Lacking alignments? The next-generation sequencing mapper segemehl revisited. *Bioinformatics* 2014 March 13; 30 doi: 10.1093/bioinformatics/btu146
- <sup>4</sup> Corcoran DL, Georgiev S, Mukherjee N, Gottwein E, Skalsky RL, Keene JD, Ohler U.: PARalyzer: definition of RNA binding sites from PAR-CLIP short-read sequence data. *Genome Biol.* 2011, 12:R79 doi:10.1186/gb-2011-12-8-r79
- <sup>5</sup> Blin K, Dieterich C, Wurmus R, Rajewsky N, Landthaler M, Akalin: DoRiNA 2.0-upgrading the doRiNA database of RNA interactions in post-transcriptional regulation. *Nucleic Acids Res.* 2015 Jan; 43(Database issue): D160-7. doi: 10.1093/nar/gku1180. Epub 2014 Nov 21
- <sup>6</sup> Daniel Maticzka, Sita J Lange, Fabrizio Costa and Rolf Backofen: GraphProt: modeling binding preferences of RNA-binding proteins. *Genome Biology* 2014, 15:R17 doi:10.1186/gb-2014-15-1-r17

## Contact:



### **Prof. Dr. Rolf Backofen**

RBC Coordinator  
 Department of Bioinformatics  
 Institute for Informatics  
 Albert-Ludwigs-University Freiburg  
 Georges-Köhler-Allee 106,  
 79110 Freiburg, Germany  
 backofen@informatik.uni-freiburg.de  
[www.bioinf.uni-freiburg.de](http://www.bioinf.uni-freiburg.de)



### **Dr. Björn Grüning**

Head of the Freiburg Galaxy Team  
 Department of Bioinformatics  
 Institute for Informatics  
 Albert-Ludwigs-University Freiburg  
 Georges-Köhler-Allee 106,  
 79110 Freiburg, Germany  
 gruening@informatik.uni-freiburg.de  
 galaxy@informatik.uni-freiburg.de  
<http://galaxy.uni-freiburg.de/>

# BRENDA

## From a database to a centre of excellence

by Dietmar Schomburg and Ida Schomburg

Enzymes are essential to almost all processes of life and vital to industrial biotechnology or medical diagnostics. 30-40% of all genes encode enzymes. They accelerate chemical reactions by up to 16 orders of magnitude, allow for precisely coordinated metabolic pathways within cells and are indispensable when it comes to defence against pathogens and other processes. Many of them are highly specific, while others are less so. The functions of enzymes are dependent on many characteristics, such as their sequence, three-dimensional structure, stability and their interactions with other molecules.

Understanding the role of enzymes in disease processes or their suitability for biotechnological applications requires the combination of many pieces of data, some of them very heterogeneous.

Only a flexible database, equipped with modern bioinformatics tools, is capable of doing this (Chang *et al.*, 2015). BRENDA has

developed from nothing more than a database into a centre of excellence for information on enzymes, where, in addition to database maintenance, bioinformatics tools are developed on a broader basis (text- and data-mining, genome annotation, data integration, functional statistics, user interface optimisation, data visualisation, machine learning, etc.).

### 27 years of financing for a biochemical infrastructure

The BRENDA enzyme database was created 27 years ago as a scientific infrastructure at the former Society for Biotechnical Research (GBF), now the Helmholtz Centre for Infection Research in Braunschweig. Consistent development, modifications in response to changing scientific requirements, extensive data integration and continuous updating have turned it into the world's largest resource for information on enzymes. Over 80,000 scientists use this database each month. The heart of the database consists of around three million hand-annotated, individual values from primary sources relating to enzyme functions, their specificity, stability, structure and incidence. There are a number of parameters

Figure 1: Data Categories in BRENDA

Nomenclature	Reaction & Specificity	Functional Parameters
Enzyme Names EC-Number	Pathway Catalysed Reaction Reaction Type	$K_M$ Value $k_{cat}/K_M$ Value $K_I$ Value
Organism-related information	Natural Substrates & Products Substrates & Products	$IC_{50}$ Value pI Value Turnover Number
Organism Source Tissue Localization	Inhibitors Cofactors Metals/Ions	Specific Activity pH Optimum pH Range Temperature Optimum Temperature Range
Isolation & Preparation	Activating Compounds Ligands  Biochemicals Reactions	<b>Kinetic ENzyme DA</b> ta
Purification Cloned Expression Renatured Crystallization	Enzyme Structure	Disease, Engineering & Application
pH Stability Temperature Stability General Stability Organic Solvent Stability Oxidation Stability Storage Stability	Sequence 3D-Structure Molecular Weight Subunits Posttranslational Modification Protein-Specific Search	Disease/ Diagnostics Engineering Application
		References
		References

Source: Prof. Dr. Dietmar Schomburg

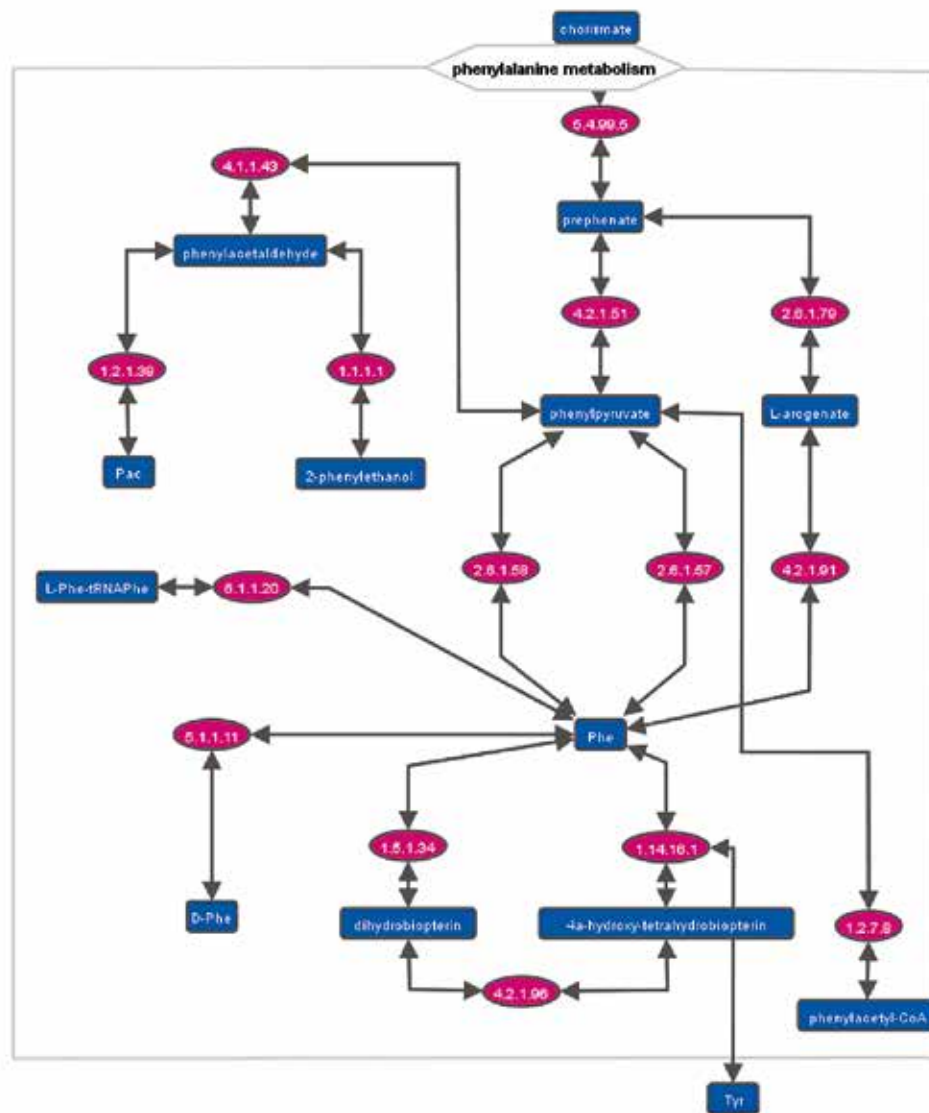


Figure 2: BRENDA Pathway of the Phenylalanine Metabolism (Source: Prof. Dr. Dietmar Schomburg).

available to characterise each one of the enzymes currently included in the database.

BRENDA has faced imminent financial collapse on a number of occasions. Full funding was only available in rare cases, so the project had to generate at least part of its funds from own sources. For instance, BRENDA received financing from GBF and was published in a book format between 1987 and 1996. For around one year, royalties from the sale of this book were the sole source of financing for database upkeep after Professor Dietmar Schomburg moved to the University of Cologne. Various forms of financing were raised between 2000 and 2012, most of them from the EU. In many cases, this involved collaborative projects with EMBL-EBI. It seemed as if the end had finally come around 2012, when the EU pulled the plug on financing scientific infrastructure. Thankfully, the German Federal Ministry of Education and Research (BMBF) and the state of Niedersachsen took over project financing. Since 2015, BRENDA is a part of the German Network for Bioinformatics Infrastructure (de.NBI). Since 2001, around 20–25% of the required funding has been raised through the sale of in-house licences to the industry.

### Data from many sources

BRENDA's data takes its structure from the enzyme classification system defined by the International Union of Biochemistry and Molecular Biology. It assigns all enzymes to six main categories. There are subcategories to define additional details such as cofactors, the substrates involved in the reaction and the type of chemical bond modified by the reaction. There are currently around 6,500 different enzyme categories.

Information on enzymes is not filed away in central repositories, but is instead buried in an unstructured form amidst millions of primary source publications. It is necessary to review and assess reference material to evaluate its significance for the characterisation of a particular enzyme, in order to make the most of this treasure. Then, the data contained in the article has to be manually extracted and carefully checked by computer programmes and scientists before it is finally integrated within the database. Only then can it be linked to other data such as sequences, protein structures or the NCBI taxonomy tree. The BRENDA team is extremely concerned to maintain high quality standards.

## BRENDA in the German Network for Bioinformatics Infrastructure

Within the German Network for Bioinformatics Infrastructure (de.NBI), BRENDA covers the large field of enzymes. Prof. Dr. Dietmar Schomburg at the Technical University of Braunschweig is responsible for its development. Linked with other de.NBI databases, it now provides a network of resources that offers users easy access to an extremely broad collection of interconnected data.

Members of the BRENDA team work in two de.NBI “special interest groups”. They run courses for the scientific community and centre employees and ensure the availability of BRENDA’s comprehensive and high-quality data for a range of scientific applications.

However, full analysis of the approximately 2.5 million relevant specialist biomedical articles is not possible. In order to ensure that BRENDA remains up-to-date, at least 5,000–10,000 new publications must be analysed each year.

The hand-annotated data is supplemented by information generated using text-mining methods. This includes medical aspects, kinetic data, the incidence of enzymes within organisms or their activity within organs (Schomburg *et al.*, 2013). Computer-assisted prediction of other data is outlined in the following section. This method produces a complex database containing over 150 million individual pieces of data displaying a variety of structures (text, numerical data, ontologies, graphical objects, networks).

### Computer-assisted prediction

A range of data is calculated using bioinformatics methods and then integrated into BRENDA. This includes the analysis of the connection between enzyme function or dysfunction and disease or the use of enzymes in the diagnosis and treatment of diseases.

To do this, machine learning methods are applied in order to combine the initial text mining of the abstract with a three-group classification system (causal interaction, treatment or diagnosis) of the paper itself (Söhngen *et al.*, 2011).

A method was developed for sequence-based prediction of enzyme function, which integrates the predictions of the most important genome-annotation websites with its own calculations to increase the accuracy of predictions.

A third area relates to the localisation of the enzyme within the cell. For example, membrane tethering and the number of transmembrane helices are predicted for enzymes based on their genetic sequence.

### Heterogeneous data require a diverse range of search options

A large number of search options have been developed to provide users with flexible access to the data. The simple entry of a key word, e.g. the name of the enzyme or metabolite, is sufficient to obtain an initial overview. It is also possible to conduct targeted searches for combinations of enzymes, for instance those that convert specific substrates or are stable at defined temperatures or pH values. Identifiers such as the EC number (enzyme classification) or reference numbers for protein structures can also be used.

BRENDA contains structures in addition to text and numerical data. Special “query engines” have been developed in order to browse these structures. Metabolites and other small molecules can be sketched using a simple structure editor, enabling searching based on their structure.

Besides ligands, metabolic pathways also possess a structure. BRENDA gives users the choice to display these pathways and to access corresponding information (see figure 2).

There are tree diagrams showing organism taxonomy or ontologies. Users can move along the tree and find relevant information displayed for each level.

The BRENDA Tissue Ontology (BTO) is a particular feature developed specifically for BRENDA (Gremse *et al.*, 2011). It represents a hierarchical classification of organs, body parts and tissue structures, each accompanied by a definition and literature reference. The BTO comprises terms from the animal and plant kingdoms.



## Data visualisation – essential for complex data structures

Search results are displayed in a tabular form. In this case, the primary unit of information is the EC number with its associated chemical reaction, organism of origin and literature reference. There is an option to conduct statistical analyses of results tables.

Within the results tables, links lead to molecule and protein structures, sequences, ontologies and tools for performing further analyses. Enzyme structures can be displayed in a freely rotatable, three-dimensional and interactive format with their key functional areas labelled. Structure diagrams show the catalysed reaction or the structures of inhibitors and activators.

The tremendous quantity of data stored in BRENDA makes it difficult to obtain an intuitive overview of an enzyme's characteristics and significance. Word maps were developed to assist here. They display the terms associated with a particular enzyme in relevant literature. Figure 3 illustrates this principle for the HIV protease and the typical biotechnological enzyme glucose oxidase, respectively.

### References:

- Chang A., Schomburg I., Placzek S., Jeske L., Ulbrich M., Xiao M., Sensen C.W., and Schomburg D. (2015) BRENDA in 2015: exciting developments in its 25th year of existence. *Nucleic Acids Res.* 43, D439-46
- Gremse M., Chang A., Schomburg I., Grote A., Scheer M., Ebeling C., and Schomburg D. (2011) The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources. *Nucleic Acids Res.*, 39, D507-513
- Schomburg I., Chang A., Placzek S., Söhngen C., Rother M., Lang M., Munaretto C., Ulas S., Stelzer M., Grote A. Scheer M., and Schomburg D. (2013) BRENDA in 2013: integrated reactions, kinetic data, enzyme function data, improved disease classification: new options and contents in BRENDA. *Nucleic Acids Res.* 41, D764-772

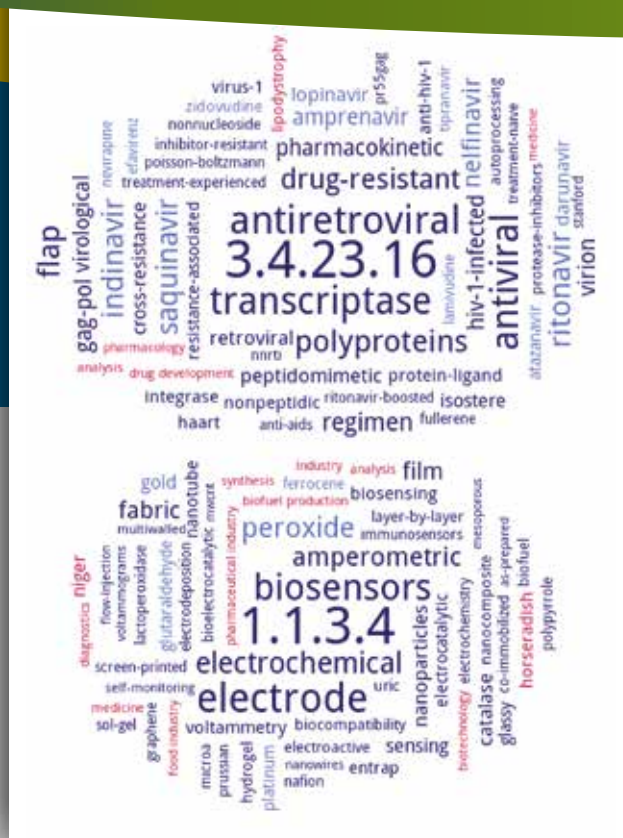


Figure 3: Word Maps of HIV Protease and Glucose Oxidase (Source: Prof. Dr. Dietmar Schomburg)

Söhngen C., Chang A., and Schomburg D. (2011) Development of a classification scheme for disease-related enzyme information. *BMC Bioinformatics*, 12, 329

### Contact:



**Prof. Dr. Dietmar Schomburg**  
Department of Bioinformatics and Biochemistry  
Institute for Biochemistry, Biotechnology and Bioinformatics  
Technical University of Braunschweig  
Braunschweig, Germany  
d.schomburg@tu-bs.de



**Dr. Ida Schomburg**  
Enzymeta GmbH  
Erfstadt, Germany  
I.Schomburg@enzymeta.de

[www.brenda-enzymes.org](http://www.brenda-enzymes.org)

# NBI-SysBio – the de.NBI data management hub

## From specialised data silo to interconnected data

by Wolfgang Müller

The goal of the de.NBI data management hub is to support the systems biology cycle in one key aspect: data management. For years now, the two groups involved, led by Wolfgang Müller from HITSgGmbH and Dagmar Waltemath from the University of Rostock, have spent their time digging into the issues of storage and reuse of data, models and simulation experiments relating to systems biology. They use the data management system SEEK (developed as part of a long-term collaboration with Carole Goble's group, University of Manchester, coordinated by Goble), as well as SABIO-RK, the database containing the kinetic data for reactions, with added model management solutions. Another important aspect of NBI-SysBio is the development of standards for the reproducible storage of models and simulation experiments.

Within the German Network for Bioinformatics Infrastructure, de.NBI, the data management hub provides all users with a diverse range of data services for the efficient and reliable management of data. In SEEK, the collected data and results can be organised, prepared and correlated with project partners both in and outside of de.NBI.

### The challenge of managing data for systems biology

Systems biology is a collaborative process in which scientists with a variety of specialisations work together to collect information on biological systems, to add further data and to use the results as the basis to create or refine models. The models are then taken to produce hypotheses that are then tested within an experimental setting.

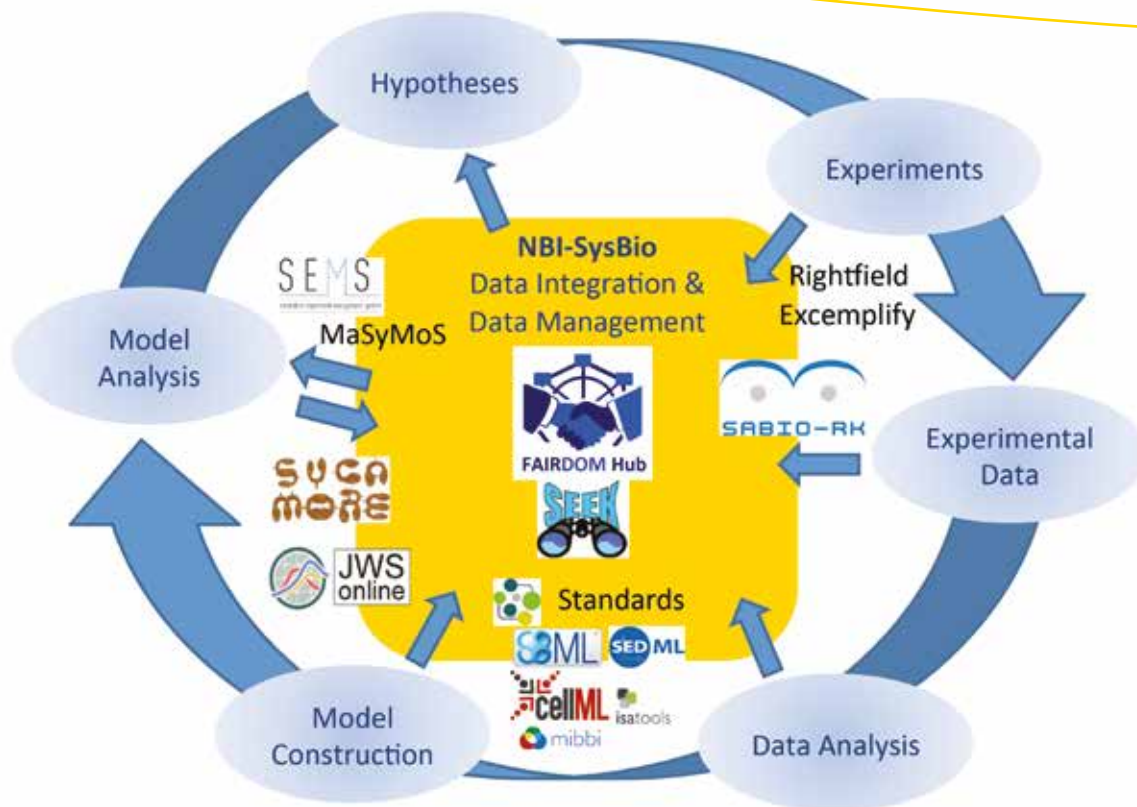
One can perceive this process as the process of enriching “Big Data” to create knowledge. Computer models are created from large experimental data sets that are then described and discussed in publications. What this produces is concentrated knowledge of processes unfolding within the biological system.

However, the term *Big Data* has acquired meaning beyond the scope of large individual data sets. It also relates to areas in which *many different* data sets and data formats require processing. And this is precisely the case in systems biology. The enormous variety of experimental methods has yielded an equally immense variety of data sets, data formats, guidelines and terminology relating to the storage of data and SOPs (standard operating procedures). [BioSharing.org](http://BioSharing.org), for instance, a project co-developed by the MIBBI and ISA teams, lists 621 relevant standards within biology and medicine.

In recent years, the diversity of biological issues and the increased quantity of available data have led to the distribution of data across a large number of specialised databases (NAR lists over 1,000 [Galperin, MY and Cochrane, GR, 2009]). This usually means that the data, models and SOPs created within a single project end up scattered. The multitude of data formats, standards and storage systems create a kind of thicket that has become virtually impenetrable to anyone apart from skilled data managers.

### What the data management hub offers within de.NBI

The data management hub seeks to untangle this thicket and to make it more useful and understandable for users. What's more, the scientists develop and maintain software, prepare specific data management services (Software as a Service, SaaS), sit on numerous standardisation committees and offer direct advice to users. They also organise training courses and even act as teachers in the individual coaching sessions.



**Figure 1:**

Through its data management services, NBI-SysBio supports collaboration with various project partners throughout the full cycle of systems biology. Experimental requirements (e.g. standard operating protocols), data or models can be catalogued in openSEEK (e.g. at <http://fairdomhub.org>) in a standardised way and linked across projects and data sources.

Additional tools such as Exemplify, Rightfield assist with the structuring of the data for later use. SABIO-RK is a specialized data source for the use by modellers and experimentalists.

The application of standards like SBML and SED-ML permits the reuse of models saved in openSEEK. Furthermore, other platforms and libraries, e.g. JWS Online, Sycamore or MaSyMos, provide users with support when locating, simulating and reproducing the simulation studies saved in openSEEK (Graph: NBI-SysBio Team).

**openSEEK Software** and its largest publicly accessible instance, **FAIRDOM Hub**, address the problem presented by the fact that structured *repositories* are usually organised according to methods or data types, but not according to projects. In openSEEK, users have a central storage facility and catalogue for their data, models and SOPs. They can also see how the project partners' data, models and SOPs interrelate, without possessing specialist knowledge of the underlying data sources. The use of standardised templates further improves the interoperability of the stored data, and makes entire studies reusable. Work conducted by the de.NBI data management hub will bring further improvements to the increasingly important search for specific models that this produces, also the versioning and traceability of data and models in openSEEK.

The standardised storage of simulation experiments within openSEEK (in **SED-ML** format) will enhance the reproducibility of models. We are continuing to develop an existing storage concept (MASYMOS), which links the simulation models and their data with their associated simulation experiments and model parameterisations, and prepares suitable combinations of models and experiments. This sounds simple, but in this field, small omissions rapidly become stumbling blocks when it comes to the reproducibility of models.

**SABIO-RK** is a database, curated by experts, which contains the kinetic data for reactions. This data is of particular importance for systems biology's quantitative models. In SABIO-RK, users can search for reactions either by means of free text or a search mask for advanced users. They receive their results as a list of "data sheets" on reactions. Prepared in this way, it is far easier to understand the information than it would be in a full paper. Data sheets contain information on kinetic data, for example on its importance for quantitative simulations along with links to other information, e.g. each protein identifier is linked to UniProt. The data can be exported from the SABIO "data sheet" as an SBML-encoded model fragment. These fragments can be expanded, combined, parameterised and finally merged into whole models.

### Services & training

**A variety of services and training courses are provided as part of the de.NBI network:**

- 1. User support on site or by telephone:** The specialists involved in data management will happily pay you a visit, ask plenty of questions about the project and develop a data management plan based on the answers you provide: What is the ideal data flow through the project? Where is there uncertainty or the need for internal standardisation or new features? Could the use of additional tools and services from the de.NBI and FAIRDOM portfolio or the wider international community make work easier? Which needs have to be prioritised?
- 2. Development collaboration:** Collaboration partners can influence our development plans or collaborate with us on the implementation of precise features.
- 3. Curating:** <http://sabiork.h-its.org> permits users to define requirements for curating: Which publications, pathway or experimental data should be added to SABIO-RK in a curated form?
- 4. Training:** There are currently a numerous different workshops and training sessions on offer for de.NBI partners. They range from the organisation of an *ICSB tutorial* in 2015 to running further training on the topic of modelling and model management at the de.NBI *Late Summer School*

and the co-organisation of an ERASysAPP workshop on *reproducible and citable data and models*. We also run open webinars on the topics of SABIO-RK data and model management and standardisation. On 30./31.5.2016 we celebrate a decade of SABIO-RK with a workshop.

**You can get in touch via the de.NBI homepage or via [nbi-sysbio@denbi.de](mailto:nbi-sysbio@denbi.de) if you want to know more about our service and/or consider a collaboration.**

---

### References:

- Galperin, MY and Cochrane, GR (2009). Nucleic Acids Research annual Database Issue and the NAR online Molecular Biology Database Collection in 2009. Nucleic Acids Res. 2009 Jan;37(Database issue):D1-4. doi: 10.1093/nar/gkn942.
- Wolstencroft K, Owen S, du Preez F, Krebs O, Mueller W, Goble CA, Snoep JL (2011). The SEEK: A Platform for Sharing Data and Models in Systems Biology, Methods in Enzymology, Volume 500: 629-655. doi: 10.1016/B978-0-12-385118-5.00029-3.

---

### Contact:



**Priv.-Doz. Dr. Wolfgang Müller**  
HITS gGmbH  
Heidelberg Institute for Theoretical Studies  
Heidelberg, Germany  
[wolfgang.mueller@h-its.org](mailto:wolfgang.mueller@h-its.org)

**For questions regarding our services and related de.NBI services you can reach us at:** [nbi-sysbio@denbi.de](mailto:nbi-sysbio@denbi.de)

**FAIRDOM Hub:** Management of systems biology data as a part of the de.NBI and the transnational FAIRDOM project  
<http://fairdomhub.org>

**SABIO-RK:** Curated database for systems biology data  
<http://sabiork.h-its.org>



# the development of software solutions for microbial bioinformatics

## Institution Portrait

### Justus-Liebig University Giessen Bioinformatics Center

by Alexander Goesmann

The Justus-Liebig University's Institute for Systems Biology in Giessen, under the leadership of Professor Alexander Goesmann, forms part of the de.NBI microbial bioinformatics service centre, in which Bielefeld University participates along with the University of Giessen. The service centre focuses on preparing and developing software solutions for the analysis of genetic material of microorganisms relevant to the fields of medicine and biotechnology. The service centre also provides a comprehensive hardware infrastructure, which is being gradually expanded as part of the de.NBI's funding programme.

When Alexander Goesmann was appointed in 2013, the new institute at Justus-Liebig University (JLU) was newly equipped with an excellent IT infrastructure, financed by the state of Hesse and the JLU. With the help of funding from de.NBI since

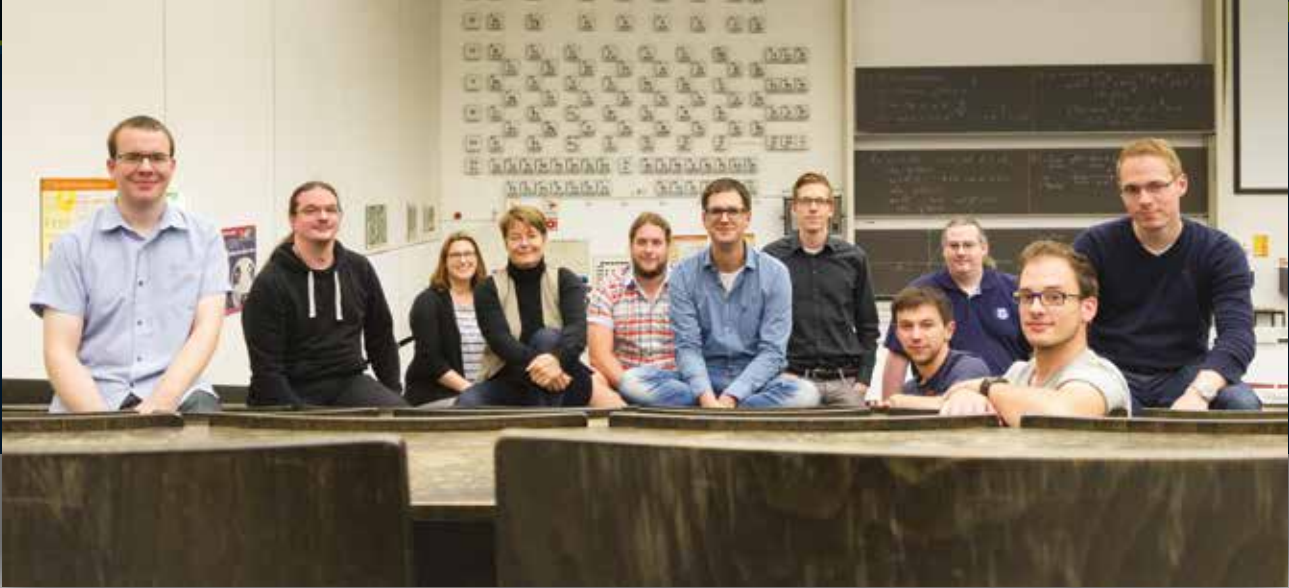
March 2015, this infrastructure is now being expanded even further and is being adapted to current needs. In addition to a computer cluster with approximately 1,000 processor cores, special hardware is also made available to collaborative partners, including a SMP server with 160 processor cores and two terabytes of working memory as well as a TimeLogic DeCypher system for accelerating sequence homology searches. This energy-efficient system allows multiple hundreds of thousands of sequences to be compared against large databases every day. Table 1 contains a list of investments planned within de.NBI.

The Department of Systems Biology possesses storage and computing capacities which are optimally suited to the specific requirements of bioinformatics. Over the coming months, some of the available resources will also be made available via a cloud-computing infrastructure, giving users an even greater degree of flexibility to develop their own tools.

**Table 1: Progressive extension of storage and computing capacity**

As part of the de.NBI funding programme, storage and processing capacities at JLU Giessen are being gradually expanded. The table shows the systems which are to be provided, listed next to their respective intended use.

<b>2015</b>	<ul style="list-style-type: none"><li>• TimeLogic DeCypher system J1</li><li>• Database server</li></ul>	High-throughput sequence comparisons Hosting of database applications
<b>2016</b>	<ul style="list-style-type: none"><li>• Expansion of storage capacity (360 TB)</li><li>• Expansion of processing capacity (480 Cores)</li></ul>	Storage of project data General data analysis & cloud computing
<b>2017</b>	<ul style="list-style-type: none"><li>• TimeLogic DeCypher system J1</li><li>• Web server</li></ul>	High-throughput sequence comparisons Hosting of online applications
<b>2018</b>	<ul style="list-style-type: none"><li>• Expansion of storage capacity (300 TB)</li><li>• Expansion of processing capacity (320 Cores)</li></ul>	Storage of project data General data analysis & cloud computing
<b>2019</b>	<ul style="list-style-type: none"><li>• Renewal of SMP-Server (2 TB RAM)</li></ul>	Storage-intensive applications such as genome assembly



**Figure 1: Group of Professor Goesmann**

The colleagues in Prof. Goesmann's group develop software solutions to analyse the genetic material of organisms relevant to the fields of medicine and biotechnology (Photo: Lukas Jelonek).

However, as part of the German Network for Bioinformatics Infrastructure, BiGi also provides interested users with numerous software packages as well as regularly updated public data collections, which are required for the bioinformatic analysis of genome and post-genome data. This involves analysis of both the sequence data for individual organisms and metagenomic data. In addition, BiGi provides users with support in using the software and holds special training courses to teach people how to apply these tools.

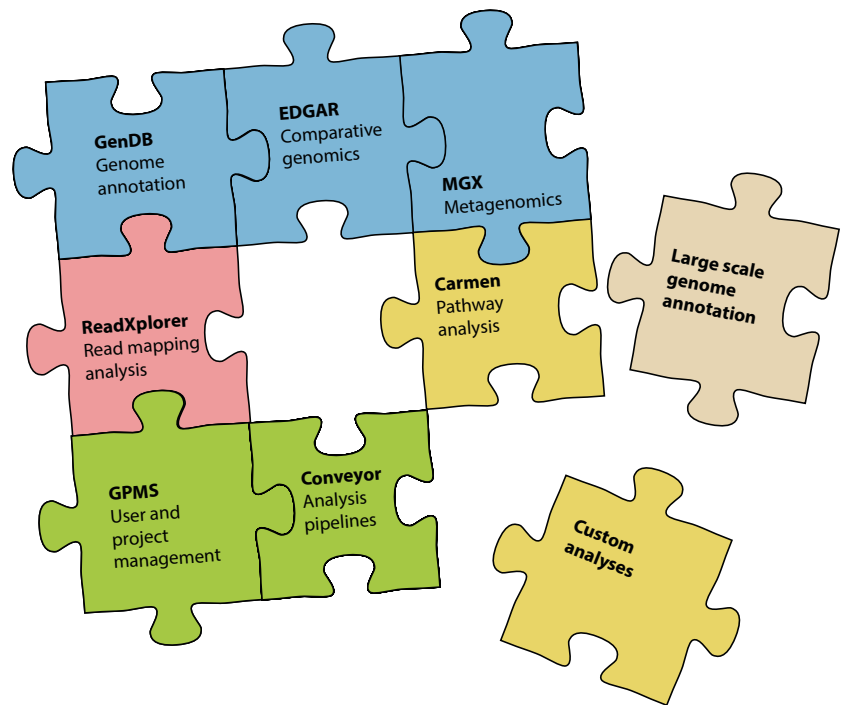
The GenDB genome annotation system is among the working group's best-known software developments (Meyer *et al.*, 2003). GenDB is a web-based platform for automatic and manual annotations used by research teams across the globe as part of over 400 projects for the community-wide validation, improvement and long-term management of annotation data for prokaryotic genomes.

The EDGAR platform (Blom *et al.*, 2009) can be used in the field of comparative genomics. It allows comparison of the genetic make-up of different bacteria in a user-friendly way. This software allows various pieces of information on genome groups to be easily extracted, their common features and differences to be identified and family relationships to be analysed from a phylogenetic perspective. With over 15,000 genome analyses from across 500 projects, EDGAR represents one of the most popular tools in the field of comparative genomics. Password-protected projects can be created upon request for data sets that are currently unpublished, as is the case for all software applications presented here. More than 45 collaborative partners have made use of this facility since de.NBI began its funding.

The freely available software ReadXplorer (Hilker *et al.*, 2014) is offered to assist with the analysis and visualisation of the data in the case of genome or transcript data from a sequence that is available as non-assembled sequence fragments (reads) and which must first be mapped onto a reference genome. The software combines an array of analysis methods and corresponding visualisation solutions to ensure that analysis of the data is as user-friendly as possible. The detailed classification of all reads, the detection of genomic variants, the identification and validation of operons or transcription starts, as well as analyses of differential gene expression, are among the most important automatic functions.

In recent years, easy access to sequence data produced by high-throughput technologies has opened up another topic area in the form of metagenomics, which focuses on recording and analysing highly complex, microbial communities. Goesmann's working group provides the MGX software as a user-friendly and highly flexible framework for the analysis of such data. Drawing on standardised Conveyor workflows (Linke *et al.*, 2011), MGX allows metagenomic data sets, e.g. those produced in connection with environmental samples or clinical isolates, to be processed in an efficient manner. Comprehensive visualisation and statistical solutions aid in the evaluation of taxonomical and functional analyses and allow easy interpretation of the obtained results.

The first de.NBI summer school at JLU Giessen was one of this past year's outstanding events. Over the course of five days, 20 international participants learned about various aspects of genomic data analysis. Theoretical knowledge acquired in the fields of basic quality control, the assembly of sequence data, genome annotation and comparative analyses, was put into



**Figure 2:** The figure shows different software developments and their area of application within the de.NBI network (Source: Lukas Jelonek).

practice using the participants' own data sets. International interest in the de.NBI summer school was as high as expected, also thanks to exceptionally high-class, external speakers such as Gene Myers, Tatiana Tatusova, Paul Kersey, Jan Gorodkin and Ursula Kummer.

Besides numerous improvements to the aforementioned software platforms, current work involves the ongoing automation of analysis workflows to enable the targeted genome analysis in the shortest possible time. An ambitious goal in this respect is to acquire the capacity to analyse 1,000 genomes per day. On the one hand, it is equally vital to allow users to specify all data and the associated working steps in the simplest possible manner. On the other hand, a modular set of analysis tools and requirement-oriented visualisation solutions must be developed to enable generation of tailor-made results reports.

Interested users can find additional information and access to available software, as well as the training courses currently on offer, at <http://bigi.computational.bio>.

**Initial queries may be sent to the following email address:**  
**[bigi@computational.bio](mailto:bigi@computational.bio)**

#### Authors:

Dr. Jochen Blom, Dr. Karina Brinkrolf, Dr. Rolf Hilker, Sebastian Jaenicke, Lukas Jelonek, Dr. Burkhard Linke, Oliver Rupp, Oliver Schwengers, Prof. Dr. Alexander Goesmann

#### References:

- Meyer, F., Goesmann, A., McHardy, A.C., Bartels, D., Bekel, T., Clausen, J., Kalinowski, J., Linke, B., Rupp, O., Giegerich, R. and Pühler, A. (2003) GenDB - an open source genome annotation system for prokaryote genomes. *Nucleic Acids Res.* 31(8):2187-95.
- Blom, J., Albaum, S., Doppmeier, D., Pühler, A., Vorhölter, F.-J., Zakrzewski, M. and Goesmann, A. (2009) EDGAR: a software framework for the comparative analysis of prokaryotic genomes. *BMC Bioinformatics* 10(1): 154, 2009.
- Hilker, R., Stadermann, K.B., Doppmeier, D., Kalinowski, J., Stoye, J., Straube, J., Winnebal, J. and Goesmann, A. (2014) ReadXplorer - Visualization and Analysis of Mapped Sequences. *Bioinformatics*, 30, 2247-2254.
- Linke, B., Giegerich, R. and Goesmann, A. (2011) Conveyor: a workflow engine for bioinformatic analyses. *Bioinformatics* 27(7), 903-911.

#### Contact:



#### Prof. Dr. Alexander Goesmann

Systems biology with a focus on Genomics, proteomics and transcriptomics  
Justus-Liebig-University Gießen  
Gießen, Germany  
[Alexander.Goesmann@Computational.Bio.Uni-Giessen.de](mailto:Alexander.Goesmann@Computational.Bio.Uni-Giessen.de)

<http://computational.bio>

# ELIXIR – keeping data flowing

## Interview with Niklas Blomberg

Since May 2013 Niklas Blomberg has been the Director of ELIXIR, a new research infrastructure on the ESFRI roadmap (European Strategy Forum for Research Infrastructures). The aim of ELIXIR is to connect individual national bioinformatics infrastructures across a Europe-wide network. At the start of 2016, 16 European countries were already members of the ELIXIR consortium.

*systembiologie.de: Dr. Blomberg, could you describe for us in a few sentences what ELIXIR is?*

**Dr. Niklas Blomberg:** ELIXIR is a Europe-wide research infrastructure connecting national bioinformatics resources, ranging from research laboratories to data centres. A vast range of services are made available to the entire scientific community. These include analytical tools for bioinformatics research and research using life sciences-related data and databases, as well as solutions for the secure exchange of data. All of these instruments are designed to help coordinate individual data collections, to ensure their quality, and to enable storage of large quantities of life-science data. Such a good network is essential – without it the flow of data would grind to a halt!

*Who is involved in ELIXIR?*

The ELIXIR Hub, ELIXIR's central control and coordination centre, is located in the United Kingdom – next door to EMBL-EBI, the European Molecular Biology Laboratory's European Bioinformatics Institute – in Hinxton, Cambridge. The ELIXIR Nodes are situated in the individual member countries and constitute the respective countries' leading institutes in the field of biosciences.

*How is ELIXIR financed?*

ELIXIR is predominantly financed through contributions by its members, the signatory countries of the ELIXIR Consortium

Agreement. For instance, the 2015 budget, which is calculated on the basis of a work programme adopted by the ELIXIR Board, amounted to around two million euros. Additional funding for specific projects is received from the European Commission. The amount of each Members' contribution is based on the economic performance of the respective country as reflected in its net national income (NNI). For example, for Germany, this equates to approximately EUR 900,000 in membership contributions for 2016. The contributions are used to support technical projects carried out by ELIXIR Nodes in member countries.

*To what extent can non-members of ELIXIR use its services? Or, to put it differently: what benefits and additional opportunities does becoming an ELIXIR member offer?*

Many of the services provided to the scientific community within ELIXIR are free and do not require ELIXIR membership. They are available to all scientists in Europe, regardless of whether their country is a member of ELIXIR or not.

However, one of the greatest challenges nowadays with respect to data-intensive research within the field of life sciences is the availability and provision of secure and effective networks for data exchange and analysis: this is not always free of charge. Only ELIXIR members have the opportunity to participate in the development of the infrastructure and to make suggestions as to the focus for future activities. As another example, members also have access to a range of specific training programmes offered by ELIXIR.

*What role does EMBL-EBI play as host to the ELIXIR Hub?*

ELIXIR uses EMBL's legal status as an international organisation, and the ELIXIR Hub is located on the Wellcome Genome Campus in Hinxton, alongside EMBL-EBI. EMBL-EBI is itself an ELIXIR Node. In this sense, ELIXIR's relationship to EMBL-EBI is similar to that with the ELIXIR member countries. EMBL-EBI





Director of ELIXIR, Niklas Blomberg, connects individual national bioinformatics infrastructures across Europe (Photo: Niklas Blomberg/ELIXIR).

possesses significant data resources, and the EMBL member countries contribute funds for their maintenance. The added bonus for EMBL-EBI in the context of ELIXIR, as for all Nodes, is the international networking of these resources with resources maintained in the respective ELIXIR member countries, and the resulting economies of scale.

*What are the advantages of this European infrastructure compared to, for example, the US-American infrastructure NCBI?*

The NCBI – the National Center for Biotechnology Information – is not an interconnected network but a monolith, similar to EMBL-EBI. Consequently it cannot be compared with an interconnected, international infrastructure like ELIXIR, which unites and coordinates national resources on an international level. ELIXIR is about creating links between existing structures. Of course we need EMBL-EBI as well as the NCBI. However, we also need links between EMBL-EBI and the many other institutions located in ELIXIR’s member countries, just as we need the link between ELIXIR and NCBI.

*Do you see synergies with other European infrastructure initiatives such as EUDAT2020, which hopes to make research data available across Europe, or the European supercomputing community PRACE?*

Yes, absolutely! Among the European e-infrastructures, GÉANT is perhaps the most significant for us, as a good flow of data is truly vital. Allow me to explain this by taking roaming as an example: At ELIXIR, we work intensively to build links between our organisation and other networks, such as for example edu-

cation roaming (eduroam). This requires little in the way of technology and many technical solutions are already well advanced. However, this still requires a large amount of coordination. The greatest challenge in this case relates to governance.

*A personal question to finish: which aspect of your job do you find particularly exciting?*

Before I came to ELIXIR I led AstraZeneca’s Computational Chemistry and Computational Biology team. It was an interesting and exciting time. With ELIXIR I am now doing work that I regard as truly important - if we succeed with ELIXIR’s mission of creating a stable public infrastructure for bioinformatics data, this will have far-reaching, positive implications for data-driven life science research and its applications across biology and to society and the economy.

*This interview was conducted by Marcus Garzón and Vera Grimm from Project Management Jülich.*

---

#### Contact:

**Dr. Niklas Blomberg**

ELIXIR Director

ELIXIR Hub

Wellcome Genome Campus

Hinxton, Cambridgeshire, CB10 1SD, UK

[niklas.blomberg@elixir-europe.org](mailto:niklas.blomberg@elixir-europe.org)

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

# Omics infrastructures for research and teaching

A concept by Leopoldina for a change in the field of life sciences

by Alfred Pühler

## Development of the Report on Tomorrow's Science

Leopoldina, the German National Academy of Sciences sees it as one of its main tasks to act in an advisory capacity to politicians and the public in scientific matters. This advice is given mainly by issuing independent opinions addressing either individual specialist disciplines or addressing the scientific system within Germany as a whole. Leopoldina has also produced a range of publications entitled “*Zukunftsreport Wissenschaft*” (“Report on Tomorrow's Science”) within the field of life sciences. The first Report on Tomorrow's Science refers to the paradigm shift in life sciences that recently took place on the basis of omics technologies. It analyses the challenges facing research and teaching and develops a concept describing how this change can be handled.

To produce the first Report on Tomorrow's Science, a working group made up of omics experts was put together, who, over the course of multiple working sessions, invited representatives from universities, non-university institutions and ministries to participate in expert discussions. Finally, this working group produced the first Report on Tomorrow's Science, focusing on the theme “Life sciences in transition, Challenges of omics technologies for Germany's infrastructures in research and teaching”. The first Report on Tomorrow's Science was approved by Leopoldina's board of directors in May 2014 following a detailed review by eight experts from Germany and abroad. It was presented during a press conference in September 2014, and finally introduced at a public event attended also by representatives from the world of politics (Fig. 1).

## Omics technologies produce paradigm shift in life science

It was the discovery of DNA as genetic material in particular that led to a rapid acceleration in the molecular understanding of cellular processes. This meant that individual genes, transcripts, proteins and metabolites became amenable to be intensively analysed from a molecular biological perspective. These isolated views of individual cellular components were not replaced by a holistic perspective until after the emergence of omics technologies. From then on, in the field of DNA sequencing in particular, it became possible to pursue technological developments which today enable the recording of all genes of a chosen organism and in doing so, render a holistic approach possible. RNA sequencing was also recently perfected, so the entirety of all transcribed genes can be included in the analysis. In addition to genomics and transcriptomics, the proteomics and metabolomics omics technologies have also been optimised, so that almost all of a cell's proteins and many of its metabolites can now be recorded in parallel. However, the development of all these omics technologies would be unachievable without the decisive contributions of user-oriented bioinformatics, as omics technologies produce large quantities of data that can only be analysed using specified bioinformatics tools. But that is in no way the only possible use for applied bioinformatics. It once again comes into play particularly with regard to the cellular processes involved in systems biology, which itself is currently undergoing development, and helps to identify the connection between cellular components from a systems biology-oriented perspective.



**Figure 1: Presentation of the Report on Tomorrow's Science at a public event in September 2014.**

All persons shown from left to right: D. Scheel (Halle), A. Pühler (Bielefeld), R. Kahmann (Marburg), M. Hecker (Greifswald) and R. Eils (Heidelberg)  
(Photo: David Ausserhofer/Leopoldina).

And thus the paradigm shift brought about by omics technologies is described in detail. It relates to the change from the analysis of individual cellular components to a holistic view of cellular processes. Omics technologies have clearly revolutionised basic research in life sciences and have since progressed as far as applied life sciences. Within medicine in particular, tailor-made diagnosis and therapy is offered which is based on biomarkers discovered through omics procedures. This development aims to establish personalised or individualised medicine. Omics procedures are also used in biotechnology to develop tailor-made bacterial production strains for industrial application. In recent years, this has involved the replacement of successfully practised mutation and selection procedures by rational strain development, which is primarily characterised by genome-based systems biology. However, omics procedures are also used in the field of plant research to develop robust and productive plants. Here the focus is placed on metabolomics, a discipline used to metabolites that could be used as metabolic markers in breeding programmes.

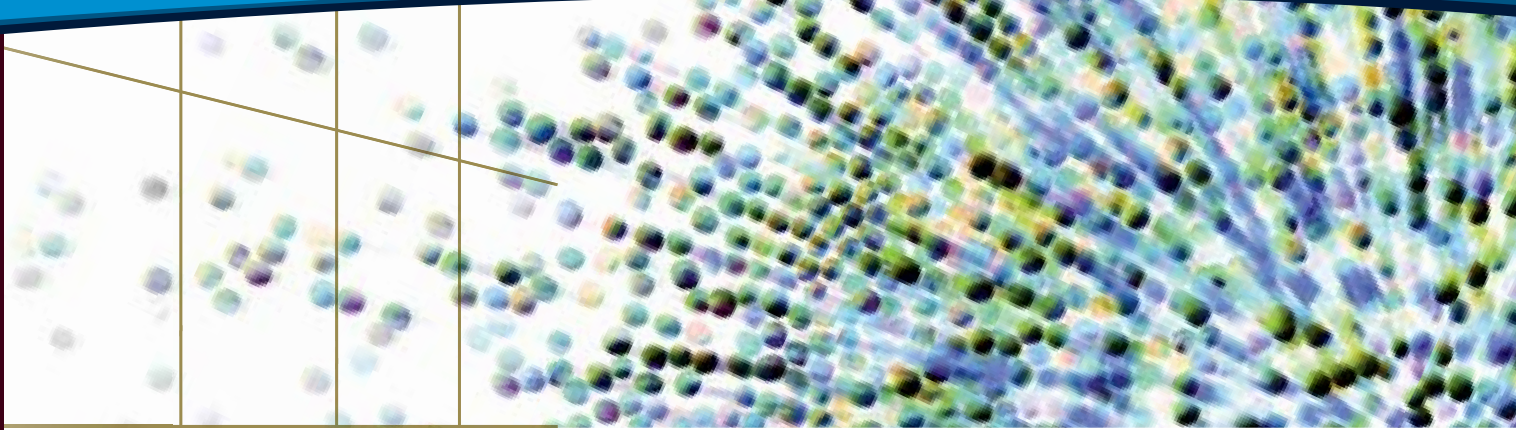
### Challenges of omics technologies with regard to research and teaching at German universities

This rapid development within the field of omics and bioinformatics technologies has led to the emergence of life sciences as a leading discipline of the 21st century. It is now a question of building upon the potential of these technologies and most importantly, integrating this into research and teaching. This calls on the universities first and foremost, as they are responsible for ensuring the education of students within the context of bachelor, master and doctoral programmes. If these

students are to receive an up-to-date education, then the German universities have an enormous amount of infrastructural tasks with which they will have to cope. They must establish the omics technologies in order to appropriately incorporate this modern branch of life sciences into teaching and research activities. However, this kind of infrastructural undertaking is currently very difficult to implement at most German universities, as financing is in the hands of regional governments, which do not possess sufficient disposable resources to do so. In this case, it is not a matter of providing start-up financing, but rather the sustainable financing of omics infrastructures, which require investment, consumables and human resources on a long-term basis. The sustainable financing of omics infrastructures is therefore the main problem at German universities. At the moment, university graduates clearly do not receive an up-to-date education with respect to the omics sector and are therefore not in the position to work on the issues at the forefront of research within the largest universities and industrial research facilities.

The working group set up by Leopoldina has, in particular, queried the state of current university education as it relates to the field of omics technologies. It has become apparent that teaching omics technologies are usually addressed by means of a purely theoretical approach, but are not taught by conducting experiments. Particularly serious is the fact that when it comes to the curriculum taught in overloaded degree courses in medicine, omics technologies are misrepresented. An especially large deficit was noted in the case of bioinformatics. Measured against future demand, notably few graduates receive





(Image rights: Sisters of Design/Leopoldina)

training in this specialised field. The expert discussions conducted revealed that students, but also teachers, did not have adequate interdisciplinary expertise: Bioinformaticians possess insufficient expertise with regard to the omics technologies sectors, while conversely, life scientists do not have enough insight into the capabilities of bioinformatics programmers. When establishing omics technologies at German universities, particular care should be taken to ensure that staff positions with attractive career prospects are created. Thought should also be given as to how permanent positions for experts can be created in addition to temporary postdoctoral positions.

### Scenarios for the organisation of an omics infrastructure in Germany

After having analysed the starting situation with regard to the state of omics technologies within the context of research and teaching at German universities, the working group of Leo-

poldina spoke out clearly in favour of pursuing significant expansion and reinforcement of omics-based research and training in life sciences. The suggestion was made to build an omics and IT infrastructure comprising a network of centres across Germany to provide cutting-edge omics technologies and to ensure that these technologies are also made available for research and teaching at other universities. A characteristic feature of these centres is that they should be home to working groups that not only use omics technologies, but also actively contribute to their further development.

Leopoldina's first Report of Tomorrow's Science developed additional models for how such a network, including a central coordination unit, could be realised. Two scenarios are mentioned, namely the DFG scenario and the Swiss scenario. The DFG scenario envisions the creation of a national omics infrastructure, supported by the DFG through the founding of a DFG

## Leopoldina's Report on Tomorrow's Science – a profile

In 2011, Leopoldina decided to examine the development of the scientific system in Germany using omics technologies as an example. To this end, a working group was put together, which in 2012 and 2013 organised the first expert discussions with relevant specialists from Germany and abroad. At the same time, a nationwide survey of life sciences and medical faculties was carried out in 2013. The working group compiled the Report on Tomorrow's Science until December 2013. Following international review it was approved by Leopoldina's board of directors in May 2014.

The Report on Tomorrow's Science focuses firstly on the paradigm shift within life sciences which has been triggered by technology. From there it proceeds to describe omics technologies and the support they have received up to now. It sheds light on structural challenges with respect to the organisation of life sciences, before developing scenarios for the establishment of a national omics and IT infrastructure for research and teaching. The Report on Tomorrow's Science can be accessed online at <http://www.leopoldina.org/en/policy-advice/standing-committees/report-on-tomorrows-science/>.

### Members of the Leopoldina working group:

**Prof. Dr. Rudolf Amann**, Max Planck Institute for Marine Microbiology, Bremen, Germany

**Prof. Dr. Roland Eils**, German Cancer Research Center, Heidelberg, Germany

**Prof. Dr. Michael Hecker**, Center for Functional Genomics, Greifswald, Germany

**Prof. Dr. Regine Kahmann**, Max Planck Institute for Terrestrial Microbiology, Marburg, Germany

**Prof. Dr. Alfred Pühler**, Center for Biotechnology, Bielefeld University, Germany

**Prof. Dr. Dierk Scheel**, Leibniz Institute of Plant Biochemistry, Halle (Saale), Germany



## Report on Tomorrow's Science



### Life sciences in transition

Challenges of omics technologies for Germany's infrastructures  
in research and teaching

Figure 2: Title page of the Report on Tomorrow's Science  
(Image rights: Sisters of Design/Leopoldina)

Senate Commission. A DFG panel for omics technologies could also be set up to deal with the financing of omics infrastructures. The Swiss scenario, however, is oriented towards the federal structure of the Swiss Institute of Bioinformatics (SIB). In this case, a legally and financially independent organisation was founded, which has been driving bioinformatics infrastructure and data analysis in Switzerland since 1998. This scenario could also serve as a model for Germany and direct the establishment of an omics and IT infrastructure through an independent organisation.

Since its publication in 2014, Leopoldina's first Report of Tomorrow's Science has attracted a great deal of interest from the daily and weekly press and has also resonated on a political level. The approaches which could be adopted in order to establish the required omics and IT infrastructure in Germany have been discussed by both BMBF and DFG in the time following the report's release. But given the urgency of this task, one cannot be satisfied by long discussions. On the other hand, and in parallel to these discussions, the establishment of the German Network for Bioinformatics Infrastructures has meant that part of Leopoldina's request has already been realised. As shown elsewhere in this special issue, the de.NBI was able to be set up in a relatively short space of time, thanks to funding by the BMBF. The establishment of the de.NBI demonstrated a feasible way of organising an omics infrastructure consisting of a network of omics centres and a coordination unit. One could even go a step further and consider using the existing de.NBI network as a hub around which an omics infrastructure could be formed. This reflection is prompted by the observation that the existing de.NBI service centres are almost all based in universities which have invested in omics technologies in recent years and are therefore well prepared to play a part in the yet to be established omics infrastructure. The proposal outlined here is certainly attractive and should be discussed soon by Leopoldina, BMBF and DFG.

### Literature:

<http://www.leopoldina.org/en/policy-advice/standing-committees/report-on-tomorrows-science/>

[http://www.leopoldina.org/nc/en/publications/detailview/?publication\[publication\]=604&cHash=83666e6055d29d2e79bb62b724aa4d7b](http://www.leopoldina.org/nc/en/publications/detailview/?publication[publication]=604&cHash=83666e6055d29d2e79bb62b724aa4d7b)

### Contact:



#### Prof. Dr. A. Pühler

Center for Biotechnology

Bielefeld University

Bielefeld, Germany

puehler@cebitec.uni-bielefeld.de

# a global initiative for cancer research

## The Pan-Cancer Analysis of Whole Genomes (PCAWG) project

by Jan O. Korbel<sup>1</sup>, Sergei Yakneen<sup>1</sup>, Sebastian M. Waszak<sup>1</sup>, Matthias Schlesner<sup>2</sup>, Roland Eils<sup>2</sup> and Fruzsina Molnár-Gábor<sup>3</sup>

The *Pan-Cancer Analysis of Whole Genomes* (PCAWG) project is an international research initiative organized within the International Cancer Genome Consortium (ICGC), which is paving the way for the widespread application of analysis methods using high-performance computing systems and cloud computing. Novel analytical tools enable the standardized examination of cancer genomes and associated data sets (e.g. transcriptome, DNA methylation and clinical data) in petabyte volumes, permitting cancer genomics and systems biology at unprecedented scale. This article presents the objectives of the PCAWG project, emphasizes its significance and potential for basic and translational research, and in this context stresses scientific, technical and normative challenges of cloud computing.

### “Big Data”: Opportunities for genome and disease research

Thanks to improvements in sequencing technology, the number of sequenced human genomes has increased dramatically in recent years. The volume of genomic data submitted to public archives is now well into the multi-petabyte range (1 petabyte corresponds to  $10^{15}$  bytes or the storage capacity of around 20,000 modern smartphones) – a previously inconceivable quantity in any academic field outside of the realm of physics. Within just five years, in the International Cancer Genome Consortium,

(ICGC; [www.icgc.org](http://www.icgc.org)), for instance, groups from 17 countries have amassed a data set in excess of two petabytes, enough data to roughly fill 500,000 DVDs (Stein *et al.*, 2015).

ICGC project focus on analyzing whole genomes from cancer patients, and its findings have already enhanced our understanding of molecular causes of specific forms of cancer. As an example, early-onset prostate cancer occurring in men younger than 50 years of age has been linked with specific DNA rearrangements affecting genes associated with the regulation of the male sex hormone androgen (Weischenfeldt *et al.*, 2013). Furthermore, in medulloblastoma, a pediatric brain tumor, DNA rearrangements affecting primarily non-coding regions of the genome can lead to oncogenic activation of genes by repositioning these genes in close proximity to *cis* regulatory elements (*i.e.*, enhancers) (Northcott *et al.*, 2014).

The cancer genome is characterized by substantial heterogeneity, *i.e.*, the abundance of cancer-specific DNA alterations differing from cell to cell within the tumor’s mass. And also when assessed across different patients’ cancer genomes show many differences. As an example, even clinically relevant gene mutations often appear with a frequency of less than 5% (occasionally even less than 1%) within tumor type cohorts (Lawrence *et al.*, 2014). Hence, in order to be able to develop personalized approaches to cancer therapy in the future, a large number of tumors need to be analyzed to uncover clinically relevant relationships. The comparative analyses of fully sequenced patient genomes within or between different tumor types is the starting point of the Pan-Cancer Analysis of Whole Genomes (PCAWG) project. The large number of approximately 2,800 cancer genomes\* (Figure 1), associated mo-

<sup>1</sup> European Molecular Biology Laboratory (EMBL), Heidelberg, Germany

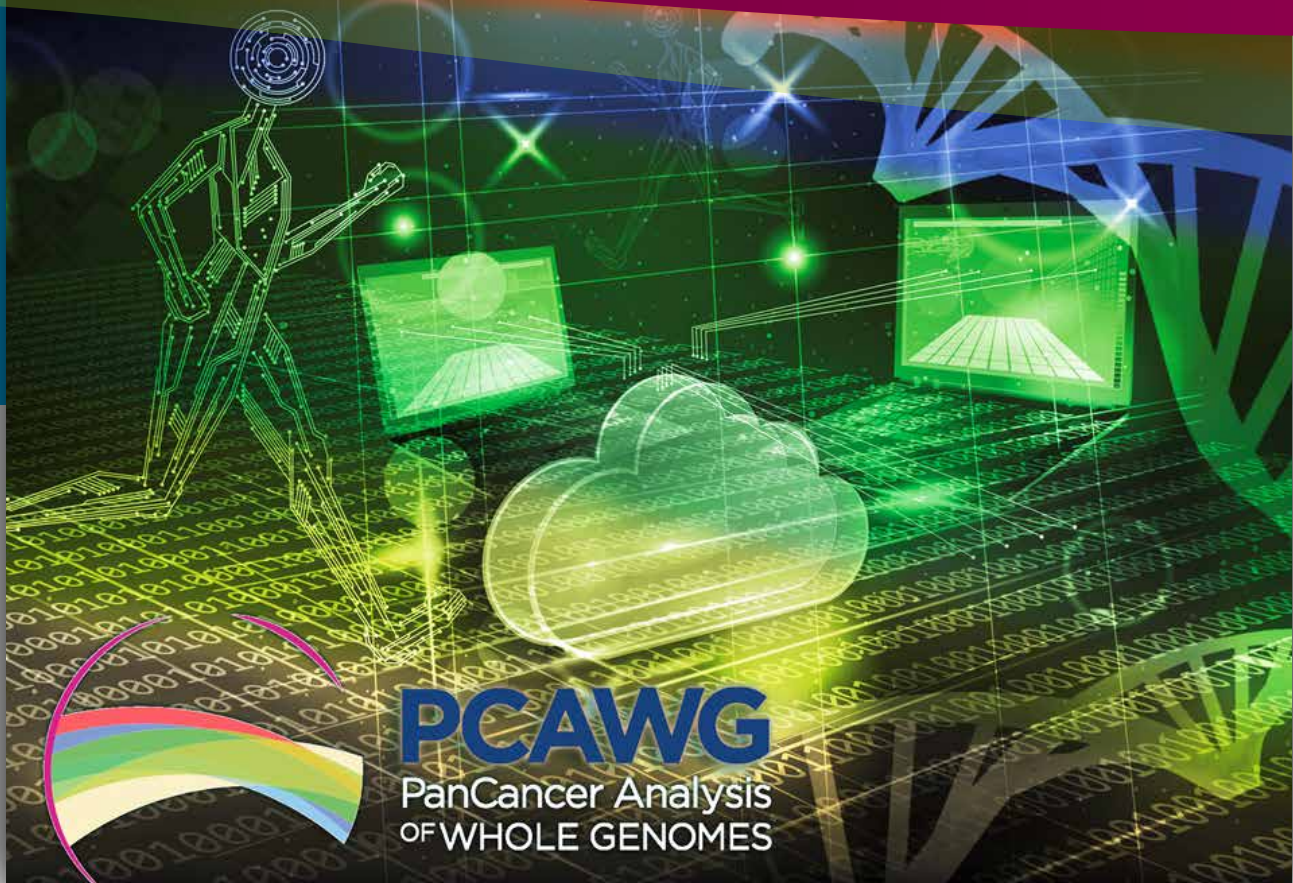
<sup>2</sup> German Cancer Research Center (DKFZ), Heidelberg, Germany

<sup>3</sup> Heidelberg Academy of Sciences and Humanities, Heidelberg, Germany

© Please address any questions to: [korbel@embl.de](mailto:korbel@embl.de)

\*The majority of these genomes have been contributed by the ICGC (68%). The remaining 32% have been provided by the US-American Cancer Genome Atlas (TCGA) project, which has largely focused on exome sequencing rather than whole genome sequencing.





The *Pan-Cancer Analysis of Whole Genomes* (PCAWG) project, an international initiative for research into common cancer mutation patterns, is paving the way for the widespread application of analysis methods using high-performance computing systems and cloud computing. Within the PCAWG, over 700 scientists across the globe compare the data of more than 2,800 cancer genomes of various tumor types. These investigations focus on the causes and consequences of somatic and germline variations in coding and non-coding areas of the genome (Copyright©EMBL; Design by P. Riedinger).

lecular information (e.g. gene expression and DNA methylation), and clinical data allows PCAWG to address a range of currently unresolved questions, such as pertaining to the abundance and relevance cancer-specific mutations and DNA rearrangements affecting *cis*-regulatory regions (Horn *et al.*, 2013; Northcott *et al.*, 2014). Larger-scale studies of the cancer genome could also provide new insights into cancer-causing viruses or bacteria. In addition, integrative analyses of genetic material, cancer mutations, and clinical data may allow us to develop a better understanding of the impact of germline genetic factors in the aetiology of cancer.

### Challenges for research and IT

Since several of these questions are already pursued in smaller scale studies, it would be entirely reasonable to ask what exactly the novelty of the PCAWG initiative entails is. Marked benefits and opportunities due to increased statistical power in large sample cohorts, but also challenges related to collaboratively sharing large data quantities at Petabyte-scale, are relevant to this question. The reality is, that using a current internet connection typically found in German universities, it would take months to transfer all of the published ICGC raw sequencing data from its repository into the researcher's local network, before any form of meaningful analysis could even start. In reality

this means that important studies making use of already published data, but using these data with a slightly different angle (e.g. focusing on mutations in non-coding regions) are hampered by limitations in technology, namely bandwidth but also computational processing power which becomes very relevant too when analyzing data at Petabyte-scale. Another factor to be considered in this regard are the hardware costs for data storage and processing: Due to the quantity of data produced by the ICGC alone, the costs associated with their storage within a professional network correspond to more than one million euros per year. Furthermore, researcher cannot restrict their analysis to already existing dataset on genomic mutations (which would require less storage space), as standardized pipelines for analyzing a cancer genome are not yet in place. The use of a range of individual analysis protocols for genome analyses result in incompatibilities of data due to missing standardization. These difficulties limit significantly the reuse of data across studies in spite of the globally available volume of genomic data sets.

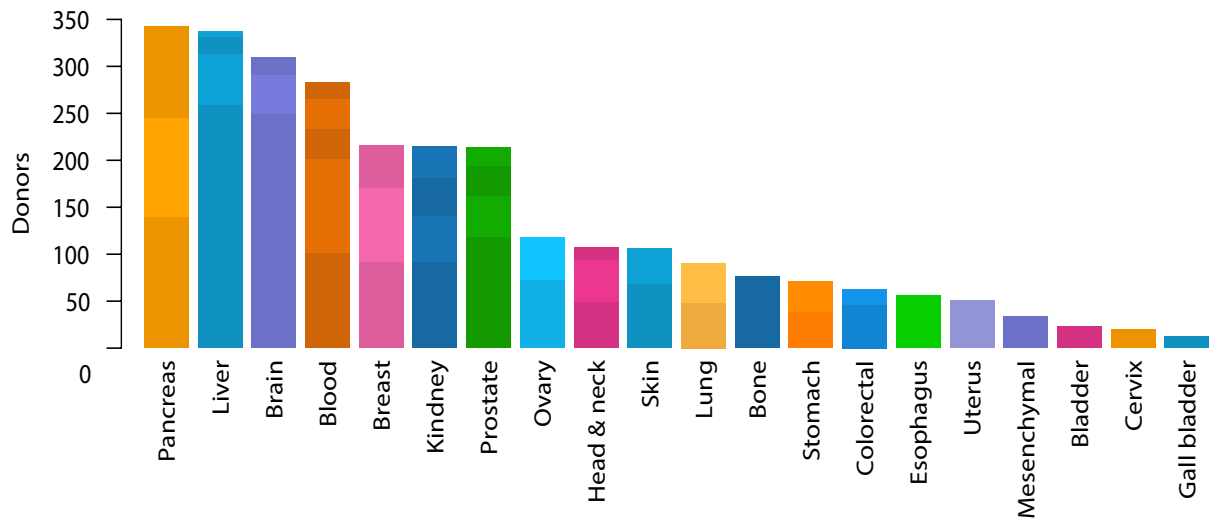


Figure 1: Breakdown of PCAWG cancer patients according to tumor type as well as ICGC projects (light/dark shading). (Source<sup>†</sup>: <https://dcc.icgc.org/pcawg/>)

In this regard, the PCAWG project (<https://dcc.icgc.org/pcawg>) has been established as a pilot project for the shared and collaborative analysis of cancer genome data and the standardization of analysis processes at an international level. The sequencing data available to PCAWG is analyzed in a uniform way using standardized pipelines for the alignment of DNA sequences as well as the detection of mutations. Processed data are immediately made available to all project partners, to allow all partners to benefit from the standardized data set.

The uniform processing of such large data quantities represents the key challenge. Altogether, seven of the project's own data repositories (in Heidelberg, London, Barcelona, as well as North America and Asia) have donated their compute capacity to perform this analysis, totaling 8,824 compute cores, 24 TB of RAM and 6 PB on hard disk storage. In addition, a commercial cloud computing provider, Amazon Web Services (AWS), has been used to analyze about 20% of the samples (Stein *et al.*, 2015). Other industrial partners (Annai Systems, Fujitsu, Intel, SAP and Seven Bridges Genomics, among others) supported individual analyses within the project, and also academic cloud providers have been involved in this project. Importantly, cloud services offer their users considerable storage and computing capacities on a pay-as-you-go basis. As cloud services are available via the Internet and thus many users may access the same hardware high data security standards must be applied.

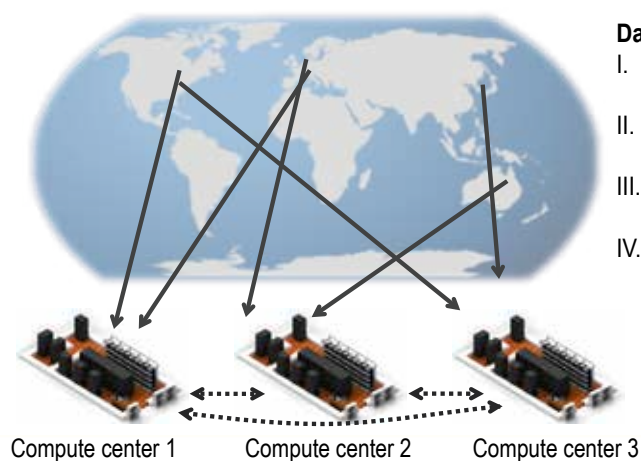
<sup>†</sup> Kindly acknowledging the PCAWG technical working group led by Lincoln Stein, as well as the PCAWG steering committee: Peter Campbell, Gad Getz, Jan Korbel, Lincoln Stein, Joshua Stuart

The processing of personal data in the cloud must not violate the data subjects' integrity and rights. The right of informational self-determination is the foundation of data protection and privacy law in numerous European countries. The use of commercial cloud providers for data analysis (and hence limited control by the user and the data subject) can result in uncertainties related to data security and the compliance with data protection regulation, *i. e.*, when using global cloud-services. As a consequence, cancer genome data from several European ICGC projects has not been analyzed with the Amazon Cloud, but instead in academic data processing centers. There are currently insufficient legal provisions to tackle new international data protection challenges such as these – both on an international level as well as in many European countries. The definition of personal data, the responsibilities for such data in clouds, transnational data processing and sharing need to be reconsidered regarding the specific requirements of cooperation in international cancer genome research as well as development of technologies enabling such collaboration allowing further advances in cancer research, technological development and personalized medicine.

### PCAWG: Standardized analysis of cancer genome data in the global context

But how do standardized analysis pipelines operate within PCAWG? A key technical challenge taken on within PCAWG in order to offer standardized genome analyses is the development of frameworks enabling to process genomic data in a decentralized way (*i. e.*, divided between different compute centers) (Figure 2). These uniform data processing pipelines contain tools for sequence alignment, error correction, du-





#### Data analysis phases

- I. Submission of data to different high-performance compute centers (e.g. clouds)
- II. Processing of data in standardized fashion (all with the same pipeline)
- III. Sharing of processed data (e.g. variant calls, RNA-seq data)
- IV. Data mining, hypothesis generation and testing

Figure 2: Approach involving shared processing divided between various compute centers and clouds within the PCAWG. Data processing takes place within different IT centers and clouds – like the EBI Embassy Cloud, which is used as an academic community cloud at the EMBL-EBI – as well as the DKFZ high-performance compute center. (Graphic: Jan Korbel and PCAWG Project).

plicate detection, and germline and somatic variant calling. At the moment, three standardized data processing pipelines are used (developed by American, British, and German PCAWG partners) to ensure that the genome analyses are carried out efficiently and robustly. After analysis, the processed data is synchronized between academic IT centers and merged to form a standardized, global data resource for the scientific community.

One relevant practice in PCAWG is the use of virtual machines or “virtual containers” such as the open-source software “Docker”, which allows the project to run standardized IT processes regardless of the respective local IT infrastructure. In order to promote further developments in the field, PCAWG will publish the standardized processes developed within the context of the project in addition to a data resource containing 2,800 cancer genomes. The aim is not only to perform reproducible cancer genome analyses, such as those carried out by the PCAWG, but also to create prerequisites for further expansion of the number of patient genomes collectively analyzed in future cohort studies that may be modelled after PCAWG.

#### Outlook

PCAWG is a pilot project for the shared use of IT resources and data processing pipelines to further cancer genome research. Numbers of analyzed cancer genomes are increasing immensely: Within the coming 5-10 years, recently developed platforms for DNA sequencing (particularly the Illumina HiSeq X Ten platform) are projected to allow over a million patient genomes to be sequenced worldwide which can be integrated

with clinical data (e.g. tumor markers, information on cancer treatment, etc.). Standardized and widely accepted data processing pipelines with corresponding scalability are required to make widespread use of these data in research and personalized medicine. High-performance compute centers which allow data access in a decentralized manner, and cloud computing in particular, will play an important role in this area. This emphasizes also the importance of normative approaches and the development of cloud computing-specific regulations as well as regional clouds that could operate together in a federated system. A future model of a regional genome cloud for research and translation, recently developed for such a context, is available online at [www.genome-cloud.de](http://www.genome-cloud.de).

#### Acknowledgements:

We would like to thank the scientists involved in PCAWG, the “Technical Working Group” in particular, who coordinated the analysis of unprecedented volume of DNA sequencing across different IT centers at an international level. Furthermore, we would like to thank Nina Habermann for her valuable assistance with the translation and preparation of this manuscript. J.O.K and F.M.G. would like to thank the Heidelberg Academy of Sciences for the support given to their normative research project.

## PCAWG research project profile

The *Pan-Cancer Analysis of Whole Genomes* (PCAWG) project is an international collaboration to identify common patterns of mutation in more than 2,800 cancer whole genomes from the International Cancer Genome Consortium (ICGC). PCAWG focuses on investigating the causes and consequences of somatic and germline variations in coding and non-coding areas of the genome. This is in contrast to previous studies within the ICGC as well as the Cancer Genome Atlas Research Project, which primarily aimed at understanding DNA alterations in protein-coding regions. Within PCAWG, over 700 scientists investigate cancer and accompanying germline genomes, comparing different types and subtypes of tumors. Data is analyzed using unique standardized data processing pipelines in local compute centers as well as in clouds – thereby ensuring a standardization of genome analyses.

### Project name:

Pan-Cancer Analysis of Whole Genomes (PCAWG)

### Project Partners involved (Germany):

EMBL (Jan Korbel, Wolfgang Huber);

DKFZ (Roland Eils, Peter Lichter, Benedikt Brors, Christoph Plass);

UKE Hamburg (Guido Sauter, Thorsten Schlomm);

University of Kiel (Reiner Siebert)

### Coordinator (Member of the executive committee and responsible for Germany):

Jan Korbel (EMBL)

### References:

Horn, S., Figl, A., Rachakonda, P.S., Fischer, C., Sucker, A., Gast, A., Kadel, S., Moll, I., Nagore, E., Hemminki, K., *et al.* (2013). TERT promoter mutations in familial and sporadic melanoma. *Science* 339, 959-961.

Lawrence, M.S., Stojanov, P., Mermel, C.H., Robinson, J.T., Garraway, L.A., Golub, T.R., Meyerson, M., Gabriel, S.B., Lander, E.S., and Getz, G. (2014). Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* 505, 495-501.

Northcott, P.A., Lee, C., Zichner, T., Stutz, A.M., Erkek, S., Kawachi, D., Shih, D.J.H., Hovestadt, V., Zapatka, M., Sturm, D., *et al.* (2014). Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. *Nature* 511, 428-+.

Stein, L.D., Knoppers, B.M., Campbell, P., Getz, G., and Korbel, J.O. (2015). Data analysis: Create a cloud commons. *Nature* 523, 149-151.

Weischenfeldt, J., Simon, R., Feuerbach, L., Schlangen, K., Weichenhan, D., Minner, S., Wuttig, D., Warnatz, H.J., Stehr, H., Rausch, T., *et al.* (2013). Integrative Genomic Analyses Reveal an Androgen-Driven Somatic Alteration Landscape in Early-Onset Prostate Cancer. *Cancer Cell* 23, 159-170.

### Contact:



#### Dr. Jan Korbel

European Molecular Biology Laboratory  
(EMBL)

Genome Biology Unit  
Heidelberg, Germany  
korbel@embl.de

<https://dcc.icgc.org/pcawg>

and

[www.genome-cloud.de](http://www.genome-cloud.de)

**17<sup>th</sup> INTERNATIONAL CONFERENCE ON SYSTEMS BIOLOGY**



**DISCOVER THE LATEST TRENDS AND NEW EXCITING RESULT IN SYSTEMS BIOLOGY:**

**Basic Systems Biology**

- Modelling Networks and Circuits
- Signalling Pathways
- Cellular decision making
- Large-Scale Networks
- SB of multicellular tissues and organs
- Image-driven Systems Biology
- Variability and Noise: from single-cells up to populations
- Evolutionary Systems Biology
- Cellular populations and ecological interactions

**Applications**

- Systems Biology of Cancer and Multifactorial Diseases
- Systems Immunology
- Systems Biology of Stem Cells
- Systems and Personalised Medicine
- Synthetic Biology and Biotechnology

**Highlighted Topics**

- Chemotaxis
- Systems Neuroscience

**KEY DATES:**

- **1 June 2016** : Abstracts submission deadline
- **4 July 2016** : Abstracts submitters notification
- **18 July 2016** : Early bird registration deadline

Organised by:



# sensitive genome data

EURAT is addressing ethical and legal issues  
in genome research

by Sebastian Schuol and Eva C. Winkler

Genome sequencing generates many opportunities as well as new ethical and legal issues. Some of the most important topics are: the handling of incidental findings, information and consent of patients or test subjects, protection of personal data and the responsibility of researchers. Scientists from the fields of life science, law and ethics are working together in Heidelberg as part of the EURAT project to address these issues and develop practical solutions for their specific context. They have released a statement in response to the questions raised, and, in light of loopholes in legal regulation, are advocating self-regulation by the scientific community.

## Genome sequencing raises questions

Since the turn of the millennium, genome research has developed into a significant area of research, producing a great deal of knowledge. Gains in this knowledge are mainly driven

by the huge progress in sequencing technology, bioinformatic data analysis and advances in digital data storage. The increasing efficiency with regard to time and money of the new “next generation sequencing” technologies have contributed decisively to broadening the view from isolated genes to entire genomes. Nowadays, genome sequencing not only plays a significant role in basic, but also in translational research. This is especially true for cancer research (Winkler *et al.*, 2013) in which, for example, genetic variations in tumour tissue are detected in comparison to the germline. These analyses aim to improve therapeutic approaches by identifying mutations that will influence the growth of the tumour or the effectiveness of treatment. Considering the tremendous benefits, it is not far fetched to speculate that the use of genome-wide analyses will soon be part of routine clinical diagnostics.

## The EURAT project is developing solutions

In 2011, a consortium of medical practitioners, scientists, bioinformaticians, legal experts and ethicists from Ruprecht Karls

The enormous progress in knowledge in genome research has enabled new application areas, raising, at the same time, new ethical and legal issues. The following topics are subject of controversial debate (Winkler and Schickhardt, 2014):

1. Repeated data analysis (follow-up research) increases the probability of encountering unexpected but medically relevant findings, so-called incidental findings. How should they be dealt with?
2. The genetic data obtained might give the patient access to information that could change their future. How should this be taken into consideration with respect to the process of patient information and consent?
3. The genomic data is identifiable and may contain sensitive information about patients. How is it possible to reconcile protection of privacy and the pursuit of research using this data?
4. The process of genome sequencing is strongly based on the division of labor, hence leading to potentially unclear answerability. Which responsibilities does each individual researcher involved in the process hold?





Microfluidic system of a sequencing instrument (Source: Marsilius-Kolleg).

University in Heidelberg, the Max Planck Institute (MPI) for Comparative Public Law and International Law, the German Cancer Research Centre (DKFZ), the European Molecular Biology Laboratory (EMBL) and the National Center for Tumor Diseases (NCT) with recognised expertise in research areas associated with genome sequencing was formed to investigate normative questions relating to genome sequencing and to develop joint proposals for practical procedures. EURAT (Ethical and Legal Aspects of Whole Human Genome Sequencing) identifies practical problems, discusses these with regard to the normative sciences – ethics and law – and responds by proposing solutions. Apart from the slower legislative process, locally encountered ethical and legal problems can be addressed in a practical and self-regulating manner. This initiates the interactions about normative challenges within the participating institutions and results in the collective preparation of documents. The following will elaborate on the latter:

A code of conduct for researchers as well as two templates for patient information and informed consent were developed within the context of the EURAT statement “*Eckpunkte für eine Heidelberger Praxis der Ganzgenomsequenzierung*” (Cornerstones for a Heidelberg Practice of Whole Genome Sequencing), first published in 2013. The researcher code addresses the issue that genome sequencing is, from an ethical and legal point of view, under-regulated, due to the fact that non-medical researchers for example, do not possess such a deeply embedded set of professional ethics as medical professionals for guidance. In addition, the German Genetic Diagnosis Act only applies to

clinical care, but explicitly excludes research (GenDG § 2, Par. 2); this legal regulation lags behind the rapid developments on the whole (Bartnek 2009).

The code of conduct establishes new rights and obligations for researchers when dealing with genetic knowledge relating to patients (and their families). Based on the work with patient genomes and its knowledge of human genetics, the individual researcher obtains a professional responsibility for action, while governing bodies are given organisational responsibility. Individual guidelines are concerned with risk mitigation with regard to the research project, the necessity of an informed consent and an ethics vote, as well as the patient’s consent for sample storage. The area of data and research findings is of crucial importance: In order to ensure the best level of patient protection possible, the data obtained must be encrypted and may only be used for areas of research approved by the patient. Unlike anonymisation, in which the link between data and patient is severed, pseudonymisation allows the patient to benefit of findings with therapeutic value, which may appear during the course of research. The researcher is obliged to report findings that presumably contribute to the patient’s welfare to the attending physician, who will then decide how to proceed. However, the researchers are not committed to actively search for such findings. The code therefore creates an institutionalised framework that imposes certain obligations on the researchers, but, within these limits, in turn provides security for their actions and thus protection.



Chemical solutions for the operation of sequencing instruments  
(Source: J. Ritzerfeld).

However, the impact of genome sequencing not only extends to the researchers, but in particular to the patients whose genomes are being sequenced. The EURAT group stresses the necessity of an informed patient consent. Hence, the group has developed templates for specific application areas to strike a balance between providing the most comprehensive information possible with respect to important implications, and the highest possible degree of comprehensibility for patients.

Incidental findings pose a particular challenge: On one hand, the patient should decide whether, and if so, of which incidental findings he would like to be informed. On the other hand, a comprehensive explanation of incidental findings is not possible, considering over 5,000 monogenic diseases (OMIM). Incidental findings should therefore be categorized (treatable, preventable, untreatable) and explained using examples. In addition, data issues are of particular importance during the information process: As the potential to identify people and obtain sensitive information about them is inherent to genomic

data, the method of pseudonymisation, the transfer of data to international and national research institutes, the opportunities and limitations of data deletion in the event of revoked consent and the accessibility of data for verification purposes will be explained, amongst other things.

### How the EURAT solutions come into effect

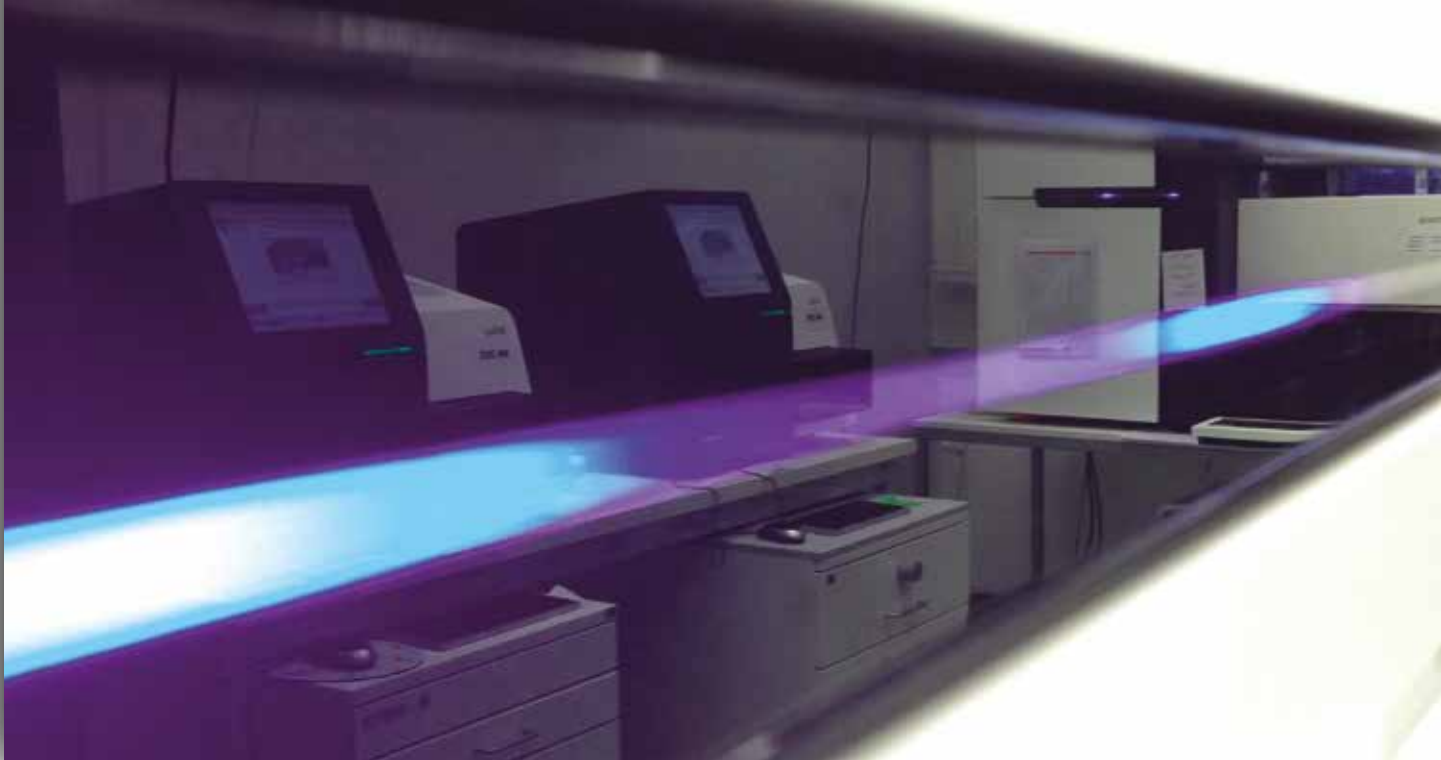
One of the main successes of the EURAT position statement on an institutional level is that shortly after its publication, the researcher code was adopted by both the Heidelberg University by means of an official senate decision and the DKFZ and has since then been binding there for all researchers in the field of genome sequencing. Breaches of the code of conduct can now also be subject to sanctioning measures, for example with regard to employment and liability law. However, the aim of the code is not sanction, but guidance in an otherwise unregulated area. The positive feedback from experts as well as public media shows that this has been successful. However, the field of genome research is continuing to develop at a rapid pace. This is why a second updated version of the EURAT position statement was published (EURAT 2016). Amongst other things, it now includes specific solutions for data privacy protection, for example an explanation of the data security concept for personal data in cancer research for research projects involving genomic data that was in the meantime developed at DKFZ in collaboration with EURAT.

It is clear that data issues are emerging as one of the most important ethical topics within the context of genome research, thus continuing to shape EURAT in future. Cloud computing is becoming an increasingly important issue in genome research, as it provides a cost-efficient solution and attends to the rising need to transfer genomic data within international research consortiums. However, data sharing across international borders affects a number of legal systems while jurisdiction remains on the cloud provider's location and protection of personal data varies. These new ethical and legal issues regarding the transfer of data in cloud solutions will require feasible solutions soon.

---

### Research project profile:

EURAT (Ethical and Legal Aspects of Whole Human Genome Sequencing) is a project on normative questions relating to



A sequencing instrument in operation (Source: Marsilius-Kolleg).

genome sequencing at the research site in Heidelberg, bringing together scientists from Heidelberg University, Heidelberg University Hospital, German Cancer Research Centre (DKFZ), European Molecular Biology Laboratory (EMBL) and Max Planck Institute (MPI) for Comparative Public Law and International Law. The EURAT project was initiated in 2011 by the Marsilius Kolleg within the framework of the Excellence Initiative.

**The members of EURAT are:**

Prof. Dr. Claus R. Bartram, Human Genetics | Prof. Dr. Roland Eils, Bioinformatics | Prof. Dr. Hanno Glimm, Oncology | Prof. Dr. Christof von Kalle, Oncology | Prof. Dr. Dr. h.c. Paul Kirchhof, Constitutional Law | Dr. Jan Korb, Bioinformatics/Genome Sequencing | Prof. Dr. Andreas E. Kulozik, Oncology | Prof. Dr. Peter Lichter, Tumour Genetics/Genome Sequencing | Prof. Dr. Peter Schirmacher, Pathology/Biobanking | Prof. Dr. Klaus Tanner, Ethics/Theology (Project Spokesperson from 2011-2013) | Prof. Dr. Stefan Wiemann, Genome Sequencing | Prof. Dr. Dr. Eva Winkler, Oncology/Medical Ethics (Project Spokesperson) | Prof. Dr. Dr. h.c. Rüdiger Wolfrum, Constitutional Law/International Law.

[www.uni-heidelberg.de/totalsequenzierung/english.html](http://www.uni-heidelberg.de/totalsequenzierung/english.html)

**References:**

Winkler, E.C., Glimm, H., Tanner, K., von Kalle, C. (2014): Ethical considerations for developing a best practice guideline for next generation sequencing in oncology. *The Ethics of Personalised Medicine – Critical Perspectives*, J. Vollmann, V. Sandow, S.

Wäscher, J. Schildmann, eds (Farnham: Ashgate), pp. 87-96.

Winkler, E.C., Schickhardt, C. (2014): Ethical challenges of whole genome sequencing in translational research and answers by the EURAT-project. *J Lab Med*, 38(4), 211–220

EURAT (2016): Position Paper: Cornerstones for an ethically and legally informed practice of Whole Genome Sequencing (Heidelberg: Nino Druck GmbH).

OMIM – Online Mendelian Inheritance in Man. [www.omim.org](http://www.omim.org)

Bartnek, T. (2009): Das Gendiagnostikgesetz: Ein lückenhafter Schutz. *Gen-ethischer Informationsdienst* 194, 50-54.

**Contact:**



**Prof. Dr. med. Dr. phil. Eva Winkler**

Spokesperson of the EURAT-Project and Head of the program “Ethics and Patient oriented Care in Oncology”

National Center for Tumor Diseases (NCT) Heidelberg, Germany  
eva.winkler@med.uni-heidelberg.de



**Sebastian Schuol, M.A.**

Coordinator of the EURAT project  
National Center for Tumor Diseases (NCT) Heidelberg, Germany  
Sebastian.Schuol@med.uni-heidelberg.de

# big data: perspectives in cancer therapy

## Impressions from an industry-based point-of-view

by Ajay Kumar

Big Data techniques are increasingly being used in industry and medical research. Decision makers in industry can optimize their existing business models and generate more value for customers. In cancer research, Big Data could enable a paradigm change! During a sabbatical in summer with the German Cancer Research Center (DKFZ) in Heidelberg, I looked at how Big Data approaches are applied in the specific use case cancer research. The open and collaborative culture at this scientific research institution made it possible to share my experiences, while at the same time the unique cross functional environment provided me with ideas I could transfer to my industry workplace.

### From industry to cancer research

Reading through the newspaper on a Sunday morning, I noticed an interesting article about usage of Big Data in cancer research. The article was an interview with two researchers from the German Cancer Research Center (DKFZ) in Heidelberg, Prof. Roland Eils and Prof. Christof von Kalle. The article explained how Big Data techniques are becoming an indispensable tool in cancer research to understand the disease at a deeper level. As an automobile manufacturing professional, I was familiar with the application of Big Data techniques in the consumer and engineering industry. The list of use cases is long, from optimizing supermarket racks to increasing efficiency of machines and processes. A sabbatical at DKFZ offered a unique opportunity to learn how technical advances in high-throughput technologies have a direct impact on cancer research and therapy.

### The four stages of industrial revolution

My interest for Big Data came from a strong trend in the production environment. Industry 4.0 or the fourth industrial revolution is perhaps one of the most frequently discussed terms today in the German industry (Kagermann *et al.*, 2013). The first three industrial revolutions came about as a result of mechanization, electrification and information technology (Figure 1). Before the first industrial revolution in the late 18<sup>th</sup> century, production work was carried out manually or using natural resources like water or wind. The invention of the steam engine and its usage for production changed the industrial landscape. Textile mills in Britain were first to experience this change. At the beginning of the 20<sup>th</sup> century, electrification and division of labor were enablers for the second industrial revolution. The production of the T-Model from Henry Ford best describes this change. It made the automobile affordable as long as everyone bought the same car. Mass production of standardized goods with low costs became reality.

In the early 1970s, electronics and information technology impacted production, launching the third industrial revolution. Complex supply chains for products could be planned using software based systems (Kagermann *et al.*, 2013). The automation level was increased with robots replacing manual work on production lines. Complex production schemes with high technical requirements could be realized using numeric control machines. Modern automobile production in Europe came into existence, producing a variety of different car models with high quality standards.

The fourth industrial revolution is based on the idea of internet of things (IoT) and services (Figure 1). In factories, machines, materials, services, products, people, suppliers and customers are connected. In a connected environment, machines interact



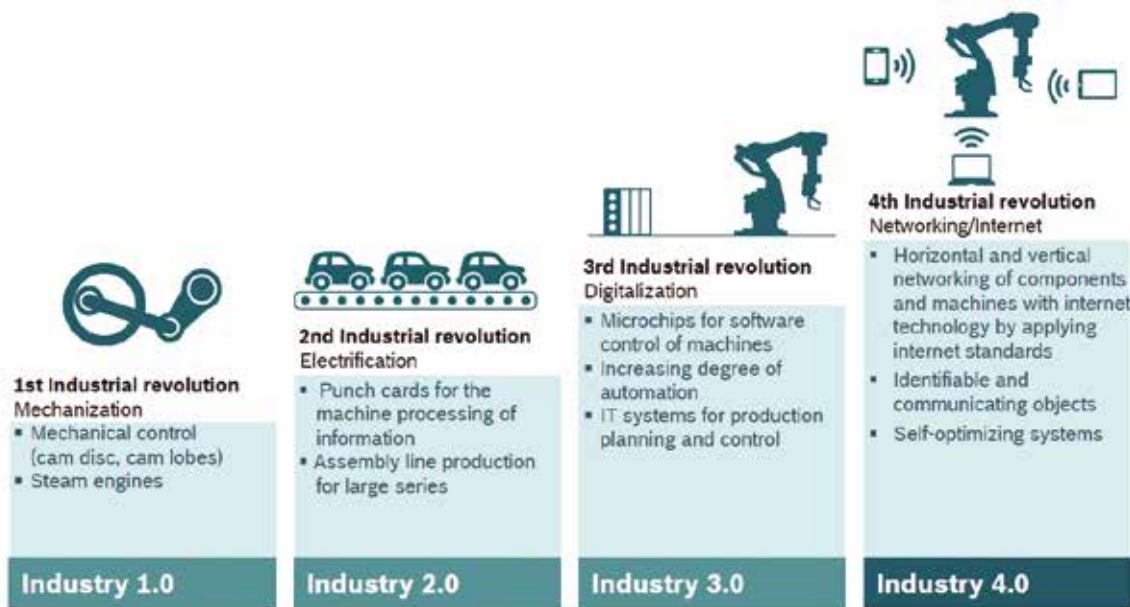


Figure 1: The four stages of industrial revolution (Source: Bosch Rexroth).

with one another, robots work with employees in so called cyber physical systems (Kagermann *et al.*, 2013) and production systems have the capacity to learn and optimize themselves. Outside the factory, customers and products are connected to the internet, opening new possibilities for businesses while at the same time disrupting existing business models.

Each industrial revolution increased the complexity of the production environment and acts as an enabler to increase the industrial output. It also has a significant impact on our society and competitiveness of the industry.

### Big Data technologies improve cancer treatment

The division of Theoretical Bioinformatics (“eilslabs”) at DKFZ and Heidelberg University is located on the Neuenheimer Feld campus in Heidelberg, which is comprised of a variety of different scientific and medical institutions.

I started my sabbatical at the data management group in eilslabs. The group comprises bioinformaticians, computer engineers and software developers. I had very little knowledge about the molecular mechanisms underlying cancer and quickly learnt that cancer occurs as a result of changes in our DNA. The identification of these underlying mutations would be a big breakthrough and a major goal in modern cancer research. A kind of mutation dictionary on cancer.

Knowledge about these genetic alterations integrated with clinical patient data enables clinical experts to work out personalized treatment plans for cancer patients. This collective and cross functional approach leads to synergies in the therapy process, thus reducing costs for treatment and, more importantly, improving patient well-being and survival.

### Enablers for Big Data: sequencing costs and the internet

I asked myself why this procedure isn’t already a standard practice in cancer care. A decade ago, high sequencing costs and bottle necks in IT infrastructure were the main hurdles for translating genomic research into the clinic. In 2001, the cost of sequencing a single human genome was about 100 Million \$ (Figure 2). Recent developments in sequencing technology have brought the costs down over the years, to 5,000 \$ per patient genome in 2014 (Hayden, 2014). By the end of this decade, I expect it should be possible to sequence a human genome for less than 800 \$.

Low cost of sequencing and availability of high performance computing nowadays enables a breakthrough for cancer research and personalized medicine. Medical doctors can consider an individual’s genomic data for developing a precise therapy. Personalized oncology for a majority of cancer patients should be reality soon!

Availability of the internet at home, on street and almost everywhere makes it possible to connect products, services and people. By 2022, analysts at Machina Research forecast 14 billion individual devices to be connected to the internet, ranging from IP enabled cars to heating systems, security cameras, sensors and production machines (Figure 3).

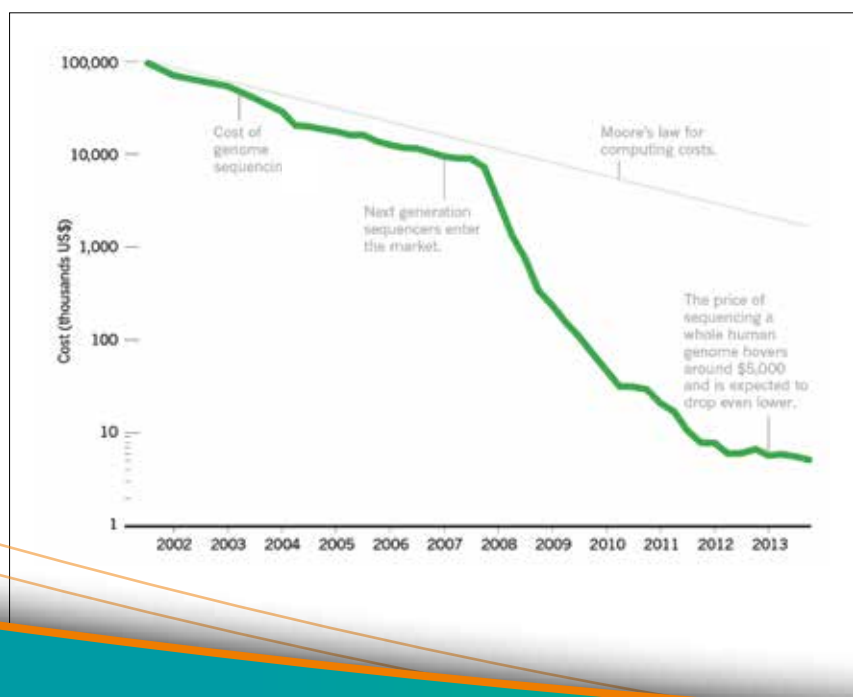
Collecting and analyzing the data that is sent from these devices over the internet, can substantially improve user experience or efficiency in factories. For example, in a factory production line, each machine can be connected to the internet or a shared network. Sensors and devices in the machine monitor the condition of the machine and the production process. The analysis of this data can be used by engineers to correct deviations in the production process or plan the maintenance of the machine to avoid an unplanned breakdown.

During my sabbatical, I became aware of many individual issues which were being dealt with for the realization of the use case at DKFZ. The three issues I experienced were: Data management, ethical issues and workflows for data analysis.

### Data management: storage and security

Handling and processing genomic data is a critical task. To emphasize the dimension of the use case, one could assume to sequence 50-80 patients every day. Sequencing data from a single human genome has a size of about 80 GB. With 2 samples, tumor and control, 8 to 12 Terabytes of data would be generated each day. This is close to the daily data volume at Twitter! For present research with about 500 patients a year, an excellent Petabyte storage is already in place. To manage the use case in coming years, many options were being evaluated, such as cloud, industry partnerships or outsourced facilities. Cloud

Figure 2: Development of sequencing costs



In the first few years after the end of the Human Genome Project, the cost of genome sequencing roughly followed Moore's law, which predicts exponential declines in computing costs. After 2007, sequencing costs dropped precipitously. (Source: Hayden, 2014)

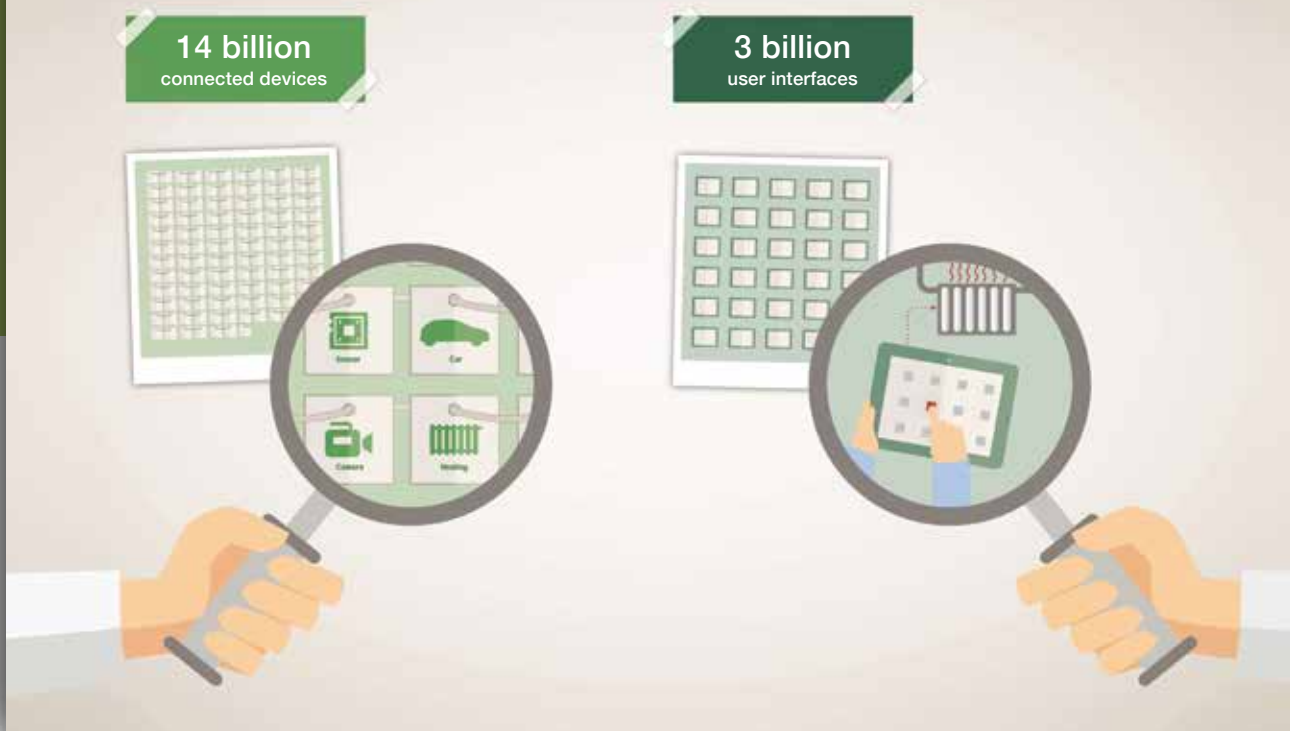


Figure 3: Connected devices by 2022 (Source: Machina Research/Bosch software innovations GmbH).

storage offers scalability but data security and connectivity need to be addressed. An onsite storage alone would imply high investments and high ownership costs but provides quick access and high security. Partnerships for the installation of data farms to store medical and IoT data with similar data protection guidelines could further reduce costs of storage solutions. Ethical issues on genomic data have to be addressed by specialized interdisciplinary project groups like EURAT (see article on page 40).

### The workflow for data analysis: *One Touch Pipeline*

The workflow for data analysis is the central or core element of the use case. The “*One Touch Pipeline*” (OTP) is an automation platform for the management of genomic data from next generation sequencing experiments, supporting data transfer, quality control, sequence alignment to the reference genome and the identification of single nucleotide variations in patient DNA. OTP is like a virtual manufacturing line. An excellent manufacturing line has stable processes, is scalable for demands, has low downtimes and is cost effective. Material flows quickly through the production steps to generate a final product. If the opposite happens, stocks pile up, scrap rates are high, quality is poor and costs are high.

OTP needs to run like an excellent manufacturing line for fast transfer of genomic data into the clinic. DKFZ would have to collaborate across functions, departments and with software companies to prepare OTP for future demands.

The data management team develops and maintains OTP. The team encounters many user requests, is facing short deadlines and has limited resources. To meet these challenges, the data management group uses Scrum, an iterative agile product development methodology, in the development of OTP. The working environment is agile, collaborative, team-oriented and has short feedback cycles. The team looks into the needs from users, prioritizes them and plans a time span of about 3 week (sprints) to solve these issues (Figure 4). I was inspired to take some of the elements of this method to my work place. Industrial environments have an organizational structure and a defined set of management levels. Goals and strategies are communicated top down. This ensures consistency but makes the organization rigid and slow. Industry 4.0 is not only a technological change but also a change for our workplaces. Flexibility and agility are important. Some principles of scrum can facilitate these needs and foster higher satisfaction for employees at work.



Figure 4: Daily stand-up meeting of the OTP scrum team. The team meets to give a brief feedback on present work issues (Source: G. Zipprich).

### The road ahead: effective cancer therapy and the connected world

Big Data technologies will act as the main enablers, both for personalized cancer care and Industry 4.0. For industry use cases, Big Data techniques will result in an improvement of our working environment, competitiveness and better customer satisfaction. In cancer research and therapy, Big Data technologies will give a strong impetus to personalized cancer care for patients. Strong partnership and cooperation is needed between different partners to implement such use cases. Partners collaborating together will have to take risks and need strong support from policy makers in the early stages of implementation.

### References:

- Kagermann, H., Wahlster, W., Helbig, J. – Industrie 4.0 working group of Industry-Science Research Alliance and acadtech National Academy of Science and Engineering (2013). Recommendations for implementing the strategic initiative INDUSTRIE 4.0.
- Hayden, E.C. (2014) Technology: The 1000 \$ Genome. Nature 507, 294-295

### Contact:



**Ajay Kumar**  
Director for production  
and operations management  
in the automotive industry  
Heidelberg, Germany  
ajaykumar@web.de

Photo: A. Kumar



# *fachgruppe bioinformatik* – representing interests with one voice

FaBI – the alliance of the bioinformatics special interest groups of five German scientific societies from the field of life sciences and informatics

by Matthias Rarey

Bioinformatics has emerged as a central element of research in life sciences. The complexity of the issues faced, combined with huge quantities of data, has created major challenges for informatics. Today's bioinformatics research is, more than ever, characterised by the interplay of application and the development of methods. Biologists and computer scientists, chemists, physicists, mathematicians and statisticians, pharmacists and physicians – scientists of every kind are involved in bioinformatics, with good reason. Although being stimulating and effective for science, this heterogeneity also implies difficulties. Scientific societies play a key role in the coordination of research, and bioinformatics is, of course, incorporated into many special interest groups. Consequently, representing the entire field is difficult, while those outside of the field usually find it hard to identify contacts. This is why in 2014, five large scientific societies from the fields of life sciences and informatics agreed to found the *Fachgruppe Bioinformatik* FaBI (joint interest group Bioinformatics) seeking to represent the interests of bioinformatics in Germany, and giving it a face and a voice.

There is a long tradition of bioinformatics in Germany. In 1985, the first conference on 'Biotechnology and Information' took place, which was then re-named to '**German Conference on Bioinformatics (GCB)**' in 1993 and has taken place annually at various locations in Germany ever since. Today, bioinformatics is advancing very dynamically, driven by both computer and life sciences. Regardless of whether one feels mainly at home in informatics or in life sciences, acting together has become crucial. This is where scientific societies come in. They play a major role in shaping specialist fields in almost every scientific discipline. For example, the support of bioinformatics by the BMBF and the DFG in the form of research centres was preceded by an important position paper with notable participation of the *Gesellschaft für Chemische Technik und Biotechnologie* (DECHEMA) (Society for Chemical Engineering and Biotechnology) and the *Gesellschaft für Informatik* (GI) (Society for Informatics). With regard to this, in 2014 the joint *Fachgruppe Bioinformatik* (FaBI) was created by five scientific societies from the fields of informatics and life sciences. FaBI now represents over 700 bioinformaticians. It advocates internationally competitive bioinformatics research and education in Germany and serves as a partner and contact for politicians and the general public.





**FaBI advisory board** (from left to right): Bertram Weiß (Bayer Pharma AG), Sven Rahmann (U. Duisburg-Essen, TU Dortmund), Oliver Kohlbacher (U. Tübingen), Matthias Rarey (U. Hamburg), Caroline Friedel (LMU München), Ina Koch (U. Frankfurt). Other advisory board members (not pictured): Thomas Engel (LMU München), Heike Pospisil (FH Wildau), Martin Vingron (MPI of Genetics, Berlin) (Source: Dechema e. V.).

FaBI has already achieved some notable success. Its priority was to create a consensus-based opinion on the state of bioinformatics in Germany as well as the improvement of visibility and transparency of bioinformatics. Based on a broad survey, a position paper was drawn up in the first half of 2015. Its core messages stress the discipline's autonomy, the creation of a sustainable infrastructure, more internationalisation and issues relating to the education and promotion of junior scientists. You can access an information portal at [www.bioinformatik.de](http://www.bioinformatik.de), which

does more than just represent FaBI's numerous activities. It provides the first directory of academic groups from bioinformatics in Germany. The search for experts and collaborative partners has become much easier, which is essential for an interdisciplinary field. All degree courses with bioinformatics-related content are presented on the website in a concise, tabular form guiding school pupils and students. News, conference announcements and job offers complete the information portfolio.

## Important activities of the *Fachgruppe Bioinformatik* in 2014 and 2015

### German Conference on Bioinformatics (GCB)

Selecting the topics of the annual international GCB conference at a German university; support with the programme design and organisation

### Website [www.bioinformatik.de](http://www.bioinformatik.de)

Operation and edition of [www.bioinformatik.de](http://www.bioinformatik.de). In addition to information on the *Fachgruppe Bioinformatik*, the GCB and bioinformatics in general, the site also provides current news, job offers and a directory of bioinformatics research groups and courses in Germany.

### Position paper, "*Bioinformatik in Deutschland – Perspektive 2015*" (Bioinformatics in Germany – 2015 Perspective)

An account of the status of bioinformatics in Germany from the point of view of the bioinformatics community, including six recommendations for further development. The position paper is based on a survey conducted among bioinformaticians and is supported by a large group of scientists (see [http://www.bioinformatik.news/images/documents/Positionspapier\\_FaBI\\_2015.pdf](http://www.bioinformatik.news/images/documents/Positionspapier_FaBI_2015.pdf))

## Expert associations represented in the FaBI, in alphabetical order

### ➤ DECHEMA:

[http://dechema.de/dechema\\_eV/en/](http://dechema.de/dechema_eV/en/)

Gesellschaft für Chemische Technik und Biotechnologie e. V.  
(Society for Chemical Engineering and Biotechnology)

### ➤ GBM:

<https://www.gbm-online.de/home.html>

Gesellschaft für Biochemie und Molekularbiologie e. V.  
(Society for Biochemistry and Molecular Biology)

### ➤ GDCh:

<http://en.gdch.de/>

Gesellschaft Deutscher Chemiker e. V.  
(German Chemical Society)

### ➤ GI:

<https://en.gi.de/startpage.html>

Gesellschaft für Informatik e. V.  
(Society for Informatics)

### ➤ GMDS:

<http://www.gmds.de>

Deutsche Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e. V. (German Society for Medical Informatics, Biometry and Epidemiology)

### Contact:



**Prof. Dr. Matthias Rarey**

University of Hamburg

ZBH – Zentrum für Bioinformatik

[rarey@zbh.uni-hamburg.de](mailto:rarey@zbh.uni-hamburg.de)

<http://www.zbh.uni-hamburg.de/en/prof-dr-matthias-rarey.html>

Speaker of the Fachgruppe Bioinformatik

(FaBI) of the DECHEMA e. V., GBM e. V.,

GDCh e. V., GI e. V. and GMDS e. V.

[sprecher@bioinformatik.de](mailto:sprecher@bioinformatik.de)

<http://www.bioinformatik.de/>

## FaBI's definition of bioinformatics

*“Bioinformatics in an interdisciplinary science. It is understood as the research, development and application of computer-based methods used to answer questions from biomolecular and biomedical research. Bioinformatics mainly focusses on models and algorithms for data at a molecular and cellular level, for example*

➤ *genomes and genes,*

➤ *gene and protein expression and regulation,*

➤ *metabolic and regulatory pathways and networks,*

➤ *structures of bio-macromolecules, especially DNA, RNA and proteins,*

➤ *molecular interactions between bio-macromolecules and between bio-macromolecules and other substances such as substrates, transmitters, messenger substances and inhibitors as well as*

➤ *the molecular characterisation of ecological systems.”*

Source: Fachgruppe Bioinformatik, <http://bioinformatik.de/de/bioinformatik-3/was-ist-bioinformatik>



# News from the BMBF

## The BMBF is making it easier for refugees to access higher education

Education is the key to successful integration. More than half of the refugees currently arriving in Germany are under 25 years of age – that is to say of an age when they need training. In the coming years, the Federal Ministry of Education and Research (BMBF) will introduce targeted measures to help higher education institutions grant places to those refugees who wish to study and have the necessary qualifications to do so. This will also foster integration.

The Federal Education Ministry has put together a comprehensive package of measures worth 100 million euros. The package consists of three components which form the basis for successful access to higher education: Measures will be taken to assess the individual competences and potential of the refugees, ensure their scholastic aptitude through preparatory specialist and language courses and support their integration at the higher education institutions.

Introducing the package of measures, Federal Education Minister Johanna Wanka said, “Education is the key to integrating refugees, particularly those who have long-term prospects for staying in Germany. The higher education institutions are of great importance in this context, among other things because they have had years of experience with foreign students. We already have suitable instruments for dealing with language challenges or different qualifications and for providing foreign students with sound counselling. We can now build on this infrastructure when integrating refugees at higher edu-

cation institutions. We intend to enable anyone wishing and able to pursue academic training to do so. Germany will also benefit from these measures.”

[www.bmbf.de/de/fluechtlinge-durch-bildung-integrieren-1944.html](http://www.bmbf.de/de/fluechtlinge-durch-bildung-integrieren-1944.html)



## More funding for education and research

2016 has seen a further rise in the BMBF’s budget, this time with an increase of 1.1 billion euros to around 16.4 billion euros. This shows the importance that the Federal Government attaches to education and research.

Around 515 million euros are being made available for vocational education and training. This represents a 15 percent increase for this funding priority compared with 2015.

The Federal Government provided the *Länder* with approximately 8 billion euros between 2007 and 2015 to fund additional university places under the Higher Education Pact. A further 2 billion euros will be provided in 2016. An additional 50 million euros will be invested in improving teacher training.

The Ministry is making around 5.5 billion euros available for institutional research funding. The 2016 budget also includes an increase in the resources available for funding research under the Federal Government’s new High-Tech Strategy.

[www.bmbf.de/en/education-and-research-priority-areas-of-federal-government-policy-1410.html](http://www.bmbf.de/en/education-and-research-priority-areas-of-federal-government-policy-1410.html)



**Focus on promoting language skills:  
Johanna Wanka with students from a  
German course.**

Source: BMBF/Hans-Joachim Rickel



### Campaign gives people the courage to learn to read and write

Around 7.5 million people in Germany are functionally illiterate. This means that they are able to read and write individual sentences but cannot read and understand continuous texts. A further 2.3 million are considered to be completely illiterate. They are unable to even write or understand individual sentences. Fear and shame prevent most of these people from actively seeking help.

The Federal Education Ministry has therefore launched a nationwide awareness campaign in association with other partners. This campaign seeks to reach out to as many people as possible and inform them about the opportunities for adults to learn to read and write. The campaign features adults talking about the pivotal moment in their lives when they plucked up courage and decided to learn how to read and write better. These moments are shown in TV spots and cinema advertising as well as on placards and post-cards.

The Federal Government and the *Länder* want to significantly improve adult reading and writing skills in Germany over the next ten years by providing more support through adult literacy courses. The Federal Education Ministry will provide up to 180 million euros for literacy projects over the next ten years and will design courses and self-learning programmes.

According to Federal Education Minister Johanna Wanka, "People who cannot read and write adequately often feel excluded because these are the basic skills for social participation. They help to secure employment in today's working world with its modern technologies and service orientation. We want to cooperate with the *Länder* and many other partners over the next ten years to ensure that more people have the courage to improve their reading and writing skills even later in life."

[www.mein-schluessel-zur-welt.de](http://www.mein-schluessel-zur-welt.de)



**What changes will there be in the ecosystem of the seas and oceans? One of the many questions posed in Science Year 2016\*17.**

Source: flysafe340 - Fotolia



**Be brave! It is never too late to take the next step. The new motto of the information campaign is intended to motivate adults to learn to read and write.**

Source: BMBF

### New Science Year devoted to marine research

Marine research is the focus of the new Science Year beginning in mid-2016. Topics range from the marine ecosystem to the significance of the oceans for our weather and climate and the societal importance of the seas and coastal regions as cultural areas, places of longing and travel destinations. Several hundred events, conferences, exhibitions and competitions will be held throughout Germany.

In the words of Federal Research Minister Johanna Wanka, "The seas and oceans are of tremendous importance for our lives. We still do not know very much about most of the world's seas. Nevertheless, they are often heedlessly exploited and polluted. For us, the seas are a source of food, an economic area and a climate machine. We want to raise people's awareness of this important topic. People must become more familiar with our planet's largest ecosystem and learn how to protect it."



The current Science Year 2015 focuses on the “City of the Future”. Scientists are working with local authorities, industry and members of the public to find local solutions to major societal challenges such as safe and secure energy supplies, climate-adapted construction and mobility.

[www.bmbf.de/en/marine-and-polar-research-2316.html](http://www.bmbf.de/en/marine-and-polar-research-2316.html)  
and  
[www.wissenschaftsjahr-zukunftstadt.de/uebergreifende-infos/english.html](http://www.wissenschaftsjahr-zukunftstadt.de/uebergreifende-infos/english.html)



### “Wendelstein 7-X” fusion device begins operations

The new Wendelstein 7-X recently began operations in Greifswald. As a national large-scale research facility it was built by the Max Planck Institute for Plasma Physics and is intended to demonstrate the ability of a stellarator-type facility to generate nuclear fusion power.

“Wendelstein 7-X stands for cutting-edge research in Germany. With Wendelstein we are treading new ground in nuclear fusion. In the long term, this can lead to power stations which are able to meet the baseload demand and produce large quantities of electricity reliably,” says Federal Research Minister Johanna Wanka.

The investment costs total 370 million euros and the overall cost of building the institute in Greifswald was approximately 1.1 billion euros. The project is co-funded by the Federal Government, the EU and the *Land* of Mecklenburg-Western Pomerania.

[www.bmbf.de/de/jetzt-wird-s-heiss-fusionsanlage-wendelstein-7-x-nimmt-betrieb-auf-2160.html](http://www.bmbf.de/de/jetzt-wird-s-heiss-fusionsanlage-wendelstein-7-x-nimmt-betrieb-auf-2160.html)



The twelve metre-long skeleton of the tyrannosaurus rex “Tristan Otto” is a great attraction. It will be on show at Berlin’s Natural History Museum for three years.

Source: BMBF/ Hans-Joachim Rickel

### Spectacular tyrannosaurus rex skeleton exhibited in Berlin

The king of all dinosaurs is the new star of a special exhibition at Berlin’s Natural History Museum. The skeleton of the tyrannosaurus rex – named “Tristan Otto” – is one of the best preserved finds world-wide. It will be on public display in Berlin for three years.

According to Jochen Vogel, Director of the Museum, “We are exhibiting 170 of the animal’s original 300 bones”. Scientists intend to examine the skeleton further in the years to come. They have already discovered, for example, that Tristan Otto had a tumour in his jaw bone and must have had terrible toothache. Researchers now want to find out more about how the dinosaur lived, how old he was and what he ate. Tristan Otto was found in Montana, USA in 2012. He is around 66 million years old.

Speaking at the opening of the exhibition, Federal Research Minister Johanna Wanka said, “As a Leibniz institute, the Natural History Museum is co-financed by the Federal Government and the *Länder*. It stands for a unique collection, excellent research and exciting exhibitions. The Museum is once again living up to its reputation by showing the spectacular skeleton of the tyrannosaurus rex Tristan Otto.”

[www.bmbf.de/de/dino-mit-zahnschmerzen-skelett-des-t-rex-fasziniert-forscher-2230.html](http://www.bmbf.de/de/dino-mit-zahnschmerzen-skelett-des-t-rex-fasziniert-forscher-2230.html)



### **BMBF is boosting the innovative strength of German small and medium-sized enterprises**

The German industrial model centres around small and medium-sized enterprises (SMEs): They provide 16 million people with employment and account for around 83 % of apprenticeships. Although many of these companies are currently doing well, global competition, digitalization and new business models mean that it is by no means certain that this will be the case in future.

The Federal Research Ministry would like to help SMEs to develop new ideas and make use of research results for themselves. "Our aim is to encourage companies which have previously not been particularly active in the field of innovation to make an effort in this area," said Federal Minister Johanna Wanka at the presentation of the new "Priority for SMEs" programme, under which the Ministry has increased its funding for SMEs to around 320 million euros per year until the end of 2017.

There is cause for concern because SMEs, unlike large companies, have not increased their expenditure on innovations in recent years. The new programme is intended to help reverse this trend. Measures are being introduced to make it easier for SMEs to access the BMBF's specialist programmes, particularly in the fields of the digital economy, healthy living and sustainable economy. The Ministry also intends to bring SMEs together with strong partners, with higher education institutions, research institutions and large companies. It wants to help to ensure that they have the necessary skilled staff at their disposal and that the younger generation receives excellent training.

[www.bmbf.de/de/fuer-einen-innovativen-mittelstand-2326.html](http://www.bmbf.de/de/fuer-einen-innovativen-mittelstand-2326.html)



### **G7 Science Ministers address global problems**

The fight against poverty-related diseases was one of the thematic priorities of the German G7 Presidency in 2015. The G7 Science Ministers agreed at a meeting in Berlin to extend the G7's research to include the entire range of neglected poverty-related infectious diseases. This now includes over 25 infectious



### **Meeting of the G7 Science Ministers in Berlin.**

Source: BMBF/ Hans-Joachim Rickel

diseases which occur predominantly in poorer or tropical countries. The Ministers agreed to survey all on-going measures on poverty-related infectious diseases and to coordinate their research funding. A joint research initiative is to be agreed at a conference in 2016.

Speaking at the meeting, Federal Research Minister Johanna Wanka said, "Science and research take high priority in all G7 states. We must contribute our knowledge and expertise to help in those areas where urgent challenges of humankind need to be solved."

Wanka also announced that Germany would be providing an additional 50 million euros in funding over the next five years for product development partnerships to encourage the targeted development of specific drugs.

[www.bmbf.de/en/the-german-g7-presidency-1428.html](http://www.bmbf.de/en/the-german-g7-presidency-1428.html)



### **Contact:**



Information on these and other interesting topics under the new High-Tech Strategy for Germany can be found at:  
[www.hightech-strategie.de/de/The-new-High-Tech-Strategy-390.php](http://www.hightech-strategie.de/de/The-new-High-Tech-Strategy-390.php)



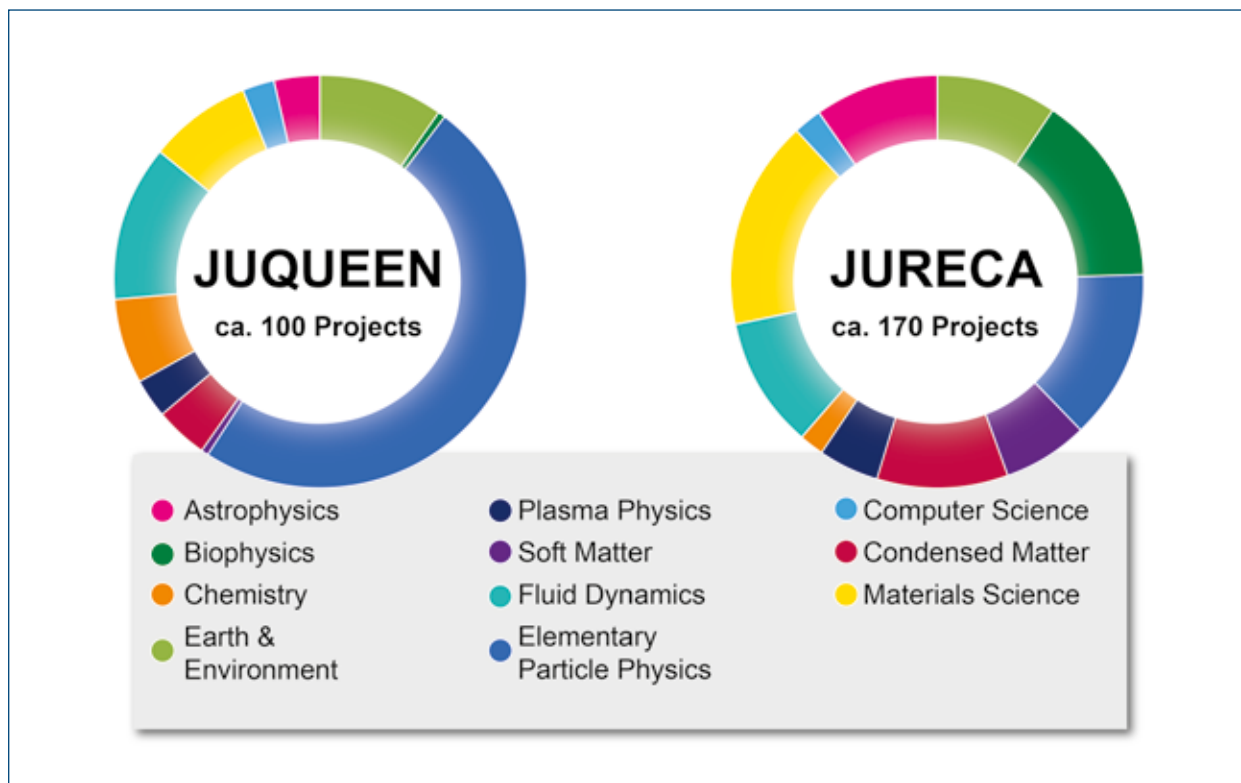
## SIMULATION LABORATORIES – SUPERCOMPUTER SUPPORT AND RESEARCH FOR COMMUNITIES

### Don't be afraid of large computers

Since 2008, the Jülich Supercomputing Centre (JSC) runs simulation laboratories. These support facilities consist of teams that actively participate in research and therefore are very familiar with the specific requirements of their respective research community concerning the use of supercomputers.

Supercomputers are capable of high performance because they spread their operations over a large number of processors, which exchange data with one another extremely quick. Such systems are therefore only able to demonstrate their computing power if the software they are running scales well, i.e. it is programmed in such a way that the speed of computation increases approximately

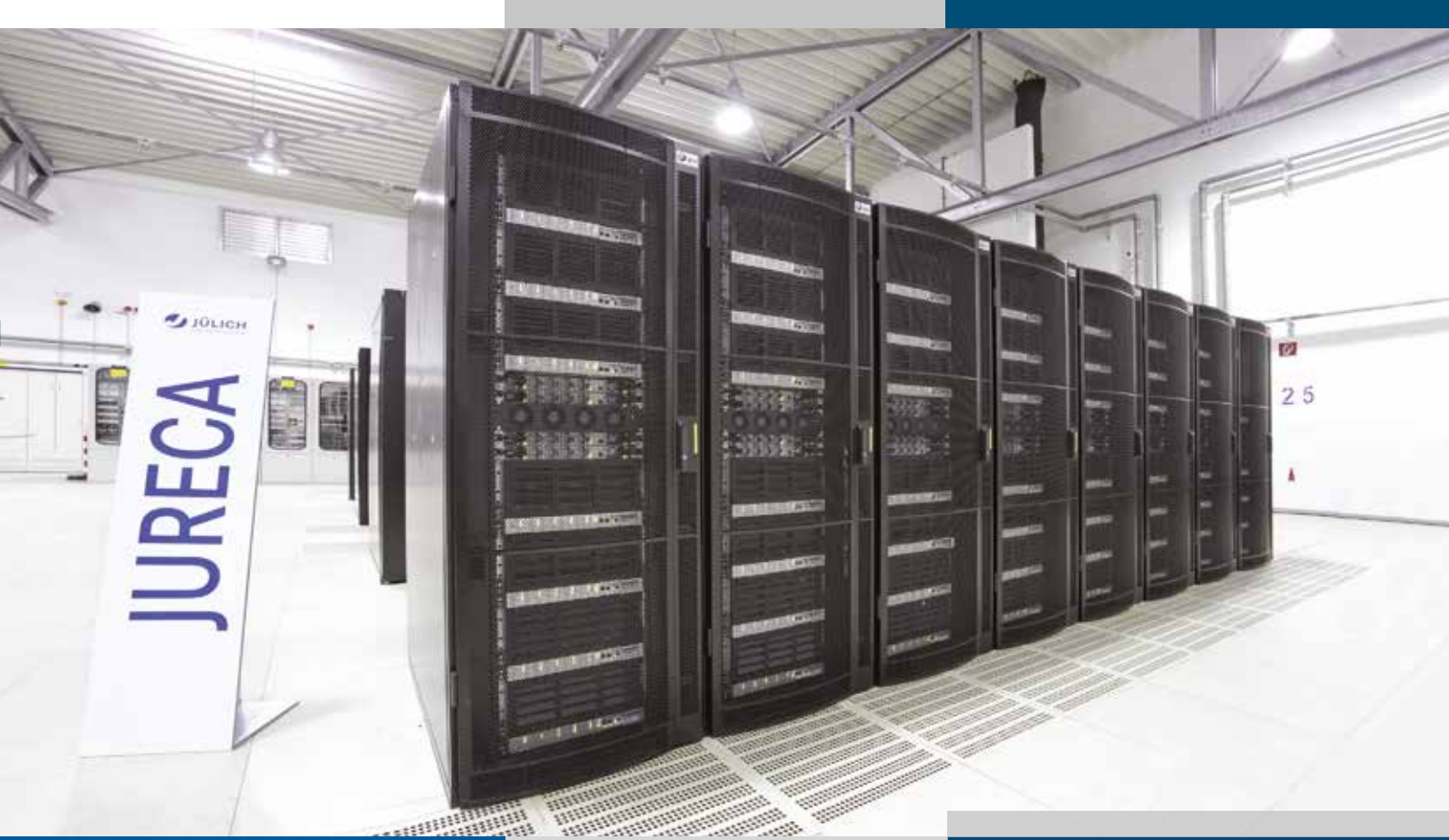
linearly with the number of processors used. In addition to code specially optimised for parallel processing, this often requires the use of completely different algorithms to those on a PC. There are currently only a few scientific groups that include a software developer with an appropriate background, so there are often reservations when it comes to working with large computers, and only a small range of potential applications are ported to supercomputers. This is exactly where the idea of simulation laboratories (SimLabs for short) comes into play: Each of the SimLabs at JSC offers a tailored range of support options concerning the development of software for supercomputers for its respective research community (climate, plasma physics, etc.).



**Figure 1:** The two Jülich supercomputers are used in many fields of science. JURECA runs a number of different programs, while JUQUEEN is specialised in highly parallel applications. Biology-related projects currently account for approximately 16% (as of Nov 2015) of the allocated computing time on JURECA.

Source: © Jülich Supercomputing Centre





**Figure 2:** The JURECA supercomputer, put into service at JSC in 2015, is capable of approximately 2 quadrillion floating point operations per second using its 3,768 processors and 174 graphics cards.

Source: © Forschungszentrum Jülich

### THE SIMULATION LABORATORY BIOLOGY

The Simulation Laboratory Biology is responsible for looking after the life sciences community's supercomputer. It supports scientists with the porting, optimisation and scaling of their programs. Its offers also include courses on parallel programming in C++ and Python, as well as on visualisation, data analysis and machine learning in Python. What's more, the SimLab Biology functions as a point of contact for supercomputer users as well as for scientist from the fields of biology, biomedicine and biotechnology who are planning to apply for computing time.

In addition to its support tasks, each SimLab maintains its own research portfolio. The SimLab Biology's research focuses on the prediction of protein structure and molecular simulation. The prediction of protein structure seeks to combine in sensible ways various bioinformatics methods such as secondary structure prediction, the detection of folds, energy calculation and model valuation in complex workflows. With molecular simulation, in contrast, the Jülich researchers use the supercomputer as a type of supermicroscope in order to gain a better understanding of the molecular details of biological processes. One of the group's specialities is the development of methods which enable simulation at atomic resolution of large changes in the conformation of protein systems, e. g. during protein folding and the aggregation of peptides. Such processes, which last hundreds of milliseconds or even longer, are beyond the scope of standard methods such as molecular dynamics. Although calculating the movement of individual atoms can be easily parallelised, a huge number of minute time steps must be calculated consecutively in the case of molecular dynamics. The folding of a protein with 92 amino acids and an experimentally measured folding duration of

1 s was first simulated in 2013 using software developed at the SimLab Biology – ProFASi –, which is based on another method, the Markov Chain Monte Carlo simulation [1]. Were a typical time

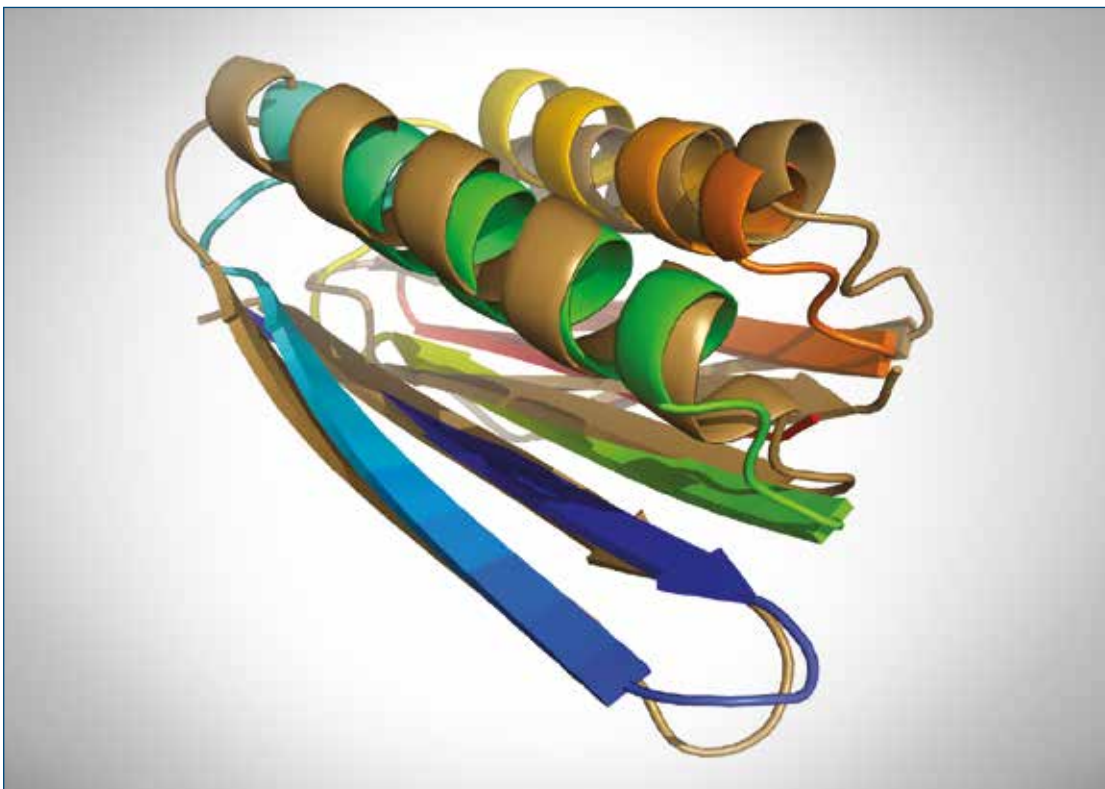
### THE JSC

The Jülich Supercomputing Centre (JSC) is one of the world's leading high-performance computing centres. It operates two of the largest supercomputers in Europe: JUQUEEN and JURECA. In addition to the operation of the computers, more than 200 staff at JSC develop new concepts for energy-efficient computer architecture, tools for software development on parallel computers, concepts for the analysis of very large quantities of data, as well as new highly parallel scientific applications for a large range of disciplines from climate research to plasma physics. A key task is to support scientific users across Europe. This ranges from the provision of computing time, allocated by means of a JSC-independent, competitive and excellence-oriented peer review process, to HPC- and subject-specific consulting for individual users.

step of 2 fs ( $=2 \times 10^{-15}$  s) to be applied, simulating the molecular dynamics of this process would require 500 billion steps and last many years. In addition to the simulations of protein folding, ProFASi is also successfully employed to investigate the aggregation of amyloidogenic peptides, which are linked to neurodegenerative diseases, as well as to simulate natively unfolded proteins (IDPs) [2], the structural preferences of which have been practically inaccessible for experimentation before now, but which play an important role in many cellular processes such as the development of cancer. In autumn 2016, interested researchers will be given a practical introduction to the Monte Carlo simulation using ProFASi at the international CECAM school organised by the SimLab Biology [3].

**Together with partners, the SimLab Biology also deals with a wide spectrum of other topics:**

- Together with computer scientists at the neighbouring RWTH Aachen, the SimLab Biology works on new analysis tools for molecular simulation data to enable statistical evaluation of up to 10 million independent protein conformations from a single Monte Carlo simulation.
- A doctoral thesis on the automatic optimisation of scientific workflows was supervised in collaboration with the University of Bonn and the JSC Federated Systems and Data department.
- The SimLab Biology was involved in the parallelisation of genome assembly methods as part of the BMBF project “NGSgoesHPC”. The partners in this case included the University of Cologne and the University of Dresden as well as the companies Bull and Intel.



**Figure 3:** A structure of the TOP7 protein (coloured), generated using the Markov Chain Monte Carlo simulation, compared with the experimentally determined structure (brown).

Source: © Sandipan Mohanty



**Figure 4:** An unfolded conformation of the TOP7 protein with all atoms depicted. The colour code along the molecule corresponds with that of Figure 3.

Source: © Olav Zimmermann

- The SimLab Biology took a step forward in the direction of cell simulation as part of a collaborative project with the Jülich Research Centre's Institute of Biotechnology 1 (IBG-1) (see portrait in *systembiologie.de* issue 9). As a cell appears gigantic from the point of view of molecular simulation, one strategy for simulating such large multi-molecule systems, is to combine multiple simulation techniques with different levels of detail. The SimLab Biology is therefore supporting the IBG-1 in the development of a 3D multi-scale model for the simulation of cellular processes. This BMBF-funded project simulates the diffusion of individual, macromolecular enzymes within the tightly packed cell interior by means of Brownian dynamics, while the far more numerous metabolite molecules are treated continuously, i. e. as a concentration field. By coupling these two techniques it is intended to simulate system sizes of practical relevance on supercomputers, thereby enabling to study e. g. the effects of immobilisation, crowding and multi-enzyme complexes on enzymatic reactions.

#### NEW OPPORTUNITIES FOR LIFE SCIENTISTS

Supercomputers are more versatile than is commonly believed and are often more energy-efficient and faster than other alternatives when performing tasks that require significant processing power. Over the coming years, the number of bioinformatics subfields which already make use of these opportunities, such as phylogenetics and protein structure prediction, will increase. In particular, it is the processes that not only search among large quantities of data, but apply complex calculations to these, e. g. training in prediction methods with the aid of machine learning methods, which could profit from the high processing power and the quick transfer of data between processors.

Future applications for extracting the currently largely unexploited information from biological and medical publications, merging them into huge semantic networks and using these resources to aid in the computer-supported development of systems biology models, for example, are already being devised. The first models of an entire cell have already been reported. However, even the creation of a model for a simple cell that within the simulation behaves like its real counterpart and hence would permit predictions on the system level, provides many challenges for both life scientists and computer scientists. But given the current progress being made within both disciplines, a model of this kind could become reality in just a few years. By this time, an exaflop computer could be in place at JSC, a machine which is capable of performing  $10^{18}$  calculations every second, 200 times more than

JUQUEEN, currently the fastest computer at Jülich. JSC is currently cooperating with computer manufacturers to conduct research into new energy-efficient computer concepts. The aim here is to enable building an exaflop computer that would also allow for simulation of an entire cell at an acceptable level of energy consumption.

#### REFERENCES:

- [1] Mohanty, S., Meinke, J.H., Zimmermann, O. (2013) Folding of Top7 in unbiased all-atom Monte Carlo simulations. *Proteins* 81:1446–1456.
- [2] Jónsson, S.A., Mohanty, S., Irbäck, A. (2012) Distinct phases of free  $\alpha$ -synuclein—a Monte Carlo study. *Proteins* 80:2169–77.
- [3] CECAM school  
Atomistic Monte Carlo Simulations of Bio-molecular Systems  
Jülich, September 19 – 23, 2016  
<http://www.cecama.org/workshop-1339.html>
- [4] Kondrat, S., Zimmermann, O., Wiechert, W., von Lieres, E. (2016) Discrete-continuous reaction-diffusion model with mobile point-like sources and sinks. *The European Physical Journal E*, 39:11.

#### FURTHER INFORMATION AND CONTACT:

Forschungszentrum Jülich GmbH  
Institute for Advanced Simulation  
Jülich Supercomputing Centre (JSC)  
52425 Jülich, Germany

[http://www.fz-juelich.de/ias/jsc/EN/Home/home\\_node.html;jsessionid=67499B626A1241B6CF20E6479C2F8E5F](http://www.fz-juelich.de/ias/jsc/EN/Home/home_node.html;jsessionid=67499B626A1241B6CF20E6479C2F8E5F)

#### CONTACT PERSON:



**Dr. Olav Zimmermann**  
Simulation Laboratory Biology  
[olav.zimmermann@fz-juelich.de](mailto:olav.zimmermann@fz-juelich.de)  
[www.fz-juelich.de/ias/jsc/slbio](http://www.fz-juelich.de/ias/jsc/slbio)

# new opportunities for medicine

## Medical informatics will strengthen research and improve patient care

### Interview with Heyo K. Kroemer

Medical data is collected during every visit to the doctor or hospital. Biomedical research is also producing an ever-increasing quantity of data. Using this predominantly digital information in the best way possible is the aim of the funding concept “Medical Informatics” presented by the German Federal Ministry of Education and Research. Heyo Kroemer, president of the German Medical Faculty Assembly, speaks about the opportunities that the unified management of this data would present.

*systembiologie.de: Herr Kroemer, what is the state of medical informatics in Germany today?*

**Prof. Dr. Heyo Kroemer:** Digitalised healthcare has great potential and is actually indispensable. However, standardized data management does not exist in Germany. Almost every hospital, every location has its own system. Some of these systems are very well developed and work wonderfully. But sometimes you encounter situations in which one clinic has a fantastic data management system and the one next door is still working with paper. This is entirely removed from current needs and modern possibilities.

*What are the current problems?*

The main problem is the lack of a standardized data management system. There is no universally accepted electronic patient file, and no long-term concept that enables the standardized management of data both within individual university clinics, and also between clinics.

*Where do you see the greatest potential in medical informatics?*

There is immense potential to improve patient care. An electronic patient file, in which every general practitioner, specialist and clinician can access a medical history with diagnostic findings and administered treatments would be fantastic progress. Research would also benefit. Vast quantities of clinical data are stored in the university clinics, but cannot be used for research purposes. If the available data were correctly integrated and analysed by research, treatment options for clinicians would also improve significantly.

*Can you give us a simple example of this?*

For instance if a patient suffers from a rare disease that affects only 50 people in Germany, the attending physician will have virtually no chance of finding other affected individuals or sharing information with colleagues who are responsible for treating them. A database interlinking individual locations would make this possible.

*A major incentive for the clinic is to improve the procedures for providing healthcare. How do you see this happening?*

Nowadays, if a patient comes to a clinic, the attending physician arranges tests which may duplicate others that have already taken place earlier. If they were to have access to an electronic patient file containing all the relevant information, many needlessly repeated tests could be avoided. This is one of the reasons why the USA has implemented these files on a more or less nationwide basis.





Heyo Kroemer is professor of pharmacology and personalised medicine at the University of Göttingen, dean of university medicine and spokesperson for the Board for Research and Teaching. In addition, he is a member of the DFG senate's committee for Collaborative Research Centres, and he is president of the German Medical Faculty Assembly, which is the association of medical training- and research facilities in Germany (Photo: Irene Böttcher-Gajewski/University of Göttingen).

*In your view, what must be done in order to improve the situation?*

The very first step has already been taken: We have recognised the problem as such. What we need now are intelligent solution strategies that extend beyond individual locations. And I believe that the BMBF's funding concept for medical informatics can contribute significantly towards this: It promotes cross-locational association of university clinics, as individual applications are no longer possible.

*To what extent are they prepared to work together on this problem?*

All parties involved are aware that there is a great deficit in this area. So there is a clear willingness to address the problem. It will be possible to resolve this issue, if parallel programs are initiated, that – unlike the current medical informatics initiative – aim at improving health care rather than predominantly benefitting research.

*What do you see as the biggest obstacles?*

They will become apparent as the programme progresses. Data protection in Germany is certainly the biggest obstacle. However, experience has shown that this obstacle can be overcome, if it is addressed at an early stage, if processes are designed transparently, and if data protection specialists are involved from the beginning.

Another obstacle is the unwillingness of politicians and individual university clinics to provide sufficient funding for medical informatics. Currently, less than one percent of the revenue of a clinic is spent on information technology (IT). This expenditure must be increased to five or six percent in order to become competitive.

*Is the timeframe suggested in the funding concept sufficient?*

The funding concept is very much long-term-oriented. And this is a good thing, as the programme will fundamentally alter procedures in the university clinics. We cannot achieve this overnight.

*In your opinion, what are the funding concept's strengths?*

The strategy's greatest strength is its aim to consequently broaden the scope of IT-strategies beyond the scope of individual clinics. This is a new idea in Germany. I believe also that more areas than just IT will benefit from the programme. In future it will simply be too expensive to maintain what we currently practice, namely a university medicine which provides all technologies in every single location. In my view, the concept offers a great incentive to address this fundamental problem.



Photo: everythingpossible – Fotolia.com

### *What effects will we see outside the university clinics?*

One of Germany's greatest weaknesses is the almost complete lack of communication across sectors. For instance if you are admitted to a care facility after a long period in a hospital, it would stand to reason that all electronic patient data are passed on. But unfortunately this is not the case. This presents another big challenge.

### *So do medical informatics affect only the sectors named earlier?*

The exchange of medical data is a sensitive area. Naturally, not only the purely technical and medical problems, but also the corresponding legal questions and ethical concerns must be discussed in medical informatics. Exactly this is addressed by the funding concept. It proposes the initiation of accompanying measures to elucidate these aspects and to trigger a broader public debate throughout society.

### *Where will Germany be in ten years' time?*

I'm optimistic. In ten years' time, we will be in a situation in which electronic patient files, linked with corresponding biobanks, are universally used in university medicine, and in which the sharing of data within and between university clinics is easily possible.

*Interview by Katja Nellissen. Editing by Marco Leuer and Bettina Koblenz.*

### **Information on the BMBF-funding scheme "Medical Informatics" can be found under:**

<http://www.gesundheitsforschung-bmbf.de/de/medizininformatik.php>

---

### **What is medical informatics?**

Medical informatics "is the science of the systematic generation, management, storage, processing and provision of data, information and knowledge within medicine and healthcare. It aims at contributing to the design of the best healthcare possible."

(Definition of medical informatics, Deutsche Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e. V. (German Association for Medical Informatics, Biometry and Epidemiology))

---

### **Contact:**

**Prof. Dr. Heyo K. Kroemer**  
Director of Research and Teaching  
Dean of the Medical Faculty  
UNIVERSITÄTSMEDIZIN GÖTTINGEN  
GEORG-AUGUST-UNIVERSITÄT  
Göttingen, Germany

[www.med.uni-goettingen.de/de/content/ueberuns/dekanat.html](http://www.med.uni-goettingen.de/de/content/ueberuns/dekanat.html)

# i:DSem – integrative data semantics in systems medicine

A new initiative by the German Federal Ministry of Education and Research promotes innovative data management in biomedical research

by Christian Rückert

The aim in systems medicine is to collate and use enormous quantities of patient-related data: It is important in order to be able to find the best possible treatment for a specific patient; the knowledge which stems from research and experience with many other patients provides this ability. However, this is still not possible today. In order to further research in systems medicine, a multitude of clinical data, which is partly unstructured and exists in varying levels of quality, as well as disease-relevant, molecular data must be structured, compiled, prepared and finally made available to the medical staff so that they can decide on the most effective treatment.

## Integrative data semantics – services for systems medicine

Medical practices and hospitals collect huge quantities of data on their patients: blood values, the list of prescribed medication, previous illnesses, X-ray, MRI or ultrasound examinations, pathological findings, notes on allergies and intolerances and much more. While some of this data is available in digital format and structured, as is usually the case with blood values, this is by no means the case for all data. The results of imaging techniques are often available in digital formats. But what they show is interpreted by the medical professionals based on their training and experience. Automated analyses on the basis of patterns are still not possible in this instance.

Information that is important for the treatment of patients is also contained within doctors' letters and reports. However, the majority of these records are not digitalised or at most, have been attached as a scan, making the content useless for IT systems. Even if the text is recognised, they are always still composed using natural speech and are therefore hardly suitable for further direct, automated processing.

Besides this clinical data, the future will see an increasing significance of molecular data. Results from high-throughput analyses, patient or tumour genome information, proteome and metabolome analyses will provide future practitioners of systems medicine with presently inaccessible clues for diagnosis and treatment. In this respect, integrative data semantics addresses a fundamental aspect of innovative data management. It makes a significant contribution towards the de facto standardisation of data sets, which is essential to all forms of data integration.

This multiform data must be made available before it can be exploited by computers. Structuring, semantic description and integration are the keywords that apply before it comes to analysis and the aggregation of information. Currently, there is also a lack of suitable integration concepts and systems. Ontologies must be developed for the various applications, and data structures need to be established.

The initiative "*i:DSem – Integrative Datensemantik in der Systemmedizin*" (*i:DSem – Integrative Data Semantics in Systems Medicine*) of the German Federal Ministry of Education and



Photo: santiago silver - Fotolia.com

Research has embraced this issue. Embedded in the “Future and interdisciplinary topics” module of the *e:Med* research and funding concept, *i:DSem* provides the opportunity to act on this need for innovation.

At the same time, integrative data semantics provides the basis for the development of technical and methodological innovations, so that the content of the existing data and, most importantly, the content of the even greater quantity of future data created within life sciences can be utilised, thereby allowing research results that could benefit patients to be used in clinical application more quickly.

### *i:DSem* consortia begin in 2016

Beginning in 2016, the German Federal Ministry of Education and Research will be providing eight research consortia with around 20 million euros in funding for up to five years. The selected projects focus on different epidemiologically significant diseases. Five consortia are working on different approaches which will make it possible to provide better medical care for cancer patients and should facilitate systems medicine-related research in this area.

The integration of sequence data in regard to genetic changes in cancer cells compared with healthy somatic cells, the investigation of cell signal networks or the usability of data by medical practitioners, will be the point of focus, depending on the core topics of these cancer-related projects. So that doctors can quickly and intuitively access the large quantity of existing knowledge when treating cancer, it is not just the stock data

which must be comprehensibly described and made available for computer systems. The development of information presentation systems for doctors which can be operated intuitively is also being pursued within the scope of these projects. These are interactive, clinical applications that allow the doctor – sourcing knowledge of genetic changes in the case of the individual patient for example – to use a computer to select the best treatment for the patient from among the range of known treatment options.

Between the semantic description of the data and its application in patient care, there is another level of processing, barely visible to the end user but still of the utmost importance: In order to display and classify bases for decision-making and potential choices, the computer applications require phenotype profiles, diagnostic classifiers, risk profiles and disease models. This essential preliminary work, as well as tools for model simulation of biological processes within systems medicine also forms part of the work being carried out within the scope of the *i:DSem* projects.

In addition to oncological diseases, one project also focuses on neuro-degenerative diseases such as dementia. The aim is to make sure that available patient data can be used for analyses which are as holistic as possible. Building on this, the newly developed methods should then contribute towards improving early recognition of dementia and provide optimised patient stratification. Another aim is to identify new prognostic factors, as well as, in an ideal case, promote development of new treatments.





Another project is dedicated to spinal cord injuries: Within the industrialised nations alone, there are around 250,000 to 500,000 patients with acute spinal cord injuries, for which there are still no successful therapeutic procedures. This is primarily due to the unmanageable quantity of unstructured knowledge which lies concealed within enormous collections of relevant research literature. The project therefore aims to develop a new information system which will provide neuroscientists with comprehensive information. New methods for automatically extracting information from scientific publications must first be developed in order to fill this with knowledge. This will be accompanied by the development of a valuation method which will access this system and assist the scientist or medical doctor in selecting promising therapeutic concepts. Based on this information system, clinical studies should then be designed which promise great improvements in the treatment of patients.

A similar approach, but one which finds application in the field of transplantation medicine, forms the basis of another project: It focuses on the data protection-compliant description and valuation of clinical data stores from a number of different sources, with the aim of using these to develop prediction models. An IT platform is also to be created, which will aid in collecting harmonised and anonymised clinical data in a central database, to be made available for future use. As the project progresses, there are plans to use this database for systems biology-related modelling activities, particularly in the area of transplantation medicine. These models will be taken to predict possible complications in cases of stem cell transplantation, thereby improving the affected individual's quality of life.

In addition to integrative data semantics works in a narrow sense, basic services which are essential for working with personal medical information must be developed and made available. Ontologies and disease models are just as much

a focus of these projects as the integration of different data sources and their optimal valuation. The systems will only become part of day-to-day clinical life if medical professionals are able to access the data at their end with little effort and it is given to them in a clearly presented way.

### From the basics to application

The translation of results has always been regarded as an important element of the *iDSem* funding measure. After three years, at the end of the development phase, the research consortia will be evaluated and will only be able to proceed to the two-year translation phase if they receive a favourable assessment. When submitting the project proposals, the scientists had to state the usefulness of their work and they had to include an initial application in the translation phase schedule as a proof of concept.

So it is not surprising that most of the consortia have clinics or software providers from the clinical sector acting as translation partners. In many cases, the information provided by colleagues working in a clinical setting is also an integral part of project activities. Symposia, workshops and courses are already scheduled. In most of the projects, medical staff, that subsequently will be using the developed systems in a clinical setting, is already contributing to the development phase.

---

### Contact:



**Dr. Christian Rückert**  
Project Management Jülich  
Forschungszentrum Jülich GmbH  
c.rueckert@fz-juelich.de

[www.ptj.de/idsem](http://www.ptj.de/idsem)

# “projects can fail in a good way”

## Interview with Olaf Wolkenhauer

Olaf Wolkenhauer was Germany's first professor for systems biology. He has taught at the University of Rostock for the last 12 years. In this interview with [systembiologie.de](http://systembiologie.de), he discusses his future research plans and recalls his first encounter with molecular biology. As he describes, when he made the career switch to systems biology, this engineer placed all of his eggs in one basket.

**systembiologie.de:** *You have been building a reputation at the University of Rostock for 12 years now. How has the systems biology scene in Germany changed over this period?*

**Prof. Dr. Olaf Wolkenhauer:** Systems biology has become a firmly established feature of the research landscape in this period. This is primarily due to the funding provided by the German Federal Ministry of Education and Research. They have actively supported interdisciplinary work.

*Looking back at how things started, how significant was systems biology seen to be back then in terms of a research approach?*

At that point in time, cooperation between the medical practitioners and biologists, and ourselves, the modellers, was still accompanied by a high level of risk. There was also serious doubt as to whether or not one would have any chance of obtaining funding of any kind. So the different disciplines found within systems biology have definitely become more integrated over the course of time. Nowadays, we no longer need to think of convincing arguments to support the creation of mathematical models. In the cases of many projects, the scientists can clearly see from the outset that cooperation is beneficial to both sides.

*So, who is more likely to approach whom? Do the biologists and medical practitioners contact the modellers or vice versa?*

In the early years it was me, the modeller, approaching the biologists – initially without any success at all. I wanted to understand cell behaviour in terms of processes, study networks and investigate the time sequences that affect cell functions. Initially, the biologists did not share my interest, though. Thankfully this has changed. Nowadays we simply come together and it works out. And that's great.

“I'll take care of the basic rules”

*You originally studied control technology. How did you end up working in the field of systems biology?*

I had always had a fundamental interest in biology. I would have liked to study biology, but my parents did not allow me to do so. “We will only support you if you study engineering”, is what they told me. In hindsight this was a good thing, as over the course of my studies I learned how to describe dynamic systems and temporal processes using mathematical models. When I referred back to my former biology books, these networks, which we now often study in systems biology, were shown as the endpoint. This means that experiments were performed to demonstrate that individual components interacted with each other and that was all. But having learned how to describe processes, I would have perceived this as the starting point from which to create mathematical models. It seemed completely logical to me. At the start of the 1990s however, this was fairly uncharted territory.

*What does your research focus on now? What issues are you currently working on?*

My favourite subject is cell communication, for example within the context of cancer research. We no longer look at cells in isolation, but at how they interact with their environment. The realisation that we will only be able to understand processes such as metastasis if we study networks was a very important



Olaf Wolkenhauer (Source: University of Rostock, ITMZ, Photo: E. Altrichter).

development for systems biology. Before now, we had spent our time zooming in closer and closer to find out more about molecules. However, in my view, the most important issue in future will be zooming out, which will allow cells to be examined in their respective contexts. I'm now going to take another radical step, putting all my eggs in one basket: I'm going to take care of the basic rules. In order to understand disease processes, we require new techniques which will allow us to zoom out. I would like to establish the link between concrete results and the overarching issue. This involves piecing together the many details to produce a complete picture.

## “My director was concerned”

*What might these basic rules you have spoken of look like?*

I want to contextualise current knowledge. This has already been done, but only in review articles. In these analyses, the resulting hypotheses are formulated verbally. I would like to do something similar using mathematical interpretations. If biologists say that this or that regulates something else, I want to translate this word “regulate” and place it in a formal, mathematical context. This could fail spectacularly, but I'm prepared for that. This means that as part of our group, we are working on projects in which we are trying to uncover the details. But I also want to obtain a bird's eye view. If there really are such regularities in biology, then general predictions are possible. Every model carries this risk. It always comes back to the question of how I can simplify things within a model without the model losing its explanatory power.

*You just spoke of how you now once again wish to place all of your eggs in one basket and take a risk, just as you did before. Were you referring to your move into the field of systems biology?*

That's exactly right. I was a control engineering graduate. The director of the institute and other older colleagues were concerned when I decided to make the move to systems biology. They were afraid I might ruin my career by shifting my focus to biology. However, unlike back then, I don't need to worry nowadays, as I am fortunate enough to be a professor. Now I have that freedom.

*Your particular interest in the disease progression aspect of biology also stems from personal experience.*

Yes, that's correct. A key moment for me was my father's illness. The doctors said that nothing could be done. A molecule contained in his blood was slowly destroying his lungs. I tried to understand this. I went into a book store, where I encountered cell biology for the first time. Prior to this, I had considered biology to be synonymous with animals and plants, and closely linked to my fascination with nature. And then all of a sudden it was cells that held fascination for me. It became clear to me: Why, I've studied this! Ultimately, it is also about dynamic processes. For me, that was the initial spark. I saw that I could also apply what I had learned to this field. Of course, I then discovered that biology is far more complicated than engineering.

## "Complexity is encouraging, not discouraging"

*Did you ever find this experience frustrating?*

No, nature is just this complex. But this diversity is also what makes it so beautiful and so exciting to study. This has always motivated me. Complexity is encouraging, not discouraging. And this was also the case when we discovered during the course of certain projects that things were even more complex than we had imagined. Of course, this is undesirable first and foremost, as it means we haven't achieved our goal. Nevertheless, if one is aware of this complexity and is able to continue, progress has been made. This is how projects can also fail in a good way. The reasons for our failure can also help to advance our knowledge.

*What would you recommend to young scientists hoping to work in systems biology?*

I don't know if it is a good idea to set up our own academic programme for systems biology. We only teach from master's level upwards and our experience of this has been very good. People first study physics, informatics, biology or medicine. Only then do they specialise and attempt to look beyond the horizon. This is not necessarily successful if decided right from the outset. So I do recommend acquiring specialist knowledge first then broaden your horizons with a masters degree or doctorate.

*You have also considered the philosophical aspect of systems biology. What was your most important insight in this regard?*

Arthur Schopenhauer is someone whose work interests me greatly. It helps me when I am reflecting on how to approach something. The philosophers examine things with a bird's-eye view and use this to watch how the sciences evolve. It may well be that in a few years' time we will laugh at many of the things we did in systems biology, such as the study of individual signalling pathways. It could be regarded as naive to believe that a subsystem viewed in isolation could be used to understand an organisation in its entirety, with all its complex interactions. And philosophy helps to provide that bird's-eye perspective. That's why I find it fruitful to work together with philosophers of science, because this shows you where things went astray in other areas such as physics and the role that modelling played in these instances. Engaging with philosophy allows one to reflect on one's own actions. Unfortunately, day-to-day life often keeps us so busy that barely any time remains for such reflection.

*This interview was conducted by Melanie Bergs and Gesa Terstiege.*

---

### Contact:



**Prof. Dr. Olaf Wolkenhauer**  
Systems Biology and Bioinformatics  
University of Rostock  
Rostock, Germany  
olaf.wolkenhauer@uni-rostock.de

[www.sbi.uni-rostock.de](http://www.sbi.uni-rostock.de)



# “the liver is the first choice of both scientists and our network”

## Interview with Peter Jansen Programme Director for the LiSyM research network – Liver Systems Medicine

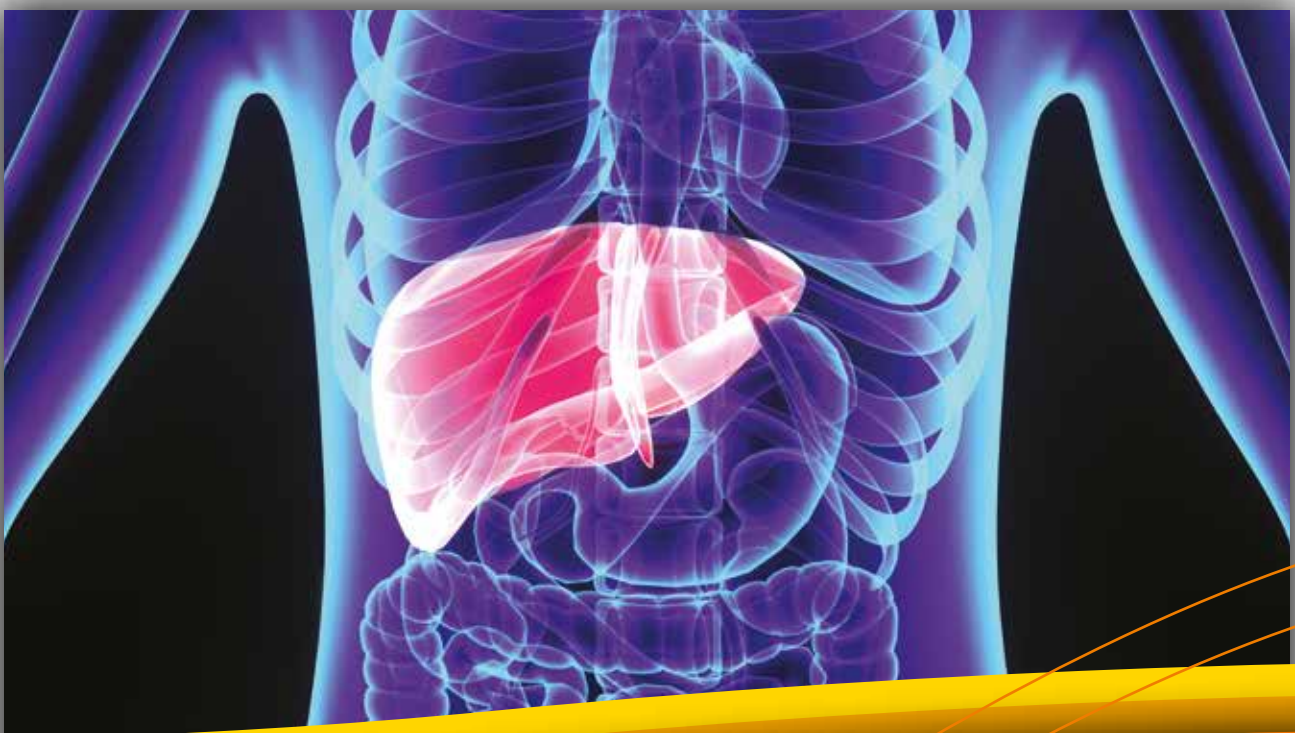
For Peter Jansen, research and working with patients go hand in hand. The liver has been the main focus of his medical professional career since the 1970s. In his role as Programme Director, the Dutchman has been acting as a representative and coordinator of the new LiSyM research network since the beginning of this year. In his interview with *systembiologie.de*, he discusses his goals for the coming years and explains his fascination with the liver.

*systembiologie.de*: What attracted you to the role of Programme Director for the LiSyM network?

**Prof. Dr. Peter Jansen:** Working as the Programme Director means that once again I'm entering new territory after decades in liver research. Like many of my colleagues, whose background is in clinical research, I am not used to collaborating with modellers. On the one hand we have *in vitro* experiments and on the other, clinical data. There is always a gap, which exists between these two aspects, as we don't know if the laboratory results can be translated in the clinic and how this can be achieved. This is the point at which the modellers are able to create a link. I'm particularly excited about this collaboration and I'm very curious to see how it will ultimately be able to benefit medical research. I'm already certain that it will trigger

---

Peter Jansen has been fascinated by the human liver since the seventies. “There is such a great quantity of knowledge and data relating to the liver that research in this area never comes to a standstill,” he says.





The Dutch professor of medicine, Peter Jansen, expects that his role as Programme Director for the LiSyM research network will provide him with completely new insights (Photo: Academic Medical Center in Amsterdam).

a fascinating scientific debate. The German Federal Ministry of Research and Education has created something entirely new.

*What are your goals for the coming years? What would you like to achieve within the network?*

My wish is for toxicologists, drug developers and clinicians to view the LiSyM network as a prototype for future medical research. Before now, transferring the results of molecular *in vitro* research or animal testing directly to the clinic – and thus to patients – was frequently very disappointing. There are big differences in the effects of a drug. That's why *in vitro* experiments and animal tests unfortunately often produce unrealistic expectations in regard to clinical success. The modellers will hopefully be able to close these gaps and contribute towards ensuring better understanding when results are passed on, both at the molecular level and that of the organism as a whole. Perhaps then in future we will not even need animal testing.

*You have been actively involved in liver research since the 70s. What is it about this organ that you find so fascinating?*

Everything becomes fascinating as soon as you learn more about it. I have now spent nearly my entire life in liver research. I therefore also possess a great deal of knowledge on the subject and it becomes more interesting with every passing day. There are so many biological pathways which can be studied in the liver. Almost the entire field of biochemistry is the result of experiments on the liver and liver cells. There is such a great quantity of knowledge and data relating to the liver that research in this area never comes to a standstill. The importance of the numerous data sets and studies on the liver is becoming apparent within the context of LiSyM as well. There is not nearly such a vast amount of knowledge available with regard to other organs. The liver is therefore the first choice of both scientists and our network.

*In your academic profession you have to bridge the gap between research and medical practice. How do these two areas of your work mutually enrich one another?*

Whenever one meets patients and studies their clinical problems, one quickly realizes the limits of medical knowledge and this provides the motivation to address these problems in research projects. The outcome of this research has to be translated back into the clinic and this interaction provides the basis of progress in medicine. For instance, I look at the patients and their laboratory data and ask myself why serum

liver enzymes are elevated or why certain treatments have failed. Some of these questions are easy to answer but others are puzzling and answers are not easily found. When I was starting out back in the 80s, the researchers in the lab had no interest in what was happening in the clinic. Thankfully this is no longer the case. Now, when we, the doctors, come together with the scientists, they understand that the clinic holds many exciting issues for them to work on. For them it is highly motivating to know that their research is able to help people. In turn, it is exciting for the doctors to get to the bottom of the mechanisms behind the diseases that they see every day. This means that researchers and doctors motivate each other.

*In addition to your work as Programme Director for LiSyM, you also regularly find yourself beside a hospital bed – are you also thinking about retirement? Or does your family ever ask you about this?*

On the contrary. I have even accepted another position. I'm now also a guest professor in London and Maastricht. My family supports me in all my ventures because they know that this is exactly what I want to do. I also find that 65 is no age for a scientist to be retiring. I can contribute by sharing my knowledge with young scientists, which is where the decades of experience really pay off. I believe I can be a good advisor. As long as I am able to do this, I will continue to enjoy it.

*This interview was conducted by Melanie Bergs and Gesa Terstiege.*

---

**Contact:**

**Prof. Dr. Peter Jansen**

Programme Director for LiSyM research network  
p.l.jansen@amc.uva.nl

[www.Hepaconsult.com](http://www.Hepaconsult.com)

## Liver Systems Medicine (LiSyM) research network

In addition to the measures in the field of systems medicine which have been announced thus far, the German Federal Ministry of Education and Research has since January 2016 supported the Liver Systems Medicine (LiSyM) research network within the scope of their overarching research and funding concept “e:Med – Paving the Way for Systems Medicine”. The research network unites national expertise from within the areas of application-oriented basic research, clinical research and systems biology for liver research. It involves 27 projects in total, including four junior groups. The development of illness-induced changes in liver function – through to organ failure – is studied at the different levels of organisation (cell-tissue-organ-organism) within four research focuses. The objective is to develop multi-scale models which will enable the key basic mechanisms of liver disease to be explained. On the basis of current progress in clinical liver research, this approach should allow patients to be assigned to risk groups as well as permit the development of optimised treatment processes. Through the implementation of this measure, the research results from “The Virtual Liver Network”, which received funding from 2010 to 2015, will be developed further for application in a hospital setting. The German Federal Ministry of Research will release 20 million euros over the next five years for this purpose.

# something to remember

## A systems biological view of the epigenetic basis of memory

by Tonatiuh Pena Centeno, Ramon Vidal, Magali Hennion and Stefan Bonn

One of the most distinguishing features that set humans apart from other organisms is their superior mental capabilities. Basic intelligence, the formation of memories, and emotions are of such importance to us that one could say they define who we really are; yet we still are not able to fully understand how they work. In the specific case of memory, aspects like the storage and maintenance of experiences are not well understood.

Nevertheless, it has become consensus within the research community that memory is linked to so-called *synaptic plasticity*, the ability to strengthen existing neuronal connections and form new ones. New studies support the hypothesis that changes of gene activities in brain cells that go beyond the genetic code, i. e. *epigenetics*, are associated with this plasticity and they seem to have a relevant role in the encoding of memory (Figure 1). This article discusses recent systems biological insights into the mechanistic role of epigenetic changes in learning and memory processes.

### Using a “Swiss Army Knife” to dissect memory

While it is clear that memory processes are governed by the strengthening and weakening of neuronal connections in networks of cells, genetic and pharmacological evidence points towards a role of epigenetic factors in learning and memory processes (Lopez-Atalaya & Barco 2014). The term *epigenetics* refers to the study of changes in the phenotype of an organism that are not caused by changes to the underlying DNA sequence. Epigenetic modifications leading to the alteration of the phenotype occur regularly and naturally during the lifespan of an organism, and can be triggered by environmental factors like age,

disease or even lifestyle; they can also be heritable or not. At least three mechanisms are currently considered to initiate and sustain epigenetic change, namely: DNA methylation, histone modification and non-coding RNA associated gene silencing. Although great strides have been made on associating the effects of epigenetic modifications with a diverse set of biological processes (embryonic development, aging, health conditions and many others), our knowledge of their involvement in memory-related processes is still at an early stage.

This “lack” of knowledge can be severely reduced by applying systems biological approaches, the “Swiss Army Knife” of biology, to the study of learning and memory formation. Using state of the art high-throughput methodology, it is possible to obtain genome wide views of DNA and histone modification changes in distinct brain areas, specific cell types, at several time points before and after learning (Halder *et al.*, 2016). The analysis of whole genome chromatin modification and expression data in combination with published structured (e. g. databases) and unstructured (e. g. research articles) data can then be used to further our understanding of how we form memories and store them.

### So what insights can be gained by unleashing systems biological approaches on the brain?

#### Attention please

Modifications of histones, the proteins around which DNA is wrapped in the cell’s nucleus, at specific residues have long been known to play important roles in the regulation of gene expression (Figure 1). Some have been linked to active gene states, some to repressive gene states, and several of them

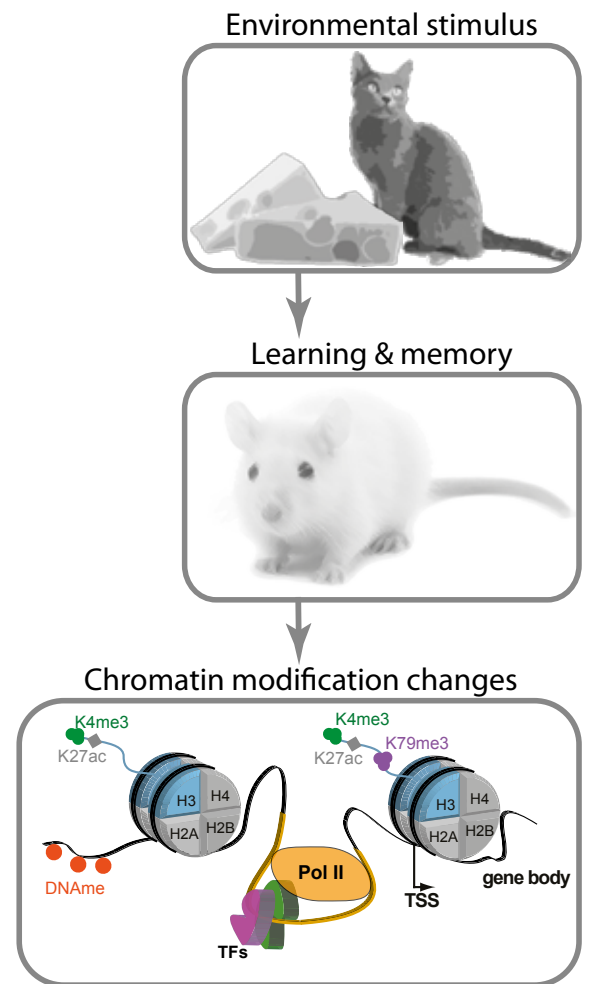


**Figure 1: Environmental cues lead to an adaptation of our behavior.** At the root of this adaptation is the ability to learn and memorize novel environmental settings. The formation and maintenance of memory requires changes in epigenetic modifications of the chromatin (Graphic: Stefan Bonn).

have been linked to learning processes. So what are the actual changes of histone modifications during learning processes and could they be a molecular correlate of memory? Surprisingly, and in contrast to most developmental data, genome-wide histone modification changes correlate rather weakly with gene expression changes during learning. On the other hand, early learning is accompanied by a “global” increase of activity-related and a “global” decrease of inactivity-related histone modifications on a large proportion of the genes. This effect is short-lived and coincides with the formation of memory (1 hour after learning) but is absent in the maintenance of memory (4 weeks after learning). Although these results would imply that histone modifications may not play a major role in memory maintenance, the observed global “activating” histone modification changes might shift genomes into an alert state. This time-dependent shift could be a molecular correlate of attention, as it might “prime” brain regions for future incoming stimuli and strengthen memory formation.

### Providing lasting impressions

While the observed learning-related histone modification changes might have caught your attention, results on DNA methylation changes in memory acquisition and maintenance might give you a lasting impression (Figure 2). While histone



modifications are rather short-lived and accompany the formation of memory, DNA methylation seems to be important to establish and keep memories for longer times. Thus, long-term DNA methylation changes in neurons are strongly linked to the differential expression and splicing of target genes. These targeted genes are involved in shaping the synaptic plasticity and wiring of neurons, giving further evidence for the functional importance of DNA methylation in learning and memory processes.

### More than just neurons and “glue”

When you think of the brain and the formation of memory, you automatically think of neurons. And yes, it is true that neurons do a lot of the heavy lifting that makes the brain work. But what about the roles of other cell types, especially glial cell populations? Are they really just the “glue” that keeps the brain to-



Research group of Dr. Stefan Bonn in January 2016 (Photo Daniel Riestler, DZNE-G).

gether? At least from an epigenetic perspective this question can be clearly negated, as gene-specific and global histone modification changes take place in non-neuronal cells to a surprisingly large amount. Especially during memory formation, non-neuronal cell types show short-lived changes in histone modifications, implying a novel yet undiscovered functional role of these cells in learning processes.

### Quo Vadis?

Systems biological approaches in Neuroscience are not mainstream and one of the reasons for this might be a healthy skepti-

cism towards “high input – low output” technologies. Recent research, however, showcases, that systems biology can deliver, i.e. by providing deep insights into the gene regulatory networks underlying memory. This understanding of the epigenetic basis of learning and memory processes in healthy subjects can now serve as a point of comparison to disease states such as Alzheimer’s Disease or natural aging. It would also be very interesting to understand the epigenetic changes within memory-forming networks of cells, by combining genetic labeling techniques with single cell sequencing and bioinformatics analyses.

Figure 2:



Memory maintenance is accompanied by DNA methylation (5mC) changes in cortical brain areas. When we learn, DNA methylation is changed at specific promoter and enhancer regions. Some of the changes persist for months, coinciding in time and space with the location of memory (Graphic: Stefan Bonn).

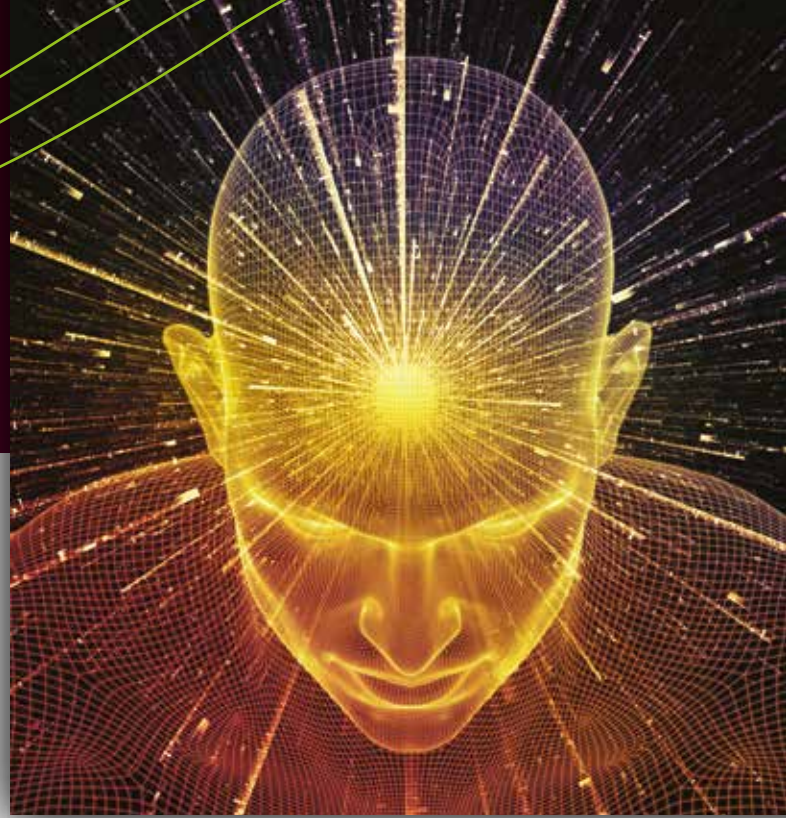


Photo: agsandrew – Fotolia.com

---

### Project partners:

#### **Dr. Bettina Schmid**

German Center of Neurodegenerative Diseases (DZNE)  
Munich, Germany  
Munich Cluster for Systems Neurology (SyNergy),  
Munich, Germany

#### **Prof. Dr. Christian Haass**

German Center of Neurodegenerative Diseases (DZNE)  
Munich, Germany  
Biomedical Center, Ludwig Maximilians University Munich,  
Munich, Germany  
Munich Cluster for Systems Neurology (SyNergy),  
Munich, Germany

#### **Prof. Dr. Andre Fischer**

Group of Epigenetic Mechanisms in Dementia  
German Center for Neurodegenerative Diseases (DZNE)  
Department of Psychiatry and Psychotherapy,  
University Medical Center Göttingen,  
Göttingen, Germany

---

### References:

Lopez-Atalaya JP, Barco A. (2014) Can changes in histone acetylation contribute to memory formation? Trends Genet. Dec;30(12):529-39.

Halder R, Hennion M, Vidal R, Shomroni O, Rahman R, Rajput A, Pena Centeno T, van Bebber F, Capece V, Garcia Vizcaino J, Schuetz A-L, Burkhardt S, Benito E, Navarro Sala M, Bahari Javan S, Haass C, Schmid B, Fischer A, Bonn S. (2016) DNA methylation changes in plasticity genes accompany the formation and maintenance of memory. Nat Neurosci. 19(1):102-10.

---

### Contact:



#### **Dr. Stefan Bonn**

stefan.bonn@dzne.de

#### **Dr. Magali Hennion,**

#### **Dr. Tonatiuh Pena Centeno,**

#### **Dr. Ramon Vidal**

Group of Computational Systems Biology  
German Center for Neurodegenerative  
Diseases (DZNE)  
Göttingen, Germany

[www.dzne.de/en/sites/goettingen/forscherguppen/bonn.html](http://www.dzne.de/en/sites/goettingen/forscherguppen/bonn.html)



# AptaBodies

## DNA aptamers as an alternative to antibodies in Western blotting

by Jasmin Dehnen and Frieda Anna Sorgenfrei

Please refer also to the Laborjournal Magazine 12/2015, pages 56-57, to view the article

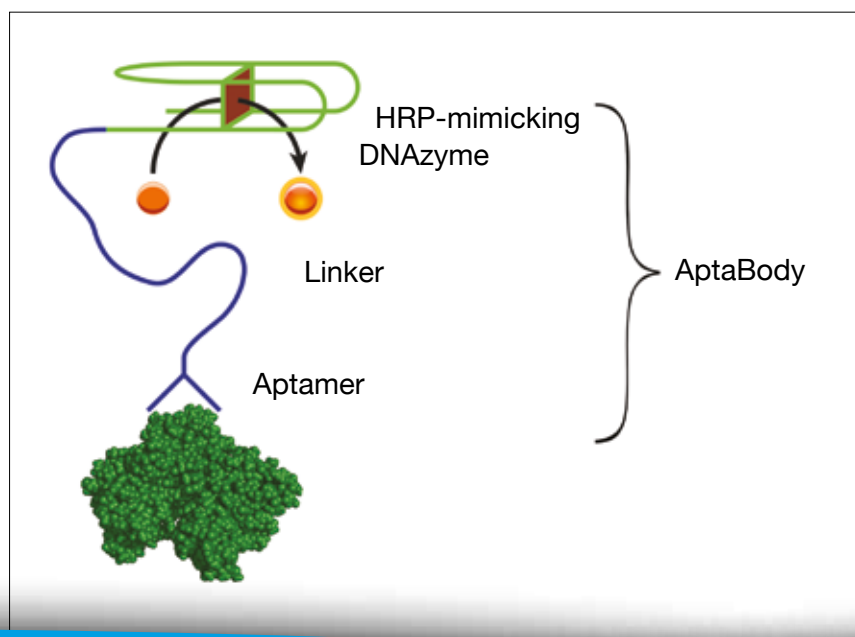
What do you get if you attach a protein-binding aptamer to a DNAzyme which mimicks the horse-radish peroxidase reaction? An AptaBody for Western blotting!

Every molecular biologist knows it, most have already asked it for advice and everyone curses it. No, we don't mean the boss – we're talking about Western blotting. This is where experimenters separate the protein in a polyacrylamide gel and subsequently transfer it to a membrane. Finally, they incubate this with a protein-specific antibody, which binds to the protein of interest. A secondary antibody that docks to the primary antibody is usually required for detection.

Since Harry Towbin introduced Western blotting as a technique for detecting proteins in 1979, it has caused us just as much sorrow as joy. Joy because the protocol is simple and the applications wide-ranging. Sorrow because it usually doesn't work the way it's described in textbooks. Over the years, various groups have refined the Western blot protocol with the aim of detecting rare proteins or evading the lengthy wash and incubation periods.

A critical point of the Western blot, which doesn't just cause headaches for life scientists, but also eats up funds, is the antibody. Almost every laboratory is familiar with situations in which the antibody fails to do what it's supposed to do and binds unspecifically or not at all. This is frustrating for

Figure 1: AptaBody



The AptaBody binds to a protein via the aptamer and using the HRP-mimicking enzyme, catalyses the chemiluminescent reaction of luminol and hydrogen peroxide (Photo: iGEM team Heidelberg).





The Heidelberg AptaBody crew with mentor Roland Eils (right front) at the iGEM finals in Boston (Photo: iGEM team Heidelberg).

researchers in two ways: Not only have they gone round and round in circles in terms of the experiment, they have also spent a lot of money buying a tube containing a useless protein solution, which in addition is taking up one of the highly contested spaces in the freezer.

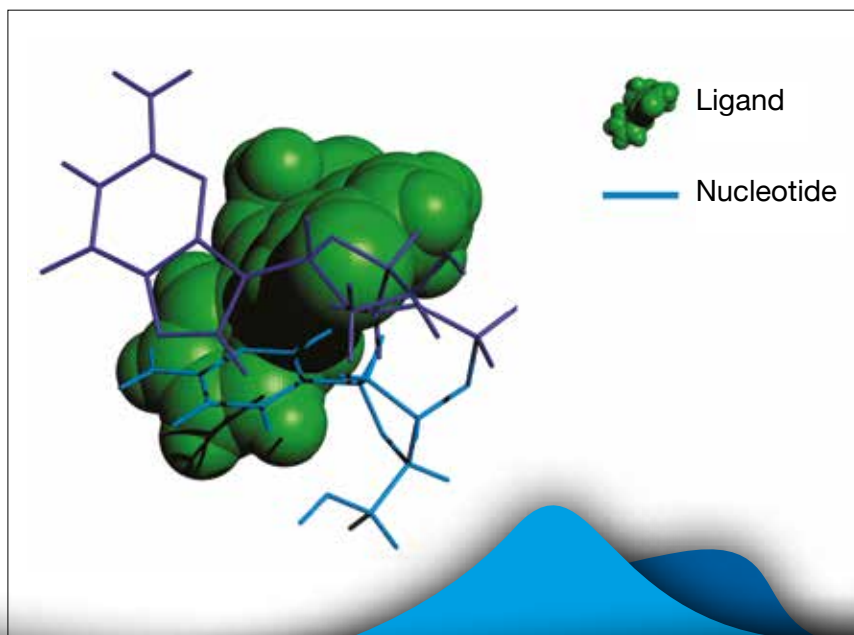
The search for a functioning antibody continues after every failed attempt and the tough procedure begins all over again, from ordering the antibody to detecting the protein on the membrane. And this is not just a case of becoming quickly strapped for cash, three years of Ph.D are gone in an instant. A marked primary antibody perhaps ensures a better night's

sleep, but the rude awakening comes at the very latest after you have seen the full bill. The whole thing is made difficult by the considerable number of proteins for which a matching antibody has yet to be developed. A cheap, quick and reliable alternative to antibodies would therefore be desirable. But what might this look like?

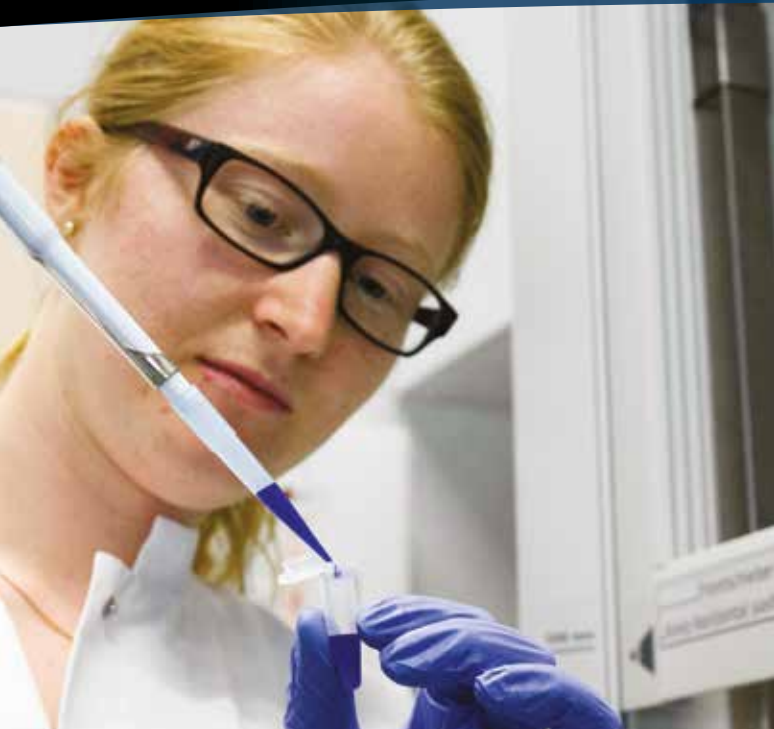
### Exotic DNAzyme

Ten students at the Heidelberg University asked themselves this question. They entered their solution in last year's iGEM (international Genetically Engineered Machine) competition in Boston. During the project identification phase, the group

Figure 2: MAWS – Calculating new aptamers



Using entropy minimization, the MAWS software calculates a new aptamer for a ligand, nucleotide by nucleotide (Photo: iGEM team Heidelberg).



Frieda Anna Sorgenfrei (Photo: iGEM team Heidelberg)

members submerged themselves deep in the world of molecules, finally focusing on functional RNA molecules, ribozymes for instance, which catalyse specific reactions. However, the Heidelberg iGEM team were quickly forced to realise that working with RNA had its drawbacks, as it can degrade very easily. Luckily biologists developed DNAzymes – DNA pendants on ribozymes – some time ago, which are significantly more stable than classic ribozymes.

The students succeeded in bagging the oddest functional molecules. The biggest haul was the 17-nucleotide-long, catalytically active DNA, the horseradish peroxidase-(HRP-) mimicking DNAzyme, discovered by Dipankar Sen's group from Simon Fraser University in Canada in 1998 (Travascio *et al.*, *Chemistry and Biology*, 9, 505–17). The classic HRP catalyses the chemilu-

minescent reaction of luminol and hydrogen peroxide, which gives off a blue light. Therefore, the enzyme is often covalently bound to the secondary antibody during Western blotting. But how does a DNA fragment, 17 nucleotides in length, manage to catalyse the same reaction as HRP?

The DNAzyme forms a G-quadruplex secondary structure, which enables hemin to bind at its centre. Finally, the complexed hemin catalyses the same reaction as HRP. So why not using the HRP-mimicking DNAzyme as read-out signal for the antibody substitute?

The question of how the new antibody substitute recognises the targeted protein remains unanswered. With this in mind, the Heidelberg team opted for another class of functional DNA molecules: aptamers. Aptamers are short nucleic acids that bind to virtually any target. To be able to use these for Western blotting, the student group naturally required protein-binding aptamers.

### Simple AptaBody protocol

They connected the HRP-mimicking DNAzyme to an aptamer, which recognises His tags. The result was a short single stranded DNA, which binds via the 5'-end to His-tagged protein and activates luminol via the 3'-end. Using this so-called AptaBody, the students were able to detect proteins in the cell lysate of *Escherichia coli*. The protocol for the use of AptaBodies for Western blots is simple and quick: Firstly, the AptaBody is boiled, so that it folds into its proper secondary structure during the subsequent cooling process after which hemin is added.

## iGEM

The leading international iGEM (International Genetically Engineered Machines) competition has been held at the renowned Massachusetts Institute of Technology (MIT) in Boston for over ten years. It is the world's biggest competition in synthetic biology and takes place every year.

To compete, teams of students and high school students spend the summer working on a research project of their choice, which they present along with the results during the "Giant Jamboree" in Boston. Here, medals and top prizes are awarded in over ten categories. The aim of the competition is to develop new biological systems that offer a solution to everyday problems or make a contribution towards basic research. For this purpose, the teams are given a combination of DNA sequences. These biological modules, so-called biobricks, can be combined in model organisms to create new systems. In addition, the participants organise events to draw the public's attention to the huge potential of synthetic biology. The iGEM competition is becoming increasingly popular. 280 teams from across North and South America, Europe, Asia and Australia took part in 2015.

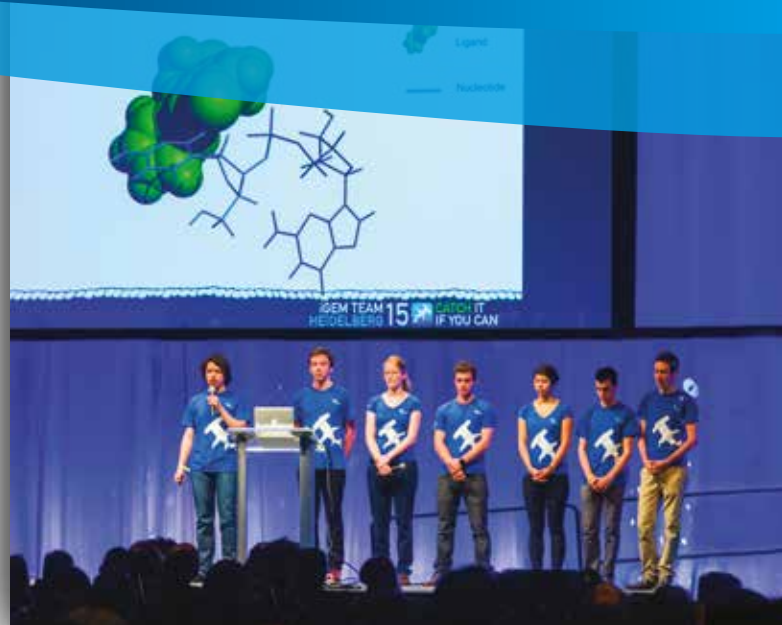
Most of the work has now already been done. All the life scientist has to do while AptaBody and membrane are incubated is to put his or her feet up and have a cup of tea. For the inveterate Western blotter, the last steps of the protocol are nothing out of the ordinary. Following a short wash cycle, which should remove the unbound AptaBodies, a luminol/hydrogen peroxide solution is pipetted onto the membrane. The AptaBody thus functions as a marked primary antibody, but it is many times cheaper.

AptaBodies would of course be of somewhat limited use if they only recognised His-tags. Ideally, one wants AptaBodies that bind to target proteins even without a His-tag. The prerequisite for this is protein-specific aptamers, which will be fused with the HRP-mimicking DNAzyme. Aptamers with a high degree of specificity and affinity for a protein ligand can be routinely selected using the SELEX process. This involves the incubation of a random DNA library with the target protein and the subsequent washing of unbound DNA sequences. All other sequences are mutated in order to optimise binding and then reintroduced to the protein. This step is repeated until aptamers with a maximum affinity for the target protein are obtained.

But let's face it, the SELEX process is laborious, time-consuming and not always successful, as it is highly dependent on the original DNA pool. Reason enough for the Heidelberg students to embark on a search for a new alternative to the traditional technique. Instead of remaining in the wet lab, they developed a software that generates the new aptamers *in silico*. The algorithm used to achieve this is based on the principle of entropy minimization. Using a known 3D structure of the target protein as a basis, it calculates the optimal aptamer candidates. It is not just down to chance that the name selected by the Heidelberg team for the software, MAWS (Making Aptamers without SELEX) is reminiscent of JAWS – the great white shark in Steven Spielberg's first blockbuster.

### All that's left is to order oligos

In practice, the whole thing should function as follows: Disappointed so far by antibodies or Western blots, the life scientists enter the structure of the target protein into MAWS. The software generates different aptamers which recognise specific protein epitopes. Finally, the researcher connects the



The AptaBody crew at the iGEM finals in Boston  
(Photo: Justin Knight, the iGEM Foundation)

generated aptamers to the HRP-mimicking DNAzyme, orders the DNA as oligo and has the ready-to-use AptaBodies on the bench the very next day.

Further experiments will be conducted to obtain a more detailed impression of the AptaBodies' function and to make them into a molecular biology tool with a wide range of applications. Western blots using DNA fragments that are a hundred times cheaper than antibodies – this sounds almost too good to be true, and yet the students are on track to achieve exactly that. It's not for nothing that the Heidelberg crew came third overall in the iGEM competition.

### Contact:



#### Frieda Anna Sorgenfrei

Student at the Heidelberg University  
Molecular Biotechnology (M.Sc)  
Heidelberg, Germany  
frieda.sorgenfrei@bioquant.uni-heidelberg.de



#### Jasmin Dehnen

Student at the Heidelberg University  
Biosciences (B.Sc)  
Heidelberg, Germany  
jasmin-dehnen@gmx.de

<http://2015.igem.org/Team:Heidelberg> and  
<http://2015.igem.org/Team:Heidelberg/Project/AB>



# events

## Computational Genomics approaches to Precision Medicine

12 – 23 September 2016, Berlin

### SPOT ON: THE COMPUTATIONAL GENOMICS SUMMER SCHOOL

Biology and medicine depend more and more on high-throughput methods with each experiment generating huge datasets. The skills needed for processing and analyzing such datasets are a discipline in itself, and it is crucial that all scientists conducting high-throughput experiments understand how the analyses are done.

The Berlin Institute for Medical Systems Biology (BIMSB) at the Max-Delbrück-Centre for Molecular Medicine (MDC) continued their Summer School Series with special focus on computa-

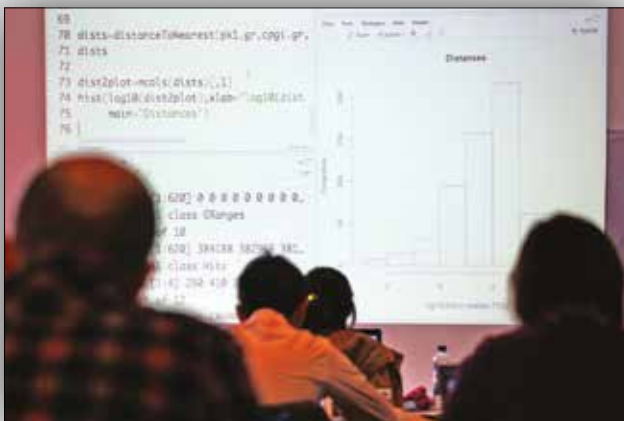
tional analysis of genomic data sets. This included lectures and exercises by renowned researchers from Memorial Sloan Kettering Cancer Center, New York University, and Zurich University, who taught approaches to analyze Next-generation sequencing data like RNA-Seq and Chip-Seq.

“Outstanding faculty... Amazingly well organized... Great people... Extremely interesting, even and especially for a wet-lab biologist” – the participants only had superlatives when describing the “Computational Genomics” Summer School.

Altuna Akalin, a BIMSB group leader and head of the bioinformatics technology platform, received support by Stiftung Charité to organize this course. He already has plans to continue this successful series by offering the next Summer School called “Computational Genomics approaches to Precision Medicine”. Again, computer scientists, biologists and especially clinicians are invited to apply for this course which will be held from September 12 - 23, 2016 at the MDC, supported by the BMBF.

For further information visit the website

<http://compngen2016.mdc-berlin.de>.



Impressions of last year's “Computational Genomics” Summer School (Pictures: Grietje Krabbe/MDC).







# COMPUTATIONAL GENOMICS APPROACHES TO PRECISION MEDICINE

## CONFIRMED LECTURERS

**ALTUNA AKALIN**  
Max Delbrück Center,  
Berlin Institute for Medical Systems Biology

**UWE OHLER**  
Max Delbrück Center,  
Berlin Institute for Medical Systems Biology

**NIKOLAUS RAJEWSKY**  
Max Delbrück Center,  
Berlin Institute for Medical Systems Biology

**NICHOLAS D. SOCCI**  
Memorial Sloan Kettering Cancer Center,  
New York, USA

**MARK ROBINSON**  
University of Zurich, Switzerland

**ROLAND SCHWARZ**  
University of Cambridge, UK

**CHRIS E. MASON**  
Weill Cornell Medical College  
New York, USA

**DAVIDE RISSO**  
UC Berkeley, USA

**12-23 SEP 2016** | Berlin  
Germany | Berlin Institute  
for Medical Systems Biology,  
Max Delbrück Center

**Application Deadline**  
**01 July 2016**

## COURSE MODULES

- Introduction to R & Bioconductor
- Statistics and Exploratory Data analysis
- Introduction to Next-gen sequencing
- RNA-seq analysis
- ChIP-seq analysis
- Variant calling and annotation
- Data integration and predictive modeling
- Metagenomics and human health
- Cancer classification based on HT-seq data



<http://compgen2016.mdc-berlin.de>

**MDC** MAX DELBRÜCK CENTER  
FOR MOLECULAR MEDICINE  
IN THE HELMHOLTZ ASSOCIATION  
**BIMSB** THE BERLIN INSTITUTE  
FOR MEDICAL SYSTEMS BIOLOGY



## Conference report

### 8th International Conference on Systems Biology of Human Disease – SBHD 2015

July 06 – 08, 2015, Heidelberg, Germany

#### ACQUIRING GREATER UNDERSTANDING OF DISEASES THROUGH SYSTEMS BIOLOGY

by Cornelia Depner

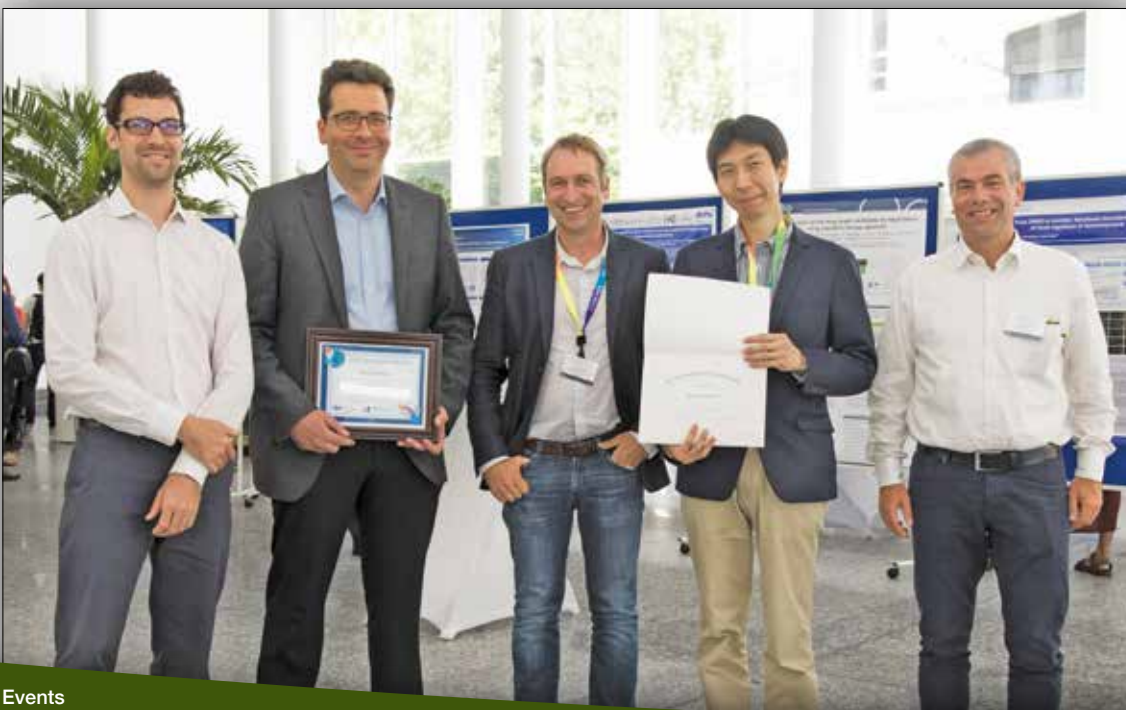
In the midst of beautiful summer weather in July 2015, the international “Systems Biology of Human Disease” (SBHD) conference was held for the third time at the German Cancer Research Center (DFKZ) in Heidelberg. The annual conference was launched a few years ago by Prof. Peter Sorger from the renowned Harvard Medical School in Boston and has developed since then into a German-American event, hosted alternately in Boston and Heidelberg. Among others, the 2015 conference was supported by the Swiss systems biology initiative SystemsX

and the e:Med systems medicine network created by the BMBF. Approximately 200 participants introduced the latest results of their research over the course of 43 talks and 81 poster presentations, with the aim of presenting and discussing the newest disease-relevant research results from the field of systems biology. The organisation committee also granted selected young researchers the opportunity to present their work in 20-minute lectures. Young researchers also gave short 5-minute talks to draw attention to their posters.

The overarching theme of the conference was the use of mathematical models and computer models for the observation and investigation of complex biological systems at all levels, from the genome to the entire organism. Through this, the conference provided broad insight into how systems biology research is applied within medicine, ranging from systematic pharmacology, computational biology and network reconstruction, up

#### SBHD 2015 award ceremony

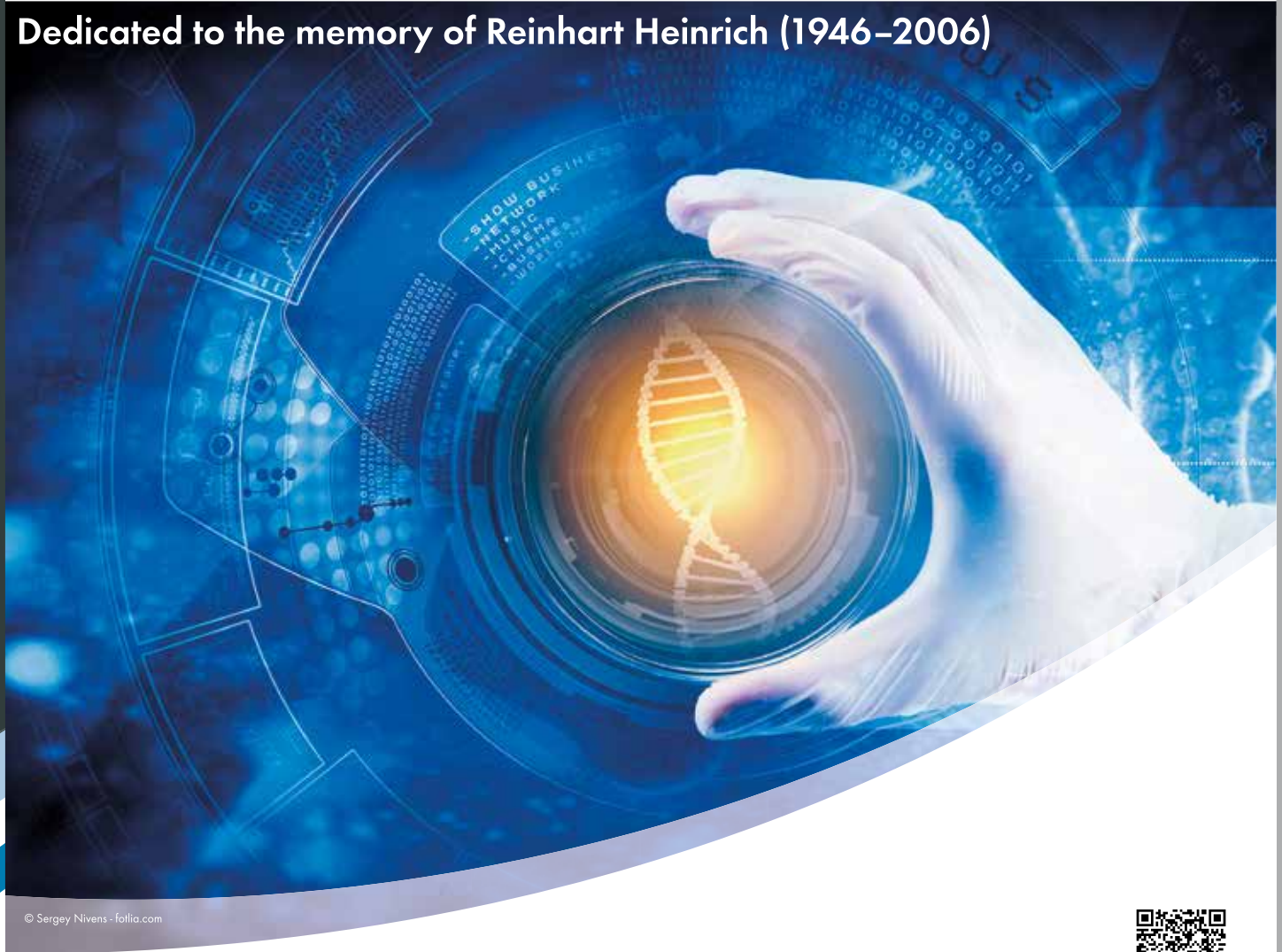
From left to right: Andreas Raue from Merrimack Pharmaceuticals, awardee Karsten Rippe (DKFZ Heidelberg), conference chair Roland Eils (DKFZ/Heidelberg University), awardee Kazuki Tainaka (University of Tokyo), Georg Draude from Chroma Technology Corp.



# ISGSB 2016 JENA

INTERNATIONAL STUDY GROUP FOR SYSTEMS BIOLOGY

Dedicated to the memory of Reinhart Heinrich (1946–2006)



© Sergey Nivens - folia.com



04–07 OCTOBER  
2016

Friedrich Schiller University of  
JENA, GERMANY

Abstract Deadline 31 May 2016

### Main themes

1. Infection modelling
2. Regulatory interactions and signalling
3. Plant physiology and development
4. Biological thermodynamics
5. Optimality principles
6. Multiscale systems medicine
7. Metabolic pathway analysis
8. Towards whole-cell models



© FSU Jena



© FSU Jena

[www.isgsb2016.de](http://www.isgsb2016.de)





Auditorium during the lectures (Photo: Editorial office systembiologie.de).



Poster presentations in the foyer of the communication centre at the DKFZ (Photo: Editorial office systembiologie.de).

to proteomics, single-cell transcription and protein analysis, as well as mathematical modelling of treatment resistance mechanisms.

Topics such as data processing and data management were also discussed. Professor Eytan Ruppim from the University of Maryland demonstrated one method with huge potential for medical research. His team developed a procedure that permits use of the increasingly large data sets acquired when analysing changes in tumour genomes. Here, a systematic comparison of data from a large number of patients can be used to predict potential new and patient-specific targets for treatments.

A special honour was bestowed upon associate professor Dr. Karsten Rippe from the German Cancer Research Center and Dr. Kazuki Tainaka from the University of Tokyo.

**Karsten Rippe** received the “**CSB2 - Prize in Systems Biology**”, sponsored by Merrimack Pharmaceuticals, for his interdisciplinary work on the structural organisation of the human genome and its pathological changes in cancer cells. Research data from Dr. Rippe’s department showed for the first time how cancer cells also misuse certain DNA repair enzymes by means of dysregulated epigenetic modifications, to once again extend the ends of chromosomes. In doing so, these cancer cells are able to divide indefinitely. In healthy cells, the ends of chromosomes, so-called telomeres, shorten a little during each cell division. Cells will no longer divide if these telomeres are not present anymore. This mechanism is deactivated in cancer cells, in which the ends of chromosomes are stabilised by the DNA repair proteins and the missing pieces of chromosomes are reattached. 29 proteins involved in the complex process of alternative lengthening of telomeres were identified by means of systems biology methods and the analysis of microscopy images and sequencing data. These proteins are now being used as a basis to identify new forms of treatment that will prevent this telomere extension.

**Kazuki Tainaka** received the “**Anne Heidenthal Prize for Fluorescence Research**”, sponsored by Chroma Technology Corp., for the development of the CUBIC method (*Clear, Unobstructed Brain Imaging Cocktails and Computational Analysis*). The CUBIC method is a newly developed process which allows extremely detailed images of the interior of individual organs and even entire organisms to be produced using a combination of chemical tissue decolourization and so-called light sheet fluorescence microscopy (LSFM). During this process, an integral part of haemoglobin, the ferrous protein complex in our red blood cells, is rinsed in amino-alcohols to prevent the absorption of light during microscopy, thereby making the tissue easier to view. The tissue can be made almost transparent by repeating this rinsing process multiple times. This CUBIC-perfusion protocol enables rapid whole-body and whole-organ imaging with single-cell resolution using light-sheet fluorescent microscopy. Eventually the technique allows scientists to acquire new understanding of organs’ three-dimensional structure and can be used for further studies at the organ and cellular level within the context of different diseases.

In addition to attending the lectures and awards, there was plenty of opportunity to socialise and network at the conference. The participants enjoyed a wonderful barbecue and music aboard a ship on the river Neckar. The highlight of the Neckar trip was a musical performance by Professor Uri Alon from the Weizman Institute in Israel, who interpreted his experiences in research to compose a song that he accompanied on the guitar to the general delight of the guests.



German Network for Bioinformatics Infrastructure (de.NBI)



## de.NBI Summer School 2016 From Big Data to Big Insights

Computational methods for the analysis and interpretation of mass-spectrometric high-throughput data

**Date:** 26<sup>th</sup>-30<sup>th</sup> September 2016

**Location:** Castle Dagstuhl, Wadern

**Website:** <https://goo.gl/ZRgxuV>



### Keynote Speakers

**Oliver Serang**

(Institute of Computer Science Metagenomics, FU Berlin)  
Protein Identification, FIDO

**Jürgen Cox**

(Max Planck Institute of Biochemistry, Martinsried)  
Protein Quantification, MaxQuant

**Juan Antonio Vizcaino**

(EMBL-EBI, Hinxton)  
PRIDE and ProteomeXchange

**Sebastian Böcker**

(Chair of Bioinformatics, Friedrich-Schiller-Universität Jena)  
Metabolomics, CSI:FingerID

**Lennart Martens**

(Department of Biochemistry, Ghent University)  
Proteomics Data Analysis, PeptideShaker

**Samuel Payne**

(Pacific Northwest National Laboratory, Richland)  
Pan-Omics, Active Data Biology

**George Rosenberger**

(Institute of Molecular Systems Biology, ETH Zürich)  
Protein Quantification, OpenSwath

### de.NBI Speakers

**Stefan Albaum**

(Center for Biotechnology - CeBiTec, Universität Bielefeld)

**Oliver Kohlbacher**

(Applied Bioinformatics Group, Eberhard Karls Universität Tübingen)

**Michael Berthold**

(Bioinformatics and Information Mining, Universität Konstanz)

**Martin Eisenacher**

(Medizinisches Proteom-Center, Ruhr-Universität Bochum)

**Robert Ahrends**

(Leibniz-Institut für Analytische Wissenschaften - ISAS - e.V.)

### Topics

To a large extent, proteomics and metabolomics are based on experimental data, for which mass spectra of a sample are assigned to specific biomolecules. Questions arising from basic and applied research can be answered, for example the differentiation of the protein composition between healthy and diseased persons or the early detection of diseases. As the omics technologies are rather wide-spread today, developed algorithms should be usable also by non-computer science experts and sustainably utilizable by them, e.g. as combinable modules in workflow systems.

## European Funding: ERACoSysMed and ERASysAPP

European funding for research and development to implement approaches in systems biology and systems medicine

The ERA-NET funding instrument is an initiative created by the European Commission aiming at developing and strengthening the coordination of important research topics through joint European activities.

### ERACoSysMed

by Sylvia Krobitsch

**The ERA-NET “ERACoSysMed – Collaboration on systems medicine funding to promote the implementation of systems biology approaches in clinical research and medical practice” is funded under the EU Framework Programme for Research and Innovation Horizon 2020 for five years with a total amount of 4.88 million euros.**



The ERACoSysMed consortium comprises 15 national funding organisations and ministries from 13 countries. The aim is to establish and strengthen systems medicine in Europe through the implementation strategy (road map) for Systems Medicine, recommended by the European consortium CASyM. Three calls are planned within the five-year duration of ERACoSysMed. The first call in 2015 received additional funding from the European Union under the ERA-NET Cofund scheme. A second call is scheduled for 2017 and a third one for 2018/2019.

Transnational consortia with research groups from at least three different countries are funded. Countries that participate in the call will fund their own research groups within a consortium. The first call focused on so-called demonstrator projects. These research projects have to demonstrate the social and economic benefit that can be achieved within a systems medicine approach by addressing concrete clinical subjects based on *P4 medicine* (*predictive, preventive, personalised and participatory*). The evaluation of the submitted project proposals was carried out in a two-stage process. Nine transnational consortia with twelve German participants were selected and recommended for funding. The requested funding on the European scale amounts to approximately 12.7 million euros. The first projects are expected to begin in the second quarter of 2016 with a duration of three years.

#### Contact:



**Dr. Sylvia Krobitsch**  
Project Management Jülich  
Forschungszentrum Jülich  
s.krobitsch@fz-juelich.de

[www.eracosysmed.eu](http://www.eracosysmed.eu)  
[www.ptj.de/eracosysmed](http://www.ptj.de/eracosysmed)

## ERASysAPP

by K. Zsuzsanna Nagy

**ERASysAPP, the ERA-NET for Applied Systems Biology was funded by the European Union for three years between January 1<sup>st</sup> 2013 and December 31<sup>st</sup> 2015 with two million euros.**



Within the ERASysAPP consortium 16 national funding organisations from 13 countries worked together. The common aim was to intensify the application of systems biology research in Europe within the areas of biotechnology, industry and health. For this purpose, the ERASysAPP partners published two transnational calls in the field of applied systems biology research in 2013 and 2014. Out of the proposals submitted, 12 transnational consortia, all including German scientists, were selected for funding for three years with about 16 million euros. In total 21 German research groups are participating in ERASysAPP projects. They are supported with 6.7 million euros by the German Federal Ministry of Education and Research.

Projects with a wide range of topics are funded: Some projects receive funding to optimise a production process in micro-organisms or plants, others are supported for the production of metal by using biotechnological techniques, to develop antiviral substances or to perform research in the field of liver cancer. Additionally, ERASysAPP organised networking workshops to strengthen the interactions between academia and industry and to continuously promote the application-oriented aspects of systems biology research. ERASysAPP also updated the strategic research agenda for systems biology from 2008 to ensure a framework for effective collaboration and networking among systems biologists all over Europe.

Another aspect was the promotion of young researchers: ERASysAPP has also made a valuable contribution with regard to the training and development of future systems biologists. Numerous workshops and courses were offered and an educational portal for systems biology was created. This portal contains links to graduate programmes, web-based educational materials and a platform for exchanging teaching materials (<https://www.erasysapp.eu/training-and-exchange/mobility-of-researchers>). Additionally, ERASysAPP – together with the ESFRI action ISBE (Infrastructure for Systems Biology in Europe) – initiated the European data and model management project FAIRDOM (<http://fair-dom.org/>). FAIRDOM offers a central platform, tools and various services to support the management and archiving of research data assets in systems biology projects. A short movie about the advantages of structured data management is available on YouTube (<https://www.youtube.com/watch?v=PWutnWBfUSw>). After the end of the ERA-NET ERASysAPP, the Horizon 2020 work programme 2016 – 2017 announced an ERA-NET Cofund call in the field of biotechnology, for which a joint application from the former ERA-NETs ERASysAPP, ERASynBio and ERA-IB has been submitted.

### Contact:



**Dr. K. Zsuzsanna Nagy**  
Project Management Jülich  
Forschungszentrum Jülich  
[k.nagy@fz-juelich.de](mailto:k.nagy@fz-juelich.de)

[www.erasysapp.eu](http://www.erasysapp.eu)  
[www.ptj.de/en/start](http://www.ptj.de/en/start)

# Welcome to *systembiologie.de!*



apops - Fotolia

Please visit our homepage if you would like to find out more about systems biology.

## What you can expect:

- Exciting stories from **everyday research** – find out more about ongoing projects
- Profiles of systems biologists – get to know the **faces** behind the research
- Extensive overview of events in systems biology – never miss another important **date**
- Information about current **funding** – stay up to date
- Get involved – suggest your own **topic**

We look forward to your visit on our homepage!



## *systembiologie.de scholae* –

### Special edition for school pupils published

Systems biology is still a young scientific field that is unknown to the majority of school pupils. The *systembiologie.de scholae special edition* was conceived to generate interest in schools for interdisciplinary research and to address future generations of students in a targeted manner. Its purpose is to provide an early glimpse into the diverse area of research that systems biology presents and to introduce the latest research results in a form that can be used directly in classes of the upper secondary school.

Therefore employees of the German Cancer Research Center and the Project Management Jülich joined with the Life Sciences Learning Lab team from Berlin to prepare selected research contributions from the *systembiologie.de* magazine, so that school pupils, but not least their teachers, could reflect upon and understand approaches to systems biology research from different angles. In addition to a general introduction to systems biology, content that related to other features of the biology curriculum, such as cancer, stem cells and epigenetics, was selected from earlier issues of the magazine. The special edition also contains two original articles from the *systembiologie.de* magazine and an interview, as well as insights into current research work by systems biologists. A list of German universities at which systems biology can be studied completes the special edition.

*systembiologie.de scholae* aims to provide insights into the field of systems biology, to promote discussion amongst school pupils and to raise awareness for scientific and medical problems, as well as to support teachers in imparting knowledge of systems biology. Each chapter contains corresponding assignments for the pupils. Possible solutions and some further notes can be found in the attached *systembiologie.de scholae/didactics* teacher's booklet.



# imprint

The *systembiologie.de scholae* special edition was financed using funds provided by the German Federal Ministry of Education and Research and the Helmholtz Association and is supplied free of charge. Individual copies and class sets can be obtained by written request from the following contact address at Berlin-Buch Life Sciences Learning Lab: [d.giese@bbb-berlin.de](mailto:d.giese@bbb-berlin.de)

The magazine can be viewed online at:

[www.systembiologie.de/de/magazin](http://www.systembiologie.de/de/magazin)



*systembiologie.de scholae* – Special edition for school pupils  
© LANGEundPFLANZ, Coverphoto: shotstudio – Fotolia.com

## *systembiologie.de* – International Edition

### The magazine for Systems Biology Research in Germany – International Edition Issue 10, June 2016

*systembiologie.de* publishes information on German systems biology research. It is published twice a year in German and once a year in English as an International Edition.

**ISSN 2191-2505**

#### **Publisher:**

*systembiologie.de* is published by the Helmholtz Association, Cross Program Topic Systems Biology and Synthetic Biology, the Virtual Liver Network/LiSyM, the German Aerospace Center (DLR) as well as the Project Management Jülich (PtJ).

#### **Editors:**

**Editor-in-Chief:** Prof. Dr. Roland Eils (DKFZ/Heidelberg University)

**Editorial Coordination:** Dr. Cornelia Depner (DKFZ Heidelberg)

#### **Editorial Team:**

Johannes Bausch (Virtual Liver Network/LiSyM, Freiburg University), Melanie Bergs (PtJ), Dr. Cornelia Depner (DKFZ Heidelberg), Dr. Jan Eufinger (DKFZ Heidelberg), Dr. Marco Leuer (DLR), Dr. Angela Mauer-Oberthür (BioQuant, Heidelberg University), Dr. Yvonne Pfeiffenschneider (PtJ), Dr. Julia Ritzerfeld (DKFZ Heidelberg) and Dr. Gesa Terstiege (PtJ).

#### **Address:**

Editorial office [systembiologie.de](http://systembiologie.de)  
c/o German Cancer Research Center (DKFZ)  
Division Theoretical Bioinformatics - B080  
Berliner Str. 41; D-69120 Heidelberg, Germany

The authors are responsible for the content of by-lined articles. Unless otherwise stated, the authors hold the copyright to the accompanying photos and illustrations. The editorial board accepts no further responsibility for the content of URLs cited by the authors in their articles.

#### **Design and layout:**

LANGEundPFLANZ Werbeagentur GmbH, Speyer ([www.LPsp.de](http://www.LPsp.de))

#### **Translations:**

Toptranslation GmbH, Germany

#### **Printed by:**

Werbedruck GmbH Horst Schreckhase, Spangenberg ([www.schreckhase.de](http://www.schreckhase.de))



#### **PEFC Certified**

This product is from sustainably managed forests, recycled and controlled sources.  
[www.pefc.org](http://www.pefc.org)

#### **Subscriptions:**

The magazine is funded by the Helmholtz Association and the German Federal Ministry of Education and Research (BMBF). It is published as part of the public relations work of the initiatives listed as "Publisher". It is supplied free of charge and must not be sold.

**For subscription please visit [www.systembiologie.de](http://www.systembiologie.de) or contact:**

Editorial office [systembiologie.de](http://systembiologie.de)  
c/o German Cancer Research Center (DKFZ) Heidelberg  
Division Theoretical Bioinformatics - B080  
Berliner Str. 41; D-69120 Heidelberg, Germany  
[abo@systembiologie.de](mailto:abo@systembiologie.de)

# about us

## Presenting the systembiologie.de editorial team

**systembiologie.de** would like to make the success of German systems biology accessible to a wider public in an illustrative way. The magazine, which is published twice per year in German and once in English, is produced jointly by the Helmholtz Association, Cross Program Topic Systems Biology and Synthetic

Biology, Virtual Liver Network / LiSyM, German Aerospace Center (DLR) and Project Management Jülich (PtJ). It is financed by the Helmholtz Association and by the German Federal Ministry of Education and Research (BMBF).

### The editorial team of systembiologie.de:

**standing, from left to right:** Roland Eils (DKFZ/Heidelberg University), Yvonne Pfeiffenschneider (PtJ), Johannes Bausch (Virtual Liver Network/LiSyM), Angela Mauer-Oberthür (BioQuant/Heidelberg University), Kai Ludwig (LANGEundPFLANZ, Speyer), Cornelia Depner (DKFZ Heidelberg), Jan Eufinger (DKFZ Heidelberg).

**seated, from left to right:** Gesa Terstiege (PtJ), Melanie Bergs (PtJ), Julia Ritzerfeld (DKFZ Heidelberg), Marco Leuer (DLR).



Photo: Tobias Schwerdt / DKFZ

# contact data

## **Helmholtz Association, Cross Program Topic Systems Biology and Synthetic Biology**

Coordinator: Prof. Dr. Roland Eils

Scientific Project Management:

Dr. Cornelia Depner, Dr. Jan Eufinger, Dr. Julia Ritzerfeld  
c/o German Cancer Research Center (DKFZ) Heidelberg

Division Theoretical Bioinformatics - B080

Berliner Str. 41; D-69120 Heidelberg, Germany

Email: [c.depner@dkfz.de](mailto:c.depner@dkfz.de), [j.eufinger@dkfz.de](mailto:j.eufinger@dkfz.de), [j.ritzerfeld@dkfz.de](mailto:j.ritzerfeld@dkfz.de)

[www.helmholtz.de/en/about\\_us/networks\\_and\\_cooperation/helmholtz\\_alliances/systems\\_biology/](http://www.helmholtz.de/en/about_us/networks_and_cooperation/helmholtz_alliances/systems_biology/) and

[www.helmholtz.de/en/about\\_us/the\\_association/initiating\\_and\\_networking/assuring\\_excellence/synthetic\\_biology/](http://www.helmholtz.de/en/about_us/the_association/initiating_and_networking/assuring_excellence/synthetic_biology/)



## **Virtual Liver Network/LiSyM – Liver Systems Medicine**

Programme Director: Dr. Adriano Henney / Prof. Dr. Peter Jansen

Scientific Project Management: Johannes Bausch

Freiburg University; Institute of Physics

Hermann-Herder-Str. 3; D-79104 Freiburg, Germany

Email: [johannes.bausch@virtual-liver.de](mailto:johannes.bausch@virtual-liver.de)

[www.virtual-liver.de](http://www.virtual-liver.de)



## **BioQuant – Heidelberg University**

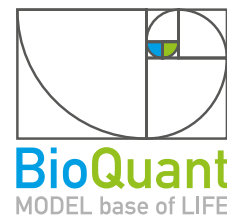
Board of Directors: Prof. Dr. Roland Eils, Prof. Dr. Hans-Georg Kräusslich,  
Prof. Dr. Robert B. Russell

Executive Management: Dr. Angela Mauer-Oberthür

Im Neuenheimer Feld 267; D-69120 Heidelberg, Germany

Email: [angela.oberthuer@bioquant.uni-heidelberg.de](mailto:angela.oberthuer@bioquant.uni-heidelberg.de)

[www.bioquant.uni-heidelberg.de](http://www.bioquant.uni-heidelberg.de)



## **Project Management Jülich**

Forschungszentrum Jülich GmbH

Life Sciences, Health, Universities of Applied Sciences

Contact persons:

Dr. Yvonne Pfeiffenschneider, Dr. Gesa Terstiege, Melanie Bergs

Department Molecular Life Sciences (LGF 2)

D-52425 Jülich, Germany

Email: [y.pfeiffenschneider@fz-juelich.de](mailto:y.pfeiffenschneider@fz-juelich.de), [g.terstiege@fz-juelich.de](mailto:g.terstiege@fz-juelich.de), [m.bergs@fz-juelich.de](mailto:m.bergs@fz-juelich.de)

[www.ptj.de/en/start](http://www.ptj.de/en/start)



## **German Aerospace Center (DLR)**

Project Management Agency

Health Research (OE20)

Contact persons:

Dr. Marco Leuer, Ursula Porwol

Heinrich-Konen-Str. 1; D-53227 Bonn, Germany

Email: [marco.leuer@dlr.de](mailto:marco.leuer@dlr.de), [ursula.porwol@dlr.de](mailto:ursula.porwol@dlr.de)

[www.dlr.de/pt/en/desktopdefault.aspx/tabid-10354/#gallery/26469](http://www.dlr.de/pt/en/desktopdefault.aspx/tabid-10354/#gallery/26469)





# EMBL 2016

## Conferences

26 - 29 JUN | EMBO | EMBL Symposium  
**Innate Immunity in Host-Pathogen Interactions**  
Z. Chen, W.-D. Hardt, N. Pariente, F. Randow  
EMBL Heidelberg, Germany

5 - 7 JUL | EMBL Conference  
**Lifelong Learning in the Biomedical Sciences**  
M. Hardman, C. Janko, C. Johnson  
EMBL Heidelberg, Germany

24 - 26 JUL | EMBL Conference  
**Microfluidics 2016**  
C. Merten, S. Quake  
EMBL Heidelberg, Germany

27 - 30 AUG | EMBL Conference  
**Transcription and Chromatin**  
D. Duboule, E. Furlong, A. Shilatifard, M. Timmers | EMBL Heidelberg, Germany

31 AUG - 3 SEP | EMBO Conference  
**Chemical Biology 2016**  
M. Köhn, J. Overington, C. Schultz  
EMBL Heidelberg, Germany

7 - 10 SEP | EMBO | EMBL Symposium  
**Actin in Action: From Molecules to Cellular Functions**  
B. Baum, J. Faix, P. Lenart, D. Mullins, F. Nedelec, C. Sykes | EMBL Heidelberg, Germany

14 - 17 SEP | EMBL-Wellcome Genome Campus Conference  
**Proteomics in Cell Biology and Disease Mechanisms**  
A.-C. Gavin, A. Lamond, M. Mann | EMBL Heidelberg, Germany

25 - 27 SEP | EMBL-Wellcome Genome Campus Conference  
**Big Data in Biology and Health**  
E. Birney, B. Grossman, J. Korbel, C. Relton  
EMBL Heidelberg, Germany

5 - 8 OCT | EMBO | EMBL Symposium  
**The Complex Life of mRNA**  
A. Ephrussi, N. Sonenberg, J. Steitz, D. Tollervay | EMBL Heidelberg, Germany

12 - 15 OCT | EMBO | EMBL Symposium  
**Organoids: Modelling Organ Development and Disease in 3D Culture**  
M. Bissell, J. Knoblich, E. Schnapp  
EMBL Heidelberg, Germany

19 - 23 OCT | EMBO Conference  
**Experimental Approaches to Evolution and Ecology Using Yeast and Other Model Systems**  
J. Berman, M. Dunham, J. Leu, L. Steinmetz  
EMBL Heidelberg, Germany

12 - 15 NOV | EMBO Conference  
**From Functional Genomics to Systems Biology**  
E. Furlong, F.C.P. Holstege, N. Rajewsky, M. Walhout  
EMBL Heidelberg, Germany

20 - 23 NOV | EMBO Conference  
**Molecular Machines: Integrative Structural and Molecular Biology**  
J. Briggs, T. Carlomagno, G. Kleywegt, D. Panne, D. Svergun  
EMBL Heidelberg, Germany

4 - 6 DEC | EMBL-Wellcome Genome Campus Conference  
**Target Validation Using Genomics and Informatics**  
E. Birney, C. Fox, S. John, M. Fergussan  
EMBL Heidelberg, Germany



## Courses

19 - 23 JUN | EMBO Practical Course  
**Computational Biology: Genomes to Systems**  
P. Bork, F. Ciccarelli, J. Korbel, R. Krause  
EMBL Heidelberg, Germany

20 - 24 JUN | EMBL Course  
**Fundamentals of Widefield and Confocal Microscopy and Imaging**  
F. Eich, J. Marquardt, S. Terjung | EMBL Heidelberg, Germany

27 JUN - 1 JUL | EMBL-EBI Course  
**Cancer Genomics**  
G. Rustici | EMBL-EBI Hinxton, UK

28 JUN - 1 JUL, 4 - 7 OCT | EMBL Courses  
**Whole Transcriptome Data Analysis**  
V. Benes, R. Calogero  
EMBL Heidelberg, Germany

3 - 8 JUL | EMBL Course  
**Advanced Fluorescence Imaging Techniques**  
F. Eich, R. Pepperkok, S. Terjung  
EMBL Heidelberg, Germany

3 - 8 JUL | EMBL-EBI-Wellcome Genome Campus Course  
**In silico Systems Biology**  
L. Emery | EMBL-EBI Hinxton, UK

4 - 5 JUL, 28 - 29 NOV | EMBL Courses  
**NGS: Whole Genome Sequencing Library Preparation**  
V. Benes, J. Dreyer-Lamm, A. Heim  
EMBL Heidelberg, Germany

11 - 15 JUL | EMBL Course  
**Quantitative Proteomics**  
J. Krijgsveld, M. Savitski  
EMBL Heidelberg, Germany

11 - 15 JUL, 21 - 25 NOV | EMBL Courses  
**NGS: Enrichment Based Targeted Resequencing**  
V. Benes, J. Dreyer-Lamm, A. Heim  
EMBL Heidelberg, Germany

28 AUG - 5 SEP | EMBO Practical Course  
**Cryo-Electron Microscopy and 3D Image Processing**  
J. Briggs, B. Boettcher, L. Passmore, C. Sachse, H. Stahlberg  
EMBL Heidelberg, Germany

29 AUG - 2 SEP | EMBL Course  
**Chromatin Signatures During Differentiation**  
J. Dreyer-Lamm, P. Grandi, K.-M. Noh  
EMBL Heidelberg, Germany

12 - 14 SEP | EMBL-EBI Course  
**Metagenomics Bioinformatics**  
H. Denise, L. Emery, A. Mitchell  
EMBL-EBI Hinxton, UK

12 - 20 SEP | EMBO Practical Course  
**Protein Expression, Purification and Characterization**  
C. Loew, R. Meijers, A. Parret  
EMBL Hamburg, Germany

19 - 23 SEP | EMBL-EBI Course  
**Structural Bioinformatics**  
T. Hancock, G. Kleywegt, C. Orengo  
EMBL-EBI Hinxton, UK

19 - 24 SEP | EMBL Course  
**Extracellular Vesicles: From Biology to Biomedical Applications**  
J. Dreyer-Lamm, A. Hendrix, E. Nolte-'t Hoen  
EMBL Heidelberg, Germany

10 - 13 OCT | EMBO Practical Course  
**RNA Sequencing Library Preparation - How low can you go?**  
V. Benes, B. Textor  
EMBL Heidelberg, Germany

17 - 24 OCT | EMBO Practical Course  
**Solution Scattering from Biological Macromolecules**  
A. Kikhney, D. Svergun  
EMBL Hamburg, Germany

17 - 23 OCT | EMBO Practical Course  
**High-Throughput Microscopy for Systems Biology**  
J. Ellenberg, D.W. Gerlich, B. Neumann, R. Pepperkok | EMBL Heidelberg, Germany

7 - 11 NOV | EMBL-EBI-Wellcome Genome Campus Course  
**Resources for Computational Drug Discovery**  
T. Hancock | EMBL-EBI Hinxton, UK

9 - 10 NOV | EMBL Course  
**Microinjection into Adherent Cells**  
J. Dreyer-Lamm, S. Stobrawa | EMBL Heidelberg, Germany

28 NOV - 2 DEC | EMBL-EBI Course  
**Biological Interpretation of Next Generation Sequencing**  
G. Rustici | EMBL-EBI, Hinxton UK

4 - 9 DEC | EMBL-EBI-Wellcome Genome Campus Course  
**Proteomics Bioinformatics**  
L. Emery | EMBL-EBI, Hinxton UK

7 - 11 DEC | EMBL Course  
**Microbial Communities: Modelling Meets Experiments**  
R. Mahadevan, K. Patil, K. Sasaki  
EMBL Heidelberg, Germany

For full event listing please visit our website  
[www.embl.org/events](http://www.embl.org/events)



@emblevents

We would like to thank the members of the EMBL ATC Corporate Partnership Programme:  
**Founder Partners:** Leica Microsystems, Olympus  
**Corporate Partners:** BD, Boehringer Ingelheim, GSK, Illumina, Thermo Fisher Scientific  
**Associate Partners:** Eppendorf, Merck, Nikon, Sanofi

