Imprint | Research Report 2015

Published by the Max Planck Institute for Molecular Genetics (MPIMG), Berlin, Germany, Oktober 2015

Editorial Board: Bernhard G. Herrmann & Martin Vingron
Coordination: Patricia Marquardt
Photos & Scientific Illustrations: MPIMG
Production: Thomas Didier, Meta Drack
Copies: 500

Contact: Max Planck Institute for Molecular Genetics Ihnestr. 63 – 73 14195 Berlin Germany
Phone: +49 (0)30 8413-0
Fax: +49 (0)30 8413-1207
Email: info@molgen.mpg.de

For further information about the MPIMG, see http://www.molgen.mpg.de

Cover image
3D-rendered light sheet micrograph of the caudal end of an E9.5 mouse embryo illustrating the generation of neuroectoderm (green cells) and mesoderm (red cells) from common progenitor cells (yellow cells, mostly located deep in the tissue) during trunk development. The neuroectoderm gives rise to the spinal cord, the mesoderm to the vertebral column, skeletal muscles and other tissues. Picture taken by Frederic Koch, Manuela Schotze, and Matthias Marks, MPIMG.
Ada Yonath, Nobel laureate in chemistry 2009 and former member of the MPIMG, at the celebratory symposium on the occasion of the institute’s 50th anniversary in December 2014.
# Table of contents

The Max Planck Institute for Molecular Genetics 7

**Department of Developmental Genetics (B.G. Herrmann)** 21

- Introduction 22
- Scientific achievements and findings 23
- General information about the whole department 37

**Department of Computational Molecular Biology (M. Vingron)** 45

- Introduction 45
- Transcriptional Regulation Group (M. Vingron) 51
- Sequencing & Genomics Group (S. Haas) 59
- Bioinformatics Group (R. Herwig) 65
- Diploid Genomics Group (M. Hoehe) 71
- Mechanisms of Transcriptional Regulation Group (S. Meijsing) 79

**Research Group Evolutionary Genomics (P. Arndt)** 85

- General information about the whole department 94

**Research Group Development & Disease (S. Mundlos)** 113

- Introduction 114
- Scientific achievements and findings 116
- Chromosome Rearrangements & Disease group (V. Kalscheuer) 127
- General information about the whole research group 133

**Otto Warburg Laboratory**

- Max Planck Research Group Epigenomics (H.-R. Chung) 153
- Research Group RNA Bioinformatics (A. Marsico) 163
- Sofja Kovalevskaja Research Group Long non-coding RNA (U.A. Ørom) 173
- BMBF Research Group Nutrigenomics & Gene Regulation (S. Sauer) 183
- Max Planck Research Group Regulatory Networks in Stem Cells (E. Schulz) 195
- Max Planck Research Group Molecular Interaction Networks (U. Stelzl) 203
- Research Group Gene Regulation and Systems Biology of Cancer (M.-L. Yaspo) 215
- BMBF Research Group Cell Signaling Dynamics (Z. Zi) 229
- Minerva Group Neurodegenerative disorders (S. Krobitsch) 235
Max Planck Fellow Group  
Efficient Algorithms for Omics Data (K. Reinert)  

Emeritus Group Vertebrate Genomics (H. Lehrach)  
  Introduction  
  Scientific methods and achievements  
  Project groups within the department  
  General information about the whole department  

Emeritus Group Human Molecular Genetics (H.-H. Ropers)  
  20 years of human genetics at the MPIMG  
  Scientific activities and results, 2012-2015  
  Epilogue  
  General information about the whole department  

Scientific services  
  Animal facility  
  Transgenic unit  
  Sequencing core facility  
  Mass spectrometry facility  
  Microscopy & Cryo-Electron Microscopy Group  
  IT Group  
  Library  

Research Support  
  Administration  
  Technical management & workshops
Foreword

It is our pleasure to present the 2015 Research Report of the Max Planck Institute for Molecular Genetics (MPIMG) in Berlin, covering the period from 2009 to mid-2015. Many changes have marked the last years. First of all, the retirement of Hans Lehrach und Hans-Hilger Ropers in late autumn of 2014 made a deep cut in our daily life at the institute. Two major departments were closed and many activities of the MPIMG are geared towards the appointment of new directors or are deferred until their arrival. On the other hand, a very positive development has taken place at the Otto Warburg laboratory with five new groups having started since 2011.

In 2013, the construction of the long expected tower 3 has been finished and the bioinformaticians and computer scientists of the institute moved in. Immediately afterwards, the refurbishment of tower 2 commenced, promising new and state-of-the-art labs to all our “wet lab” scientists. All refurbishment work as well as the construction of tower 3 has taken place during full research operations, which led to considerable disturbances, in particular of vibration-sensitive experimental work.

Founded in 1964, the MPIMG marked its 50th anniversary in 2014 with a large celebratory symposium, bringing together many former and current scientists of the MPIMG. Today, the institute is established as a centre for research at the interface of genome research and genetics, concentrating on genome analysis of humans and other organisms to elucidate cellular processes and genetic diseases. Due to current developments, prompted by the search for two new directors within the institute as well as by the technological advances of the last years, the focus of research at the MPIMG shifts towards the systematic study of gene regulation from a whole-genome viewpoint, leading to the study of gene regulatory networks and their role in development and disease. With this report, we want to give a broad summary of the scientific work of the MPIMG during the last years and describe the ongoing changes. We hope that it will provide a clear impression of the institute and the work we are doing here.

September 2015

Bernhard G. Herrmann & Martin Vingron
Figure 1: Organization chart of the MPIMG, September 2015
The Max Planck Institute for Molecular Genetics

Structure and organization of the institute

At the time of writing this report, the Max Planck Institute for Molecular Genetics (MPIMG) is undergoing a major transition. In late autumn 2014, Hans Lehrach, head of the Department of Vertebrate Genomics, and Hans-Hilger Ropers, head of the Department of Human Molecular Genetics, retired and their departments had been closed. The institute together with the respective committee (the Stammkommission) of the Max Planck Society are actively searching for new directors. However, for the time being, the MPIMG lacks two departments. Although we have tried to counteract the loss in staff through the establishment of research groups, staff numbers have fallen from around 460 in 2009 to about 250 in 2014/2015 (see Table 1).

The MPIMG currently consists of two active departments, headed by Bernhard Herrmann (Dept. of Developmental Genetics) and Martin Vingron (Dept. of Computational Molecular Biology), respectively. The departments are complemented by the research group Development & Disease headed by Stefan Mundlos, who also holds a Chair of Medical Genetics and Human Genetics at the Charité – Universitätsmedizin Berlin, and a number of (temporary) independent research groups, collectively referred to as the Otto Warburg Laboratory (OWL). The OWL offers space and var-

Table 1: Number of employees and fellowship holders at the MPIMG

<table>
<thead>
<tr>
<th>Year</th>
<th>Externally funded</th>
<th>MPIMG funded</th>
</tr>
</thead>
<tbody>
<tr>
<td>2009</td>
<td>307</td>
<td>156</td>
</tr>
<tr>
<td>2010</td>
<td>290</td>
<td>155</td>
</tr>
<tr>
<td>2011</td>
<td>283</td>
<td>130</td>
</tr>
<tr>
<td>2012</td>
<td>265</td>
<td>95</td>
</tr>
<tr>
<td>2013</td>
<td>232</td>
<td>62</td>
</tr>
<tr>
<td>2014</td>
<td>207</td>
<td>45</td>
</tr>
<tr>
<td>2015</td>
<td>208</td>
<td>43</td>
</tr>
</tbody>
</table>
ious resources over a longer, but typically limited period of time (usually between five and nine years) to excellent junior scientists so that they can build up their own groups and work on their own scientific programs. The groups are financed by different sources. In addition to the Max Planck Research groups funded by the Max Planck Society, groups funded by the German Research Foundation (DFG, e.g. Emmy Noether program), the Humboldt Foundation (Sofja Kovalevskaja), the German Ministry for Education and Research and many more are also welcome. Current members of the OWL are Ho-Ryun Chung (Epigenomics), Annalisa Marsico (RNA Bioinformatics), Ulf Ørom (Long non-coding RNAs), Edda Schulz (Regulatory Networks in Stem Cells), Marie-Laure Yaspo (Gene Regulation & Systems Biology of Cancer), and Zhike Zi (Cell Signaling Dynamics). Sylvia Krobitsch (Neurodegenerative Disorders) has left the Institute early in 2015, and Sascha Sauer (Nutrigenomics/Gene Regulation) and Ulrich Stelzl (Molecular Interaction Networks) are currently in the process of moving to other research institutions.

In 2014, the MPIMG succeeded in appointing Knut Reinert, Professor of Algorithmic Bioinformatics at Freie Universität (FU) Berlin, as Max Planck Fellow to the institute. Reinert’s research focuses on providing efficient tools for the analysis of genomic data; he cooperates with many groups at the MPIMG already since his appointment to Freie Universität Berlin in 2002. In addition, he is co-speaker of the International Max Planck Research School for Computational Biology and Scientific Computing (IMPRS-CBSC), a joint graduate program of the MPIMG and the FU Berlin (FU) that has been established in 2006 (see below).

In addition, the scientific groups of the MPIMG are supported by a number of scientific service groups that maintain a range of core technologies, and the general administration/research support (see Figure 1).

**Scientific concept**

Molecular genetics generally deals with the study of how life processes function as a consequence of the genetic make-up of an organism. Over the years, genome research, the systematic study of genes and genomes, has changed the way in which research in molecular genetics is pursued. Today, genomic technologies allow posing scientific questions broadly in order to determine all parts of a genome that influence the process under study. When applied to humans, this is particularly important for the understanding of disease processes. The MPIMG works at the interface of genome research and genetics, concentrating on genome analysis of humans and other organisms to elucidate cellular processes and genetic diseases.

The ongoing transition prompted by the search for two new directors builds on this foundation. Modern sequencing and functional genomics methods
have made it possible to pose and answer questions concerning all genes of an individual organism, or an organism’s entire genome. The listing of all (protein-coding) genes of an organism is a largely solved problem, although many of these genes possess still unknown functions. More importantly, though, whole genome sequencing has made available the large, non-charted territory of intergenic regions. Embedded in these regions, there is much more regulatory information than previously known. Due to novel insights combined with improved genomic technologies, it is now feasible to study regulatory interactions in a systematic manner. With this capability, the prospect of eventually understanding the inner workings of a cell is coming considerably closer.

Based on the technological advances, genetic analysis has entered a new era in its ability to provide the clues towards interpreting individual genome sequences. However, the current rapid progress in the mapping of disease mutations also shows that not all mutations affect protein-coding regions. Many mutations exert their phenotypic effect through changes in non-coding, regulatory elements. Therefore, genome-based molecular genetics will turn more of its attention towards the regulatory relationships among the genes as a basis also for medical studies. In addition, genetic differences between individuals have been recognized as important modulators of disease. The role of genetic variants in regulatory networks involved in developmental, regenerative, and disease processes adds another level to the analysis of genotype-phenotype relationships. Given this background, the MPIMG aims at combining the genomic approach to genetics, as presently established at the institute, with an increased focus on the systematic study of gene regulation from a whole-genome viewpoint, leading to the study of gene regulatory networks and their role in development and disease.

Scientific highlights

The main results of the scientific work of the MPIMG during the last three to six years are described in detail in the research reports of the individual departments. At this point, we wish to present some of the most important and interesting results to give a general impression of the research performed at the institute.

For many years and mainly driven by Hilger Ropers, the genetic causes of intellectual disability have been a main topic at the MPIMG. In 2011, Ropers and his co-workers described the identification of 50 hitherto unknown genetic causes of intellectual disability by using new next generation sequencing technologies. Their findings demonstrated the enormous genetic diversity of intellectual disabilities and subdivided them into different monogenetic defects. This will not only help families to obtain a reliable diagnosis, but also provide a model for the explanation of related
disorders, such as autism, schizophrenia and epilepsy (Najmabadi et al., Nature 2011). In this context, an intensive and still ongoing collaboration has been established between Vera Kalscheuer and Stefan Haas, senior scientists and heads of project groups at the Dept. of Human Molecular Genetics and the Dept. of Computational Molecular Biology, respectively. They designed and implemented an analysis pipeline for next generation sequencing data and use it to uncover mutations, especially sequence variations, involved in neurodevelopmental disorders (e.g., Hu et al., Mol Psychiatry 2015). With the retirement of Hilger Ropers, Kalscheuer moved to the research group Development & Disease, where she continues her work.

Another field with contributions from many MPIMG groups is cancer genomics, a topic of special interest for Hans Lehrach and Marie-Laure Yaspo. After the retirement of Hans Lehrach, Yaspo continues her work as head of an independent research group of the OWL. Lately, Yaspo, in the context of an international consortium, succeeded in decoding the molecular characteristics of a subtype of pediatric acute lymphocytic leukemia (ALL). The scientists decoded both the genome and the transcriptome of the cancer cells and compared the molecular pathology between two different leukemia types, both disrupting one allele of $TCF3$. They found that the interplay between a fused $TCF3$-$HLF$ oncogenic protein, additional DNA changes, and an altered gene expression program leads to a re-programming of leukemic cells to an early, stem-cell like, developmental stage, although the phenotypic appearance of the cells remains similar (Fischer et al., Nat Genet 2015).

Michal Schweiger, who accepted a Lichtenberg Professorship on Functional Epigenomics in Cologne in 2014, has been working on the molecular characterization and the epigenetic regulatory mechanisms in different cancer types (e.g., Röhr et al., PLoS one 2013; Börno et al., Cancer Discov 2012). Ralf Herwig is involved in the analysis of specific tumors, for example non-small cell lung cancers (Hülsmann et al., Lung Cancer 2014). A complementary project of Stefan Haas in cooperation with Roman Thomas, University of Cologne, who work about small lung cancer, is still ongoing (e.g. Fernandez-Cuesta et al., Genome Biology 2015; Fernandez-Cuesta et al., Nat Commun 2014; Peifer et al., Nat Genet 2012).

Bernhard Herrmann and his team have demonstrated that long-non-coding RNAs, by acting as epigenetic control factors, can play essential roles in organogenesis. They discovered a lncRNA termed $Fendrr$, which is transiently expressed in the lateral mesoderm and, when lacking, causes malformations in organs derived from this tissue, the heart and the ventral body wall, ultimately leading to embryonic death. Organ dysfunction in the mutants emerged with several days delay after $Fendrr$ expression occurs in wild-type progenitor cells, highlighting the important role of $Fendrr$ as epigenetic regulator exerting immediate as well as long-term
effects. Herrmann and his team also showed that *Fendrr* alters the gene activity of several important transcription factors by anchoring the Polycomb Repressive Complex-2 to their promoters, thereby changing their histone modification state. Thus, *Fendrr* highlights the impact of a new class of regulators in mammalian organ development (Grote et al., Dev Cell 2013).

Ulf Ørom together with Annalisa Marsico, both from the Otto Warburg Laboratory, developed a method for studying an early step of the biogenesis of microRNA molecules in living cells. They showed that the activity of the microprocessor complex is the most important step during the biogenesis of miRNA and can be used as a reliable measure for the amount of mature miRNAs in the cell (Conrad et al., Cell Rep 2014).

Stefan Mundlos and his team have been able to describe a novel disease mechanism, which involves “rewiring” of enhancer-promoter interactions due to alterations of higher order genomic structures. These ectopic interactions that are induced by large scale rearrangements can result in gene misexpression and consecutive malformations. The group developed a modification of the CRISPR/Cas9 method to reproduce large structural rearrangements of the human genome in mice. The strength of the new “CrisVar” methodology was demonstrated by producing deletions, duplications, and inversions at several loci thereby recapitulating human disease-associated rearrangements (Kraft et al., Cell Rep 2015). Mundlos and his colleagues also used the CrisVar method to study genomic rearrangement at the EPHA4 locus *in detail*. By studying the molecular pathology of three limb malformation syndromes, they showed that the deletion of CTCF-associated boundaries that normally separate neighboring topologically associated domains (TADs) is necessary and sufficient to disrupt

![Table 2: Number of publications with at least one co-author from the MPIMG. Publications with contributions from more than one MPIMG department or group are shown in light green.](image)
TAD structures. This can lead to ectopic activation of genes, misexpression and disease. Their studies provide the first evidence that TADs and their boundaries are of biological and medical relevance (Lupianez et al., Cell 2015).

Altogether, MPIMG scientists have (co)authored 1,021 publications during the reporting period (2009-07/2015). About 15% of these had contributions from more than one department or group, thus demonstrating the continuous cooperation between the individual MPIMG groups (see Table 2).

**Scientific awards**

In appreciation of their scientific achievements, members of the MPIMG have been awarded with several prizes. Among others,

*Katerina Kraft* has been honored with the Peter Hans Hofschneider Prize for Molecular Medicine of the Max Planck Society for the Advancement of Science in 2015;

*Dario Lupianez* received an ESHG Young Scientist Award from the European Society of Human Genetics, as well as the best lecture award from the Deutsche Gesellschaft für Humangenetik [German Society of Human Genetics] in 2015;

*Bruno Pereira* gained a Young Researcher Poster Price at the EMBO Workshop: Embryonic-Extraembryonic Interfaces in Göttingen in 2015;

*Stefanie Schöne* received a fellowship of the Christiane Nüsslein-Volhard Foundation and the L’Oreal-UNES-
CO for Women in Science Program in 2015;

*Ralf Herwig* obtained the PerMediCon Award (3rd place) for the project EPITREAT Personalized lung cancer treatment based on epigenetic biomarkers at PerMediCon in Cologne in 2014 and the 31st Animal Protection Research Prize of the German Federal Ministry of Food and Agriculture (BMEL) in 2012;

*Daniele Ibrahim* received the Lecture Award of the Deutsche Gesellschaft für Humangenetik [German Society of Human Genetics] in 2014;

*Stefan Mundlos* has been appointed as member of the Berlin Brandenburg Academy of Sciences in 2014;

*Hans-Hilger Ropers* received the EURORDIS Scientific Award 2014 of the European Organization for Rare Diseases in 2014, and the Honorary medal and honorary membership of the German Society for Human Genetics in 2009;

*Martin Vingron* has been elected as a member of the Academia Europaea and named among the Thomson-Reuters Highly Cited Researchers, both in 2014. In addition, he has been selected as an ISCB Fellow in the Fellows Class of 2012;

*Björn Fischer-Zirnsak* won the Robert Koch Prize of the Charité in 2014;

*Malte Spielmann* won the ESHG Young Scientist Award of the European Society of Human Genetics in 2012;

*Ulf Andersson Ørom* obtained a Sofja Kovalevskaya Award by the Alexander von Humboldt Foundation in 2012;

*Irina Czogiel* won the Gustav-Adolf-Lienert prize of the Biometrical Society (German Region) in 2012;

*Markus Ralser* received an ERC European Starting grant (2011), a Wellcome-Beit Prize of the Wellcome Trust, UK (2011) and an EMBO Young Investigator Award (2012);

*Nana-Maria Grüning* has been awarded with the Nachwuchswissenschaftlerinnen-Preis 2012 of the Forschungsverbund Berlin;

*Marcel Holger Schulz* has been awarded an Otto Hahn Medal of the Max Planck Society in 2011;

*Rosa Karlic* has been honoured with the L’Oreal Adria-UNESCO National Fellowship “For Women in Science” in 2011;

*Lars Bertram* obtained a Special Award of the Hans-und-Ilse-Breuer Foundation for Research in Alzheimer’s in 2010 and has been recipient of the 2009 Independent Investigator Award, NARSAD;

*Eva Klopocki* received a Finalist Trainee Award of the American Society of Human Genetics, an ESHG Young Scientist Award of the European Society of Human Genetics, and the Vortragspreis of the Deutsche Gesellschaft für Humangenetik (German Society for Human Genetics), all in 2009.
Appointments of former members of the MPIMG

Several people have left the MPIMG and took up positions in other institutions or universities in Germany and abroad. Some of the most important appointments have been (ordered by year)

**2015**

*Hao Hu*: Professor of Genetics, Zhongshan School of Medicine, Sun Yat-sen University, and Director of the Department of Molecular Diagnostics, Guangzhou Women and Children’s Medical Center, Guangzhou, China

*Ulrich Stelzl*: (Full) Professorship on Biopharmaceutica und Proteomics, University of Graz, Austria

**2014**

*Lars Bertram*: Professorship (W2) on Genome Analytics, University of Lübeck

*Philip Grote*: Head of Research Group “LncRNAs in Cardio-pulmonary Development”, Goethe University Frankfurt

*Matthias Heinig*: Head of Research Group “Genetic and Epigenetic Regulation”, Helmholtz Center Munich

*Annalisa Marsico*: Professorship (W1) on RNA Bioinformatics, Freie Universität Berlin

*Michal-Ruth Schweiger*: Lichtenberg Professorship on Functional Epigenomics, University of Cologne

*Sigmar Stricker*: Professorship (W2) for Biochemistry and Genetics, Freie Universität Berlin

*Reinhard Ullmann*: Head of Research Group, Bundeswehr Institute of Radiobiology affiliated to the University of Ulm, Munich

*Christoph Wierling*: Head of Research Group, Alacris Theranostics GmbH, Berlin

**2012**

*James Adjaye*: Professorship (W3) on Stem Cell Research and Regenerative Medicine and Director of the Institute for Transplantations Diagnostics and Cell Therapies (ITZ), Heinrich-Heine-Universität Düsseldorf

*Jonathan Göke*: 2012 Postdoc at Genome Institute of Singapore (GIS); 2014 GIS Fellow

*Tim Hucho*: Professorship (W2) on Anaesthesiology and Pain Research, University Hospital of Cologne

*Eva Kloppoki*: Professorship (W2) for Human Genetics, University of Würzburg

*Uwe Kornak*: Professorship (W2) for Functional Genetics, Charité - Universitätsmedizin Berlin

*Bodo Lange*: Managing Director, Alacris Theranostics GmbH, Berlin

*Markus Ralser*: Head of Research Group, Cambridge Systems Biology Centre and Dept. of Biochemistry, Cambridge, UK

*Peter Robinson*: Professorship (W2) for Medical Bioinformatics, Charité - Universitätsmedizin Berlin

*Harald Seitz*: Head of Research Group, Fraunhofer Institute for Biomedical Engineering, Potsdam

*Ewa Szczurek*: 2012 ETH Postdoctoral Fellowship; 2015 Assistant Professor, University of Warsaw, Poland

*Morgane Thomas-Chollier*: Associate Professor on Computational Systems Biology, École Normale Supérieure, Paris, France
Recruitment of new group leaders

A range of new group leaders joined the Otto Warburg Laboratory in recent years, thus bringing not only new themes, but also new ideas and impetus to the institute.

In December 2011, Ho-Ryun Chung established a new Max Planck Research Group on Epigenomics at the MPIMG, shortly followed by Ulf Ørom early in 2012, who has been awarded with a Sofja Kovalevskaja Group on Long non-coding RNA by the Humboldt Foundation. In 2014, Annalisa Marsico took up a position as Assistant Professor at Freie Universität (FU) Berlin and established her own group on RNA Bioinformatics, funded by the FU Berlin and the MPIMG in the context of a joint Dahlem International Research Network. At the same time, Zhike Zi joined the MPIMG to build up a Junior Research Group on Cell Signaling Dynamics, funded by...
the German Ministry for Education and Research. The latest group so far has been established by Edda Schulz. Her Max Planck Research Group on Regulatory Networks in Stem Cells started in January 2015.

In addition, David Meierhofer has been recruited as head of the new mass spectrometry service group, also in 2012.

**Support of junior scientists at the MPIMG**

In July 2015, 60 students pursued their PhD studies at the institute. All students are part of an institute-wide graduation program that was established in 2008. It encourages the exchange of knowledge and the development of skills throughout the disciplines. For one, this is achieved by intense interdisciplinary courses held in turn by the different departments, the so-called PhD week. It offers insight and hands-on experience in fields that go beyond the scope of the respective PhD theses. Since 2008, 18 courses have provided supplementary experience especially to junior students in their first and second year, and have initiated new cooperation and fresh perspectives for many PhD projects. Another central aspect of the PhD program is an annual PhD retreat, organized by the Student Association (STA). This self-organized student representation organ was founded in 2001 to act for the interests of all students at the MPIMG by addressing the directors, group leaders, or the staff association. In addition, it provides a platform for organizing social and networking events to foster interdepartmental scientific discussion and exchange of technical and scientific knowledge between the students. Furthermore, each head of department encouraged the implementation of a thesis advisory committee (TAC) for each PhD student. The TAC consists of three scientists, the main supervisor at the MPIMG together with a colleague from the university and a third member from a related, but not exactly the same field. The committee meets once a year and is meant to support and guide the student in a scientific as well as in a practical way.

Most PhD students working in the field of computational molecular biology pursue their thesis work in the context of the International Max Planck Research School for Computational Biology and Scientific Computing (IMPRS-CBSC). This graduation program was established in 2006 by Martin Vingron and Knut Reinert, Faculty of Mathematics and Informatics at Freie Universität Berlin (FUB) and also Max Planck Fellow at the MPIMG since 2014. Its strength is to capitalize on the synergies of MPIMG and university in the area of biology-related computer science. To date, six group leaders from MPIMG, five from FUB, and six from associated institutes are members of the IMPRS-CBSC faculty. Up to now, the IMPRS-CBSC had 59 students, 38 of which have already successfully submitted and defended their thesis during the last years. More than 54% graduated with summa cum laude, while publishing over four papers in average. The curriculum consists of scientific training and research, career
support through education in so-called professional or soft skills, and tutoring. Excellent supervision is guaranteed through TACs and a regulation manifest to define the cornerstones for training structure, funding and supervision quality. A PhD coordinator assists in issues of the students, not only from the IMPRS-CBSC, but for the whole institute. Following the pending appointments of new directors, the existing IMPRS-CBSC shall be extended to a new graduate school with a broader focus. In line with the scientific concept of the institute, the future IMPRS is intended to bridge the gap between natural and formal sciences. In accordance, the faculty is supposed to grow and will contain not only experts in bioinformatics and computer sciences from MPIMG and FUB, but also biochemists and more „wet lab“ scientists from both institutes and other Berlin institutions. All students of the MPIMG will have the chance to become part of the school, hence it will unite all existing structures to make cooperation, scientific as well as transferable training, and social networking even more efficient and fruitful.

In 2015, the Max Planck Society approved new guidelines for junior scientists and the reorganization of the financial support structures. Most importantly, since July 2015 support contracts have to be awarded to all graduate students instead of fellowships. Max Planck Society support contracts allow combining the scientific freedom offered by a grant with the social security of an employment contract. They specify that the doctoral candidate applies all his/her working efforts on their scientific project and make sure that additional scientific work for the Institute must directly contribute to the scientific project of the doctoral candidate and must not unnecessarily prolong the time required for the dissertation.
Material resources, equipment and spatial arrangements

A unifying feature for all MPIMG’s research groups is the genomic approach to biology. The institute houses a range of large-scale equipment like sequencing systems, mass spectrometers, transmission electron or laser scanning microscopes, a new light sheet microscope, and a large IT infrastructure. Most of it is maintained by the institute’s service groups who operate the equipment and provide high-level support for all in-house scientists and many external collaborators. Details can be found in the reports of the individual service groups.

A major challenge during the last years (and still ongoing) has been the construction of the new tower 3, followed by the complete refurbishment of the neighboring towers 1 and 2. In 2013, tower 3 has been finished and handed over to the institute. The new tower connects tower 1 and 2 with the formerly separated tower 4, thus creating a coherent building for the MPIMG for the first time. It comprises the new main entrance for the whole institute, as well as three seminar rooms for conferences and events. The Department of Computational Molecular Biology and other theoretical research groups, as well as the IT service group occupy the upper floors. A large server room is located on the second floor. The new building enables the MPIMG to host seminars and small conferences up to 150 people in its own premises for the first time.

After tower 3 came to operation, tower 2 had to be cleared out completely, which meant that all departments and groups had to be distributed to
the remaining towers. In October 2013, the structural and technical refurbishment of tower 2 commenced, starting with the complete renewal of the façade including the replacement of windows and doors and a new insulation of the building shell. Tower 2 is expected to be completed and handed over to the MPIMG in spring 2016. The completely new labs will be equipped with state-of-the-art technical media for electrical energy, heating, cooling, water, ultra-pure water, steam, gas, and compressed air supplies, as well as for ventilation and air conditioning, fire detection technology, communication and building automation. Subsequently, tower 1 will be completely refurbished, too; its completion is scheduled for 2018.

Public relations work

The MPIMG has continued its public relations activities to inform about its work and discuss the implications of modern genome research with the public. Its regular communication with different target groups includes

- the distribution of press releases about scientific results and other themes of interest for the public;
- giving interviews to the press on scientific questions and topics of overall interest;
- a visiting program for school children to visit a lab, discuss with the scientists and perform simple experiments like DNA isolation, cell staining, or microscopy on their own;
- participation in the “Lange Nacht der Wissenschaften” (Long Night of Sciences) every second year. At this event, about 70-80 universities and research institutions all over Berlin and Potsdam open their doors for one night and invite the general public to visit their labs, learn about the work that is done here and discuss it with the scientists.
- participation in the Girls’ Day – Future Prospects for Girls, a large nationwide campaign, in which a wide range of professions and activities is presented to girls of 10 years upwards. Each year, a selected group of girls is invited to visit the institute for a whole day and to try out themselves at areas where usually only few females are active.

In addition, two major events have tied many forces at the institute during the last years. In October 2013, the inauguration of tower 3 has been celebrated with a ceremony followed by a Science Slam, conducted by the PhD students of the institute. The ceremony with short welcoming speeches from the State secretaries of the Land Berlin and the Federal Ministry of Education and Research, as well as of the vice presidents of the Max Planck Society and the Freie Universität Berlin has been attended by about 200 people, many of which did also attend the subsequent Science Slam.

In 2014, the MPIMG celebrated its 50th anniversary. After an official ceremony with a keynote lecture by Hans-Jörg Rheinberger, director emeritus of the Max Planck Institute for the History of Science and former PhD
student of the MPIMG, a celebratory symposium took place, covering the scientific fields, in which MPIMG researchers have been active during the last 50 years. One of the speakers was Ada Yonath, Nobel laureate in chemistry 2009 and now at the Weizmann Institute of Science, Israel, who worked at the MPIMG from 1979 to 1984, and then headed a Max Planck Working Group on Ribosome Structure at the German Electron Synchrotron (DESY) in Hamburg. Other speakers included Paulo Tavares, UPR CNRS 3296, France, Jörn Walter, Universität des Saarlandes, Saarbrücken, Han Brunner, Radboud University Nijmegen, The Netherlands, Peter Schuster, University of Vienna, Austria, Davor Solter, Institute of Medical Biology, A-STAR, Singapore, and Leroy Hood, Institute for Systems Biology, Seattle, Washington, USA. Many MPIMG alumni and friends of the institute came to the symposium and used the opportunity to meet former friends and colleagues and visit the MPIMG again.

In addition to the ceremony and the symposium, the MPIMG published a brochure aiming to consider the development of the MPIMG during the last 50 years in the scientific context of the times. Due to several requests, the brochure originally printed and published in German, has been translated to English and has been published online only in 2015, too.
Department of Developmental Genetics
Established: 11/2003

Head
Prof. Dr. Bernhard G. Herrmann
Phone: +49 (0) 30 8413-1409
Fax: +49 (0) 30 8413-1130
Email: herrmann@molgen.mpg.de

Secretary
Christel Siegers
(since 02/09, part time)
Phone: +49 (0) 30 8413-1344
Fax: +49 (0) 30 8413-1130
Email: siegers@molgen.mpg.de

Scientific assistant
Marta Caparros Rodriguez
(since 02/06, part time)
Phone: +49 (0) 30 8413-1344
Fax: +49 (0) 30 8413-1130
Email: caparros@molgen.mpg.de

Senior scientists
Lars Wittler (since 04/09)
Heinrich Schreve (since 09/04)
Hermann Bauer (since 10/03)
Phillip Grote (11/08-08/14)
Martin Werber (01/05-06/14)

Scientists
Lisette Lange (01/15-12/15)
Matthias Marks (01/15-12/15)
Jesse Veenhuijzen* (since 10/14)
Alexandra Amaral (since 03/14)
Sandra Schmitz (since 06/13)
Pavel Tsaytler* (since 09/12)
Frederic Koch* (since 11/11)
Tracie Pennimpede* (09/10-09/14)
Smita Sudheer (06/10-12/13)

PhD students
Jana Sie (since 06/15)
Bruno Pereira (since 11/14)
Jinhua Liu (since 11/13)
Lisette Lange (10/10-12/14)
Matthias Marks (10/10-12/14)
Constanze Nandy (02/09-12/14)
Arica Beisaw (06/10-10/14)
Karina Schöfisch (09/09-11/13)
Sabrina Schindler (08/08-11/12)

Master students
Pedro Gamez Rodriguez (since 06/15)
Dennis Schiffnerl (since 02/15)
Gesa Loof (05/14-10/14)
Stefania Chiochetti (02/14-07/14)
Max Meuser (05/13-05/14)
Theoidea Melissantis (07/13-12/13)
Ioana Weber (04/13-09/13)
Sophie Saussenthaler (10/12-06/13)
Vivien Vogt (10/12-06/13)
Marcel Thiele (10/12-04/13)
Stefanie Wolter (12/12-07/13)
Marina Knauf (09/12-06/13)

Bachelor students
Mareen Lüthen (06/14-08/14)
Anna Anurin (03/14-09/14)
Maximilian Pfeiffer (06/14-11/14)

Technicians
Maria Walther (12/14-12/15)
Sandra Währisch (since 03/09)
Gaby Bläß (since 02/09)
Karol Macura (Engineer, since 12/04)
Andrea König (since 09/04)
Manuela Scholze (since 09/04)
Jürgen Willert (since 11/03)
Barica Kusecek (09/08-10/14)

* externally funded
Introduction

The development of a complex organism such as a mouse or a human being from a single cell is still a “mystery” in many respects. The complexity of cellular interactions and responses and the resulting morphogenesis of a 3-dimensional structure consisting of different organs and tissues in a functional organism are still far from being fully understood. After a long phase of single gene analyses identifying a large number of essential regulators that are indispensable for particular processes, tissues, or organs, the developmental genetics field is now starting to take up genomics approaches, looking at regulatory mechanisms on a genome-wide scale. The broader view on transcriptomes and epigenomic changes during differentiation is accompanied by higher resolution with respect to the samples investigated. Previously, we used to look at the whole embryo or particular embryonic regions or tissues; now, we analyze cell types or even single cells marked by tissue-specific reporters. Therefore, the field has reached a new level of precision in analyzing the response of the genome to particular triggers or to the lack of a particular function in the context of a developmental process. In parallel, imaging tools have become more versatile, allowing the observation of even live mouse embryos, while embryogenesis is in progress. However, while we are beginning to understand the role
of transcription factors in regulatory networks controlling differentiation processes, a new class of regulatory RNAs, acting on the epigenetic level, has added another level of complexity to the picture. It is quite likely that understanding the role of noncoding RNAs in regulatory networks controlling embryogenesis will keep another generation of geneticists busy.

The major research interest of the Department of Developmental Genetics is focused on understanding the gene regulatory networks controlling trunk development from axial stem cells. We investigate genetic and epigenetic control mechanisms comprising important transcription factors as well as noncoding RNAs involved in trunk development by utilizing state-of-the-art technology.

A second line of research is focused on understanding a particularly intriguing phenomenon of non-mendelian inheritance called “transmission ratio distortion” (TRD). It allows a “selfish” genetic element to be transmitted at a high ratio from generation to generation, and is based on the complex interplay of several quantitative trait loci (QTLs) involved in the control of sperm motility. Our interest is centered on revealing the molecular mechanism causing this phenomenon, and on applying the properties of TRD to farm animal breeding.

**Scientific achievements and findings**

**Control mechanisms of mesoderm formation and trunk development in the mouse**

Cell differentiation requires differential expression of transcriptional regulators and their target genes. Important regulators of cell differentiation and organogenesis indeed frequently display a very specific pattern. Gene expression analysis of candidate genes therefore is a simple and very effective approach for identifying genes playing an important role in gene regulatory networks (GRN), controlling lineage choice, commitment, and differentiation. Previously, we have utilized *in situ* hybridization of individual genes in whole mount embryos (WISH) as a tool for the identification of important control genes. Currently, the transcriptome analysis of embryo parts or purified cell types, complemented by WISH of selected candidates is a faster and more effective tool for the identification of important genes. With the advent of the CRISPR/Cas technology for genome editing, the functional analysis of candidate genes is also facilitated. Therefore, tackling fundamental questions of developmental biology has become a lot more feasible.
The tissue-specific transcriptomic landscape of the mid-gestational mouse embryo

We have analyzed the transcriptomes of six major tissues dissected from mid-gestational (TS12) mouse embryos (Figure 1). Approximately one billion reads derived by RNA-seq analysis provided extended transcript lengths, novel first exons and alternative transcripts of known genes. We have identified 1,375 genes showing tissue-specific expression, providing gene signatures for each of the six tissues. In addition, we have identified 1,403 novel putative long noncoding RNA gene loci, 439 of which show differential expression.

Our analysis provided the first complete transcriptome data for the mouse embryo. It offers a rich data source for the analysis of individual genes and
gene regulatory networks controlling mid-gestational development (Werber et al., Development 2014).

The transcriptome data are complemented by expression data obtained by whole mount in situ hybridization of individual genes in E8.5-11.5 embryos. Approximately 8,000 genes have been analyzed, of which nearly 2,000 displayed differential expression in the embryo. The data on the latter are accessible in the MAMEP (Molecular Anatomy of the Mouse Embryo Project) database (http://mamep.molgen.mpg.de/).

The tissue-specific lncRNA *Fendrr* is an essential regulator of heart and body wall development in the mouse

In the transcriptome analysis of E8.25 mouse embryos we have identified a large number of differentially expressed long noncoding RNAs (lncRNAs). We isolated several by cDNA cloning and identified their expression pattern in mid-gestation embryos by WISH analysis. One of these showed a restricted expression in the nascent lateral mesoderm. We started a functional analysis of this gene, which we finally named *Fendrr* (Fetal-lethal non-coding developmental regulatory RNA). We targeted both alleles of *Fendrr* in ES cells and analyzed the outcome of the *Fendrr* loss-of-function mutation (Figure 2).

Figure 2: The lncRNA *Fendrr* is transiently expressed in nascent lateral plate mesoderm of mouse embryos and essential for heart and body wall development

(A) Schematic of the genomic region of *Fendrr* and *Foxf1*. (B) Whole-mount in situ hybridization of mouse embryos at E8.25 and E9.5. While *Foxf1* expression is maintained in differentiating lateral mesoderm, *Fendrr* expression is transient. (C, D) Transverse histological sections of E12.5 wild type and mutant embryos at the mid-trunk (C) or chest level (D). The body wall (C) and heart ventricle wall (D) of mutant embryos are thinner than in wild type. Abbreviations: rv, right ventricle; is, interventricular septum; cu, atrio-ventricular endocardial cushion; la, left atrium; lv, left ventricle. Scale bar: 200 µm. (E) The occupancy of the PRC2 component EZH2 at the promoters of the *Fendrr* target genes *Foxf1*, *Pitx2* and *Irx3* is strongly reduced in caudal end tissue lacking *Fendrr*. Figure adapted from Grote et al., Dev Cell 2013.
We showed that *Fendrr* is essential for proper heart and body wall development in the mouse. Embryos lacking *Fendrr* displayed a strong omphalocoele phenotype and heart dysfunction at E13.75 followed by embryonic death. Several transcription factors controlling lateral plate (giving rise to the body wall) or cardiac mesoderm (forming the heart) differentiation were upregulated. How does that come about?

Many long non-coding RNAs (lncRNAs) can bind to the histone modifying complexes PRC2 or TrxG/MLL, which set repressive or active marks, respectively, at the chromatin. These marks have an immediate effect on gene expression and/or become effective in the descendant of the cells, in which they have been set. We showed that *Fendrr* can bind to either of these complexes. We also identified three direct target genes of *Fendrr*, *Foxf1*, *Irx3* and *Pitx2*, and showed that the PRC2 occupancy at their promoters is drastically reduced in *Fendrr* mutants, going along with decreased H3K27 trimethylation (the repressive mark set by PRC2) and an increase in expression. We concluded that in wild type embryos *Fendrr* anchors PRC2 to regulatory target sequences. It thereby changes the epigenetic landscape by increasing the repressive mark, which causes down-regulation of target gene expression. Thus, we identified a lncRNA that plays an essential role in the regulatory networks controlling the fate of lateral mesoderm derivatives. (Grote et al., Dev Cell 2013; this article has been selected for F1000Prime).

We confirmed by ChIRP analysis that in differentiating cells *Fendrr* indeed binds to its target promoters at *Foxf1* and *Pitx2*. Our data suggest triplex formation as binding mechanism. Intriguingly, there is a delay of six days between the time of brief *Fendrr* expression in cardiac precursors within the lateral mesoderm at E6.5-7.0, and the heart dysfunction taking effect at E12.5 due to lack of *Fendrr* in precardiac mesoderm. From this observation we suggest that the function of *Fendrr* in cardiac precursor cells is to “imprint” the epigenetic landscape of its target regulatory elements. Perturbation of this imprint interferes with proper heart development (Grote & Herrmann, RNA Biol 2013).
What is the role of long noncoding RNAs in organogenesis?

Many lncRNAs have been shown to interact with histone modifying complexes and/or transcriptional regulators. Via such interactions, many lncRNAs are involved in controlling the activity and expression level of target genes, including important regulators of embryonic processes, and thereby fine-tune gene regulatory networks controlling cell fate, lineage balance, and organogenesis. Intriguingly, an increase in organ complexity during evolution parallels a rise in lncRNA abundance (Figure 3). The current data suggest that lncRNAs support the generation of cell diversity and organ complexity during embryogenesis, and thereby have promoted the evolution of more and more complex organisms (Grote & Herrmann, Trends Genet 2015).

SRF is essential for mesodermal cell migration during embryonic axis elongation

Mesoderm formation in the mouse embryo initiates around E6.5 at the primitive streak and continues until the end of axis extension at E12.5. It requires the process of epithelial-to-mesenchymal transition (EMT),
wherein cells detach from the epithelium, adopt mesenchymal cell morphology, and gain competence to migrate. It was shown previously that, prior to mesoderm formation, the transcription factor SRF (Serum Response Factor) is essential for the formation of the primitive streak. To elucidate the role of murine SRF in mesoderm formation during axis extension, we conditionally inactivated SRF in nascent mesoderm using the T(s):Cre driver mouse. Defects in mutant embryos became apparent at E8.75 in the heart and in the allantois. From E9.0 onwards, body axis elongation was arrested. Using genome-wide expression analysis, combined with SRF occupancy data from ChIP-seq analysis, we identified a set of direct SRF target genes acting in posterior nascent mesoderm, which are enriched for transcripts associated with migratory function. We further show that cell migration is impaired in SRF mutant embryos (Figure 4). Thus, the primary role for SRF in the nascent mesoderm during elongation of the embryonic body axis is the activation of a migratory program, which is a prerequisite for axis extension (Schwartz et al., Mech Dev 2014).

Figure 4: SRF-deficiency in mesodermal cells affects cell morphology and cytoskeletal arrangement

(A) Representative images of caudal end explants of control and mutant embryos subjected to ex vivo migration assays. Images on the right show magnifications of the boxed areas in the middle panels. Note the cell shape change of mutant cells.

(B) Representative fluorescence images from cells at the border of the explant culture after 48 hours, with Actin visualized in the migratory cells using phalloidin (green) and nuclei with DAPI (blue). Scale bar represents 50μm. (C) Immunofluorescent staining of focal adhesions (Vinculin, red) and markers of mesenchymal cells (Vimentin, non-membrane bound E-cadherin, red). Counterstaining was performed for Actin (phalloidin, green) and nuclei (DAPI, blue). Scale bar represents 50μm. In mutant cells stress fibers are reduced and focal adhesions are impaired. Figure adapted from Schwartz et al., Mech Dev 2014.
The \( t^{18} \)-lethal mutation reveals an essential role of phosphatase 2A in mesoderm formation

Until the 1980s, the \( t \)-complex on mouse chromosome 17 has been regarded as the major genomic region controlling embryonic development. Sixteen mutations affecting early mid-gestational development had been isolated from natural mouse populations. However, until last year, only one of them, the \( t^{e5} \)-lethal had been molecularly identified. We have decided to identify \( t^{e18} \), one of the most interesting \( t \)-lethals affecting early mesoderm formation. In \( t^{e18} \) homozygous embryos the epiblast cells are unable to execute the transition from an epithelial to a mesenchymal cell type (EMT). Instead, they switch back to proliferation, fill the preamniotic cavity and cause embryonic arrest. The switch mechanism between the epithelial and the mesenchymal cell state is fundamental for many embryonic processes and also for the metastasis formation of tumors. Therefore, we decided to elucidate the molecular nature of the \( t^{e18} \)-lethal. The \( t^{e18} \)-lethal is caused by a large deletion of some 4 Mb in size. However, we succeeded in identifying the causative gene locus and verified by single gene knockout and transgenic rescue that the loss of the phosphatase 2A scaffold protein PPP2R1A is causing the \( t^{e18} \) phenotype. We showed by RNA-seq that two signaling pathways, WNT and Nodal/TGFb, pivotal regulators of mesoderm formation, are affected by loss of PP2A function. Proteome analyses identified several candidates of PPP2R1A interactors, which might directly be involved in the EMT switch. Further characterization of these candidates is still ongoing.

SOX2 and BRACHYURY competitively control the lineage choice of neuro-mesodermal progenitors in the trunk

Genetic data have previously provided evidence that Brachyury (T), Wnt3a and the FGF receptor Fgfr1 are involved in a regulatory loop controlling axial elongation from the mid-trunk to the tail end. They presumably act in axial stem cells, which give rise to neuro-mesodermal progenitors (NMPs). The latter are able to form both the neuro-ectodermal and mesodermal lineages. Recently the pluripotency and neural differentiation factor Sox2 has been shown to be co-expressed with T in NMPs (Figure 5). This led to the suggestion that Sox2 and T determine the two lineages. However, several publications have contradicted this view by suggesting that Sox2 and thus the neural lineage is antagonized by Tbx6 (Tbx16 in zebrafish), driving NMPs to the mesodermal lineage, rather than by T. We have utilized T- and Sox2-reporters for FACS sorting of NMPs and their descendants, and subjected them to RNA-seq and ChIP-seq analysis. Our data suggest that indeed in vivo T is counteracting Sox2 in NMPs, undergoing lineage decisions towards the neuro-ectodermal or mesodermal fate. T represses important control genes of the neural lineage, while it activates mesoder-
mal genes. From several sets of data we conclude that T and Sox2 establish the neural and mesodermal lineages from NMPs, respectively, while Tbx6 acts downstream of T and promotes the formation of paraxial mesoderm.

The role of the Mediator subunit Med12 during mouse development

Multiple signaling pathways responsible for homeostasis, cell growth, and differentiation converge on the Mediator complex through transcriptional regulators that target one or more of its subunits. Mediator senses signals and then delivers a properly calibrated output to the transcriptional machinery. Various subunits are required for the interaction with specific transcription factors; others have a broader role, required for all functions of the complex. The MED12 subunit participates in many of the general functions, but also in gene-specific processes.

As Med12 null mouse embryos show lethality during early embryogenesis, we used a conditional Med12 line to clarify functions of Med12 at later
stages of development with cre deleter lines. Heterozygous females that express Med12 in a mosaic fashion display neural tube closure defects and various other planar cell polarity (PCP) related defects, including misorientation of sensory hair cells in the inner ear, cleft palate, and open eyelids at birth, suggesting that a potential general PCP transcriptional regulator could act via Med12 to control all PCP-related processes. Glia-specific Med12 deletion leads to terminal differentiation defects of myelinating glia caused by the disability of Sox10 to activate its targets in the absence of Med12. Embryos lacking Med12 in mesenchymal cells of the limb buds fail to activate Sox9 target genes and show severe chondrogenesis defects.

The Ørom laboratory identified MED12 in association with long ncRNAs to control gene expression. Interestingly, MED12 mutants involved in the Intellectual Disability (ID) syndrome Opitz-Kaveggia cannot bind these long ncRNAs and fail to induce transcription of target genes. We are currently establishing mouse cell lines with MED12/ID mutations to study the regulatory activity of these mutant forms.

**Planned developments**

We will continue to investigate the gene regulatory networks controlling mesoderm formation and trunk development, in particular the identification of axial stem cells, the switch from the stem to the progenitor state, lineage choice and differentiation, with a particular focus on long non-coding RNAs. We will try to establish an organoid-like culture system for axial stem cells allowing the analysis of trunk development *in vitro*, thereby supporting and accelerating the analysis of the principal mechanisms acting in these processes.

**Transmission Ratio Distortion: elucidating the molecular strategies of a “selfish” genetic element**

Not all genes follow Mendel’s rules. There are “selfish” genetic elements promoting their own transmission to the next generation on the expense of the “sister” allele. The mouse *t*-haplotype, a segment of some 40 Mb on chromosome 17, is an example of such a “selfish” genetic element existing in the mouse. From heterozygous (*t/+*) males it can be transmitted to up to 99% of their offspring. It contains a “selfish” gene, called “responder” (*Tcr*), and several auxiliary factors, which additively assist the responder in getting the *t*-haplotype passed on to the next generation at a high frequency. The auxiliary factors are genetic variants; quantitative trait loci (QTL) called “distorter”. The distorters impair the motility behaviour of all spermatozoa produced by a *t/+* male. The responder is able to rescue this malfunction of the sperm. However, it does so only in half of the spermatozoa, those that carry the *t*-haplotype and the responder gene. That’s
how the latter gain an advantage in reaching the egg cells first and fertilize them. A clever trick: the \( t \)-haplotype manipulates the sperm motility control mechanism in a way to make sure that the sperm carrying it wins the race for life.

We have isolated and analyzed several key components causing TRD, the responder (encoding the kinase SMOK\(_{TCR}\)) and the distorters \( Tcd1a \) (Tagap), \( Tcd2a \) (Fgd2) and \( Tcd2b \) (Nme3). The latter code for regulators of Rho small G proteins, known to control cell motility of slow migrating cells such as fibroblasts. The GAP (GTPase activating protein) TAGAP1 is a negative regulator of RHOA and possibly of RAC1, the GEF (GDP/GTP exchange factor) FGD2 is a positive regulator of CDC42. NME3 is a nucleoside diphosphate kinase, converting GDP to GTP required for activating RHO proteins. They all are involved in controlling the activity of

![Figure 6: Model of the molecular mechanism causing transmission ratio distortion.](image)

The \( t \)-haplotype encodes the distorter genes, Tagap\(_{Tcd1a}\), Tiam2\(_{Tcd1b}\), Fgd2\(_{Tcd2a}\) and Nme3\(_{Tcd2b}\), which are shared between spermatids (arrows crossing cells connected in a syncytium), and later in spermatozoa act as control factors of RHO small G proteins regulating the protein kinase SMOK through activating and inhibitory pathways. SMOK controls the directional movement of spermatozoa. The enhanced activation and reduced inhibition of SMOK by increased (red upward arrow) or reduced (blue downward arrow) activity of the distorters together results in hyperactivation of SMOK, and thus in the impairment of axonemal function (red downward arrow) and sperm motility. This deleterious effect of the distorters is rescued by the dominant-negative kinase SMOK\(_{TCR}\) encoded by the responder; however, this happens exclusively in \( t \)-sperm. Thus \( t \)-sperm have an advantage over \( + \)-sperm in fertilizing the egg cells resulting in transmission ratio distortion in favor of the \( t \)-haplotype. Black arrows indicate activation, blocked lines inhibition; red upward pointing arrows indicate up-regulation, blue down-pointing arrows down-regulation; the green down-pointing arrow symbolizes rescued, the dark-red down-pointing arrow impaired flagellar motility.
the protein kinase SMOK, which appears to regulate the flagellar behavior of spermatozoa. A model how the interaction of all components results in TRD is shown in Figure 6.

The GEF TIAM2 has a dual role in TRD

Recently we have identified another distorter, Tcd1b, acting on TRD. It encodes the GEF TIAM2, which is known to activate RAC1. Tiam2 is expressed in two mRNA variants, Tiam2L and Tiam2S. Tiam2L encodes the full-length protein, Tiam2S a truncated protein retaining the catalytic DH/PHc domain. Tiam2L expression from the t-allele is reduced, while Tiam2S expression is enhanced compared to wild-type. By transgenic approaches we showed that Tiam2L has a negative effect on t-haplotype transmission, while Tiam2S promotes t-transmission. Thus, differential promoter usage at the t-allele has evolved in a manner that is optimal for t-transmission. The wild-type allele, in contrast, shows the opposite pattern counteracting the t-allele.

Perspectives

We will continue research to establish a direct link between the distorters, SMOK and the axoneme, in order to prove that TRD is indeed established via sperm motility control. We will also try to obtain proof that excessive activity of the distorters causes male sterility, as proposed by Lyon. In terms of translational approaches we are trying to establish sex ratio distortion in the mouse and to apply the system to farm animals.

Mechanisms of intestinal tumor formation

A tumour-specific gene signature of mouse intestinal adenoma identified by DNA-methylome analysis is partially conserved in human colon cancer

Aberrant CpG methylation is a universal epigenetic trait of cancer cell genomes. However, human cancer samples or cell lines preclude the investigation of epigenetic changes occurring early during tumour development. We have used MeDIP-seq to analyze the DNA methylome of APC\textsuperscript{Min} adenoma as a model for intestinal cancer initiation, and we present a list of more than 13,000 recurring differentially methylated regions (DMRs) characterizing intestinal adenoma of the mouse. We showed that Polycomb Repressive Complex (PRC) targets are strongly enriched among hyper-methylated DMRs, and several PRC2 components and DNA methyltrans-
34

ferases were up-regulated in adenoma. We further demonstrated by bisulfite pyrosequencing of purified cell populations that the DMR signature arises *de novo* in adenoma cells rather than by expansion of a pre-existing pattern in intestinal stem cells or undifferentiated crypt cells. We found that epigenetic silencing of tumour suppressors, which occurs frequently in colon cancer, was rare in adenoma. Quite strikingly, we identified a core set of DMRs, which is conserved between mouse adenoma and human colon cancer, thus possibly revealing a global panel of epigenetically modified genes for intestinal tumours. Our data allow distinguishing between early conserved epigenetic alterations occurring in intestinal adenoma and late stochastic events promoting colon cancer progression, and may facilitate the selection of more specific clinical epigenetic biomarkers (Figure 7, Grimm et al., PLoS Genetics 2013).

**Figure 7:** A model for stepwise formation of cancer cell CpG epigenomes.
CpG methylation is uniform within the normal cellular hierarchy of the intestine (blue, to the left). Upon tumour initiation, recurring CpG methylation patterns form, guided by an instructive mechanism that is linked to PRC2 for hypermethylated sites (blue to green). Further CpG methylation changes occur slowly, probably in a stochastic manner. A fraction of these bestow tumour cells with a selective advantage and are subject to clonal expansion during tumour progression (green to red). Figure adapted from Grimm et al, PLoS Genetics 2013.

Oncogenic BRAF induces loss of stem cells in the mouse intestine and is antagonized by β-catenin activity

Colon cancer cells frequently carry mutations that activate the β-catenin and mitogen-activated protein kinase (MAPK) signaling cascades. Yet how oncogenic alterations interact to control cellular hierarchies during tumor initiation and progression is largely unknown. We found that oncogenic BRAF modulates gene expression associated with cell differentiation in colon cancer cells. We therefore engineered a mouse with an inducible oncogenic BRAF transgene, and analyzed BRAF effects on cellular hierarchies in the intestinal epithelium *in vivo* and in primary organotypic culture. We demonstrated that transgenic expression of oncogenic BRAF in the mouse strongly activated MAPK signal transduction, resulted in the rapid development of generalized serrated dysplasia, but unexpectedly also induced depletion of the intestinal stem cell (ISC) pool (Figure 8). Histological and gene expression analyses indicate that ISCs collectively con-
verted to short-lived progenitor cells after BRAF activation. As Wnt/β-catenin signals encourage ISC identity, we asked whether β-catenin activity could counteract oncogenic BRAF. Indeed, we found that intestinal organoids could be partially protected from deleterious oncogenic BRAF effects by Wnt3a or by small-molecule inhibition of GSK3β. Similarly, transgenic expression of stabilized β-catenin in addition to oncogenic BRAF partially prevented loss of stem cells in the mouse intestine. We also used BRAFV637E knock-in mice to follow changes in the stem cell pool during serrated tumor progression and found ISC marker expression reduced in serrated hyperplasia forming after BRAF activation, but intensified in progressive dysplastic foci characterized by additional mutations that activate the Wnt/β-catenin pathway. Our study suggests that oncogenic alterations activating the MAPK and Wnt/β-catenin pathways must be consecutively and coordinately selected to assure stem cell maintenance during colon cancer initiation and progression. Notably, loss of stem cell identity upon induction of BRAF/MAPK activity may represent a novel fail-safe mechanism protecting intestinal tissue from oncogene activation (Riemer et al., Oncogene 2014).

Cooperation within the institute

**Research Group Development and Disease**

- Sigmar Stricker: *The role of Osr1 during embryonic development*
- Malte Spielmann: *Generation of genetically altered mouse lines and embryos for the analysis of genomic rearrangements at the Pitx1 locus.*
- Katerina Kraft: *Generation of genetically altered mouse lines and embryos carrying CRISPR/Cas9 engineered structural variants.*
- Dario Lupianez: Establishment of genetically altered mouse lines, embryos and ES-Cell lines to analyze gene-enhancer interactions within topological chromatin domains.

**Department of Computational Biology**
- Ralf Herwig: Analysis of molecular mechanisms of tumor suppression by genetic variants.

**Otto Warburg Laboratory**
- Marie-Laure Yaspo: TREAT20 Consortium “Tumor REsearch And Treatment - 20 Patient Pilot” funded by the BMBF
- Ulf Ørom: The function of IncRNA-Mediator complexes in neurological disease

**Sequencing Core Facility/Bernd Timmermann**
- Transcriptome sequencing of embryonic tissues derived from E8.25 mouse embryos
- Genome sequence and assembly of the mouse t-haplotype t^5 on the C57BL/6 background
- Genome sequence and assembly of the Mus m. musculus strain PWD/Ph.
- RNA-seq analysis of staged mouse testes
- RNA-seq and ChIP-seq of various libraries from FACS purified embryonic cells
- RNA-seq analysis of intestinal organoids and spheroids

**Mass Spectrometry Facility/David Meierhofer**
- Identification of phosphorylation sites on SMOK and its target proteins
- Proteome analysis of wild-type and t-sperm from mouse

**Microscopy and Cryo Electron Microscopy Group/Thorsten Mielke**
- Localization of SMOK and its interactions partners in spermatozoa by high-resolution immunocytochemistry

**Special facilities and equipment**

The Department of Developmental Genetics provides personnel and expertise to the Transgenic Unit of the institute. It produces transgenic embryos and mice from genetically modified ES cells for various groups at the institute. It also provides expertise in handling and culturing ES cells, and gives advice in vector construction and generation of knock-out and other genetically modified mice.

Members of the department supervise the Fluorescence Activated Cell Sorter (FACS) of the MPIMG. Another important piece of equipment is a 2-Photon Laser Scanning Microscope of the institute, which is housed in the department.
General information about the whole department

Complete list of publications (2009 – 2015)
Department members are underlined.

2015


2014


2013


2012


Krautzberger AM, Kosiol B, Scholze M & Schreve H (2012). Expression of Vasorin (Vasn) during embryonic development of the mouse. Gene Expression Patterns, 12, 167-171


Pennimpede T, Proske J, König A, Vidigal JA, Morkel M, Jesper B Bramsen, Herrmann BG & Wittler L (2012). In vivo knockdown of Brachyury results in skeletal defects and urogenital malformations resembling...


2011


2010


Rocha PP, Scholze M, Bleiß W, Schreve H (2010). Med12 is essential for early mouse development and for canonical Wnt and Wnt/PCP signal-


2009


**Scientific honours**

Bruno Pereira: *Young Researcher Poster Price*, EMBO Workshop: Embryonic-Extraembryonic Interfaces, Göttingen 2015

Pedro Rocha: *Best talk at the NucSys Marie Curie Research Training Network*, Wageningen, 2010

**Appointments of former members of the Department**

Phillip Grote: appointment as *head* of the research group "LncRNAs in Cardio-pulmonary Development", Goethe University Frankfurt, Institute of Cardiovascular Regeneration, Center for Molecular Medicine, September 2014

**PhD theses**


Arica Beisaw: *Investigating the role of Brachyury (T) in the control of chromatin state in early mesodermal cells*, Freie Universität Berlin, 2014

Matthias Marks: *Investigations into expression and function of the murine Fam181 gene family*. Freie Universität Berlin, 2014


Karina Schöfisch: *Untersuchungen zur Haploidspezifität des t-Komplex Responder*. Freie Universität Berlin, 2013


Eun-ha Shin: *Transcriptome analysis of Bmp4-induced mesoderm formation in vitro*. Technische Universität Berlin, 2011

Marc Leushacke: *Functional Characterisation of genes that regulate Intestinal Tumor Progression*. Freie Universität Berlin, 2011

Pedro Rocha: *Med12 is an essential coordinator of gene regulation during mouse development*. Freie Universität Berlin, 2011

Joana Vidigal: *An inducible RNAi system for the functional dissection of genes in the mouse*. Freie Universität Berlin, 2011

Anja Michaela Mayer: *Analysis of the expression pattern and knock-out phenotype of Slit-like 2 (Slitl2) in the mouse*. Freie Universität Berlin, 2010

Nathalie Verón: *Untersuchungen zu den molekularen Grundlagen der nicht-mendelschen Vererbung in der Maus*. Freie Universität Berlin, 2009

Solveig Müller: *Funktionelle Charakterisierung regulatorischer Gene bei der Bildung der Wirbelsäule der Maus*. Freie Universität Berlin, 2009

### Student theses


Anna Anurin: *Characterization of the early differentiation of mouse embryonic stem cells from the naïve pluripotent state towards the primed epiblast stem cell state*. Freie Universität Berlin, Bachelor thesis, 2014


Maximilian Pfeiffer: *Transcriptome assembly and discovery of IncRNA’s as potential pluripotency markers in*
the mouse embryo, Freie Universität Berlin, Bachelor thesis, 2014


Lisette Lange: Functional characterization of the Brachyury Interacting Protein METTL2, Master Thesis, Humboldt University of Berlin, 2010

Mathias Marks: Characterization of genes involved in somitogenesis, Master Thesis, Humboldt University of Berlin, 2010


Teaching activities

Bernhard Herrmann is Professor for Genetics and Head of the Institute for Medical Genetics at the Charité-University Medicine Berlin, Campus Benjamin Franklin, and involved in the teaching of medical students. All scientists within the department participate in the teaching of students. We teach embryology in a practical course and a seminar to medical students, and provide a 3-week-long practical course with lectures and tutorials to students of the Masters Program “Molecular Medicine” at the Charité-University Medicine Berlin. In addition, we offer a 4-week-long practical course with lectures and tutorials to students of the Biology and Biochemistry curricula at the Free University Berlin. We also offer a 1-week lab course with lectures to students of the MIPMG. These activities bring both early and later career scientists from the department in contact with students and help them develop their teaching skills. At the same time students of the Life Sciences are introduced to state-of-the-art developmental biology concepts and techniques used in stem cell biology and developmental genetics in the mouse, which is not taught anywhere else in Berlin. In this way we hope to induce an interest into the fascinating field of stem cell and developmental biology.

In addition all scientists of the department are involved in teaching individual students taking courses in the department, or doing their practical work for a Bachelor or Master degree.
Department of
Computational Molecular Biology
Established: 10/2000

Introduction

Computational biology studies biological questions with mathematical and computational methods. In the area of molecular biology and genomics, the possibility to apply such formal methods, of course, comes from the availability not only of genome sequences, but also of large amounts of functional data about cellular processes. Computational molecular biology encompasses both development and adaptation of methods in the areas of mathematics, statistics, and computer science, as well as pursuing biological questions applying these tools and close collaborations with experimentalists. The research interest of the Computational Molecular Biology Department lies in understanding gene regulatory mechanisms in the
context of structure and evolution of the eukaryotic genome. To this end, mathematical, computational, and also experimental approaches are being developed and employed. Within the MPI, computational approaches have also become an integral part of many of the research projects pursued.

**Structure and organization of the department**

The department comprises scientists with backgrounds ranging from mathematics, statistics and computer science via physics to biology and genetics. They are organized in several project groups, the largest of which is the *Transcriptional Regulation Group* headed by Martin Vingron. The work of this group focuses on theoretical concepts in gene regulation, regulatory networks and epigenetic aspects of regulation. The project group *Sequencing & Genomics* headed by Stefan Haas focusses on the analysis of sequencing data in human genetics and cancer genomics. Sebastiaan Meijsing heads up a project group *Mechanisms of Transcriptional Regulation*, where experimental and computational work is being combined in order to study the glucocorticoid receptor as a model transcription factor. Peter Arndt heads an independent research group on *Evolutionary Genomics*, which works on developing models how the DNA in humans and other non-human primates has evolved.

Two project leaders from the Lehrach department joined the Vingron department after the retirement of Hans Lehrach. Ralf Herwig focusses on genome analysis, mostly in the context of cancer genomics and disease bioinformatics. With the interest in cancer genomics, he is close to the work of Stefan Haas and also to the group of Marie-Laure Yaspo from the Otto Warburg Laboratory. Together, these groups are forming an emanating research cluster within the institute. Margret Hoehe, who also came from the Lehrach department, has for many years worked on human haplotyping and is continuing this work now mostly through computational analysis.

**Scientific methods and findings**

During the last couple of years, the efforts on prediction of transcription factor binding sites have been continued and techniques have matured to a state, where we are actually offering a web server and software package for transcription factor target prediction, motif enrichment, and ChIP-seq peak analysis. The methods have been widely applied, both within many of our own projects and in various collaborations. Most notably, the application of the target prediction methods has led to the discovery, in the context of an international collaboration, of a network of inflammation-related genes regulated by the transcription factor Irf7, with genes from this network contributing the Type I diabetes risk in humans.
Chromatin structure and epigenetic modifications also play major roles in transcriptional regulation and we have spent increasing efforts on integrating these aspects into our view of gene regulation. These efforts have led to a model for predicting gene expression from the histone modifications that are observed in promoter regions of genes. This connection is remarkable, because transcription factors do not play a role in this analysis, seemingly contradicting the traditional understanding. It reinforces the point that transcriptional control is exerted on many levels and those levels appear to be highly interrelated. This work has received considerable attention in the community and has given rise to many similar analyses.

While the above analyses have been purely computational, the Meijsing group works primarily experimentally, namely on the glucocorticoid receptor as a model system for transcriptional regulation. The experiments in this group are guided by the department’s understanding of the possibilities of computational analysis and the feedback loop between experiment planning, experimentation, and analysis has become very short.

Evolutionary analyses of genomes and of genetic regulation in the Arndt and Vingron groups have yielded new insights into forces shaping our genome and its regulatory networks. Many of these findings rest on Arndt’s detailed study of the Cytosine methylation-deamination process, as it can be extracted from the differences between complete, very similar genomes. Given the well-known regulatory role of CpG-islands, many people in the department have been in one way or the other working on the evolution or transcriptional functions of these genomic elements.

In an intensive collaboration with the Human Genetics Department of Hilger Ropers, project group Kalscheuer, Stefan Haas and co-workers have designed and implemented an analysis pipeline for next generation sequencing data, in particular RNA-seq data and ChIP-seq data. The RNA-seq pipeline is particularly geared towards identification of mutations in sequenced genomes of patients with particular disease phenotypes. Within this collaboration, it was employed, among others, to uncover mutations involved in intellectual disability. The ChIP-seq pipeline serves primarily the analysis of gene regulatory networks. After the retirement of Hilger Ropers, the work will be continued in collaboration with Vera Kalscheuer moving to the research group Development & Disease (Stefan Mundlos).

As can be seen from this short summary, the methods employed in the department are, with the exception of the experimental work of Sebastiaan Meijsing, mostly theoretical. Mathematics, statistics, and computer science supply the basis for the bioinformatics analysis performed. Projects may either approach biological questions through theoretical analysis, or may be collaborative projects, where frequent interaction and feedback between experimentalists and theoreticians drive the work forward.
Material resources, equipment, and spatial arrangements

The theoretical work of the Computational Molecular Biology Department relies heavily on powerful computers. The MPIMG commands a powerful compute cluster and several individual servers with 48 or 64 processors and containing 256 GB or 512 GB of RAM, respectively. This architecture serves the classical sequence analysis, the numerical calculations, and the analysis of next generation sequencing data. The cluster is not accessed by the researchers directly, but through a queuing system. Storage space on hard disks in the institute comprises approximately 7 PB and the department participates in this. The computer set-up is maintained by the institute’s IT group. Sebastiaan Meijsing, who does experimental work, has laboratory space in tower 4.

Teaching

Department members contribute substantially to the bioinformatics curriculum at Freie Universität Berlin. Over the last years, courses and practical sessions have been taught by Peter Arndt, Alena van Bömmel, Juliane Perner, Mahsa Ghanbari, Mohammad Sadeh, Matt Huska, Stefan Haas, and Martin Vingron. We teach courses like Algorithmic Bioinformatics, Population Genetics, or Network Biology. Students can do internships, practical courses, and thesis work with members of the department. This brings many bright, young students to the department and at the same time allows the university to show the students a much larger spectrum of bioinformatics than would normally be possible in the university framework. In cooperation with the university, the department has established the International Max Planck Research School for Computational Biology and Scientific Computing (IMPRS-CBSC, http://www.imprs-cbsc.mpg.de/), which is now approaching the end of its second 6-year funding period. In the context of the IMPRS, we organize, e.g., a summer school. In the future, a new IMPRS shall be applied for which is then to bridge from computational biology to experimental genomics. Martin Vingron is also a co-author on the latest edition of the standard German textbook on molecular genetics entitled “Molekulare Genetik” and edited by Alfred Nordheim and Rolf Knippers.
Cooperation with national and international research institutions

Very strong ties exist between our department and the Bioinformatics Group of Knut Reinert at Freie Universität Berlin. We have many joint IMPRS students, who move freely between the MPIMG and the university. Reinert develops NGS analysis algorithms, which are employed at the MPIMG, and the needs that arise at the MPIMG influence the algorithm development at the university. The strong link is also reflected in the Max Planck Fellow status that was conferred to Knut Reinert in 2014.

Department members have also established many long-standing, fruitful ties to other colleagues and institutions. Vingron collaborates with Norbert Hübner from the Max Delbrück Center for Molecular Medicine (MDC) Berlin-Buch on functional and quantitative genetics. A senior postdoc, Matthias Heinig, has acted as liaison between the groups, a model, which was exceedingly successful and has led to many high-ranking publications. At the same time, the two groups have been members of two successive EU consortia. Several other collaborations started through visits of PhD students to other laboratories. In that way, very successful collaborations with Ewan Birney (EBI), with Huck-Hui Ng (Genome Institute of Singapore), and Wolfgang Huber (EMBL) started.

Stefan Haas collaborates intensively with the group of Roman Thomas at the Medical School in Cologne. This connection has led to a thread of high-level publications on cancer genomics. The project is funded by the Deutsche Krebshilfe [German Cancer Aid].

During the reporting period, we have further been involved in a number of national and international projects and collaborations. We are part of the the European BLUEPRINT consortium for creating epigenetic maps. On the German level, the counterpart to this is the German Epigenome project DEEP, where the groups of Ho-Ryun Chung and Martin Vingron are members. Vingron is also a part of the BMBF funded CancerEpiSys project to study epigenetics of chronic lymphocytic leukemia (CLL). In terms of DFG funding [Deutsche Forschungsgemeinschaft, German Research Foundation], Marsico and Vingron are part of the Transregional Collaborative Research Center “Innate Immunity of the Lung” (SFB-TR 84) and of a DFG-funded graduate school, the Research Training Group “Computational Systems Biology” (DFG-Graduiertenkolleg 1772). Sebastiaan Meijsing has one DFG grant and a grant from a private foundation (Else Kröner-Fresenius-Stiftung). The Collaborative Research Center for Theoretical Biology (Sonderforschungsbereich, SFB 618) has reached its maximal twelve year duration during the reporting period and has come to an end. Likewise the EU project EURATRANS has run out during this reporting period.
Planned developments

Gene regulation, evolution, and disease bioinformatics are the overarching questions pursued in the department. With the recognition of the importance of epigenetics in gene regulation, we are now trying to understand both of these levels of regulation as well as their interplay. All this is, of course, reflected in the architecture of the genome, which we are trying to understand from an evolutionary as well as from a mechanistic point of view. Our work relies on the plethora of data that is available these days. Also genome analysis is instrumental in understanding congenital diseases and cancer genomics. We are thriving to bring the knowledge about regulatory interactions to bear on the questions of disease bioinformatics as well, because we suspect that many so-far unrecognized disease mechanisms are related to regulatory changes in the genome. Much effort goes into finding or developing the best methods to extract answers to our biological questions from the genomic and functional data. At the same time, we are intensifying the collaborations between theoreticians and experimentalists in order to continuously test our hypotheses, and to come up with experimental procedures that produce the most information gain.
Transcriptional Regulation Group
Established: 10/2000

Head
Prof. Dr. Martin Vingron
Phone: +49 (0)30 8413-1150
Fax: +49 (0)30 8413-1152
Email: vingron@molgen.mpg.de

Scientists and postdocs
Robert Schöpflin (since 9/14)
Xinyi Yang (since 07/13)
Lilian Villarin* (since 01/13)
Alena van Bömmel* (since 10/12)
José Maria Muino Acuna* (since 07/12)

Matthias Heinig* (10/10-05/15)
Navodit Misra (01/11-12/14)
joint with Haas group)
Brian Caffrey* (04/13-08/14)
Guofeng Meng (09/11-6/2014)
Annalisa Marsico* (09/10-05/14)
Roman Brinzanik (01/07-10/13)
Julia Lasserre* (04/08-08/13)
Tomasz Zemojtel (03/04-12/12)
Alexander Bolshoy* (09/11-02/12; 09/12)
Morgane Thomas-Chollier* (04/09-08/12)
Ewa Szczurek (04/11-06/12)
Ho-Ryun Chung (06/05-12/11)

PhD students
Emmeke Aarts (since 10/14)
Edgar Steiger (since 09/14)
Anna Ramisch (since 02/14, IMPRS)
Mohammad Hossein Moeynizadeh (since 09/13, IMPRS)
Wolfgang Kopp (since 09/12, IMPRS)
Mahsa Ghanbari* (since 10/11)
Matthew Huska (since 09/11, IMPRS)
Juliane Perner* (12/10-09/15, IMPRS)
Alena van Bömmel (10/09-10/12)
Jonathan Göke (09/07-09/12)
Rosa Karlic (10/07-12/11, IMPRS)
Ewa Szczurek (10/06-04/11, IMPRS)

Scientific overview
The field of transcriptional regulation has gone through a rapid development over the last couple of years. This is due to the plethora of whole-genome sequence data and the functional genomics data on gene expression, DNA-binding proteins, and epigenetics, which have become available (e.g., the ENCODE data). The group works on exploiting this data for the purpose of gaining a better understanding of transcriptional regulation in eukaryotes. The main questions lie in the identification of regulatory sequence motifs and the interplay between epigenetic marks and regulation. To this end, we develop methods and analyse particular data sets. The ultimate goal is to unravel biological networks and pinpoint possible transcriptional mechanisms behind the interactions.

* externally funded
Sequence-based gene regulation

[Morgane Thomas-Chollier, Matthias Heinig, Jose Muino, Meng Guofeng, Jonathan Göke, Annalisa Marsico]

For many years, our group has worked on developing methods to predict transcription factor binding sites from sequence motifs (the TRAP method, Roider at al., Bioinformatics 2007) and to find common motifs among co-expressed promoters (the PASTAA method, Roider et al., Nucleic Acids Res 2009, together with Stefan Haas). This has also led to an interest in regulatory mutations, the recognition of which is the goal of the sTRAP method (Manke et al., Hum Mutat 2010). The software that has resulted from these efforts has been summarized in a Nature Protocol paper (Thomas-Chollier et al., Nat Prot 2012) and is available via a web server. A number of collaborative projects have profited from our expertise in this area, most notably the project with Norbert Hübner, Max Delbrück Center for Molecular Medicine (MDC), Berlin, on eQTLs (Heinig et al., Nature 2010) and within the department the projects with Sebastiaan Meijsing (see his report). Jose Muino has analysed regulatory circuitry in plants (e.g., Schiessl et al., PNAS 2014).

For many years, there had been the hope that the sequence-based prediction of regulation would allow us to understand a large part of the transcription factor-target relations. However, the increasing amount of tissue- and condition-specific functional data have made it apparent that cellular processes are much more dynamic and that sequence alone may (in part) be a prerequisite for regulation, but is by no means sufficient to explain gene expression. While this trivial realization has led to an increased push to understand epigenetic regulation, we have also used gene expression data in addition to sequence patterns for prediction of transcription factor-target relationships (Meng & Vingron, Bioinformatics 2014). Likewise, in collaboration with Huck-Hui Ng, Genome Institute of Singapore (GIS), careful analysis of gene expression patterns together with transcription factor binding has led to the unravelling of a network of different ERK signalling pathways (Göke et al., Mol Cell 2013).

In the context of a DFG-funded collaboration (Transregio SFB) with Bernd Schmeck (initially at Charité and recently at Marburg University), we are studying miRNAs, their regulation, and their targets. In particular, in the case of miRNAs it is very difficult to tell where the promoter lies, because even in the sequencing data, the 5’ end of the gene is usually not visible due to the processing of the miRNA. Again, other information beyond the usual promoter elements needs to be taken into account. Annalisa Marsico developed a promoter recognition method for miRNAs that utilizes CAGE data in conjunction with machine learning algorithms to extract the promoters of the miRNAs (the PROmiRNA method, Marsico et al., Genome Biol 2013). This was within a collaboration with Ulf Ørom from the Otto
Warburg Laboratory, whose group tested predictions experimentally. Annalisa Marsico has meanwhile become an Assistant Professor at the Freie Universität (FU) Berlin. Within a cooperation between MPIMG and FU, she now leads her own group at the MPIMG.

**Combinatorial regulation and co-occupancy networks**

[Jonathan Göke, Alena van Bömmel]

Massively parallel sequencing technology (also known as Next Generation Sequencing, NGS) has revolutionized not only genomic sequencing (see report by Stefan Haas), but also functional genomics. Microarrays have largely been made superfluous and today most techniques are being reduced to sequencing. In gene regulation, the primary technique is sequencing of the DNA that was precipitated in a Chromatin Immunoprecipitation (ChIP) assay. The combination of ChIP and sequencing is called ChIP-seq. The binding sites of a transcription factor are subsequently determined by mapping of the sequence reads to the genome. Large amounts of ChIP-seq data for many transcription factors and chromatin marks in many cell lines are now publicly available, e.g., in the ENCODE project.

Figure 1: The figure shows a network of predicted interactions among transcription factors in the hematopoietic system. Red nodes indicate transcription factors that are known to be functional in this context. Green nodes indicate transcription factors, whose mRNA is found to be expressed in the respective cell line. Coloured edges refer to either direct (red) or indirect (yellow) known protein-protein interactions.
Many expression experiments yield groups of genes which are co-expressed under a set of conditions or time-points. This allows learning about regulatory mechanisms from the ChIP-seq data by analysing, which transcription factor binding sites are shared among the promoters of co-expressed genes. This is the logic of the PASTAA algorithm mentioned above. However, one can also aim at the level of the interactions among transcription factors by asking, which ones co-occur in promoters. This logic was introduced in an early paper from our group (Manke et al., J Mol Biol 2003) at a time, when the logic could only be tested on yeast data. With the availability of mammalian data, applying this logic became a challenge of its own. Jonathan Göke studied transcription factors, which in mouse embryonal stem cells co-occupy regulatory elements, and derived networks of transcription factor interactions (Göke et al., PLoS Comp Biol 2011). In a first project, Alena Mysickova (now van Bömmel) studied transcription factor binding motif occurrence and co-occurrence in promoters of genes that are expressed in specific tissues (Mysickova & Vingron, BMC Genomics 2012). In an ongoing project, she has derived interaction networks from the binding sites that can be found in regions, which are accessible in specific tissues (as measured by DNAase hypersensitivity). Such a network is shown in Figure 1. We call networks like the ones described here “co-occupancy networks” to emphasize that the information stems from proteins interacting within regions, or co-occurring genomics features in general.

Co-occupancy networks and epigenetic regulation

[Rosa Karlic, Julia Lasserre, Juliane Perner, Xinyi Yang, Jose Muino]

Eukaryotic DNA is organized into nucleosomes. Each nucleosome is made up of DNA wrapped around a histone octamer. This structure is also connected to gene regulation and transcription, as reflected in posttranslational modifications of histones, shortly called the histone modifications. Based upon antibodies against differently modified histone tails, the ChIP-seq technology is also being applied for determining the histone modifications that can be found along the genome in cells of different tissues or states. Transcription factors regulate gene expression in accordance with chromatin structure, histone modifications, DNA methylation patterns, and maybe other players like non-coding RNAs. We showed that it is possible to predict expression of a gene from only the histone modifications in its promoter region (Karlic et al., PNAS 2010). This is not to say that histone modifications directly influence the level of transcription, but the interpretation of this mathematical relationship is rather that histone modifications reflect the transcriptional state of a promoter. Lead authors of this paper were Rosa Karlic, a PhD student, and Ho-Ryun Chung, then a postdoc in the department and since 2011 an OWL research group leader.
Julia Lasserre, a postdoc in the group, in collaboration with Ho-Ryun Chung, has subsequently examined the correlation structure among histone modifications. This data is of a similar structure as the transcription factor occupancy data discussed above. For each promoter, ChIP-seq experiments yield a number of reads that map to that particular promoter, indicating the existence of a particular histone modification on the histone tails of the nucleosomes in this region. Thus, the co-occurrence of different modifications in promoter regions allows us to study correlations among histone modifications. The challenge lies in identifying direct interactions and distinguishing those from such co-occurrence that is only a consequence of a biochemical relationship. To this end we could build on classical statistical theory, which distinguishes between the well-known correlation coefficient and the partial correlation coefficient. The latter accounts for common influences on two variables (or by the two variables) in order to determine direct interactions. Interestingly, partial correlations coefficients can be computed by inverting the sample variance-covariance matrix of the data. Wherever one observes a value in the inverse of the variance-covariance matrix to be near zero, one assumes that there is no connection between the variables. In our algorithm we applied this to rank-ordered data due to the non-normality of the data, and further utilized cross-validation to determine the non-zero edges robustly. We call our algorithm SPCN for Sparse Partial Correlation Network (Lasserre et al., PLoS Comp Bio 2012). Figure 2, taken from that paper, shows the resulting network for histone modifications in CD4+ T-cells.

Based on additional publicly available ChIP-seq data, Juliane Perner, a PhD student in the group, in collaboration with Ho-Ryun Chung developed the co-occupancy network further to study signal transduction from chromatin modifying enzymes to histone modifications (Perner et al., Nucleic Acids Res 2014). She models this process in layers and uses sparse regression methods to describe the influence, which the chromatin modifiers have on the histone modifications. Figure 3 shows the chromatin modifiers in oval shapes and the histone modifications in square shapes. One can, e.g., see the activation-associated modifications H3K4me3 and H3K27ac, which are positively influenced by the modifiers CHD1 and PHF8, among others, and are repressed by the antagonistic action of SETDB1 and HDAC6.
On the top of the network one sees the components of the polycomb complex (SUZ12, EZH2, CBX2, CBX8), which are responsible for the (repressive) trimethylation of H3K27.

An overlay of this network with the SPCN results for the same variables significantly reduces the number of edges and leads to clear hypothesis about physically interacting modifiers. The group of Ho-Ryun Chung has used immunoblotting of the precipitated DNA to show that the proteins, which are thought to interact, indeed tend to co-localize on the same DNA segments. Currently, as part of the BLUEPRINT project, we are working on similar analysis, although not restricted to promoters, and for much larger data-sets. In a related effort, Xinyi Yang, postdoc in the group, is working on characterizing subgroups of promoters through not only sequence patterns, but through the promoter-resident combination of transcription factors and modifications. Jose Muino uses networks to study alterations in the epigenetic regulation in cancer cells.

**Gene network reconstruction**

[Julia Lasserre, Mahsa Ghanbari, Edgar Steiger, Navodit Misra, Ewa Szczurek]

Network reconstruction as reported above has become a central task in computational biology today. Mostly, though, it is not done for promoter-occupancy data like in our case, but for gene expression data. From the point of view of network reconstruction algorithms, the SPCN method is a variant of a so-called Gaussian Graphical Network, which is frequently used in gene network reconstruction. Constructing a gene network from expression data is considerably more challenging, because the number of genes – the nodes in the gene network – is so much higher than the number of measurements – the expression experiments on different cells or conditions. This difficult situation is reflected in the shape of the data matrix: It is a very wide matrix with many columns, on which the network is built. For the promoter occupancy networks, the nodes correspond to the smaller dimension of the matrix, and the large dimension of the matrix corresponds to the many promoters, for which we have data. So for promoter co-occupancy networks we are in a data rich situation, while for gene network reconstruction we are in a data-poor situation.
As a consequence for the construction of gene expression networks, we need to utilize additional data beyond the gene expression matrix. Therefore, PhD student Mahsa Ghanbari has developed a method that can account for the confidence we would have in particular edges of a network. The confidence values typically stem from a ChIP-seq experiment that links a transcription factor to its targets, or from protein-interaction data. Our new algorithm again builds on the inverse variance-covariance matrix, which is used in the context of the PC-algorithm for Bayes Network construction. This method tests for the presence of edges and we modify it to sort the tests from low-confidence edges to high-confidence edges. In another ongoing effort, Edgar Steiger is exploring how time-series expression data can be used to remedy the poor-data situation.

In the context of cancer genomics, Navodit Misra applies Bayesian Networks to model the co-occurrence of mutations in tumors. He combined ideas from phylogeny with heuristics for Bayesian Network construction to design an efficient method for inferring “mutation phylogenies” from large numbers of sequenced samples of one and the same tumor type (Misra et al., Bioinformatics 2014). In another cancer-related study, Ewa Szczurek identified pairs of genes, which are rarely found to mutate jointly in a tumor. Such mutually exclusive pairs might be indicators of “weak spots” in a tumor because the viability of the tumor appears to be decreased when the genes are both mutated (Szczurek et al., Int J Cancer 2013).

**Sequence algorithms**

[Szymon Kielbasa, Marcel Schulz]

Several collaborative efforts were devoted to the development of sequence analysis algorithms. Together with Martin Frith from AIST, Japan, Szymon Kielbasa developed a fast database searching program (LAST), where the speed is not hampered by the frequently occurring repeat structures in genomic sequences (Kielbasa et al., Genome Res 2011).

Marcel Schulz, who had contributed significantly to the Science paper on the eukaryotic RNA-seq technology in 2008, developed a program OASES together with Daniel Zerbino from the EBI (Schulz et al., Bioinformatics 2012). OASES can assemble transcript sequences *de novo* from RNA-seq data. By summer 2015, the OASES paper had been cited 440 times according to google scholar, which makes it the highest cited paper out of the department in the last five years.
Perspectives

Ongoing projects continue the work on networks and network reconstruction algorithms. At the same time, we are searching for predictors for various genomic elements, like CpG islands, active/inactive/bivalent promoters, active enhancers, etc. Such predictors can draw on sequence information as well as on dynamic, epigenetic or structural information. In the long run, the challenge lies in integrating all this information to form a coherent image of the transcriptional status of a dynamic genome.

Selected publications

Group members are underlined.


Sequencing & Genomics Group
Established: 01/2001

Scientists and postdocs
Mohammad Sadeh (since 10/13)
Navodit Misra (01/11-12/14)
Ruping Sun (10/09-06/13)
Yasmin Aristei (10/10-09/11)

PhD students
Peng Xiao (09/13-09/14)
Sergio Torres Sanchez (01/14-07/14)
Anne-Katrin Emde (10/08-04/13, IMPRS)
Michael Love (05/10-01/13, IMPRS)

Head
Dr. Stefan Haas
Phone: +49 (0)30 8413 1164
Fax: +49 (0)30 8413 1152
Email: haas@molgen.mpg.de

Scientific overview
The emergence of next generation sequencing (NGS) technologies opened new avenues by allowing for the first time an unbiased, sensitive and genome-wide sequencing of entire transcriptomes or genomes. This high resolution data is especially important, when studying human diseases, where the comprehensive screening for disease-causing mutations is a crucial step towards the development of diagnostic tools and, in the long term perspective, potential therapies. However, with the concomitant massive increase of sequencing data, the efficient handling and processing of NGS data becomes challenging. Due to this technological progress and the concomitant shift in research focus, we renamed the former “Gene structure & Array design” group to “Sequencing & Genomics”. The group focuses on the development of computational tools for the large-scale analysis of individual genomes/transcriptomes related to inherited diseases or cancer. The group was among the first developing efficient tools for basic NGS processing steps, but also for the comprehensive detection of various kinds of mutations in exome or genome-sequencing data. In collaboration with experimental scientists, all algorithms were applied and optimized on
large-scale projects involving data of hundreds of individuals. The tight link between computational method development and experimental validation led to the successful discovery of mutations causing X-linked intellectual disability.

Detection of disease-causing mutations

[Anne-Katrin Emde, Michael Love, Ruping Sun]

Traditionally, the screening for disease-causing mutations was time-consuming and usually limited to a small number of candidate genes or genomic target regions. With the advances in NGS technologies, the exome/genome-wide detection of genetic variations became feasible, but also required the development of adequate computational tools for this new type of data.

In collaboration with Vera Kalscheuer, formerly Department of Human Molecular Genetics (H.-H. Ropers) and now Research Group Development & Disease (Stefan Mundlos), we set up a computational pipeline to perform an efficient analysis of exome and chromosome re-sequencing data with the aim to unravel candidate mutations causing X-linked intellectual disability (XLID). Given the potential future use of such mutations in diagnostics, we put special emphasis on the comprehensive and reliable detection of sequence variations (SV). In order to address the different types of SV and sequencing technologies, we developed dedicated algorithms (SplazerS, Emde et al., Bioinformatics 2012; snpStore) for calling small SVs as well as short insertions/deletions using read alignments. Additionally, we implemented ExomeCopy (Love et al., Stat Appl Genet Mol Biol 2011), one of the first methods to recover large deletions/duplication in exome sequencing data by comparing read distributions across samples.

As an important requirement to pinpoint candidate disease-causing variations, we filter out common variants unlikely to be pathogenic and prioritize variants by their potential functional impact. Besides the basic evaluation of gene features like coding sequences or splice sites, we also use the database of known disease variations (HGMD) and the recently published CADD score to rank candidate mutations found in the XLID patients. With the increasing power of contrasting our SV data to the growing amount of other sequencing projects, we could detect on average 3-4 candidate mutations per individual. All these variants potentially causing XLID were subjected to experimental validation. Provided sufficient functional support as well as consistent co-segregation within the respective family, the results of the computational analysis are used as basis for genetic counselling of parents of the affected children.
Lung cancer genomics

[Ruping Sun, Navodit Misra, Mohammad Sadeh]

In a collaboration with Bayer HealthCare on cancer transcriptome sequencing, we were responsible for setting up an appropriate computational infrastructure for processing large datasets. This infrastructure was further optimized and expanded in a recent collaborative project with the group of Roman Thomas, Cologne University, entitled “Comprehensive genomic analysis of small cell lung cancer” allowing us to apply and test our computational tools on small cell lung cancer (SCLC) as well as additional lung cancer types. In the course of this project, whole-genome sequencing of about 100 SCLC samples as well as matched normal tissue was performed. In addition for most tumor samples RNAseq data was generated. As a basis for subsequent screens for potential tumor-driving mutations, the project focuses on a comprehensive detection of different kinds of somatic variations (copy number variations, SNVs, genomic rearrangements). Since artificial gene fusions are known cancer driving events, we put special emphasis on the development of the software (TRUP; Fernandez-Cuesta et al., Genome Biol 2015) for sensitive detection of expressed fusion transcripts from RNAseq data. TRUP predicts candidate breakpoints based on inconsistencies in the mapping of paired-end reads, and on the partial mappings of single reads crossing a breakpoint. To maximize sensitivity, TRUP uses these reads together with unmapped reads to assemble a putative breakpoint region. After an iterative process of computational prediction and experimental validation, we could successfully apply TRUP on a variety of lung tumor samples, where we detected novel, potentially tumor-driving gene fusions in lung adenocarcinomas. Interestingly, CD74-NRG1 fusions were detected exclusively in the invasive mucinous subtype, thus potentially offering a therapeutic opportunity for this so far untreatable tumor type.

The evolution of cancer is to a major extent driven by mutations in genes involved in key pathways, whose dysregulation is eventually reflected in increased cell division, reduced apoptosis, or defects in DNA repair and signaling. However, during tumor progression tumor cells usually acquire large numbers of mutations, including few driver mutations fixated during consecutive rounds of selection. The observed set of somatic mutations is thus the result of a mutational path through a complex fitness landscape. While for different tumor samples the mutation patterns are likely to be different, mutational paths including driver genes might be shared. In our Bayesian Mutation Landscape method (Misra et al., Bioinformatics 2014) we use the mutational recurrence across samples to delineate the potential evolutionary history of mutation events. This tool is able to take unobserved mutation states into account and allows the efficient analysis of large gene and sample numbers. When applied to lung adenocarcinoma
samples (TCGA), we could successfully recover two different mutational paths known from literature involving the genes EGFR and KRAS, respectively. Since basic functional properties of genes often play an important role in tumor progression, we categorized genes by protein domain. For lung carcinoids this strategy allowed us to detect potential mutational paths involving a variety of histone modifying genes, where otherwise the recurrence of mutated genes is low.

Surprisingly, in case of small cell lung cancer we discovered recurrent mutations in the two tumor suppressor genes TP53 and RB1 in nearly all samples analyzed with exception of two samples. This is in contrast to other tumor types, in which usually only one of the genes may be mutated in a subset of samples. Further analyses performed by our collaboration partners pointed towards the existence of additional driver genes in subgroups of SCLC samples. Inspired by recurrent copy number variations observed in the genomic locus of tumor suppressor TP73, we performed an in-depth analysis of CNVs on splicing. Intriguingly, a subgroup of 13% of the SCLC samples showed aberrant splicing caused by genomic rearrangements. All of these splicing events caused in-frame exon-skipping affecting different exons, most of them leading to an N-terminal truncation of TP73.

Closer inspection of the two SCLC samples, for which neither a TP53 nor RB1 mutations were detected, revealed a striking difference in the number of putative genomic breakpoints. While most SCLC samples showed a small number (<40) of breakpoints, we predicted a few thousands for these two samples. Surprisingly, in both samples only chromosomes 3 and 11 were heavily subjected to genomic rearrangements, thus reflecting cases of chromothripsis that lead to the same phenotype as samples with the combined TP53/RB1 mutations. RNAseq analysis showed a substantial overexpression of the cell-cycle regulator CCND1 in the chromothripsis cases thus pointing towards a specific regulatory mechanism driving tumor progression. In contrast, the non-chromothripsis cases showed differential expression of a large number of genes involved in cell-cycle regulation and DNA repair.

Given the comprehensive genomic and transcriptomic data on lung cancers we will in future especially focus on a detailed analysis of tumor-specific splice variants that may not only change gene function by affecting protein domains, but also by modifying specific protein-protein interactions. We hope that further in-depth investigation of the genetically diverse set of SCLC samples will allow us to shed light on basic processes leading to the formation of cancer.
Selected publications

Group members are underlined.


Bioinformatics Group
Established: 02/2001 (Dept. of Vertebrate Genomics), since 12/2014 Dept. of Computational Molecular Biology

Head
Ralf Herwig, PhD
Phone: +49 (0)30 8413-1126
Fax: +49 (0)30 8413-1152
Email: herwig@molgen.mpg.de

Scientists and postdocs
Moritz Beber* (since 02/14)
Axel Rasche* (since 09/09)
Christopher Hardt* (since 11/08)
Atanas Kamburov* (02/12 - 01/13)
Felix Dreher* (08/06 - 12/12)

PhD student
Matthias Lienhard*
(IMPRS, since 07/12)

Reha Yildirimman* (07/06 - 12/12)
Kerstin Neubert* (07/11 - 06/12)
Marcus Albrecht* (09/07 - 01/12)
Lukas Chavez* (01/11 - 12/11)
Anja Thormann* (06/10 - 12/11)

* externally funded
Scientific overview

Research of the group covers (i) the development of computational methods for the analysis of molecular data, in particular high-throughput sequencing (HTS) data, derived from complex phenotypes and (ii) the integration and interpretation of these data in the context of biological networks. To achieve this, the bioinformatics group develops methods for genome analysis, based on multivariate statistics and information theoretic concepts, and for network analyses of molecular data, which appeared a helpful, integrative concept for data interpretability and elucidation of genotype-phenotype relationships.

Ralf Herwig is member of several international consortia, for example the 1000 Genomes Project (1000GP Consortium 2010, 2012), the SEQC consortium and the COST action SeqAhead funded by EC’s Framework 7. The group has developed computational methods for HTS applications, in particular for exome sequencing, RNA-seq and MeDIP-seq and works on the integrative analysis of these data in order to elucidate the interplay of methylation, gene expression and genome structure that are operative in human (disease) processes, for example related to cancer progression, drug toxicity, renal dysfunction and stem cell development. We have published 69 scientific publications in the reporting period.

Exome sequencing of lung cancer patients

Figure 4: (A) Results from NSCLC genome analysis. Xenograft tumor (outer ring) and corresponding primary lung tumor (inner ring) correspond highly in copy number and mutational landscape (with M. Schweiger) Red: copy number increase; green: copy number decrease in tumor vs. normal tissue. (B) Combinatorial treatment of H1299 xenografts with the AMPK activating compound A-769662 and erlotinib decreased tumor growth. Upper panel: H1299 xenograft mice were treated for 5 days (day 7–11). Subcutaneous tumor growth (each treatment group with five replicates) was measured in vivo during 34 days, and finally surgically removed. Student’s t-test showed significant reduction of tumor growth in combinatorial treated mice (p = 0.03) versus single drug and control group. Lower panel: Resected tumors (three replicates closest to median size) from different treatment regimens were illustrated for growth response.
Ralf Herwig coordinated the PREDICT project funded by the BMBF “Medical Systems Biology” program (2009-2012; five partners), in which we analysed the mutational landscape of non-small cell lung cancers (NSCLC as opposed to SCLC, which is studied by S. Haas) in order to improve therapy regimens for the patients using computational methods. The tumors were removed from the patients in the clinic and transplanted onto immune-deficient nude mice (xenograft mice). Molecular characterization included targeted exome sequencing as well as gene and protein expression. Mutational landscapes of xenografted tumors showed high similarity with those of the corresponding primary tumors (Figure 4A); however, similarity depended largely on tumor purity. Xenografted tumors enabled us to test targeted drugs (and drug combinations) for the particular tumors in vivo and thus to analyse resistance mechanisms. Using a combined analysis of mutations, gene and protein expression data identified, for example, a loss of AMP-activated protein kinase (AMPK) expression in non-responders of erlotinib / cetuximab treatment. In order to compensate this effect, we tested reactivation of AMP kinases with a combinatorial treatment consisting of erlotinib and an AMPK activating compound, which resulted in a significant improvement of therapy (Figure 4B; Hülsmann et al., Lung Cancer 2014).

Methylation and gene expression in colon cancer patients and mouse models

In cooperation with the Department of Developmental Genetics (B. Herrmann), we investigated molecular data of a mouse model for colon cancer (APCMin mouse) and human tissues using our MEDIPS software for genome-wide enrichment-based methylation analysis (Chavez et al.,...
Genome Res 2010). Methylation and gene expression patterns allowed identifying candidate markers that were conserved between mice and human representing earlier and later tumor stages respectively (Grimm et al., PLoS Genet 2013). In particular, genome-wide methylation turned out as a very stable molecular signature for cancer progression (Figure 5).

**Genome analysis tools**

We developed a new method ARH - alternative splicing robust prediction by entropy - for the prediction of alternative splicing events from microarrays (Rasche & Herwig, Bioinformatics 2010) based on the information theoretic concept of entropy. An extension of the method, ARH-seq, has been adapted to high-throughput sequencing data recently. Essentially, ARH-seq is based on the evaluation of exon and exon-junction count differences between two experimental conditions, which are translated into a probability distribution and subsequently evaluated with statistical entropy. The method was intensively tested and evaluated with benchmark data (Figure 6A, 6B; Rasche et al., Nucleic Acids Res 2014) and we were able to show that the ARH-seq method is highly competitive compared to existing approaches.

Furthermore, we developed the MEDIPS software for analysing enrichment-based genome-wide methylation sequencing data, as for example generated by the MeDIP-seq approach, and first introduced the method by identifying differentially methylated regions between human embryonic stem cells and differentiated cells (Chavez et al., Genome Res 2010).
MEDIPS is a full pipeline consisting of QC features and methods for data pre-processing and statistical analysis. MEDIPS has been made available for the community with an R/Bioconductor package (Figure 6C; Lienhard et al., Bioinformatics 2014).

**Analysis of interaction networks and pathway signatures**

We developed the ConsensusPathDB – a resource for human, mouse and yeast molecular interactions (Kamburov et al., Nucleic Acids Res 2009, 2011, and 2013). ConsensusPathDB integrates diverse heterogeneous interaction types such as protein-protein, signaling, metabolic, drug-target and gene regulatory interactions and builds, with 155,930 unique physical entities and 435,360 interactions, the largest collection of human interactions worldwide. Furthermore, we developed the web servers IMPaLA for the joint network analysis of metabolites and genes (Kamburov et al., Bioinformatics 2011), and IntScore for the confidence assessment of protein-protein interaction data (Kamburov et al., Nucleic Acids Res 2012). We use these resources to guide genome-wide analysis and to predict functional consequences from high-throughput data. A specific focus of the work is in toxicogenomics, where we investigate the toxicity of drugs and chemical compounds. We have developed a method for computing pathway responses from omics data and showed that such signatures derived from pathways are far more stable than those derived from gene expression patterns (Yildirimman et al., Toxicol Sci 2011). This approach has been applied to predict the carcinogenic hazard of chemicals in human ES cell-derived hepatocyte cells and was awarded with the 31st Animal Protection Research Prize of the German Federal Ministry of Food and Agriculture (R. Herwig, December 2012).

**Planned developments**

We plan to further improve our genome analysis methods for predicting therapy responses based on the molecular characterization of tumors. R. Herwig coordinates the EPITREAT project funded by the e:BIO program of the BMBF (since 2013; six partners) that investigates the epigenetic impact on therapy response mechanisms in NSCLCs. In particular, the interplay between DNA stability, mutations, methylation differences and gene expression regulation will be investigated.

Furthermore, the group continues its efforts in drug toxicity prediction and participates in a large EC-funded project that aims to elucidate the molecular mechanisms underlying liver and heart toxicity induced by drugs (HeCaToS; since 2013). We will further develop our pathway- and network-based analysis approaches in order to identify predictive networks for drug induced liver injury as well as mechanistic models for risk assessment.
Selected publications

Group members are underlined.


Diploid Genomics Group
Established: 02/2002 (Dept. of Vertebrate Genomics), since 12/2014 Dept. of Computational Molecular Biology

Head
Dr. Margret R. Hoehe
Phone: +49 (0)30-8413 1614
Fax: +49 (0)30-8413 1152
Email: hoehe@molgen.mpg.de

Scientists and postdocs
Thomas Hübsch (since 08/08)
Gayle McEwen* (02/10 – 02/12)
Eun-Kyung Suk* (07/05 – 08/11)
Sven Stengel* (04/10 – 04/11)
Roger Horton* (09/08 – 03/10)
Dorothea Kirchner* (09/09 – 12/09)

Guest scientists
Eun-Kyung Suk (09/11-12/13)
Gayle McEwen (03/12-12/12)
Thomas Sander (07/09-06/10)

Technicians/Engineers
Stefanie Palczewski* (06/08 – 12/11)
Sabrina Schulz* (07/08 – 05/11)

Administration
Barbara Gibas* (09/10 – 03/11)
Britta Horstmann* (07/09 – 07/10)

Students
Matthias Linser* (07/09 – 09/11)
Anne-Susann Musahl* (07/09 – 11/10)
Jorge Duitama * (01/10 – 07/10)

Scientific overview
Human genomes are diploid by nature. Beyond identifying and cataloguing genetic variants, it is therefore essential to determine their distribution between the two homologous chromosomes. The phase of variants is key to fully understand human biology and link genotype to phenotype. Thus, the group focuses on the analysis of haplotypes and diploidy as a funda-

* externally funded
mental feature of the human genome. Work includes (i) the development of novel molecular genetics and bioinformatics approaches and methods to haplotype-resolve whole genomes, and their application to data production; (ii) the analysis and annotation of haplotype-resolved genomes at the individual and population level; (iii) the establishment of public resources to advance integration of phase information at the gene and genome level and prepare the ground for ‘phase-sensitive’ personal genomics and individualized medicine.

In summary, we have established an independent haplotype sequence data production and analysis pipeline, generated the most comprehensively haplotype-resolved genome to date, produced about half of the worldwide available body of molecularly haplotype-resolved genomes, performed the first population level analysis including 386 phased genomes to identify key features characterizing the diploid nature of human genomes, and made ~2.5 Terabytes of sequence data and haplotype resources publicly available. Moreover, key results unveiling global patterns of phase and diplotype have been corroborated and expanded in 1,092 genomes from the 1000 Genomes database. The group has shifted focus fully to data analysis to mine the huge body of available and prospective molecular data plus 1000 Genomes haplotype data and, complementary, functionally informative data bases. The aim is to address key questions concerning the diploid nature of human genomes and its functional implications computationally on a larger scale. The group has been part of the Lehrach Department until its closing down 11/30/14, and then joined Martin Vingron’s Department of Computational Biology.

A fosmid pool-based next generation sequencing (NGS) approach to haplotype-resolve whole genomes and its application to data production

As the basis, we have established a world-wide unique ‘Haploid Reference Resource’ of 100 human fosmid libraries. Principle and molecular genetics and bioinformatics components of fosmid pool-based NGS have been outlined in the MPIMG Research Report 2012. The specific fosmid-based wet lab protocols and/or bioinformatics algorithms, which we developed de novo to establish an independent, integrated NGS haplotype data production and analysis pipeline and assemble haplotype-resolved genomes, have been described in detail (Suk et al., Genome Res 2011, highlighted in the Nature Methods Special ‘Method of the Year 2011’; Duitama et al., Proc 1st ACM Int Conf Bioinf Comp Biol 2010; Nucleic Acids Res 2012). We have applied this method to haplotype-resolve genome “Max Planck One” (MP1) (Suk et al., Genome Res 2011), HapMap Trio child NA12878 (Duitama et al., Nucleic Acids Res 2012), and 15 more genomes (Hoehe et al., Nat Commun 2014). These genomes represent about half of
the production worldwide (~30 total haplotype-resolved by clone-based approaches). Also, we have developed several fosmid-based protocols to capture, sequence, and assemble the MHC region. This line of work was funded with ~2.8 Mio Euro to Margret Hoehe as the principal investigator by the BMBF in the NGFN-Plus Program.

Completeness and accuracy of phasing, molecular versus statistical data

MPI1 represents still the most completely haplotype-resolved genome to date, with >99% of all heterozygous SNPs phased and over 3.3 Mio SNPs assigned to their homologous chromosomes, >50% of the genome in contigs >1 Mb and a maximum contig length of 6.3 Mb (see also MPIMG Research Report 2012). The power and accuracy of fosmid-based phasing were demonstrated by resolving HapMap trio child NA12878, for which trio sequencing data were released from 1000 Genomes (1000G). Phasing data were identical, where available from both approaches. Fosmid-based phasing, however, resolved ~20% more of the heterozygous SNPs (>98% total) (Duitama et al., Nucleic Acids Res 2012). Comparing a set of 12 molecularly haplotype-resolved genomes to statistical haplotype data from 1000G showed a phase discordance of 3.6% for exome data, 5.3% for transcript data, and 5.9% genome-wide (Hoehe et al., Nat Commun 2014). These results motivated the complementary use of 1000G statistical haplotype data to up-scale analyses and test hypotheses on a larger scale.

Diploype architecture of individual genomes

We have systematically dissected an individual’s ‘diplotype’ on the example of MPI1 (Suk et al., Genome Res 2011), and did so analogously in the other genomes (Hoehe et al., Nat Commun 2014). We found that the extent and nature of nucleotide differences between the two homologous chromosomes was very similar in each individual. Thus, over 77% of all autosomal genes (primary transcripts) had two different molecular haplotypes; so did ~90% of the genes, when upstream regions were included. Consistently ~20% of the genes were found to encode two different proteins, and ~3% had two or more potentially perturbing mutations, which existed in cis in ~60%, in trans in ~40% of the observations. The importance of phase for gene function, disease risk and treatment response was comprehensively evaluated (see also Research Report 2012).
First population level analysis of haplotype-resolved genomes

We have performed the first systematic analysis of diplotype architecture at the population level (Hoehe et al., Nat Commun 2014). We used a set of 14 molecularly haplotype-resolved genomes, complemented and expanded by up to 372 statistically resolved genomes from 1000G. The analysis of multiple haplotype-resolved genomes allowed addressing the following key questions: (i) What is the entirety/diversity of haploid and diploid gene forms that constitute the ‘true’ molecular toolbox underlying cellular and organismal processes and their variation in population samples of defined size? (ii) Do certain classes of genes preferentially encode two different forms of the protein? This will provide insight into the potential functional importance of diploidy. (iii) What is the distribution of cis versus trans configurations at the gene and whole genome level? Can we distinguish common patterns of phase?

Figure 7: Diversity of unique gene and protein haplotypes and diplotypes

Overall scheme: The numbers of unique, different haplotypes (red colors) and pairs thereof, diplotypes (blue colors) presented relative to increasing numbers of haplotype-resolved genomes, drawn from the 14 molecularly resolved (14G) and statistically resolved genomes of European ancestry from the 1000G database; (a,b) gene haplotypes and diplotypes; (c,d) protein haplotypes and diplotypes; (a,c) haplotypes and diplotypes expressed as whole genome counts, and (b,d) as global averages ‘per gene’. a, Decreasing curves present fractions of unique gene haplotypes, and diplotypes, relative to all measured haplotypes, and diplotypes; increasing curves their absolute numbers. Data points correspond to 5, 10 and 14 genomes from 14G and 5, 10, 14, 57, and 372 genomes of European ancestry derived from 1000G database. An additional data point, 200 genomes (1000G) was integrated to anchor graphs. b, Unique gene haplotypes and diplotypes presented as global averages ‘per gene’ for given sample sizes; data shown for 14G and subsets thereof, and 1000G-derived sets of genomes. c, Unique protein haplotypes and diplotypes analogous to a. d, Average numbers of unique protein haplotypes and diplotypes ‘per gene’ analogous to b. For detailed information see Hoehe et al., Nat Commun 2014.
In summary, we found immense diversity of both haploid and diploid gene forms, up to 4.1 and 3.9 million corresponding to 249 and 235 per gene on average (Figure 7). These numbers were still far from approaching a plateau. Evaluating the haplotypes and diplotypes by frequency of occurrence (FoO) showed that genes, that had one major haplotype with a FoO of at least 50% (category 1), represented by far the smallest fraction of all autosomal genes, 13-15% (Figure 8), challenging the concept of a ‘wild-type’ form of the gene. The rule rather than the exception is that each gene represents the equivalent of multiple different forms, which accounts for only limited fractions of all haplotypes observed. Thus, roughly one third of all genes had at least one common haplotype with a FoO > 20% (category 2), and over half of all genes had uncommon forms only (category 3). Classifying genes by their diplotype spectra unveiled an even higher complexity constituting diploid gene function (Figure 8).

Moreover, we identified a ‘common diplotypic proteome’, a distinctive subset of over 4,000 genes preferentially encoding two different proteins, allowing gene functions to be differentially exerted and/or diversified. Enrichment and pathway results suggested an important role of diploidy for preserving flexibility of receptor-mediated cell-cell communications, immune-related processes and transcriptional regulation. We showed moreover a significant abundance of cis configurations of mutations in each of the 386 genomes, with an average cis/trans ratio of ~60:40 (Figure 9). Cis configurations of mutations, leaving one form of the gene unperturbed, would be expected to occur more frequently in an individual genome to

Figure 8: Categorization of autosomal genes
Pie charts show classification of autosomal genes into three categories based on frequency of occurrence (FoO) of their unique haplotypes (red colors) and diplotypes (blue colors). For instance, Category 1 includes all genes that have one predominant haplotype, or diplotype, accounting for ≤ 50% of all measured haplotypes, or diplotypes; the definitions for Category 2 and 3 genes are analogous. Fractions of these categories (%) relative to the total of autosomal RefSeq Hg18 genes assigned. Data shown for the sets of 14 molecularly haplotype-resolved genomes (14G), and 57CEU and 372EUR statistically resolved genomes derived from 1000G database. From Hoehe et al., Nat Commun 2014.
preserve organismal function. Furthermore, distinguishable classes of cis-versus trans-abundant genes were observed. With this work, we have identified key features characterizing the diploptypic nature of human genomes and provided a conceptual and analytical framework, rich resources and novel hypotheses on the functional importance of diploidy. And we provided some original answers to important questions about the true nature of genetic variation in genome sequences, as was also recognized by the reviewers of the manuscript.

**Cis versus trans configurations of mutations in 1,092 genomes**

We have now confirmed significant abundance of cis configurations of protein-altering mutations with an average cis/trans ratio of ~60:40 (Figure 9) in 1,092 genomes from 1000G, including four major populations and a total of 14 subpopulations. With the exception of three, each one of the 1,092 genomes had more cis than trans configurations (99.7%). Almost identical results were obtained when analyzing the entirety of nsSNPs. A subset of ~2,400 genes encoding either cis or trans configurations of perturbing mutations was found shared by all populations, quasi the global phase-sensitive part of the genome. These results suggest that mutations are not randomly distributed between the homologous chromosomes, but do exist in global patterns, which preferably affect certain functional classes of the genes.

Figure 9: Cis abundance of protein-altering mutations in human genomes

Two or more mutations in a gene can reside on the same chromosome, that is, in a cis configuration, leaving the second protein intact (upper half), or they can reside on opposite chromosomes, in a trans configuration, affecting structure and function of both proteins (lower half). Cis configurations exist significantly more frequently in human genomes with an average cis/trans ratio of ~60:40. Modified from Max Planck Research 4.2011, © Art 4 Science.
The ‘Max Planck Haplome Resource’

Rich resources including NGS data and variant files, chromosomal haplotypes, multiple diplotypic gene sets and phase-alternate mutations, haploid landscapes and algorithms have been made available to the Scientific Community via public data bases and our homepage. This resource will be significantly expanded in collaboration with George Church, Harvard Medical School & MIT, and Radoje Drmanac, Complete Genomics/BGI, through access to another ~200 molecularly haplotype-resolved genomes. Additional modules will be integrated such as to provide the molecular basis for the analysis of transcriptional regulation, which is inherently phase-sensitive, or for instance a ‘Pharmacogenomics HaploBase’ and molecular haplotypes of disease genes and their regions. This resource will provide useful haplotype information for all aspects of genome biology and functional genomics, disease gene discovery, individualized medicine and pharmacogenomics.

Planned developments

We will continue to follow up key results obtained from our population level study, by use of computational tools and appropriate data bases, in collaboration with Ralf Herwig and other members of the Vingron Department, and international partners. We will assess (i) the generality of cis abundance in 2,500 individuals (1000G) and other diploid species; (ii) the translation of the phase-sensitive part of the genome into function, through analysis of corresponding RNASEq/transcriptome data, with particular focus on allele-specific and mono-allelic expression, in conjunction with transcription regulatory motifs; (iii) the impact of alternative splicing on phase configurations. We will continue to perform enrichment and pathway analyses to specify the gene classes and functions that are mediated by different phase configurations. We will integrate data sets from disease tissues. The global importance of the ‘common diplotypic proteome’ will be analyzed analogously.
Selected publications

Group members are underlined.


Mechanisms of Transcriptional Regulation
Group
Established: 09/2009

Scientific overview
Normal development and homeostasis require the correct temporal and spatial expression of genes. Furthermore, genes are not simply turned on or off but rather their expression can be fine-tuned to meet the individual demands of a cell. This is, at least in part, warranted by transcription factors (TFs) that orchestrate gene expression by binding to regulatory DNA sequences associated with their target genes. We study transcriptional regulation using the glucocorticoid receptor (GR). GR is a member of the steroid hormone receptor family, whose activity is strictly hormone dependent, thus effectively providing the experimenter with a TF with an on-off switch. Although GR is expressed throughout the body, the genes regulated and the genomic loci bound by GR show little overlap between cell-types. My long-term goal is to understand how a single transcription factor can regulate vastly different sets of genes depending on the cell type and to elucidate processes that influence the expression level of individual target genes.
Towards this objective, we use integrative experimental, structural and computational modeling approaches, with the ultimate goal to qualitatively and quantitatively explain the genome-wide transcriptional consequences of GR signaling.

**Role of genetic and epigenetic landscape in guiding GR to specific genomic loci**

Binding-site motifs of eukaryotic transcription factors typically only have a handful of constrained nucleotide positions. Hence, these sequences are ubiquitously found in the genome, whereas binding only occurs at a small subset of putative binding sites. Consequently, the binding site motif provides insufficient information and additional inputs, for example the chromatin state, are needed to specify, which of the potential binding sites found in the genome are actually bound. Furthermore, many bound regions appear not to contain known GR recognition sequences indicating that GR can bind additional sequences.

To identify sequences responsible for recruitment of GR to individual loci, we turned to a modified chromatin immunoprecipitation assay, which combines ChIP with an exonuclease (ChIP-exo) digestion step. This approach gives “footprints” of binding rather than bound regions, and hence bound sequences can be identified at individual loci rather than by statistical overrepresentation, as is the case with conventional genome-wide profiling approaches. We found that ChIP-exo footprints are protein- and
recognition sequence-specific signatures of genomic GR association that can be used to discriminate between direct and indirect (tethering to other DNA-bound proteins) DNA association of GR (Figure 10). Furthermore, these studies show that the absence of classical recognition sequences at GR-bound regions can be explained by direct GR binding to newly identified GR recognition sequences. In addition, our footprint-based approach shows that GR binds to highly degenerate sequences that would not be uncovered using conventional methods based on statistical overrepresentation of sequence motifs at bound regions.

Most of the scientific focus has been on the identification of positive sequence signals that recruit TFs to specific genomic loci to regulate transcription. We reasoned that another way to specify, where TFs bind, might be conferred by sequence signals that prevent TF recruitment. If such signals exist, they would be depleted at genomic regions, where GR binds when compared to unbound regions. Accordingly, bioinformatical analysis of genome-wide GR binding identified several candidate sequences. Subsequent experiments showed that these candidates indeed interfere with GR binding and studies in zebrafish demonstrated that the mechanisms mediating their effect are conserved across species and active in most tissues. Contrary to our expectation, the candidates exert their effect by mechanisms other than chromatin accessibility, possibly involving anchoring to sub-nuclear regions that may be less permissive to TF binding. Together, our studies highlight that the joint influence of positive and negative sequence signals partitions the genome into regions where GR can bind, and those where it cannot.

In addition to sequence, chromatin features play a key role in specifying the genomic binding pattern of TFs. Therefore, we mapped genome-wide GR binding in several cell types with a well-characterized “epigenome” (e.g. DNA methylation, DNase-I sensitivity, >20 histone modifications; data from ENCODE and epigenome roadmap). Subsequent hierarchical modeling, which allowed comparisons to be made between experiments (individual histone modifications e.g.) and between cell lines, resulted in the identification of both shared and cell-type specific chromatin features that correlate with GR binding. One interesting and surprising finding we made was that various histone modifications associated with promoter regions negatively correlate with GR binding. Ongoing experiments, for example using cell lines lacking the enzymes that deposit these promoter-associated histone modifications, are aimed at understanding the mechanisms underlying the observed correlations and a possible role of depleted GR binding in promoter regions in facilitating cell-type specific regulation of GR target genes.
DNA as an allosteric modulator of GR structure & activity

GR’s activity is modulated by several signaling inputs including ligand chemistry, combinatorial interactions with other transcription factors at genomic response elements, post-translational modifications and the interactions with proteins, RNA and of particular interest with DNA. These signals are not encountered in isolation, but simultaneously and combinatorically, thus allowing the same protein to produce different outputs depending on the inputs it encounters.

To reduce the complexity of the signaling cross-talk, we initially focused on a single input: the sequence of the DNA bound by GR, which can modulate the structure and function of GR (Figure 11). To study how information is transferred from the DNA:protein interface to influence the activity of other regulatory surfaces of GR, we compared two GR isoforms that differ by a single amino acid insertion in the lever arm, a domain that adopts DNA sequence-specific conformations. We reasoned that the structural changes induced by the arginine insertion in the lever arm would allow us to study its role in transmitting signals from the DNA:protein interface to the rest of the protein. A comparison of the genome-wide binding, transcriptional regulation and structural studies (NMR, in collaboration with Lisa Watson, UCSF) showed that these isoforms differentially regulate gene expression levels through two mechanisms: differential DNA binding and altered communication between GR domains. This suggests that the lever arm relays DNA-induced structural changes to remote functional domains to regulate GR’s activity. These domains in turn interact with co-regulators that either directly or indirectly influence the recruitment of RNA polymerase II. In collaboration with Ulrich Stelzl (MPIMG, Otto Warburg Laboratory), we have identified candidate DNA sequence-specific co-regulators of GR that ultimately might yield clues about how GR binding site (GBS) variants nucleate the assembly of different regulatory complexes at individual binding sites.
To determine if GBS-variants influence GR activity in a chromosomal context, we adopted a system we developed to study regulatory DNA in an isogenic chromosomal setting using zinc-finger nuclease-driven transgenesis. Subsequent testing of GBS-variants showed that sequence variants induce different levels of activation, which did not require higher GR occupancy, consistent with the idea that the increased levels of activation are a consequence of sequence-induced changes in GR conformation. Furthermore, computational analysis of genome-wide GR binding and transcriptional regulation showed that the bases flanking the consensus DNA binding site can modulate GR’s activity towards its target genes. Since these bases are not directly contacted by GR and did not result in a changed affinity of GR for the GBS, we (in collaboration with Remo Rohs, USC) assayed their effect on the DNA structure. This analysis showed that the “flanking sites” change the topology of the DNA, and structural studies (NMR) done by Marcel Jurk, a postdoc in my lab, indicate that the topological changes in the DNA are propagated to change the structure of GR. Together, these studies indicate that DNA topology plays a role in fine-tuning the expression of individual endogenous GR target genes.

**Selected publications**

*Group members are underlined.*


Research Group Evolutionary Genomics
Established: 10/2003

Head
Dr. Peter Arndt
Phone: +49 (0)30 8413 1162
Fax: +49 (0)30 8413 1152
Email: arndt@molgen.mpg.de

Scientists and postdocs
Misha Sheinman (since 09/13)
Ercan Kuruoglu
(04/13 - 10/14, 04/15 - 06/15, Alexander von Humboldt Fellow)
Irina Czogiel (10/11 - 01/14)
Brian Cusack (06/08 - 12/11)

PhD students
Katharina Imkeller* (since 11/13)
Florian Massip* (since 04/12)
Barbara Wilhelm
(09/09 - 06/14, IMPRS)
Yves Clement
(10/08 - 12/12, IMPRS)
Paz Polak (10/06 - 03/11, IMPRS)
Federico Squartini
(10/05 - 01/10, IMPRS)

Students
Judith Abecassis* (02/13 - 06/13)
Carina Mugal* (07/11 - 12/11)
Brian Lunt (08/09 - 11/09)
Thomas Engleitner (08/09 - 09/09)

* externally funded
Scientific overview
Evolutionary biology studies the processes responsible for the generation of the diversity of life on our planet. Clearly, these processes acted on the genomes of species and left distinctive marks on them. The ubiquitous availability of genomic sequence data makes it nowadays possible to extensively apply computational methods in evolutionary biology. Under these premises, the Evolutionary Genomics Group in the Department of Computational Molecular Biology at the MPIMG focuses on the utilization of mathematical concepts and methodology to understand complex phenomena in evolutionary biology.

A prevalent topic of interest within the group is the evolution of eukaryotic genomes, especially the modelling of processes that change the DNA of a species. These processes encompass single nucleotide exchanges, insertions and deletions of short segments of DNA (due to slippage), insertions of repeats, segmental duplications, and whole genome duplications. All these processes left distinct marks in present day genomes, which we set out to investigate. For such an analysis different data sources are at our disposal. In case we are interested in the action of such processes over evolutionary time scales, i.e. over millions of years, we may use available data on the divergence of genomic sequences between species, as well as data on variations between individuals within one population. However, evolution can also be studied as it happens also on much smaller time scales, for instance in the development of a tumor out of healthy tissue. Furthermore the repertoire of antibodies in an immune system is constantly changing in response to detected antigens. Due to advances in next generation sequencing technologies and the availability of vast amounts of sequence data in public databases this analysis is possible with more power and precision than before.

The impact of segmental duplications on genome evolution
[Florian Massip, Misha Sheinman]

Evolving genomes are shaped by a multitude of mutational processes, including point mutations and segmental duplications. These mutational processes can leave distinctive qualitative marks in the statistical features of genomic DNA sequences. One such feature is the match length distribution (MLD) of exactly matching sequence segments within an individual genome or between the genomes of related species. These distributions have been observed to exhibit characteristic power law decays in many species. We showed that a simple dynamical models consisting solely of duplication and mutation processes can already explain the characteristic features of MLDs observed in genomic sequences. We found that these features are largely insensitive to details of the underlying mutational pro-
cesses and do not necessarily rely on the action of natural selection. Our results demonstrate how analyzing statistical features of DNA sequences can help us reveal and quantify the different mutational processes that underlie genome evolution.

**Statistical properties of Yule trees**

[Misha Sheinman, Florian Massip]

A Yule tree is the result of a branching process with constant birth and death rates. Such a process serves as an instructive null model of many empirical systems, for instance, the evolution of species leading to a phylogenetic tree. However, often in phylogeny the only available information are the pairwise distances between a small fraction of extant species representing the leaves of the tree. We therefore studied the statistical properties of the pairwise distances in a Yule tree. Using a method based on a recursion, we derive an exact, analytic, and compact formula for the expected number of pairs separated by a certain time distance. This number turns out to follow an increasing exponential function. This property of a Yule tree can serve...
as a simple test for empirical data to be well described by a Yule process. To make our results useful for realistic scenarios, we explicitly take into account that the leaves of a tree may be incompletely sampled and derive a criterion for poorly sampled phylogenies. We were able to show that our result can account for empirical data, using families of bird species.

Cancer type-specific distributions for consecutive somatic mutation distances

[Erkan Kuruoglu, Jose Muino]

Specific molecular mechanisms may affect the pattern of mutation in particular regions, and therefore leave a footprint or signature in the DNA of their activity. The common approach to identify these signatures is studying the frequency of substitutions. However, such an analysis ignores the important spatial information, which is important with regards to the mutation occurrence statistics. We propose that the study of the distribution of distances between consecutive mutations along the DNA molecule can provide information about the types of somatic mutational processes. In particular, we have found that specific cancer types show a power-law in inter-occurrence distances, instead of the expected exponential distribution dictated with the Poisson assumption commonly made in the literature. Cancer genomes exhibiting power-law inter-occurrence distances were enriched in cancer types, where the main mutational process is described to be the activity of the APOBEC protein family, which produces a particular pattern of mutations called Kataegis.
Spatial and temporal dynamics of the immune repertoire in mice

[Katharina Imkeller, Irina Czogiel]

In a joint project with immunologists from the German Cancer Research Center (DKFZ) in Heidelberg (Hedda Wardemann, Christian Busse) we try to accurately quantify the overall size, clonality, and histoanatomical distribution of the immune globulin gene repertoire in mice. Our wet-lab partners have developed a novel experimental approach for acquiring the necessary data that moves away from sequencing bulk isolated B cells. B cells will be isolated from different histoanatomical locations (i.e. from all lymphoid tissues and several non-lymphoid tissues) so that the acquired dataset will contain information of previously inaccessible detail. For the first time, we will be able to quantify the diversity of the immune globulin gene repertoire on a monoclonal level. Moreover, we will develop a model for the underlying evolutionary phylodynamics of the B cell populations that will increase our understanding of the selection processes that constantly shape the antibody repertoire during B cell development and differentiation.

Comparative analysis of nucleotide substitutions

[Yves Clement, Paz Polak]

Mammalian genomes show large-scale regional variations of GC-content (the isochors), but the substitution processes responsible for this structure are poorly understood. We have shown that meiotic recombination has a major impact on substitution patterns in humans and mice driving the evolution of GC-content. Furthermore also other cellular processes have an influence on nucleotide substitutions on a more local scale.

A regional analysis of nucleotide substitution rates along human genes and their flanking regions allowed us to quantify the effect of mutational mechanisms associated with transcription in germ line cells. Our results revealed three distinct patterns of substitution rates. First, a sharp decline in the deamination rate of methylated CpG dinucleotides, which is observed in the vicinity of the 5’ end of genes. Second, a strand asymmetry in complementary substitution rates, which extends from the 5’ end to 1 kbp downstream from the 3’ end, associated with transcription-coupled repair. Finally, a localized strand asymmetry, i.e. an excess of C->T over G->A substitution in the non-template strand confined to the first 1-2 kbp downstream of the 5’ end of genes.
Mathematics aspects of evolutionary models

[Barbara Wilhelm, Federico Squartini]

Markov models describing the evolution of the nucleotide substitution process are widely used in phylogeny reconstruction. They usually assume the stationarity and time reversibility of this process. Although corresponding models give meaningful results, when applied to biological data, it is not clear, if the two assumptions hold and, if not, how much sequence evolution processes deviate from them. To this end, we introduced two sets of indices to quantify violations of the above two assumptions using the Kolmogorov cycle conditions for time reversibility.

In the future, we will try to answer questions about the limitations of parameter estimations in comparative genomics. Especially we want to explore whether the addition of more species improves the estimation of nucleotide substitution rates along a given branch in a phylogeny.

Phenotypic mutations

[Brian Cusack]

Recent studies have hinted at the importance of “phenotypic mutations” (errors made in transcription and translation) in molecular evolution. These are thought to facilitate positive selection for adaptations that require multiple-substitutions, but the generality of this phenomenon has yet to be explored.

Our research in this area focuses on the importance of phenotypic mutations to negative selection and to the maintenance of genomic robustness by selective constraint. We initially approached this in the context of Nonsense Mediated Decay (NMD)-based surveillance of human gene transcription. We have discovered a pattern of codon usage in human genes that compensates for the variable NMD efficiency by minimizing nonsense errors during transcription. Our future work will focus on whether phenotypic mutations due to other types of mis-transcription constitute a similar selective force.
General information about the whole group

Complete list of publications (2009 – 2015)

Research group members are underlined.

2015
Berglund J, Quilez J, Arndt PF & Webster MT (2015). Germline methylation patterns determine the distribution of recombination events in the dog genome. Genome Biology and Evolution, 7(2), 522–530


2014


2013


2011
Clement Y & Arndt PF (2011). Substitution patterns are under different
influences in primates and rodents. *Genome Biology and Evolution*, 3, 236–245


2010


2009


### Selected invited talks (Peter Arndt)

**Neutral evolution of duplicated DNA – an evolutionary stick-breaking process.** Systems Biology Lecture, Max Delbrück Center for Molecular Medicine, Berlin, 2014

**Modelling the evolution of DNA sequences.** Multiscale Integration in Biological Systems, Institute Curie, Paris, France, 2014

**Natural evolution of duplicated DNA – An evolutionary stick-breaking process.** Santa Fe Institute, Santa Fe, USA, 2014

**Neutral evolution of duplicated DNA – An evolutionary stick-breaking process.** Simons Institute, Berkeley, USA, 2014

**Neutral evolution of duplicated DNA – An evolutionary stick-breaking process causes scale-invariant behaviour.** Mathematical and Computational Evolutionary Biology (MCEB2013), Hameau de l’Etoile, France, 2013

**The evolutionary fate of duplicated neutral DNA - breaking sticks on evolutionary time scales.** Memorial Sloan-Kettering Cancer Center, New York, USA, 2013

**The evolutionary fate of duplicated neutral DNA - breaking sticks on evolutionary time scales.** INRA, Jouy-en-Josas, France, 2013

**The complexity of neutrally evolving genomes.** European Conference on Complex systems (ECCS2012), Brussels, Belgium, 2012

**Breaking sticks on evolutionary time scales.** 3rd International Conference
on the Genomic Impact of Eukaryotic Transposable Elements, Asilomar, USA, 2012


Mutagenic processes and their association with transcription. Colloquium at the Dahlem Centre of Plant Sciences, Berlin, 2011


Nucleotide substitution models - mathematical definitions and genomic applications. Molecular Evolution Meeting, Orange County, Coorg, India, 2009

Evolutionary signatures of mutagenic processes associated with transcription. Expert conference on “Future of Computational Biology” Berlin/Potsdam, 2009

PhD theses


Federico Squartini: Stationarity and reversibility in the nucleotide evolutionary process. Freie Universität Berlin, 05/2010

Student thesis


Teaching Activities

Peter Arndt & Ho-Ryun Chung: Bayesian Analysis, Freie Universität Berlin, Dept. of Mathematics and Computer Science, winter term 2013/14

Peter Arndt: Population Genetics, Universite Pierre et Marie Curie, Paris, France, October/November 2010

Peter Arndt: Dynamical Models Describing Genomic Nucleotide Substitutions, OIST Summer School on Quantitative Evolutionary and Comparative Genomics, Okinawa, Japan, May/June 2010

Peter Arndt: Population Genetics and Evolutionary Game Theory, Freie Universität Berlin, Dept. of Mathematics and Computer Science, winter term 2008/09

Organization of scientific events

Peter Arndt is one of the organizers of the Otto Warburg International Summer School series:

Otto Warburg International Summer School and Workshop on Comparative and Evolutionary Genomics, PICB Shanghai, September 14-21, 2014

Otto Warburg International Summer School and Workshop on Next Generation Sequencing and its Impact on Genetics, Zuse Institute Berlin, August 19-26, 2013


Otto Warburg International Summer School and Workshop on Regulatory (Epi-)Genomics, Harnack House Berlin, August 29 - September 6, 2009
General information about the whole department

For Ralf Herwig and Margret Hoehe, only the publications and other general information since 2015 are listed here. The publications and information about former years are listed in the report of the Dept. of Vertebrate Genomics (H. Lehrach).

Complete list of publications (2009-2015)

Department members are underlined.

2015
Berglund J, Quilez J, Arndt PF & Webster MT (2015). Germline methylation patterns determine the distribution of recombination events in the dog genome. *Genome Biology and Evolution*, 7(2), 522–530


2014


Misra N, Szczurek E & Vingron M (2014). Inferring the paths of somatic
evolution in cancer. *Bioinformatics*, 30(17), 2456-2463


Schiessel K, Muino JM & Sablowski R. Arabidopsis JAGGED links floral organ patterning to tissue growth by repressing Kip-related cell cycle inhibitors. *Proceedings of the National Academy of Sciences of the USA*, 111(7), 2830-2835


logy. American Journal of Human Genetics, 95(6), 729-735


2013

Bolshoy A (2013). Modeling of DNA curvature: comment on “Sequence-dependent collective properties of DNAs and their role in biological systems” by Pasquale De Santis and Anita Scipioni. Physics of life reviews, 10(1), 73-74; discussion 82-84


Goke J, Chan YS, Yan JL, Vingron M & Ng HH (2013). Genome-wide kinase-chromatin interactions reveal the regulatory network of ERK signaling in human embryonic stem cells. Molecular Cell, 50(6), 844-855


Mammana A, Vingron M & Chung HR (2013). Inferring nucleosome positions with their histone mark annotation from ChIP data. Bioinformatics, 29(20), 2547-2554


Mugal CF, Arndt PF, Ellegren H (2013). Twisted signatures of GC-biased gene conversion embedded in
an evolutionary stable karyotype. *Molecular Biology and Evolution*, 30(7), 1700–1712


2012


2011

Clement Y & Arndt PF (2011). Substitution patterns are under different influences in primates and rodents. *Genome Biology and Evolution*, 3, 236-245

Cusack BP, Arndt PF, Duret L & Roest Crollius H (2011). Preventing dangerous nonsense: selection for robustness to transcriptional error in
human genes. *PLoS Genetics*, 7(10), e1002276


Schraders M, Haas SA, Weegerink NJ, Oostrik J, Hu H, Hoefsloot LH,


Wiedenhoeft J, Krause R & Eulenstein O (2011). The plexus model for the inference of ancestral multido-


2009


Marcel Holger Schulz: Otto Hahn Medal, Max Planck Society, 06/2011


Selected invited talks
(Martin Vingron)

„Knippers“ Molekulare Genetik – 2015, Workshop, Tübingen, 04/2015

Université Pierre et Marie Curie, Paris, France, 03/2015

Bioquant-Seminar, Deutsches Krebsforschungszentrum Heidelberg, 02/2015

Human Genome Meeting 2014, Geneva, Switzerland, 04/2014

Symposium on Medical Epigenetics, Universitätsklinikum Freiburg, 04/2014

Bioinformatics, Optimization and Graphs, Workshop, Tel Aviv University, Israel, 12/2013


College of Life Sciences, University of Dundee, Scotland, 10/2013

EMS Autumn School on Computational Aspects of Regulation, Bedlewo, Poland, 10/2013

German Conference on Bioinformatics 2012, Jena, 09/2012

Computational Biology Symposium, University of Southern California, USA, 04/2012

University of New York, USA, 04/2012

JOBIM (Journées Ouvertes en Biologie, Informatique et Mathématiques) 2012. Rennes, France, 07/2012

Biozentrum, University Basel, Switzerland, 11/2011

Scientific honours

Stefanie Schöne: Fellowship of the Christiane Nüsslein-Volhard Foundation and the L’Oreal-UNESCO for Women in Science Program, 2015

Martin Vingron: named among the Thomson-Reuters Highly Cited Researchers, 2014

Martin Vingron: Elected member of the Academia Europaea, 2014

Irina Czogiel: Gustav-Adolf Lienert Prize, Biometrical Society, German Region, 03/2012

Martin Vingron: Elected Fellow of the International Society for Computational Biology (ISCB), 07/2012

Marcel Holger Schulz: Otto Hahn Medal, Max Planck Society, 06/2011

Appointments of former members of the Group

Hugues Richard: Assistant Professor at University Pierre & Marie Curie (UPMC), Paris, France, 2009

Morgane Thomas-Chollier: Associate Professor at Ecole Normale Superieur, Paris, France, 2012

Matthias Heinig: group leader “Genetic and Epigenetic Regulation”, Helmholtz Center Munich, 2014

Ho-Ryun Chung: Max Planck Research Group Leader, MPIMG, Berlin, 2011

Annalisa Marsico: Assistant Professor, Freie Universität Berlin, 2014

Marcel Schulz: 2010 Lane Postdoctoral Fellow at Carnegie Mellon University; 2013 group leader at the Max Planck Institute for Informatics, High-throughput Genomics & Systems Biology group, Saarbrücken

Ewa Szczurek: 2012 ETH Postdoctoral Fellowship; 2015 Assistant Professor University of Warsaw, Poland

Jonathan Göke: 2012 Postdoc at Genome Institute of Singapore (GIS); 2014 GIS Fellow

PhD theses


Stephan S. Starick: DNA sequence and chromatin landscape regulate genomic binding of the glucocorticoid receptor. Freie Universität Berlin, 2015

Jonas Telorac: DNA encoded signals regulate genomic binding of transcription factors. Freie Universität Berlin, 2015

Anne-Katrin Emde: Next-generation sequencing algorithms. Freie Universität Berlin, 2013


Rosa Karlic: Influence of histone modifications on mRNA abundance and structure. Freie Universität Berlin, 2011


Akdes Serin: Biclustering analysis for large scale data. Freie Universität Berlin, 2011


Holger Klein: Co-occurrence of transcription factor binding sites. Freie Universität Berlin, 2010

Marcel H. Schulz: Data structures and algorithms for analysis of alternative splicing with RNA-seq data. Freie Universität Berlin, 2010

Stefan Bentink: Transcriptional profiling of aggressive lymphoma. Freie Universität Berlin, 2009


**Student theses**


Valentin Schneider: Estimating the number of transcription factor DNA-
binding events from ChIP-seq experiments. Freie Universität Berlin, Bachelor thesis, 2013


**Teaching activities**

The Department of Computational Molecular Biology contributes to teaching in the Bioinformatics curriculum of Free University both at the Bachelors and at the Masters level. Every other year in the fall semester we teach Algorithmic Bioinformatics, which is a 3rd year senior undergraduate course. Each year we share the teaching of a Masters course, Network Biology, with Prof. Bockmayr. In addition, many department members give practical courses, teach recitals, hold seminars, and supervise Bachelor and Master theses.

Sebastiaan Meijsing participated in the teaching of Pharmacogenetics and of Anatomy and Physiology for Pharmacology students at FU Berlin. He also taught “Computational analysis of cis-regulatory elements” as part of an M2 master course organized by Morgane Thomas-Chollier & Denis Thieffry at Institut de biologie de l’école normale supérieure (ENS), Paris, France. In addition, he gave a “Genomes & Phenotypes” course for M1 master students, organized by Hugues Roest-Crollius and Marie-Anne Félix, Institut de biologie de l’école normale supérieure (ENS), Paris, France.

Peter Arndt teaches a course in population genetics at FU, and in fall 13/14 together with Ho-Ryun Chung taught a course on Bayesian data analysis. Peter Arndt has been the main organizer of the biannual Otto Warburg Summer School for the last years.
Organization of scientific events

Martin Vingron was co-chair of the ISMB/ECCB conference in Berlin in 2013. He further was Chair of the Steering Committee of the RECOMB conference series from 2009-2015.

Every other year, the department organizes the International Otto Warburg Summer School and Research Symposium, which last about one and a half week and bring together several well-known researchers and PhD students from different backgrounds to discuss recent advances varying fields of computational molecular biology. The themes of the last summer schools have been:

- Transcription factors and their role in establishing and maintaining cellular identity (2015)
- Comparative and evolutionary genomics (2014)
- Next Generation Sequencing and its impact on genetics (2013)
- Genes, metabolism and systems modelling (2012)
- Evolutionary genomics (2011)
- Regulatory (epi)-genomics (2009)

For more details, please see http://ows.molgen.mpg.de/.

Every year, Martin Vingron, together with BioTOP Berlin-Brandenburg, organizes the Treffpunkt Bioinformatik, a local, German-speaking workshop with high-ranking speakers that allows students and local actors to meet each other and get an overview about the scientific activities in the Berlin-Brandenburg area on various aspects of computational molecular biology. The last Treffpunkt Bioinformatik have been focused on systems biology (2009), structural biology (2010), evolutionary biology (2011), RNA technologies (2012), and glyco-logy (2014). In 2013, the series celebrated its 10th anniversary with an extended meeting about development and trends in bioinformatics during the last decade.

In addition to these regularly events, members of the department also organized the following national and international workshops.

- Workshop on Mathematical and Statistical Aspects of Molecular Biology (MASAMB), Berlin, 04/2012 (organizers: Martin Vingron, Julia Lasserre, Alena Mysickova)
- Meeting of Section 2 (Information Sciences) of Leopoldina, Berlin, 02/2012 (organizer: Martin Vingron, together with Martin Grötschel, Zuse Institute Berlin, and Thomas Lengauer, Max Planck Institute for Informatics)
- Bioinformatics Summerschool, EU-TRACC Consortium, Bergen, Norway, 07/2009 (co-organizer Christine Steinhoff)
Research Group Development & Disease
Established: 05/2000

Head
Prof. Dr. Stefan Mundlos
Phone: +49 (0) 30 8413-1449
Fax: +49 (0) 30 8413-1385
Email: mundlos@molgen.mpg.de

Secretary
Cordula Mancini (part time)
Phone: +49 (0) 30 8413-1691
Fax: +49 (0) 30 8413-1960
Email: mancini@molgen.mpg.de

Scientists and postdocs
Lila Allou (since 09/15)
Martin Mensah* (07/15)
Daniel Ibrahim* (since 05/14)
Björn Fischer-Zirnsak* (since 04/14)
Guillaume Andrey* (since 02/14)
Wing Lee Chan* (since 07/12)
Dario Lupianez* (since 05/12)
Francisca Real-Martinez (since 04/12)
Malte Spielmann* (since 09/10)
Peter Krawitz* (since 01/09)
Uwe Kornak* (since 03/03)
Peter Robinson* (since 10/00)
Sigmar Stricker* (09/02-06/14, guest since 07/14)
Jochen Hecht* (10/01-02/15)
Claus Eric Ott* (03/05-12/14)
Mateusz Kolanczyk* (07/03-10/14)
Sandra Dölken* (06/08-07/14)
Jirko Kühnisch* (11/11-05/14)
Johannes Egerer* (09/07-03/14)

PhD students
Alexandra Despang (since 04/15)
Christina Paliou (since 03/14)
Arunima Murgai* (since 11/13, guest)
Björn Kragesteen* (since 10/13)
Katerina Kraft* (since 03/13)
Ivana Jercovic* (since 10/12)
Mickael Orguer* (since 09/12, guest)
Denise Emmerich* (since 08/12)
Sinje Geuer* (since 01/12)
Anja Will* (since 12/11)
Jürgen Stumm* (since 10/11, guest)
Pedro Vallecillo-Garcia* (since 01/11, guest)
Martin Franke* (since 10/10)
Naeimeh Tayebi* (09/13-07/15)
Magdalena Steiner* (02/15-06/15)
Saniye Yumlu* (01/09-05/14)
Daniel Ibrahim* (08/09-04/14)
Björn Fischer-Zirnsak* (04/08-03/14)
Julia Grohmann* (01/13-12/13)
Silke Lohan* (01/09-12/13)
Hendrikje Hein* (01/11-10/13)
Wing Lee Chan* (01/07-06/12)

Diploma students
Georgeta Leonte* (09/14-02/15)
Jenny Viebig* (04/12-07/14)
Felix Wiggers* (03/13-08/13)
Verena Kappert* (09/12-07/13)
Krzysztof Brzezinka* (03/12-09/12)

* externally funded
Introduction

Structure of the research group

The research group Development & Disease is part of and works in close collaboration with the Institute for Medical and Human Genetics (IMHG) at the Charité-Universitätsmedizin Berlin. The research group focuses on fundamental questions regarding normal and abnormal development with particular interest in the molecular pathology of human disease associated mutations. The IMHG provides clinical and diagnostic genetic service within Germany and worldwide. Medical doctors in training get scientific education at the Development & Disease group and scientists from the MPIMG have the opportunity to specialize in Medical Genetics at the IMHG. Thus, the IMHG covers the clinical aspects, genome analysis and diagnostics of human genetic disease, whereas the research group is more focused on the functional analysis and basic biology of genomic alterations. Shared infrastructures, exchange of technical achievements and expertise, as well as common research goals ensure a successful interdisciplinary approach to study the mechanisms of genetic disease.

Development and regeneration are related and it is generally believed that developmental pathways get re-activated during healing processes. To synergistically use our expertise in the molecular control of cell differentiation and development with new advances in regenerative medicine, we collaborate closely with the Berlin-Brandenburg Center for Regenerative Medicine (BCRT), which is funded by the BMBF.

With the retirement of Hans-Hilger Ropers in October 2014, Vera Kalscheuer and her co-workers joined the Development & Disease group and continue their ongoing projects.
Research concept

By combining developmental biology, genetics, and clinical medicine we aim at generating in-depth knowledge of human genetic disease. Through clinical and diagnostic services as well as collaborations, patient cohorts are generated that are analyzed for genetic defects. Besides patient recruitment, this involves expert phenotyping, data management and analysis, mutation detection, and functional analysis of disease genes/variants. Within this setting, the Development & Disease group focuses on analysis of normal and abnormal developmental processes in model systems. Our particular interest is directed towards conditions that are related to abnormal development, growth, and aging of the musculoskeletal system. The skeleton is a particularly informative model system for our phenotype driven approach, due to an almost unlimited number of distinct phenotypes, the accessibility of structures, the easy assessment of phenotypes and the wealth of knowledge about the molecular mechanisms of limb development.

The focus of the Development & Disease group has moved from gene function to mechanisms of gene regulation during development and the analysis of mutations that alter this process. The identification and characterization of regulatory mutations is an emerging field in genetics. In contrast to the coding genome, the function of non-coding DNA cannot be predicted from the sequence alone. A number of technologies including chromosome conformation capture and the identification of epigenetic marks have helped to identify and characterize the regulatory potential of genomic regions. Furthermore, the folding of the genome has been shown to be directly involved in gene regulation by facilitating DNA-DNA contact e.g. between promoters and enhancers. How long range regulation functions in this setting and how mutations interfere with this process is our topic for the future. With the establishment of the CRISPR/Cas9 methodology and its derivatives, we are now able to manipulate the genome in any possible way. This opens the opportunity to test human mutations for their regulatory potential. At the same time, we can dissect specific loci and investigate how genomic structure and sequence influence the regulatory function of the genome.

It is our aim to translate the knowledge gained in clinical practice, mainly by developing novel test systems, as well as algorithms and methods for better data analysis, thereby improving the diagnostics for genetic diseases.
Scientific achievements and findings

Over the last years, we have been successful in the detection and characterization of skeletal defects on a genetic, molecular, and developmental level. We identified the molecular cause of a range of conditions, often in conjunction with detailed functional analysis. We developed a novel algorithm for next generation sequencing (NGS)-based disease gene testing (PhenIX), which incorporates the phenotype in the analysis. One of our major achievements has been the in detail description of a novel disease mechanism, which involves “rewiring” of enhancer-promoter interactions due to alterations of higher order genomic structures. These ectopic interactions that are induced by large scale rearrangements can result in gene misexpression and consecutive malformations.

Long range regulation

Development requires tight spatial and temporal control of gene expression. Most of this appears to be achieved via long range regulation involving DNA looping and DNA-DNA contacts between e.g. enhancers and promoters, a process mediated by protein complexes that recognize certain DNA sequences and/or modifications. Technologies based on chromosome conformation capture (3C and its derivatives) can be used to quantify these contacts and relate them to gene activity and specific regulatory sequences.

Figure 1: Schematic of Topologically Associated Domains (TADs). Triangels represent DNA-DNA interactions within one domain, e.g. enhancer-promoter contacts for gene regulation. Activities of neighboring TADs are shielded from each other by boundaries, indicated as wall between TADs.
(enhancers). These data have helped to understand the genome as a 3-dimensional molecule, in which folding is a prerequisite for the regulation of gene expression. Furthermore, it has been shown that the genome is divided into bins of regulatory activity that promote DNA-DNA contacts within one region, but shield activity against neighboring regions (Figure 1). These sections of DNA have been called Topologically Associated Domains (TADs). TADs appear to provide a general scaffold for the genome that determines 3D-folding thereby controlling the size and position of regulatory regions. TADs are conserved between species and remain remarkably stable during cell differentiation.

By screening cohorts of patients with limb malformations via array-CGH, we have identified a series of structural variations (deletions, duplications, inversions, translocations) involving non-coding conserved elements (CNEs) that are located in the vicinity of developmentally important genes. This includes structural variations at the BMP2 (brachydactyly type A2), SHH (mirror image polydactyly), SOX9 (Cooks syndrome), IHH (craniosynostosis with syndactyly), MSX2 (cleidocranial dysplasia), EPHA4 (F-syndrome, brachydactyly, polydactyly), and a number of other loci. Our findings show that structural variations (SVs) can result in phenotypes and diseases, even if only non-coding regions of the genome are affected. Furthermore, the entire genomic region has to be taken into consideration, if the structure of a TAD is altered due to possible regulatory effects on neighboring genes. We described these novel disease mechanisms in a hy-

Figure 2: Breaking TADs. A dysfunctional boundary, caused by e.g. loss due to deletion or misplacement due to a duplication or inversion, can result in rewiring of enhancer promoter contacts and consecutive misexpression of genes.
Research Group  
Development & Disease

P basis paper (Spielmann & Mundlos, Bioassays 2013) and were now able to demonstrate the predicted pathogenesis of rearrangement-induced genomic misregulation.

In a rare genetic condition affecting the elbow/wrist/hands (Liebenberg syndrome) we identified deletions of a gene (H2AFY) located 300 kb away from PITX1, a transcription factor that determines hind limb identity. We show that the deletion results in the activation of a nearby enhancer, which in turn results in the ectopic expression of PITX1 in the forelimb. This results in a homeotic arm-to-leg transformation with the elbow acquiring morphological characteristics of the knee (Spielmann et al., Am J Hum Genet 2012). The ectopic expression of PITX1 is caused by the removal of a TAD boundary between PITX1 and the enhancer. Figure 2 shows the proposed mechanism of gene activation by a loss of boundary activity. The disease mechanism and how Pitx1 is normally regulated is currently being investigated using mouse mutants.

![Figure 2: Proposed mechanism of gene activation by a loss of boundary activity.](image)

Detailed studies of genomic rearrangement were carried out at the EPHA4 locus. We studied three limb malformation syndromes (brachydactyly, syndactyly, polydactyly) caused by various types of SVs that interfere with

![Figure 3: Ectopic interaction and gene activation depends on presence/absence of CTCF-associated boundaries.](image)
either the telomeric or the centromeric boundary of the large EPHA4-associated TAD. We can show that the SVs induce misexpression of nearby genes through the ectopic activation of an Epha4-associated enhancer cluster (schematically shown in Figure 3). The induction of this misexpression depends on the presence/absence of CTCF-associated boundary elements that normally shield neighboring regions of activity. Figure 3 shows the Epha4 locus, the proposed boundaries and the CRISPR/Cas-induced deletions. 4C analysis demonstrates ectopic interaction, which results in gene misexpression and limb malformation. These studies provide the first evidence that TADs and their boundaries are of biological and medical relevance (Lupianez et al., Cell 2015).

Other SVs that do not disrupt TADs or their boundaries can, however, also result in disease. Interesting disease mechanisms in this respect are intra-TAD duplications like those described by us near BMP2 or IHH. We have been investigating this mechanism in a series of duplications involving the regulatory domain of IHH. The duplications include regulatory sequences that drive Ihh expression. We are investigating this mechanism in detail by dissecting the Ihh regulatory TAD.

To understand the mechanisms of disease and how TADs control gene expression, we have developed an adaption of the CRISPR/Cas9 technology that allows us to efficiently produce large rearrangements in mice (Kraft et al., Cell Rep 2015). This methodology – CrisVar – is based on inducing two double strand breaks in ES cells by CRISPR that correspond to the SV breakpoints (Figure 4). The repair mechanism in the cell results in re-joining of the fragments, but alternatively, also in deletion, duplication or inversion of the fragment. ES cell are screened for the desired SV and used
to produce mice. We have been dissecting several loci producing deletions, duplications and inversions that recapitulate human disease-associated rearrangements or to study basic mechanisms of genomic regulation.

**Mechanisms of limb development**

We use the limb as a model system to study molecular disease mechanisms of developmental defects (malformations). The limb/skeleton is particularly suited for this purpose because of the extensive knowledge about the molecular basis of limb development, the easy accessibility in model systems and the efficient readout of phenotypes. One focus has been on normal and abnormal digit development, mainly based on our clinical interest in brachydactylies. Using a disease and phenotype-driven approach, we identified the BMP pathway as the major player in digit and joint formation and were able to correlate mutations and their mechanisms with disease phenotypes. The in-depth analysis of mechanisms of digit development led to basic concepts of joint formation and digit elongation.

The function of Hox genes has been another major interest in the lab. We showed that Hox genes determine the shape and identity of limb bones and that their inactivation causes a homeotic transformation of long bones (metacarpals) into round bones (carpals). The latter process involves the Wnt pathway and in particular Wnt5a thereby regulating cell polarity and, consequently, the shape of bones (Kuss et al., Dev Biol 2014). Together, our findings show that Hox genes are essential modifiers of shape and gestalt of the limbs by controlling stem cell differentiation into chondrocytes or osteoblasts.

More recently, we expanded our interest towards other phenotypes of the limbs including polydactyly, limb deficiencies and other conditions. Phenotypes produced by large scale rearrangements or other mutations were generated by CrisVar and analyzed in detail using our extensive knowledge in basic mechanisms of limb development.

**Transcription factors in bone/limb development**

Transcription factors (TFs) regulate gene expression. How mutations in TFs result in specific phenotypes has been a study subject in the lab. A focus has been on TFs involved in skeletal development such as Runx2, Pitx1, and the 5’ Hox genes from the D and A cluster. We expanded our functional analysis of TFs by establishing methodologies to identify TF targets and binding sites within the genome. We adapted the technology of chromatin immunoprecipitation followed by deep sequencing (ChIP-Seq) to analyze TFs and their sequence variants in a standardized in vitro system. We use chicken limb bud mesenchymal cells, grow them in micro-mass and express tagged versions of the TFs using an avian-specific virus.
Based on this technology, we have been able to create a genomic binding profile for various TFs that play important roles in bone/limb development.

Furthermore, we tested mutations identified in our patient screen. One of these mutations (Q317K) is located in the DNA-binding homeobox domain of HOXD13. The mutated glutamine (Q) is conserved in most homeodomains, a notable exception being bicoid-type (Pitx1) homeodomains that have lysine (K) at this position. Our results show that the mutation results in a shift in the binding profile of the mutant toward a bicoid/PITX1 motif (Figure 5) and thus in a partial conversion of HOXD13 into a TF with bicoid/PITX1 properties (Ibrahim et al., Genome Res 2013). In another project we systematically analyze the genomic binding sites for 5’ Hoxd and Hoxa genes with the aim to identify the mechanisms of Hox gene target identification.

Osteoporosis and mechanisms of aging

Reduced bone mass, measured as bone mineral density, results in increased fracture risk. Bone mass is regulated by a complex network of pathways, which we have been investigating mainly by studying human mutations and the related disease genes and their function. A focus has been on conditions with segmental progeria, i.e. conditions that show signs of ageing in specific organs/tissues, such as skin and bone. One prime example for these conditions is Gerodermia osteodysplastica, a recessive disease with skin laxity and osteoporosis. We created a mouse model by inactivating Gorab and show that these mice recapitulate the human phenotype. A deficiency of Gorab results in altered skin matrix composition in particular in regard to the modification of proteoglycans. This in turn results in elevated TGF-β signaling, oxidative stress, and the induction of cellular senescence. Treatment with an antioxidant rescued the osteoporosis phenotype.

Figure 5: A homeobox mutation results in HOXD13 to PITX1 conversion. Left: severe limb malformation due to Q317K mutation in HOXD13. Middle: Analysis in micromass cultures after ChIP-seq identifies the known wt Hoxd13 motif and a similarity of the mutant with Pitx1 motif. Right: overexpression of wt Hoxd13, Q317K mutant and Pitx1 in chicken limb buds. Note infected wing buds lack a forelimb-specific pad, have a straightened wrist, and develop a fourth digit (arrowheads), similar to Pitx1 overexpression.
Increased oxidative stress-related cellular senescence appears to be an important trigger for age-related changes in skin and bone, offering novel therapeutic options. Several overlapping disorders are under investigation including \textit{Pycr1}- and \textit{Atp6v0a2}-related Cutis laxa, which show an overlapping pathophysiology indicating novel functional connections between the secretory pathway and mitochondria.

Uwe Kornak, Hardy Chan, Björn Fischer, and Magdalena Steiner are in charge of this project.

**Muscle and connective tissue development**

Coordinated development of the different tissues forming the musculoskeletal system ensures the emergence of a functional unit. The cell-cell communication involved in these developmental events as well as their recapitulation and modification in adult repair processes are far from being understood. In both contexts we are interested in the cross-talk of resident connective tissue progenitors with myogenic cells. We have identified a zinc-finger transcription factor as novel marker for muscle connective tissue precursors, thereby describing the first embryonic lineage of so-called fibro-adipogenic progenitors (FAPs). We found that a mouse mutant for this factor displays local defects in muscle patterning that, via RNA-Seq transcription profiling, was traced down to deregulation of both chemokine signaling as well as extracellular matrix modeling. In short, embryonic FAPs thereby provide a local microenvironment necessary for correct muscle patterning. At present we are analyzing the role of this factor and the cell population marked by it during muscle repair. In another project...
we are focusing on specific secreted signaling molecules and their role in connective tissue-muscle crosstalk. Furthermore, we have previously shown a role for Neurofibromin (Nf1) in muscle development. We are now dissecting this role using tissue-specific conditional inactivation.

Sigmar Stricker, Pedro Vallecillo-Garcia, Jürgen Stumm, Mickael Orgeur, and Arunima Murgai are in charge of this project.

**Bioinformatics**

The Human Phenotype Ontology (HPO) was developed in order to enable computational analysis of the clinical abnormalities observed in human disease. The HPO has been adopted worldwide by programs such as the UK 100,000 Genomes Project, the Sanger Institute’s DECIPER/DDD, the NIH Undiagnosed Diseases Network, etc. It has been used to develop phenotype-driven algorithms for the analysis of exome and genome sequences for the identification of novel disease genes as well as for clinical diagnostics. The HPO was originally developed mainly in the field of rare diseases, but has recently been extended to over 3,000 common diseases, and used for system-level phenotypic analysis of complex disease. The HPO has been used to develop a novel program for diagnostic purposes combining variant with phenotype analysis (Zemojtel et al., Sci Transl Med 2014). More recently, algorithms have been developed to perform prioritization of non-coding mutations in whole-genome sequencing (Ibn-Salem et al., Genome Biol 2014). The bioinformatics group additionally develops pipelines and bespoke algorithms for other areas of genomics include ChIP-seq and T-cell receptor profiling by next generation sequencing.

Peter Robinson leads the bioinformatics group.

**Genome analysis for disease variant identification**

With exome and whole genome sequencing the identification of disease causing variants has entered a new stage. Sequencing and variant calling has become a routine, but the interpretation of the results remains challenging. This is particularly true for non-coding variants because their effect is very difficult to predict. In the long range project we aim at improving the interpretation of data in regard to structural variants and their effect on gene regulation. The interpretation of point mutations and the prediction of their pathogenicity, in contrast, are still in the beginning. The focus of our mutation identification efforts is therefore directed towards non-coding variants.

Variants have to be tested for their pathogenicity. This is particularly important for rare non-coding variants as their effect is difficult to predict.
and because of the rarity of these conditions. We have therefore advanced our testing tools by implementing mouse models as a variant testing system. Using CRISPR/Cas, we can quickly create a mutation and test its pathogenicity in mice embryos, even without creating individual mutant lines. This has been successfully applied to several variants with unclear significance. Using this approach we verified mutations in a family with ectrodactyly, lower limb deficiency, and hypoventilation.

This project is interdisciplinary and involves several departments at the Charité and elsewhere, clinicians for sampling, diagnosing, and phenotyping as well as bioinformaticians for the analysis of phenotypic and sequence data, and sequencing technology for mutation identification. The continuous supply of patient material provides us with a constant flow of novel genes and mutations.

Large scale sequencing produces many variants of unknown significance that can make a diagnosis difficult. The bioinformatics group has developed an algorithm combining variant analysis with phenotype analysis based on the Human Phenotype Ontology. This program (PhenIX) has been tested by us in a patient cohort with unknown diagnosis. It was shown to be very effective and identified the correct diagnosis in approx. 30% (Zemojtel et al., Sci Transl Med 2014).

**Cooperation within the institute**

Cooperation over the past years have been with the Department of Bernhard Herrmann on mouse transgenic technology, CRISPR/Cas-based genome editing in ES cells, and the analysis of mutant mice. Cooperation with the Vingron Department has focused on computational analysis of ChIP-Seq data and the bioinformatic analysis of chromosome conformation capture (4C, capture-C) data. Intense collaborations exist with the mouse and the sequencing facilities.

**Special facilities and equipment**

The research group as well as the IMG is equipped with the standard facilities for research into genetics, developmental biology, cell biology, and molecular biology. Special equipment includes the histology unit for the MPIMG and a sequencing facility for the Charité/BCRT.
Planned development

The non-coding genome and its role in gene regulation and disease will be the major focus of the next years. There are a number of important questions that need to be addressed in this respect. First, we will use whole genome data for mutation analysis with a focus on non-coding DNA. The interpretation of these variants is difficult and requires stringent procedures regarding the selection of patient material, the calling of variants, their bioinformatic and clinical interpretation and the testing in model systems. We are cooperating with several groups to cover certain aspects of this pipeline; nevertheless, it is our aim to have the expertise and to be at the forefront of development regarding the majority of above mentioned aspects.

Our results show that regulatory elements can ectopically activate other genes. However, not all enhancers are able to do this and not all genes appear to be responsive. Thus, there appears to be a degree of enhancer-promoter specificity, which argues against a general promiscuity and exchangeability of elements. A prediction of possible targets and/or activation potential of enhancers are important also for the interpretation of SVs. We will address this question using in vitro and in vivo model systems and genome editing. Other important questions in this respect are related to the organization of TADs and how mutations/SVs can influence their function and configuration. The function of boundaries needs to be studied in detail. We will address this question by altering boundary structures, TF binding sites and the placement of boundaries in other genomic contexts. The function of CTCF in TAD formation and DNA looping will be another important question.

Selected publications

Group members are underlined.


Spielmann M & Mundlos S (2013). Structural variations, the regulatory landscape of the genome and their alteration in human disease. *Bioessays*, 35(6), 533-543
Chromosome Rearrangements & Disease group

Established: 07/1995 (Dept. of Human Molecular Genetics), since 11/2014 Research Group Development & Disease

**Head**
Dr. Vera M. Kalscheuer  
Phone: +49 (0)30 8413-1293  
Fax: +49 (0)30 8413-1383  
Email: kalscheu@molgen.mpg.de

**Scientist**
Melanie Hambrock* (since 06/14)

**PhD students**
Friederike Hennig (since 04/14)  
Melanie Hambrock (12/09 - 05/14)

**Master student**
Friederike Hennig (05/13-03/14)

**Technicians**
Astrid Grimme (since 02/04, part time)  
Ute Fischer (since 08/95, part time)

**Scientific overview**

The overall goals of the group are to elucidate the causative genetic defects of human neurodevelopmental disorders (NDD) by using state-of-the-art genetics and genomics strategies, and to understand the functional relevance of disease-relevant mutations and the cell protein signaling networks the respective proteins are embedded in. The results will provide insight into normal and disturbed gene functions and into molecular pathways, and they will considerably improve our understanding of the biological and cellular mechanisms underlying normal brain development.

* externally funded
Our group is actively searching for genetic causes of NDD by systematic mapping of disease-associated chromosome breakpoints (previously in collaboration with W. Chen and R. Ullmann, MPIMG; presently as part of the International Breakpoint Mapping Consortium (IBMC), led by N. Tommerup, Copenhagen, Denmark) and massively parallel sequencing. These activities are complemented by in vitro and in vivo studies on previously discovered and newly identified NDD genes and proteins.

Identification of novel genes and genetic defects for X-linked intellectual disability

During the last years, we have discovered or made major contributions to the discovery of numerous novel genes for NDD. Employing massively parallel sequencing of all X-chromosome exons in index males from 470 families with X-linked intellectual disability (XLID) collected by the European MRX consortium and associated groups, followed by computational analysis of our data set in collaboration with Hao Hu from our department and the department of Computational Molecular Biology (Martin Vingron), we found 14 novel genes being involved in syndromic or non-syndromic XLID (e.g. KIAA2022, ZC4H2, (Figure 7), CNKSR2, CLCN4, FRMPD4, KLHL15, LASIL, RLIM, THOC2) (e.g., van Maldergem et al., Hum Mol Gen 2013; Hirata et al., Am J Hum Gen 2013; Willemsen et al., J Med Gen 2014; Vaags et al., Ann Neurol 2014; Hu et al.

Figure 7: Zc4h2 knockdown in zebrafish causes compromised swimming. Superimposed images of tail movement in control (left) and zc4h2 mutant (right) zebrafish showing weak swimming contraction in the mutant. This defect could be rescued with wild-type zebrafish zc4h2, but not with constructs carrying the missense mutations identified in the families (adapted from Hirata et al., Am J Hum Genet, 2013).
The corresponding proteins are implicated in diverse cellular processes, including transcription, mRNA export and translation, and belong to pathways and networks with established roles in cognitive function and intellectual disability in particular (Figure 8). Further, our results suggest that systematic resequencing of all X-chromosomal genes in a cohort of patients with genetic evidence for X-chromosome locus involvement may resolve up to 58% of fragile X-negative cases. We also re-investigate selected families with XLID, in which we could not identify the disease-relevant mutation in the exons of genes.

**Functional studies on previously identified X-linked intellectual disability genes**

Significant progress was also made to better understand disorders through genes identified in previous years, including the polyglutamine binding protein 1 (PQBP1) gene for Renpenning syndrome and the cyclin-dependent kinase-like 5 (CDKL5) gene for a variant of Rett syndrome.

Mutations in PQBP1 cause intellectual disability, microcephaly, short stature and other midline defects with variation in severity of the phenotypes, both within and between families (Kalscheuer et al., Nat Genet 2003). In the patients all frameshift mutations result in the production of a truncated PQBP1 protein (Musante et al., Hum Mutat 2010). Therefore, it is highly likely that the clinical phenotype is caused by the loss of wild-type PQBP1-associated functions and the presence of defect PQBP1 protein. To get more insight into the possible functions of PQBP1, we established a PQBP1 interaction network. Analysis of PQBP1 complexes revealed
that PQBP1 interacts with RNA-binding proteins, including the fragile X mental retardation protein (FMRP) and subunits of the intracellular transport-related dynactin complex, and that PQBP1 protein complex formation is dependent on the presence of RNA. In primary neurons PQBP1 co-localized with its interaction partners in specific cytoplasmic granules. In further studies, we showed that oxidative stress caused relocalization of PQBP1 to stress granules (SGs) and that the cellular distribution of PQBP1 plays a role in SG assembly. Together these data demonstrate a role for PQBP1 in the modulation of SGs and suggest its involvement in the transport of neuronal RNA granules, which are of critical importance for the development and maintenance of neuronal networks, thus illuminating a route by which PQBP1 aberrations might influence cognitive function (Kunde et al., Hum Mol Genet 2011). Also, Pqbp1-hypofunction in mouse neural stem progenitor cells caused microcephaly and we found that among others especially anaphase promoting complex subunit 4 (Apc4) is a critical downstream target of Pqbp1 for microcephaly (Ito et al., Mol Psychiatry 2015).

**CDKL5** mutations primarily affect girls and cause severe infantile encephalopathy with early onset, mostly untreatable epileptic seizures, usually starting within the first 6 months after birth. CDKL5 patients develop severe ID with absent speech and stereotypic behavior (Kalscheuer et al., Nat Genet 2003; Tao et al., Am J Hum Genet 2004; Cordova-Fletes et al., Clin Genet 2010; Rademacher et al., Neurogenet 2011). Likewise, we have shown that truncation of the Netrin G1 (NTNG1) gene caused a highly similar clinical phenotype in a patient with a balanced translocation involving chromosomes 1 and 7 (Borg et al., Eur J Hum Genet 2005). Given the overlapping clinical features between CDKL5 patients and the girl with truncated NTNG1, we hypothesized that CDKL5 and Netrin G1 could function in the same molecular pathway. In subsequent work we found that CDKL5 localizes at excitatory synapses and contributes to correct dendritic spine structure and synapse activity, and that CDKL5 interacts with the specific Netrin G1 receptor NGL-1, which plays a crucial role in early synapse formation and maturation. Next, we investigated if NGL-1 could be a phospho-substrate of CDKL5 and we indeed could show that CDKL5 phosphorylates NGL1 on a serine residue (S631) close to the cytoplasmic C-terminal PDZ binding domain. This phosphorylation is necessary (i) for reinforcing the interaction between CDKL5 and NGL-1 and (ii) for promoting a stable association between NGL-1 and PSD95, a major protein of the postsynaptic density, which plays an important role in synaptic plasticity. Accordingly, phospho-mutant NGL-1 was not any longer able to induce synaptic contacts while its phospho-mimetic form bound PSD95 more efficiently and partially rescued the CDKL5-specific spine defects. We also investigated the phosphorylation level of NGL-1 in humans using a fibroblast cell line derived from a girl, who carried a balanced chromo-
some translocation that truncated CDKL5 and due to inactivation of the normal X-chromosome lacked functional CDKL5. Importantly, these cells when compared to normal control fibroblasts, which express endogenous CDKL5, exhibited reduced levels of phosphorylated NGL1 with respect to the total amount of NGL1 protein and this level could be increased by overexpression of CDKL5. Together, our findings suggest a critical regulatory role for CDKL5 in the formation of excitatory synapses by coupling, through NGL-1 phosphorylation, the Netrin G1-NGL-1 adhesion with the recruitment of PSD95 and thereby provide important molecular insights into the pathophysiology of RTT (Ricciardi et al., Nat Cell Biol 2012).

I have been instrumental in formation of an ERA-NET-Neuron consortium around our attempt to further study the pathophysiology of CDKL5 in the brain. In that consortium, which I am coordinating, we undertake a multidisciplinary approach to link CDKL5 and two other genetic syndromes with cognitive impairment (Opitz/BBB-G syndrome and Rett syndrome) with mTOR activity and brain function. We use in vitro (e.g. primary hippocampal neurons, Figure 9) and in vivo model systems and combine biochemical, cellular, electrophysiological and mouse behavior approaches in order to answer the fundamental question of how mTOR-dependent protein synthesis translates into synaptic plasticity and learning and memory. Within the project, we hope to gain insight into the role of mTOR signaling in brain function and the influence of its deficiency in the pathogenesis of ID. These data will be fundamental for the differential diagnosis of ID and in the future may also open new avenues for the development of novel therapeutic strategies for patients with ID syndromes and other related disorders (in collaboration with Vania Broccoli, Milano, Susann Schweiger,
Mainz, Rainer Schneider, Innsbruck, Yann Herault, Strasbourg and David Meierhofer, Mass spectrometry facility, MPIMG).

In parallel, functional analyses of other genes and proteins are being performed in collaboration with specialized groups. Also, we currently work on establishing the introduction of missense changes into mouse ES cells and zygotes using the CRISPR/Cas9 system (in collaboration with Judith Fiedler, Lars Wittler and Stefan Mundlos). This tool will allow testing of missense variants of unknown disease causation identified in patients and families.

**Selected publications**

Group members are underlined.


General information about the whole research group

For Vera Kalscheuer, only the publications and other general information since 2015 are listed here. The publications and information about former years are listed in the report of the Dept. of Human Molecular Genetics (H.-H. Ropers).

Complete list of publications (2009-2015)

Research group members are underlined.

2015


Flöttmann R, Knaus A, Zemojtel T, Robinson PN, Mundlos S, Horn D & Spielmann M (2015). FGFR2 mutation in a patient without typical features of Pfeiffer syndrome - The emerging role of combined NGS and phenotype based strategies. European Journal of Medical Genetics, 8(8), 376-380


**2014**


2013


of bone fragility. *American Journal of Human Genetics*, 92(4), 565-574


**2013**


Spielmann M & Mundlos S (2013). Structural variations, the regulatory landscape of the genome and their alteration in human disease. *Bioessays*, 35(6), 533-543


2012


Heinrich V, Stange J, Dickhaus T, Imkeller P, Krüger U, Bauer S, Mundlos S, Robinson PN, Hecht J & Krawitz PM (2011). The allele distribution in next-generation sequencing data sets is accurately described as the result of a stochastic branching process. Nucleic Acids Research, 40(6), 2426-31


niosynostosis. *American Journal of Human Genetics, 88*(1), 70-75


in siblings. European Journal of Medical Genetics, 54(4), e441-e445


2010


duplications of the HOXD locus on chromosome 2q. European Journal of Human Genetics, 18(12), 1310-1314


Robinson PN (2010). Whole-exome sequencing for finding de novo mutations in sporadic mental retardation. Genome Biology, 11(12), 144

Robinson PN & Mundlos S (2010). The human phenotype ontology. Clinical Genetics, 77(6), 525-534


ylated dishevelled via Ror2. *FASEB Journal*, 24(7), 2417-2426

Witte F, Chan D, Economides AN, Mundlos S & Stricker S (2010). Receptor tyrosine kinase-like orphan receptor 2 (ROR2) and Indian hedgehog regulate digit outgrowth mediated by the phalanx-forming region. *Proceedings of the National Academy of Sciences of the USA*, 107(32), 14211-14216

2009


Tuysuz B, Mizumoto S, Sugahara K, Celebi A, Mundlos S & Türkmen S
Omani-type spondyloepiphyseal dysplasia with cardiac involvement caused by a missense mutation in CHST3. *Clinical Genetics, 75*(4), 375-383


## Scientific honours

Dario Lupianez: *ESHG Young Scientist Award*, European Society of Human Genetics, 2015

Dario Lupianez: *Vortragspreis (best lecture award)*, Deutsche Gesellschaft für Humangenetik, 2015

Katerina Kraft: *Peter-Hans Hofschneider Prize*, Max Planck Society for the Advancement of Science, 2015

Daniel Ibrahim: *Vortragspreis (lecture award)*, Deutsche Gesellschaft für Humangenetik, 2014

Björn Fischer-Zirnsak: *Robert Koch Prize of the Charité*, 2014

Stefan Mundlos: *Appointed member of the Berlin Brandenburg Academy of Sciences*, 2014

Malte Spielmann: *ESHG Young Scientist Award*, European Society of Human Genetics, 2012


Uwe Kornak: *Ulmer Dermatologiegut-preis*, University of Ulm, 2011

Eva Klopacki: *Finalist Trainee Award*, American Society of Human Genetics, 2009

Eva Klopacki: *ESHG Young Scientist Award*, European Society of Human Genetics, 2009

Eva Klopacki: *Vortragspreis (lecture award)*, Deutsche Gesellschaft für Humangenetik, 2009

## Selected invited talks

(Stefan Mundlos)

*Long Range Regulation, Genomics and Genome Editing. EMBO Nuclear Structure and Dynamics, Avignon, France, 10/2015*

*Structural variations, the regulatory landscape of the genome and their alteration in human disease. II Retreat of the Department of Molecular Medicine University of Pavia, Pavia, Italy, 03/2015*

*Chromatin associated regulatory domains of the genome and their alteration in disease. Weizman Institute of Science, Rehovot, Israel, 03/2015*

*Genetische Diagnostik seltener Erkrankungen - eine Herausforderung. Opening Ceremony of the Zentrum für Seltenen Erkrankungen Aachen (Centre for Rare Diseases), Uniklinik RWTH Aachen, 11/2014*

*The regulome - the Next Frontier in Human Genetics. Meeting of the Polish Society for Human Genetics, Poznan, Poland, 09/2013*

*Regulatory mutations in human disease. Turkish Society for Human Genetics, Izmir, Turkey, 09/2013*

*Limb malformations as a model to study genetic defects of development. Newlife Birth Defects Research Centre (BDRC), University College London, London, UK, 10/2012*

*Clinical relevance of copy number variation. British Human Genetics Conference, University of Warwick, UK, 09/2012*
Regulatory CNVs - phenotypes and mechanisms. British Society for Human Genetics, Norwich, UK, 09/2012

Regulatory mutations – the next frontier in Human Genetics. European Society for Human Genetics, Erlangen, Germany, 06/2012

Regulating skeletal development – lessons to be learned from rare disease. Paris Descartes University Hôpital Necker, Paris, France, 03/2012

Regulatory CNVs, genomic disorders 2012. The Genomics of Rare Disease, Sanger Center, Hinxton, UK, 03/2012

HOX genes sculpture our bones. Keynote lecture at the Day of Clinical Research of the Department Clinical Research at the University of Bern, Bern, Switzerland, 11/2011

Structural variations of the human genome and their role in congenital disease. Sanger Center, Hinxton, UK, 10/2011

The molecular basis of skeletal disease. Spanish Society for Genetics, Murcia, Spain, 09/2011

Far, far away – long range regulation in skeletal development and disease. Gordon Research Conference Bone & Teeth, Les Diablerets, Switzerland, 06/2011

The role of Hox genes in limb development and bone formation. 3rd joint Meeting of the European Society of Calcified Tissues & the International Bone and Mineral Society, Athens, Greece, 05/2011

Digit development, a model for skeletal morphogenesis. Extracellular Matrix in Health and Disease, Boston, USA, 04/2011

Phenotypes and the regulome. Wilhelm Johansen Symposium: The Impact of Deep Sequencing on the Gene, Genotype and Phenotype Concepts, Copenhagen, Denmark, 03/2011

Defects of long range regulation. Lausanne Genomic Days, Lausanne, Switzerland, 02/2011

Genetics of limb malformations. Italian Society for Human Genetics, Florence, Italy, 10/2010

Far reaching consequences - mechanisms and problems of long range control. European Human Genetics Conference 2010, Gothenburg, Sweden, 06/2010

Chondrogenic development and disease models. Current Concepts in Regenerative Orthopaedics, Düsseldorf, Germany, 06/2010

Chondrogenesis and patterning. International Bone and Mineral Society, IBMS Davos Workshop, Davos, Switzerland, 03/2010

Genetics of limb malformation. 8th World Symposium on Congenital Malformations of the Hand and Upper Limb, Hamburg, Germany, 09/2009

 Syndromes with segmental progeria as models for the ageing bone and skin. 9th International Skeletal Dysplasia Society, Boston, USA, 07/2009

Bone development and dysplasias. 2nd Joint Meeting of the British Society for Matrix Biology and Bone Research Society, London, UK, 06/2009


Appointments of former members of the group

Sigmar Stricker: Professorship (W2) for Biochemistry and Genetics, Freie Universität Berlin, 2014

Eva Klopocki: Professorship (W2) for Human Genetics, University of Würzburg, 2012
Uwe Kornak: Professorship (W2) for Functional Genetics, Charité – Universitätsmedizin Berlin, 2012

Peter Robinson: Professorship (W2) for Medical Bioinformatics, Charité – Universitätsmedizin Berlin, 2012

Katrin Hoffmann: Professorship (W3) for Human Genetics, Martin-Luther-Universität Halle, 2011

Petra Seemann: Professorship (W1) for Model Systems for Cell Differentiation, Berlin-Brandenburg Center for Regenerative Therapies, 2009

Habilitationen/State doctorates

Katharina Dathe: Molekulare Ursachen isolierter Handfehlbildungen am Beispiel des BMP-Signalwegs und von SHH. Freie Universität Berlin, 2010

Sigmar Stricker: Molekulargenetik und funktionelle Analyse embryonaler Extremitätenfehlbildungen. Charité – Universitätsmedizin Berlin, 2010

PhD theses

Johannes Grünhagen (Dr. rer. medic.): Nicht kodierende RNAs in der Knochenentwicklung. Charité – Universitätsmedizin Berlin, 2015

Julia Grohmann (Dr. rer. nat.): The role of the tumour suppressor Nfl in growth and metabolism of skeletal muscle cells. Technische Universität Berlin, 2014

Ibrahim, Daniel (Dr. rer. nat.): ChIP-seq reveals mutation-specific pathomechanisms of HOXD13 missense mutations. Humboldt-Universität zu Berlin, 2014

Saniye Sprenger (Dr. rer. nat.): The role of Pycr1 in the pathomechanism of autosomal recessive cutis laxa. Technische Universität Berlin, 2014

Laure Bosquillon de Jarcy (Dr. med.): Brachydaktylie Typ E und ZNF521. Charité – Universitätsmedizin Berlin, 2014


Hendrikje Hein (Dr. rer. nat.): Funktionelle Analysen von Transkriptionsfaktoren mit einer Rolle in der chondrogenen und osteogenen Differenzierung mittels ChIP seq. Freie Universität Berlin, 2013

Silke Lohan (Dr. rer. nat.): Analyse von genomischen Aberrationen mit hochauflösender Array-CGH bei Patienten mit Fehlbildungen der Extremitäten. Freie Universität Berlin, 2013

Sebastian Bauer (Dr. rer. nat.): Algorithms for knowledge integration in biomedical sciences. Freie Universität Berlin, 2012

Wing Lee Chan (Dr. rer. nat.): Molecular basis of Gerodermia Osteodysplastica, a premature ageing disorder. Freie Universität Berlin, 2012

Stefanie Forler (Dr. rer. nat.): Effekte von Polymorphysmen auf die geschlechtsspezifische Proteinexpression in gesunden und hypertrophierten Herzen. Freie Universität Berlin, 2012

Sebastian Köhler (Dr. rer. nat.): Phenotype informatics: Network approaches towards understanding the deseaseome. Freie Universität Berlin, 2012

Gao Guo (Dr. rer. nat.): Fibrillin-1 and elastin fragmentation in the pathogenesis of thoracic aortic aneurism in Marfan syndrome. Freie Universität Berlin, 2011

Jirko Kühnisch (Dr. rer. nat.): The ANK protein: pathologies, genetics and intracellular function. Freie Universität Berlin, 2011

Christian Rödelperger (Dr. rer. nat.): Computational characterization of
Research Group
Development & Disease

Wibke Schwarzer (Dr. rer. nat.): Phenotypic variability in monogenic disorders involving skeletal malformations. Freie Universität Berlin, 2010

Michael Töpfer (Dr. med.): Der Transkriptionsfaktor Osrl in der Extremitätenentwicklung. Charité – Universitätsmedizin Berlin, 2010

Aikaterini Dimopoulou (Dr. med.): Investigation of the genetical basis of autosomal recessive Cutis Laxa. Charité – Universitätsmedizin Berlin, 2010

Kim Ryong (Dr. rer. medic.): Assoziationsstudie zur klinischen Variabilität bei Patienten mit dem Nijmegen-Breakage Syndrom. Charité – Universitätsmedizin Berlin, 2010

Wenke Seifert (Dr. rer. nat.): Pathology of Cohen syndrome: Expression analysis and functional characterization of COH1. Freie Universität Berlin, 2010

Florian Witte (Dr. rer. nat.): Analyse der Ror2-Funktion in vivo und in vitro - Die Ror2 W749X-Maus als Modell für humane Brachydaktylie Typ B. Freie Universität Berlin, 2009

Ulrich Wilkening (Dr. rer. nat.): Funktionelle Analyse von in der Skelettentwicklung differentiell regulierten Genen. Freie Universität Berlin, 2009

Chayarop Supanchart (Dr. med. dent.): Characterization of an Osteopetrosis mouse model. Charité – Universitätsmedizin Berlin, 2009

Friederike Kremer (Dr. med.): Nonsense-mediated mRNA decay in collagen X. Charité – Universitätsmedizin Berlin, 2009

Anja Brehm (Dr. rer. nat.): Molekulare Pathologie und Embryologie von Hoxd13-assoziierten Fehlbildungen der Extremitäten. Freie Universität Berlin, 2009

Charlotte Wilhelmina Ockeloen (Dr. med.): Split hand/split foot malformation: determining the frequency of genomic aberrations with molecular-genetic methods. Charité – Universitätsmedizin Berlin, 2009

Student theses


Beuth University of Applied Sciences, Berlin, Bachelor thesis, 2014


Nina Günther: Funktionelle Charakterisierung aktiverierter RAS-Mutanten im Modellsystem Gallus gallus. Beuth University of Applied Sciences, Master thesis (Biotechnology), 2010

Susanne Mathia: Analyse des Knockdowns von Osr1 und Osr2 in Primärzellkulturen. Beuth University of Applied Sciences, Master thesis (Biotechnology), 2010


**Teaching activities**

The Institute for Medical and Human Genetics (IMHG) together with the research group at the MPI-MG run the entire teaching for human genetics at the Charité. Furthermore, we also are involved in teaching Bioinformaticians at the Freie Universität.
Otto Warburg Laboratory
Max Planck Research Group Epigenomics
Established: 12/2011

Secretary of the OWL
Cordula Mancini
Phone: +49 (0)30 8413-1691
Fax: +49 (0)30 8413-1960
Email: mancini@molgen.mpg.de

Scientists
Xian Yang (since 04/15)
Alisa Fuchs (since 02/13)
Sarah Kinkley (since 01/13)
Na Li (since 04/12)
Felipe Leal Valentim (06/14 - 12/14)
Guofeng Meng (08/12 - 02/14)

PhD students
Marcos Torroba (since 02/14)
Johannes Helmuth (since 04/12, IMPRS-CBSC)
Alessandro Mammana (since 09/11, IMPRS-CBSC)

Technician
Claudia Schade (since 06/15)
Ilona Dunkel (12/11 – 10/14)

Scientific overview

The nucleosome is the fundamental repeating unit of eukaryotic chromatin. It forms by the association of two copies each of the four core histones H2A, H2B, H3 and H4. Nucleosomes form along the complete length of eukaryotic chromosomes and thereby impact on all biological processes that require access to the DNA. Apart from the genetic information encoded in the DNA sequence, chromatin is heavily modified constituting an additional layer of epigenetic information. Chromatin modifications convey information about the past decisions taken during development and differentiation, i.e. they constitute epigenetic memory. Chromatin modifications include DNA 5-methyl-cytosine methylation, histone acetylation and methylation of Lysine residues to name a few. They are written, read and erased by chromatin-modifying enzymes, which in addition communicate with biochemical processes to facilitate and/or regulate genome function. Not all chromatin signaling pathways are involved in epigenetic memory. However, epigenetic memory requires chromatin signaling to establish and maintain epigenetic states and moreover to enable adequate responses.
The OWL group Epigenomics aims at unraveling
- patterns of chromatin modifications and their localization with respect to functional genomic elements;
- chromatin-signaling pathways that are associated with these patterns;
- dynamics of chromatin modifications;
- mechanisms that establish and/or maintain epigenetic silencing of gene expression;
- the role of “epi-mutations” in disease and trans-generational epigenetics

by combining computational/bioinformatic analyses with the development and application of existing and novel experimental approaches for ChIP-seq, ATAC-seq, bisulfite-seq etc.

Scientific achievements and findings

Chromatin segmentation and localization of nucleosomes using histone modification ChIP-seq data

The last years have seen an explosion of ChIP-seq data for many different cell lines/types and many histone modifications. The core histones H2A, H2B, H3 and H4 are an integral part of the nucleosome indicating that ChIP-seq reads against modifications of these proteins should harbor information about the localization of nucleosomes. Following this idea, we developed NucHunter that uses the data from ChIP-seq experiments directed against many histone modifications to infer positioned nucleosomes. NucHunter annotates each of these nucleosomes with the intensities of the histone modifications. These annotations can be used to infer nucleosomal states with distinct correlations to underlying genomic features and chromatin-related processes, such as transcriptional start sites, enhancers, elongation by RNA polymerase II and chromatin-mediated repression.

Figure 1: EpiCSeg explains a larger proportion of the epigenome than ChromHMM.
Next, we developed EpiCseg that combines several histone modification maps for the segmentation and characterization of cell-type-specific epigenomic landscapes. By using an accurate probabilistic model for the read counts (the negative multinomial distribution), our method provides a useful annotation for a considerably larger portion of the genome (Figure 1), shows a stronger association with genetic elements such as promoters and enhancers as well as gene expression, and yields more consistent predictions across replicate experiments when compared to existing methods.

**Chromatin-signaling pathway inference and function**

Chromatin segmentation reveals recurring patterns of histone modifications that are associated with regulatory elements such as promoters and enhancers. These recurring patterns may be due to distinct nucleosomal modification states, with a certain combination of histone modifications. Alternatively, these patterns may result from chromatin signaling, where nucleosomal modification states are generated and removed in a temporal fashion – i.e. the observed pattern is in fact the time average over the dynamic chromatin signaling activity, where certain combinations of histone modifications may not coexist on the same nucleosome at the same time.

Figure 2: Chromatin-signalling network. Graphical representation of the interactions between CMs (circles) and HMs (squares). Red lines indicate positive and blue lines negative interactions. The continuous lines indicate interactions with literature support; the dashed lines indicate interactions without supporting evidence. The stars indicate the two interactions confirmed in this study.
In a first study, we used existing biochemical evidence to construct a signaling network that relates histone modifications to the progression of RNA polymerase II (Pol II) through the transcription cycle taking place at the promoter. We test predictions from this network using whole genome in vivo data. One emerging property of this network is positive feedback between pre-initiation complex (PIC) formation and transcription initiation, which is required to define a competent promoter that allows Pol II to initiate transcription. This feedback loop is in partial competition with another pathway, which enables Pol II to proceed from initiation to elongation. The outcome of this competition determines whether Pol II can proceed to synthesize a full-length transcript or not.

In a second study, we applied computational methods to recover interactions between chromatin modifiers and histone modifications from genome-wide ChIP-Seq data. These interactions provide a high-confidence backbone of the chromatin-signaling network (Figure 2). Many recovered interactions have literature support; others provide hypotheses about yet unknown interactions. We experimentally verified two of these predicted interactions, leading to a link between H4K20me1 and members of the Polycomb Repressive Complexes 1 and 2. Our results suggest that our computationally derived interactions are likely to lead to novel biological insights required to establish the connectivity of the chromatin signaling network involved in transcription and its regulation.

A novel sequential ChIP-seq approach identifies widespread bivalency in differentiated primary human central memory T-cells

Identification of nucleosomes modified by two distinct histone modifications remains a technical challenge. To tackle this challenge we have developed a novel re-ChIP-seq approach that enriches for nucleosomes carrying two distinct histone modifications in an unbiased and genome-wide manner. We illustrate the utility of our approach by identifying nucleosomes bivalently modified by H3K4me3 and H3K27me3 in primary human central memory T-cells (TCMs). We unravel widespread true bivalency of many CpG-island promoters, whose DNA is unmethylated (Figure 3, red box). Genes driven by these promoters function during development and differentiation and...
remain repressed during the activation of TCMs, but are involved in lineage decisions. Thus, bivalency in TCMs is a repression mechanism for CpG island promoters, where H3K4me3 represses DNA methylation and H3K27me3 represses transcription.

The tale of two tails

Nucleosomes can be in three states with respect to a specific histone modification: (i) un-, (ii) hemi- and (iii) full modified. We demonstrate that discrimination of three nucleosomal states for H3K4me3 is possible, enabling the characterization of state-specific downstream consequences. H3K4me3 ChIP-seq peaks follow a bi-modal distribution, where the lower mode is the square root of the higher mode predicted from a quantitative model involving hemi- and full H3K4me3 methylated nucleosomes. Thus, hemi H3K4me3 methylated nucleosomes are characterized by a ChIP-enrichment, which is the square root of the enrichment of full H3K4me3 modified nucleosomes, a relationship, which we refer to as “square root rule”. H3K4me3 present on at least one H3-tail represses DNA methylation, while transcription initiation coincides with full H3K4me3 methylated nucleosomes. Bivalent nucleosomes are likely to be hemi H3K4me3 methylated in line with the idea that bivalency requires an asymmetrically modified nucleosome, with H3K4me3 on one and H3K27me3 on the other H3-tail. All three states elicit distinct downstream responses showing for the first time that epigenetic effectors distinguish between unmodified, hemi- and full H3K4me3 methylated nucleosomes, revealing an additional layer of complexity in epigenetic regulation.

Planned developments

Dynamics of histone modifications during the maternal-to-zygotic transition in Drosophila melanogaster

Cell cycle 14 during Drosophila melanogaster embryogenesis provides a unique time window to study temporal aspects of chromatin signaling. During the 60 minutes of cell cycle 14, the cells replicate their DNA, maternal transcripts are degraded and the embryo starts transcription from its own genome. Morphologically, the embryo transits from the syncytial to the cellular blastoderm stage, i.e. the progression of cellularization can be used as a proxy for time.

We want to use this highly stereotypic stage of development to unravel dynamics of histone modifications in single embryos and, using morphological characteristics, to determine a temporal ordering. To this end, we want to employ an experimental strategy that involves indexing single embryo chromatin with DNA barcodes first, pool chromatin from many
embryos to perform ChIP and sequence the resulting ChIP material, where the barcodes can be used to distinguish between the embryos. In addition, this strategy alleviates two important limitations of ChIP: (i) by pooling chromatin from many embryos more material can be used for ChIP and (ii) technical variation during ChIP gets minimized such that the resulting ChIP-seq data is much more comparable than in independent ChIP-seq experiments. This approach allows for a “temporal resolution” that is limited only by the number of embryos, resulting in both high spatial and temporal resolution. Moreover, if successful, such an approach together with strategies to enrich preselected genomic regions may pave the road for the establishment of a high throughput ChIP-assay to interrogate the presence of a histone modification in many samples at a small number of preselected loci required for clinical diagnosis.

**Maintenance of epigenetic states after DNA replication**

Epigenetic states, in particular those defined by histone modifications, face a challenge during DNA replication, where parental nucleosomes are as a whole randomly distributed to the two daughter strands and the “holes” are filled with new nucleosomes. Recent studies have shown that this mode of histone distribution after DNA replication is dominant for nucleosomes containing the canonical, replicative histones H3.1/H3.2. However, a substantial fraction of H3.3-containing nucleosomes is split during replication, such that each nucleosome obtains one old and one new copy of H3.3. H3.3-containing nucleosomes could therefore be the basis for a semi-conservative mechanism of histone modification inheritance. We have developed bioinformatic and experimental tools to test this hypothesis: (i) the square root rule is the basis to test whether certain histone modifications are semi-conservatively inherited. In case of a semi-conservative mode, the ChIP-seq signal after replication should be the square root of the one before replication. In case nucleosomes are as a whole randomly placed, we expect that the signal after replication should be half of the signal before replication. Thus, based on the relationship between the ChIP-seq signal before and after replication it is possible to distinguish between these scenarios. (ii) Using our sequential ChIP-approach, we could directly test for nucleosomes containing an old H3.3 (tagged say with a FLAG tag) and a new one (tagged say with a HIS tag). Combining both approaches will yield insight into the mode of inheritance of histone modifications.
The role of “epi-mutations” in disease

Recent years have seen a wealth of genetic data showing the causal involvement of genetic mutations in disease. We reasoned that in addition to genetic variability, also epigenetic variability is involved in disease. Epigenetic variability can be acquired during life and may depend on the life-style/environment, providing a mechanistic understanding of certain risk factors associated with diseases. The OWL group Epigenomics is, as part of the German Epigenome Program (Deutsches Epigenom Program, DEEP), involved in the generation of histone modification maps for cells involved in chronic inflammation, such as rheumatic arthritis and Crohn’s disease. Moreover, we are also participating in the analysis of all DEEP data that includes cells involved in metabolic diseases, such as liver steatosis and adipositas. We will develop bioinformatic tools and new experimental assays to unravel “epi-mutations” associated with the disease state. In particular, we are interested in changes in the epigenome that may phenocopy deletions and duplications of genetic regions. For example, the expansion of a heterochromatic domain into an euchromatic one may be functionally equivalent to a deletion. By uncovering such events, we hope to unravel the chromatin signaling pathways that cause these changes, providing in the long run targets for the treatment of these diseases.

Cooperation within the institute

Within the institute, the OWL group Epigenomics cooperates with the following people and their groups: Peter Arndt, Hans Lehrach, Annalisa Marsico, Sebastiaan Meijsing, Bernd Timmermann, and Martin Vingron.

Cooperation outside the institute

Outside the MPIMG, we cooperate with the following labs:

- Ann Ehrenhofer-Murray, Humbold University, Berlin, Germany
- Herbert Jäckle, Max Planck Institute for Biophysical Chemistry, Götttingen, Germany
- Thomas Manke & Thomas Jenuwein, Max Planck Institute of Immunobiology and Epigenetics, Freiburg, Germany
- Asifa Akhtar, Max Planck Institute of Immunobiology and Epigenetics, Freiburg, Germany
- Alf Herzig, Max Planck Institute for Infection Biology, Berlin, Germany
- The DEEP consortium (see http://www.deutsches-epigenom-programm.de)
General information

Complete list of publications (2011-2015)

Group members are underlined.

2015


2014


2013


2012


Selected invited talks
(Ho-Ryun Chung)

Chromatin segmentation with a joint model for reads explains a larger portion of the epigenome. Regulatory Genomics Special Interest Group (RegGenSIG) at ISMB 2015, Dublin, Ireland, 07/2015

The tale of two tails, Humboldt-Universität zu Berlin, Germany, 05/2015

The tale of two tails. Center of Advanced European Studies and Research (caesar), Bonn, Germany, 03/2015

Mapping chromatin accessibility. EpiGeneSys-IMB Workshop: Epigenomics as a Discovery Tool in Current Biology, Mainz, Germany, 09/2013

Nucleosome positioning and histone octamer sequence preferences. Nucleosome positioning, chromatin structure and evolution, Haifa, Israel, 05/2012

Student thesis


Teaching activities

Lecture on Probability Theory, Freie Universität Berlin, winter term 2013/2014

Lecture on Functional Genomics, Freie Universität Berlin, winter term 2011/2012
Otto Warburg Laboratory
Research Group RNA Bioinformatics

Established: 06/2014

Head
Prof. Dr. Annalisa Marsico
(since 06/14)
Phone: +49 (30) 8413-1843
Fax: +49 (30) 8413-1960
Email: marsico@molgen.mpg.de

Scientist
Bian Caffrey* (since 06/14)

PhD student
Sabrina Krakau (since 10/14, IMPRS)

Master students
Ria Peschutter (since 04/15)
David Heller (since 02/15)
Lisa Barros de Andrade e Sousa
(since 06/14)

Student assistant
Stefan Budach (since 11/14)

Introduction
The RNA Bioinformatics group at the Max Planck Institute for Molecular Genetics (MPIMG) has been established in the context of a joint Dahlem International Research Network, which aims at strengthening scientific cooperation between the MPIMG and the Freie Universität (FU) Berlin.

Annalisa Marsico has taken up a position as Assistant Professor for High Throughput Genomics at the Institute of Bioinformatics, FU Berlin, in

* externally funded
June 2014. The professorship was jointly established by the FU Berlin, Faculty of Mathematics and Computer Science, and the MPIMG. Annalisa Marsico teaches in the Bioinformatics course of studies offered by the Faculty of Mathematics and Computer Science, but her group is physically located at the MPIMG.

Scientific overview

Background

In the cell, genomic DNA is transcribed into various types of RNAs, but not all RNAs are translated into proteins. Over the past few years it has been observed, thanks to high-throughput sequencing technologies, that a big portion of the human genome is transcribed in a tissue- and time-specific manner. Most of the detected transcripts in mammals and other complex organisms are non-coding RNAs (ncRNAs), RNAs that do not encode for proteins and whose functional consequences are still controversial and not fully understood.

Non-coding RNAs in higher eukaryotes can be classified based on their size, ranging from small RNAs of length between 20 and 90 nt (e.g. microRNAs, piwi-RNAs, snoRNAs, tRNAs) to long non-coding transcripts of length higher than 200 nt (e.g. long non-coding RNAs). Different classes of RNAs have been shown to participate in different cellular processes, for example protein synthesis (rRNA, tRNA) and RNA maturation (snRNA, snoRNA). In particular, many newly discovered ncRNAs (such as microRNAs and long non-coding RNAs) have been suggested to constitute a hidden layer of gene regulation that is necessary to establish complex regulatory programs in higher organisms.

In our group, we are especially interested in the regulation and function of those ncRNAs (e.g. microRNAs and long non-coding RNAs), which act as regulators of gene expression, and their coordinated action with transcription factors, RNA-binding proteins and epigenetic mechanisms. High-throughput experiments provide a valuable source of different genomic data, such as RNA expression profiles, chromatin modifications, RNA-binding protein profiles and genetic variations that, if analyzed with the proper computational tools, can boost the effort towards the annotation and functional characterization of non-coding transcripts genome-wide. In our group we develop statistical models and predictive approaches to integrate and interpret different genomic data in order to address the questions

- how microRNAs are regulated at transcriptional and post-transcriptional level,
- how we can predict IncRNA function by integrating different genomic data.
In my newly formed group I would like to address these questions from a computational point of view and experimentally through collaborations with experimental groups inside and outside of the MPIMG.

Transcriptional regulation of microRNA genes

The non-coding RNA revolution gained huge momentum in the early 2000s with the discovery of microRNAs and their ability to modulate gene expression in eukaryotic organisms by binding to the 3'-UTRs or coding regions of target genes. MicroRNAs are associated with many biological processes, including cancer and immunological response. Given the contribution of microRNAs to gene expression, gene regulatory networks have been expanded to include both transcription factors (TFs) and microRNAs. However, the knowledge of which TFs regulate certain microRNAs in a certain tissue or cellular state is the missing link in gene regulatory networks. Reliable TF binding site predictions for microRNA genes have been hindered by the scarcity of annotated microRNA promoters in databases and by the difficulty of experimentally detect the transcription start sites (TSSs) of transient microRNA primary transcripts. To fill the gap, in my previous work I have developed PROmiRNA, a semi-supervised machine learning method to predict microRNA promoters by integrating publicly available high-throughput sequencing data (namely deepCAGE data from
the FANTOM4 Consortium) and specific sequence features (Figure 1). The application of PROmiRNA to the human genome has enabled us to discover previously uncharacterized tissue-specific intronic promoters for more than half of the annotated intragenic microRNAs. In addition, we were able to describe for the first time the properties of different microRNA promoter classes and show that (i) intronic microRNA promoters are TATA-box-like rather than CpG-like promoters; (ii) different sets of TFs bind intronic and intergenic microRNA promoters; (iii) tissue-specific intronic promoters decouple the expression of intragenic microRNAs from the expression of their hosting transcripts.

PROmiRNA has been applied in a (still-ongoing) project to study the activation of microRNAs in inflammatory processes, such as L. pneumophila bacterial infection (collaboration with Bernd Schmeck at the University of Marburg).

**Post-transcriptional processing of microRNAs**

Global mature microRNA expression is not only controlled at transcriptional level, but several post-transcriptional steps influence the final microRNA expression level. In detail, microRNA initially generated as long primary transcripts are processed in the nucleus by the Microprocessor complex (Drosha/DGCR8) to produce stem-loop structured precursors, which are then further processed in the cytoplasm by Dicer. Altered microRNA processing correlates with several cancer or other disease phenotypes. In this project, in collaboration with the experimental group of Ulf Ørom (Long non-coding RNA group) we have investigated the question, how Microprocessor is able to distinguish microRNA hairpins from random hairpin structures along the genome and efficiently process them. To answer this question, we have performed high-throughput RNA sequencing experiments of nascent transcripts associated to the chromatin fraction in different cell lines and defined a quantitative measure of processing efficiency called Microprocessing Index (MPI) by statistical modeling of the read count density around microRNA hairpins. Efficient Drosa cleavage is reflected in significant drops in the read coverage in the microRNA hairpin region (Figure 2).

Our analysis has highlighted three major findings: (i) microRNA processing happens co-transcriptionally, while the microRNA primary transcript is still associated to the chromatin; (ii) microRNA processing patterns are, in most of the cases, conserved among cell lines, suggesting an important role of invariant features, such as the primary sequence, in determining efficient processing; (iii) in silico statistical modeling based on nucleotide k-mer counts identifies three sequence motifs, namely the GNNU motif at the 5’-end of microRNA hairpins, the CNNC motif downstream of the 3’ hairpin arm, and the GC dinucleotide at the base of the stem loop, which are
significantly associated with efficient microRNA processing. The CNNC and GC motifs were experimentally validated by means of mutagenesis experiments in the lab of Ulf Ørom.

Figure 2: Genome Browser view of the genomic regions around microRNAs hsa-miR-100 (A) and hsa-miR-573 (B). Normalized read coverage from small RNA-Seq experiments in HeLa cells is shown around the microRNA loci. The significant drop in read coverage at the miR-100 precursor indicates that this microRNA is efficiently processed (A), while miR-573 is not (B).

Transcriptional regulatory functions of long non-coding RNAs (lncRNAs)

The biological functions and mechanisms of lncRNAs are numerous and diverse; new lncRNA are discovered every month and new categories and paradigms are proposed annually. In addition, lncRNA can carry out different functions, depending on the tissue and developmental stage they are active in.

In this project we focus on long non-coding intergenic RNAs (lincRNAs), which have been proven to have regulatory roles in gene transcriptional control. Many of these lincRNAs have been shown, based on their chromatin marks, to emanate from enhancer-like regions and to enhance gene expression in cis (near their site of synthesis) or in trans across different
chromosomes. Enhancer-like functions of lincRNAs are often carried out via chromatin looping mechanisms, and it has been shown that in many cases enhancer-associated lincRNAs can modulate enhancer activity by altering chromatin accessibility and/or structure. In line with previous studies which try to find significant associations between lincRNAs and nearby genes based on co-expression analysis, in this (still-ongoing) project we integrate, in a statistical framework, chromatin looping data (ENCODE ChIA-PET data) and correlation analysis of lincRNAs expression and protein-coding gene chromatin accessibility in order to pinpoint significant associations. Our preliminary results indicate that the association between lincRNA expression and chromatin accessibility of nearby genes is stronger for those pairs which are physically linked at chromatin level.

Outlook

Regulation of microRNA genes

We are currently working on a new version of the PROmiRNA software, able to scale with the increased sequencing depth of the FANTOM5 deepCAGE libraries. The goal is to be able to screen for microRNA promoters genome-wide and across thousands of different tissues and cell types, and provide a basis for individual studies on miRNA regulation.

The knowledge of the location of microRNA promoters genome-wide, together with the exact genomic positions of microRNA hairpins, will enable us to study and model additional aspects of microRNA regulation. For example, it will allow us to study the impact of both, chromatin modifications and genetic variants along the full primary microRNA gene on microRNA expression in a specific tissue or condition.

In a first (ongoing) project in the lab we are modeling microRNA expression in Hela and Imr90 cells from publicly available ENCODE sequencing data, chromatin changes and PROmiRNA predicted microRNA promoters. Our first results indicate a crucial role of DNA methylation in both promoters, as well as microRNA hairpin flanking genomic regions in regulating miRNA expression.

In cooperation with the group of Ho-Ryun Chung (Epigenomics) we are going to apply our model to different cell types, mainly primary cells such as hepatocytes, adipocytes, macrophages and different T-cell populations, for which the group has already generated RNA-sequencing data and CHIP-seq data for several histone marks. The goal is to define a global histone-code for microRNA expression and processing, as well as highlight cell-type specific differences, which produce distinct microRNA expression profiles.
In a second (ongoing) project, in collaboration with Dr. Matthias Heinig (Helmholtz Zentrum München), we are investigating the effect of genetic variation on microRNA transcription. The goal of this project is to build an *in silico* model to classify single nucleotide polymorphisms (SNPs) into silent versus eQTL (expression Quantitative Trait Loci able to alter microRNA expression) based on the location of such SNPs with respect to microRNA promoters and other regulatory elements.

**Interplay between lincRNAs and RNA-binding proteins**

Besides predicting significant associations and target interactions between lincRNAs and genes, we would like to move a step further in the direction of the mechanisms, by which lincRNAs target specific genomic regions. It has been shown that several lincRNAs carry on their regulatory function by associating with chromatin-modifying complexes and transcriptional regulatory proteins in the nucleus.

In recent years, the development of crosslinking immunoprecipitation (CLIP)-Seq technologies has enabled the investigation of the interactions between RNA-binding proteins (RBPs) and their target RNAs, including coding and non-coding RNAs, in a quantitative and high-resolution manner. Despite the great potential of such technologies, we have realized, by investigating publicly available datasets that the current methods for identification of RBP sites from CLIP-Seq data do not take into account all possible biases, which might lead to the identification of many false positives (e.g. high PCR duplicate rates, sequencing and PCR errors, lack of appropriate controls). In order to infer robust conclusions from the data and correctly interpret the results, there is a huge need for *in silico* methods for reliable RBP site calling.

In our group we would like to develop robust *in silico* methods for CLIP-Seq data analysis and, in collaboration with the Ørom group, we will apply this methodology to investigate the binding and function of new candidate RBPs through iCLIP experiments.

**Comprehensive functional classification of lincRNAs**

Most lincRNAs show little evidence for evolutionary conservation and therefore their function cannot be inferred from sequence alone based on homology, as it is often done for protein-coding genes. The group of Knut Reinert (FU Berlin & MPIMG) has developed an algorithm for accurate multiple alignment of RNA structures. In collaboration with the Reinert group and through a joint PhD student, we would like to extend this algorithm to explore structural motifs in lincRNAs. As the structure of a RNA
is related to its function and structures evolve slower than sequences, by grouping the lincRNAs into functional groups (e.g. lincRNAs expressed in the same tissue in human and mouse or lincRNAs binding the same RBP) we can apply multiple structural alignment methods to identify conserved structural motifs associated to lincRNA function. This would represent the first step towards a comprehensive functional classification of known and novel lincRNAs.

**Cooperation within the Institute**

Within the institute, the RNA Bioinformatics group closely cooperates with the following people and their groups: Martin Vingron, Ulf Ørom, Ho-Ryun Chung, Knut Reinert.

**Cooperation outside the Institute**

Outside the MPIMG, we cooperate with the following labs:

- Benedikt Beckmann, Humboldt University, IRI Life Sciences, Berlin, Germany
- Heike Siebert, Department of Mathematics and Computer Science, Freie Universität Berlin, Germany
- Knut Reinert, Department of Mathematics and Computer Science, Freie Universität Berlin, Germany
- Bernd Schmeck, Institute for Lung Research, Universität Marburg, Germany
- Matthias Heinig, Helmholtz Zentrum München, Germany
- Hughes Richard, Pierre and Marie Curie University, Paris, France
General information

Complete list of publications (2013-2015)
Group members are underlined.

2015

2014

2013


Selected invited talks
(Annalisa Marsico)


Statistical modeling of NGS data to investigate the biogenesis and function of non-coding RNAs. Workshop ‘Bringing Math to Life’, Naples, Italy, 2014


Teaching activities

Algorithmische Bioinformatik (lectures and tutorials, shared with Prof. Martin Vingron), Bioinformatics Bachelor studies, Freie Universität Berlin, winter term 2014/15

Seminar course RNA Bioinformatics, Bioinformatics master studies, Freie Universität Berlin, winter term 2014/15

Applied Machine Learning (lectures and tutorial, shared with Bernhard Renard, from the Robert Koch Institute Berlin), Bioinformatics master studies, Freie Universität Berlin, summer term 2015

Seminar course Machine Learning, Bioinformatics master studies, Freie Universität Berlin, summer term 2015

Single lecture and tutorial Promoter Evolution, Summer School ‘Programming for Evolutionary Biology’, University of Leipzig, 03/2012
Otto Warburg Laboratory

Sofja Kovalevskaja Research Group
Long non-coding RNA

Established: 01/2012

Head
Dr. Ulf Andersson Ørom*
Phone: +49 (30) 8413-1664
Fax: +49 (30) 8413-1960
Email: oerom@molgen.mpg.de

Secretary of the OWL
Cordula Mancini
Phone: +49 (0)30 8413-1691
Fax: +49 (0)30 8413-1960
Email: mancini@molgen.mpg.de

Scientists
Julia Liz (since 01/15)*
Evgenia Ntini, PhD (since 11/13)*
Thomas Conrad (since 09/12)*

PhD students
Verônica Rodrigues de Melo Costa (since 09/15)*
Annita Louloupi (since 11/14)*
Dubravka Vucicevic (04/12-11/15)
Anne-Susan Musahl (04/12-04/15)

Master students
Alexander Kiefer (03/15-08/15)
Maja Gehre (08/13-04/14)
Antonia Hilbig (06/13-12/13)
Masha Kenda (01/13-06/13)*

Student assistant
Leonie Chiara Martens (03/13-07/13)

Technician
Sabrina Schlüngel (09/13-02/14)

* externally funded
Scientific overview

We are interested in the molecular mechanisms of how enhancer-like long ncRNA can mediate gene activation. While this functional group of macromolecules has received a lot of attention, their mechanism and how widespread their functions are is not known. To identify and characterize enhancer-like long ncRNAs, my lab is looking at enhancer marks associated with annotated long ncRNAs in the human genome, large-scale transcriptomics assays and RNA structure.

Scientific methods and achievements/findings

To identify tissue-specific enhancers across 11 cell lines, we modified the PreSTIGE enhancer prediction approach (Figure 1A). PreSTIGE is a method that predicts enhancers by first finding PCGs with tissue-specific increased expression, and, based on the assumption that these are targets of tissue-specific enhancers, interrogates H3K4me1 domains in the vicinity. The predicted enhancers are therefore based on the presence of H3K4me1 domains in proximity to genes with tissue-specific expression. Predicted enhancers based on H3K4me1 can therefore be linked with the tissue-specific elevated expression of the putative target gene. While the results presented suggest that long ncRNA transcription is linked to tissue-specific
enhancers (Figure 1B), they do not tell us how they are mechanistically involved. The long ncRNAs expressed at enhancers have been suggested to be directly involved in mediating the enhancer activity on the regulated genes (Lai et al., Nature 2013), while other studies have predominantly found evidence for a correlation in expression. How many of these enhancer-transcribed long ncRNAs are mediating the enhancer activity and how many are expressed as a consequence of enhancer activation should be addressed in future experiments. We find, however, that tissue-specific enhancer-associated expression of long ncRNAs is characteristic for a subset of the predicted enhancers, implying a functional relationship (Vucicevic et al., Cell Cycle 2015).

To extend these studies and obtain a more general understanding of how long ncRNAs associate with various aspects of enhancers such as DNA hypomethylation, chromatin-interaction to protein-coding gene promoters, and histone modifications, we have incorporated these data in MCF-7 cells. We have been able to expand the catalogue of enhancer-like long ncRNAs from a few validated examples to several hundred strong candidates, and categorize enhancers dependent on the histone modifications, DNA methylation, and expression of long ncRNA using k-means clustering.

We have used these data to identify transcription factors regulating the expression of enhancer activity, through modulating long ncRNA expression. We have identified FOXC1 in breast cancer cells as a regulator of expression of a subset of enhancer-like long ncRNAs.

While the dynamics of chromatin conformation has been shown not to be the determining factor of enhancer activity, the dynamics of DNA methylation could be a contributing factor. We have assessed the dynamics of DNA methylation following estradiol treatment, which induces transcriptional changes of many enhancer-like long ncRNAs. While we do observe dynamic DNA methylation, it does not happen at the majority of enhancer-like long ncRNA promoters. We therefore suggest that, while both chromatin conformation and DNA methylation status are important factors for identifying enhancers and enhancer-like long ncRNAs, the long ncRNA expression is the principal dynamic component of the associated enhancer function.

My previous research has identified the Mediator complex as an interacting partner of candidate enhancer-like long ncRNA (Lai et al., Nature 2013), with this interaction being important for enhancer function. This study focused on a few candidate enhancer-like long ncRNAs. My lab is currently expanding these studies to transcriptome-wide interactions using iCLIP and RNA sequencing and will extend these experiments in mouse models of neurodegenerative disease in collaboration with the lab of Heinrich Schrewe at the Max Planck Institute for Molecular Genetics, to obtain a more complete understanding of the importance of this interaction in vivo.
Large-scale transcriptomics assays for RNA processing

We have developed a method for transcriptome-wide pri-miRNA processing based on RNA-sequencing. We show that the endogenous Microprocessor activity towards individual pri-miRNAs can be determined using RNA sequencing (Conrad et al, Cell Reports 2014). We identify the Microprocessor cleavage signature, a pronounced dip in the coverage of reads obtained by RNA sequencing (Figure 2), and define the MicroProcessing Index (MPI) based on relative reads surrounding and spanning the miRNA processing site, respectively, as a measure for processing efficiency. We provide experimental evidence that processing efficiency is one of the major determinants for expression of individual miRNAs at the mature level. We show that the processing between cell lines is very stable, suggesting a major influence of primary sequence on the pri-miRNA transcript, and a minor contribution from cell type specific protein interactions modulating the activity of Microprocessor. Large-scale RNA sequencing based assays for transcription and processing are emerging as very important tools to understand these processes better. (Conrad & Ørom, Oncotarget 2015). My lab is currently establishing nascent RNA processing assays based on BrU incorporation and RNA sequencing to study processing differences transcriptome-wide between different classes of RNAs as well as upon modulation of regulatory factors such as m6A RNA methylation.
Transcriptome-wide identification of chromatin-associated RNA-binding proteins

To identify chromatin-associated RNA-interacting proteins we have optimized and extended the interactome capture technique published by the Landthaler and Hentze labs to focus on chromatin-associated RNA binding proteins. Using this approach we have identified 64 nuclear proteins not previously described to be RNA-binding proteins (Figure 3A). We have used iCLIP to identify the RNAs interacting with the histone lysine demethylase KDM5A involved in chromatin modification and transcriptional regulation to get an understanding of the role of RNA interaction with KDM5A. Unexpectedly, we observe a translation effect on specific transcripts encoding secretion factors interacting with KDM5A, with an enrichment of the KDM5A protein binding to the 3' UTR of regulated transcripts (Figure 3B). As a result we see decreased secretion of e.g. the angiogenesis factor VEGFA. RNA-dependent co-IPs of KDM5A suggest an ER-associated histone demethylase-independent function and an association with the translation apparatus.

Figure 3: A) Overview of the functional groups of all proteins identified using chromatin and nuclear interactome capture. The majority (315) proteins are known to be RNA binding. 134 have known roles at chromatin or in transcription with most of these (124) being known RNA binders. 66 proteins have not been assigned to RNA binding of which only two have been previously identified with similar methods. B) The upper panel shows how the KDM5A binding at 3' UTRs are distributed across a relative 3' UTR and the lower panel shows the same results for an IgG control.
Cooperation within the Institute

Within the institute, the group is collaborating with
- Annalisa Marsico, RNA Bioinformatics Group
- David Meierhofer, Mass Spectrometry facility
- Sascha Sauer, Nutrigenomics & Gene Regulation Group
- Bernd Timmermann, Sequencing facility
- Heiner Schreve, Department of Developmental Genetics

Planned developments

The mechanisms with which long ncRNAs are involved in enhancer function through sequence- and structure-specific contributions remains a challenging open question for the field. Recent developments in CRISPR/Cas9 technologies, both to target DNA and RNA, seems promising for addressing this question in vivo. The extension of my research, as described below, will involve systematic in vivo perturbations of DNA and RNA regions to address the specificity and functionality of long ncRNAs in enhancer function. To establish a better understanding of how RNA structure and processing determine the functions of long ncRNA my lab is focused on methodologies to unravel these aspects to get a broad and mechanistically informative insight into their molecular function.

Understanding long ncRNAs in enhancer function.

Both the upstream and downstream regulators and effectors of RNA-dependent enhancer functions would reveal much about the mechanism of gene regulation. My lab has successfully identified the FOXC1 transcription factor as a regulator of long ncRNA expression and enhancer function in MCF-7 breast cancer cells. Based on the large set of enhancer-like long ncRNAs we have identified, we will study the upstream regulatory network using transcription factor prediction, ChIP-seq, and CRISPR-based approaches for experimental identification and validation of the regulatory factors. This might identify transcription factors with an unappreciated function in enhancer activity, mediated through the expression of enhancer-like long ncRNAs. We will use a system with MCF-7 derivative cells with different properties in cancer-related processes such as cell cycle and invasion, to incorporate a disease-relevant interpretation of the functional output of differential regulation of expression of long ncRNAs from enhancers. This system will also be incorporated into the research plans described hereafter to have a functional angle in a more biological context on the mechanistic studies planned. An outstanding question is how to identify the most important aspects of long ncRNA in enhancer function.
maintaining expression in \textit{cis} with regard to their target genes and incorporating distance between enhancer and promoter, a question that has so far only been addressed using reporter assays. With a repertoire of functional enhancer-like long ncRNAs we will address this at a general level using the novel possibilities provided by CRISPR-mediated manipulation of long ncRNA \textit{loci}, disrupting their expression or changing their sequence.

**RNA sequencing-based assays for transcription and RNA processing**

We have established a method to study miRNA processing and are in the process of applying BrU-seq to study transcriptional and processing aspects of RNA, with a focus on long ncRNAs, dependent on RNA modifications. Also methods as NET-seq, to replace the more challenging GRO-seq, is a method we are currently incorporating to study the transcriptional consequences of long ncRNA manipulation and their upstream regulation.

**Understanding RNA processing, structure and function**

We aim to use the different approaches established in the lab to understand structure, processing and functionality of RNA better. As sequence conservation for long ncRNAs is limited, it is believed that structural aspects are very important for their function. A timely question in my laboratory addressing the regulation and functionality of RNA is how the RNA modification m6A can regulate RNA structure leading to altered RNA function and processing. We are using the optimized interactome capture to identify proteins binding to m6A \textit{in vivo} in a cellular department (by cellular fractionation) and time-resolved (by BrU pulse-chase of nascent RNA) manner. The specific effects of interacting proteins on processing and export of mRNAs and long ncRNAs directly binding the proteins as determined by iCLIP will be addressed. Finally, we are setting up established methods in the lab to study transcriptome-wide structure of RNAs in different conditions and dependent on m6A RNA modifications, to integrate the structural aspect with protein binding, processing and function of RNA.
General information

Complete list of publications (2012-2015)

Group members are underlined.

2015


2014


2013


2012


Awards

Ulf Ørom: *Sofja Kovalevskaja Award 2012*, Alexander von Humboldt Foundation and German Ministry of Research and Education, 2012
Selected invited talks
(Ulf Ørom)

Chromatin-associated functions of (nc)RNA. IRI symposium, Berlin, Germany, 03/2015

In vivo determination of primary miRNA processing. DPG annual meeting, Magdeburg, Germany, 03/2015

Long noncoding RNA and enhancers. CECAD Symposium on non-coding RNA, Cologne, Germany, 04/2014

Long noncoding RNA and enhancers. 25th Annual Meeting of the German Society of Human Genetics, Essen, Germany, 03/2014

Bidirectional expression of long noncoding RNAs and protein-coding genes. EMBO Meeting on Nuclear Structure and Dynamics, L’isle sur la Sorgue, France, 10/2013 (oral presentation selected from submitted abstracts)

Enhancer activity and chromatin conformation regulated by long non-coding RNAs. Conference on Gene Regulation and Information Theory, Martin Luther University, Halle, Germany, 04/2013

Long noncoding RNAs in chromatin dynamics. Gordon Research Conference on Chromatin Structure and Function, Lucca, Italy, 05/2012

PhD theses


Dubravka Vucicevic: Diverse regulatory functions of IncRNAs. Freie Universität Berlin, 2015

Student theses


Scientific overview

Physiological processes are controlled by complex molecular mechanisms. These interconnected mechanisms are required to respond to daily changing environmental factors such as nutrition. In order to prevent health decline and thereby prolong the quality of life, we aim to identify causal connections between diet and disease.

The group started with financing from the German Ministry of Education and Research (BMBF) for a Junior Research Group on nutrigenomics. It explores health implications of the interaction between nutrition and genomics. The regulation of genes plays an important role in various molecular processes of metabolic disorders such as insulin resistance or atherosclerosis.

One emphasis of our research lies in analyzing genome-wide the modulation of gene and protein expression in cellular processes that are relevant during adipocyte or macrophage cell differentiation or may lead to phys-

* externally funded
iological deregulation in metabolic target tissues. These gene regulatory processes can be significantly influenced through the interaction between genes and naturally occurring compounds. Consequently, as the second emphasis of our research group, we study the capability and mechanisms of natural products to interact with genes and gene products. In order to identify active natural products, we screened and systematically characterized natural substances derived from small molecule libraries that featured large structural variability.

**Scientific achievements**

Given worldwide increases in the incidence of metabolic diseases such as obesity, type 2 diabetes or atherosclerosis, alternative approaches for preventing and treating these disorders are required. We focus our research on the analysis of metabolic deregulation on the cellular and the physiological level. We further aim to beneficially influence physiological deregulation at an early stage by intervening with specific natural products derived from dietary resources. Such basic and applied nutrigenomics research heavily relies on systems biology approaches (Figure 1).

![Figure 1: Nutrigenomics – from molecular aspects of nutrition to prevention of (age-related) diseases](image-url)
Amorfrutins and LXR ligands

In collaboration with partners from Cornell University and the company Analyticon Discovery, we discovered a family of natural products that bind to and activate specifically the ligand-dependent transcription factor PPARγ (Figure 2). These compounds, the amorfrutins, were derived from edible parts of two legumes, *Glycyrrhiza foetida* and *Amorpha fruticosa*. The natural amorfrutins are structurally new powerful anti-diabetics with unprecedented effects for a dietary molecule (Weidner et al., Diabetologia 2013; Sauer, Chembiochem 2014; Sauer, Trends Pharmacol Sci 2015). We further developed new synthetic derivatives with a large potential for further application (de Groot et al., J Med Chem 2013; Aidhen et al., Org Lett 2015). Notably, we recently also observed anti-inflammatory effects derived from modulating PPARγ by amorfrutins (Fuhr et al., J Nat Prod 2015).

Moreover, we identified a new ligand for transcription factor or nuclear receptor LXRα, which is subtype-specific - in contrast to any other known ligands of LXRs. This specific LXRα ligand is in particular active in lipid-loaded foam cells that are involved in atherosclerotic plaque formation. This LXR ligand will be explored as a chemical tool as well as for potential drug-ability (Feldmann et al., Nucleic Acids Res 2013). Furthermore, in collaboration with the company Steigerwald, we identified synergistic activation of nuclear receptors such as PPARs by compound pools derived from phytotherapeutic products as a strategy to inhibit metabolic disease (Weidner et al., PLoS one 2013, Weidner et al., Mol Nutr Food Res 2014). Our results showed that selective activation of nuclear receptors by (diet-derived) ligands constitutes a promising approach to combat metabolic disease.

Figure 2: Amorfrutins isolated from liquorice are highly selective agonists of the nuclear receptor PPARgamma
Transcriptomics, proteomics and metabolomics

To further analyse the effects of dietary intake and molecular interventions in vivo, we developed a bunch of integrated genomics, proteomics and metabolomics pipelines to decipher perturbations of molecular pathways underlying disease. Thereby we discovered amongst others molecular evidence for physiologically important (side) effects of drug treatments (Meierhofer et al., Mol Cell Proteomics 2013; Meierhofer et al., J Proteome Res 2014), especially on the protein level resulting in varying production of key metabolites. Furthermore, we observed protein networks required for metabolically relevant cell-cell communication, for example to better understand crosstalk between fat cells and macrophages. Our results provided insights in the metabolic adaptation of these metabolic target cells, which are involved in the development of insulin resistance (Freiwald et al., Proteomics 2013). Further work was done to provide data analysis tools to dissect causative enzymes from mass spectrometry-based data of ten thousands of post-translational modifications (our software to solve this problem is termed PHOXTRACK; Weidner et al., Bioinformatics 2014). Moreover, we developed and applied methodologies for “proteomics informed by transcriptomics” approaches to analyze the interaction of microbial species such as phytoplasma with host plants (Luge et al., Proteomics 2014; Sauer & Luge, Proteomics 2015).

Gene regulation during mild stress response

The natural product resveratrol is a widely known, even famous molecule found in red wine. Multiple health-beneficial and striking anti-aging effects have been reported, as well as a number of potential target proteins and underlying mechanisms have been suggested. However, the mechanism of action of “magic resveratrol” remained essentially elusive. Based on novel mass spectrometry-based assays we analyzed compounds for potential activation of target proteins that are potentially involved in anti-aging processes. Therefore, we developed a novel functional high-throughput mass spectrometry assay to screen and characterize natural products interacting with protein-modifying enzymes such as genome-regulating deacetylases including sirtuin 1 (Sirt1) or acetyl transferases like p300 (Holzhauser et al., Angew Chem Int Ed Engl 2013).

We further showed explicitly that the beneficial cellular effects of resveratrol can be explained by its chemical degradation in physiological buffers (all containing bicarbonate ions), leading to production of reactive oxygen species derived from reacting resveratrol, subsequent genome-wide remodelling of Nrf2-triggered transcriptional pathways to boost cellular defence, resulting in modulated metabolic profiles (Figure 3). A concerted action of Nrf2-driven cellular defence mechanisms, including reduction of
cellular redox environment due to increased endogenous pools of glutathione, may explain mechanistically many reported aspects of resveratrol on the cellular and physiological level. Based on our data, we propose a hormesis model of the action of resveratrol. This line of research led to intensified collaboration with the nutritional and cosmetics industry including companies such as Unilever.

Transcriptional networks of foam cells

Atherosclerosis is an important global health problem and a leading cause of cardiovascular disease. Adaptation of macrophages to physiological stimuli as lipid overload or elevated levels of cholesterol requires dynamic regulatory molecular networks. We deciphered the LXRα-dependent gene-regulatory architecture of atherosclerotic foam cells, as well as key networks triggered by LXRα-modulation for treating efficiently atherosclerosis, by using integrated genome-wide analysis of LXR-alpha. Functional analyses integrating genetic variation disease association data revealed cholesterol induced disease gene expression and suggest avenues for treating systematically foam cell development and atherosclerotic plaques, for example via specific LXRα-activation of the APOC-APOE gene cluster (Feldmann et al., Nucleic Acids Res 2013). This work laid the ground for our current endeavours to better understand physiological plasticity of stressed cells such as macrophages by using single-cell approaches.
Single cell genomics

Using, amongst others, microfluidics and flow cytometer-based techniques combined with Illumina sequencing, our group pioneered in the institute workflows for sequencing transcriptomes of single cells. We focussed on analyzing metabolically and otherwise stressed macrophages to gain insights in the development of cellular subpopulations and general cellular strategies to respond to varying environmental challenges (Figure 4). Thereby, we could decipher surprising interactions of transcriptional regulators as well as unexpected mutual activities of transcriptional pathways. Targeted functional analyses are currently being done to gain further mechanistic insights in these molecular activities to make cell populations robust versus environmental changes. This work is being done in collaboration with ETH Zürich (R. Zenobi). The output expected from this research seems to have the potential to fundamentally change our current view on cellular organisation.

Additional scientific achievements between 2009-2012

In this time period, as previously reported in more detail, our group set up a number of approaches for our nutrigenomics research program. One important and publically highlighted publication was the first paper dealing with above mentioned amorfrutins as potent antidiabetic dietary natural products (Weidner et al., PNAS 2012).
Furthermore, our group collaborated with the company Bruker Daltonics since 2006 to establish mass spectrometry-based methods for detecting bacteria and profiling patients. These technologies, based on protein pattern detection algorithms, have in the meantime revolutionized microbial diagnostics for first-line identification of bacteria, and are widely applied as certified methods in the clinics (Freiwald & Sauer, Nat Protoc 2009; Sauer & Kliem, Nat Rev Microbiol 2010; Kliem & Sauer, Curr Opin Microbiol 2012; Figure 5). Our pattern-dependent tools developed for classification of bacteria have in a sense also stimulated the above-described methodologies to monitor in metabolic target tissues complex physiological outcomes of dietary interventions or drug treatments.

**Development of research infrastructures**

**Externally**

Based on our long-standing experience in genotyping (development of the GOOD assays) and further methods development in genome research, the Sauer group was involved in various collaborative genome research initiatives such as the National Genome Research Network in Germany (NGFN). For example, it was recently noted by Snyder et al. that our work presented in (Burgtorf et al., Genome Res 2003) “laid the groundwork for massively parallel implementations” to genome-wide haplotyping - after about 12 years of the original publication… (see Snyder et al., Nat Rev Genet 2015).

Nowadays, Sascha Sauer coordinates the European Sequencing and Genotyping Infrastructure (ESGI, www.esgi-infrastructure.eu; Figure 6), which
pools leading European genomics and bioinformatics facilities to provide
the larger scientific community with access to new genomic technologies
and the latest bioinformatics tools. The aim of ESGI is to enable scientists
across all disciplines to use emerging sequencing technologies to deci-
pher the complex functions of genes, without breaking the bank. About
29 scientific projects have been finished or are currently still active, with
a particular focus in the areas of (epi)genetics and functional genomics
of complex diseases, and on general mechanisms of gene regulation and
chromatin biology.

Selected publications of ESGI projects coordinated by us were for exam-
ple:

- 't Hoen PA et al., Nat Biotechnol 2013 Nov; 31(11):1015-1022
- Lappalainen T et al., Nature 2013 Sep 26; 501(7468):506-511

Internally

The mass spectrometry-based proteomics pipelines described above are
additionally being used for a number of internal collaborative projects to
complement ongoing research in the institute. Meanwhile, the MPIMG has
installed a mass spectrometry service group, and our responsible postdoc
at that time, David Meierhofer, has been recruited as head of this facility.

Selected publications of this collaborative work were for example:

Cooperation within the institute

- Department of Vertebrate Genomics (Hans Lehrach)
- Otto Warburg Laboratory (Ulrich Stelzl, Ulf Ørom)
- Mass Spectrometry Facility (David Meierhofer)
- Sequencing Facility (Bernd Timmermann)

Special facilities/equipment of the group

- Nano-HPLC LTQ Orbitrap XL (EDT) ESI Mass Spectrometer (Thermo) *
- Cap-LC HCT ultra mass spectrometer (Bruker)
- Genome Analyser IIx (Illumina) *

* externally funded

General information

Complete list of publications (2009-2015)

Group members are underlined.

2015


2014


Sauer S (2015). Ligands for the nuclear peroxisome proliferator-activated receptor gamma. Trends in Pharmacological Sciences, accepted

Meierhofer D, Weidner C & Sauer S (2014). Integrative analysis of transcriptomics, proteomics, and metabolomics data of white adipose and liver tissue of high-fat diet and rosiglitazone-treated insulin-resistant mice identified pathway alterations and
molecular hubs. *Journal of Proteome Research*, 13(12), 5592-5602


2013


2012


2011


2010


*S. Sauer. C. Weidner, M. Kliem: Natural PPAR-Ligands as pharmaceuticalaly active agents*, EP13165716.5

*S. Sauer, S. Holzhauser, R. Feldmann, A. Geikowski: Foam cell specific Liver X Receptor (LXR) alpha agonist, SIRT1 inhibitors as well as p300 inhibitors as pharmaceutically active agents*, EP13153142.8
Selected invited talks
(Sascha Sauer)

Population dynamics and regulatory circuits during macrophage polarization. Systems Biology of Infection Symposium, Ascona, Switzerland, 09/2015

Food, genes and diseases. Interdisciplinary Dialogue, IRI Life Sciences, Berlin, Germany, 01/2015.

Analysing transcriptional landscapes of dietary molecules. Summer School on Nutrigenomics, University of Camerino, Camerino, Italy, 09/2014.

Dissecting the molecular effects of (hormetic) stress response to counteract disease and aging processes. MRC Clinical Centre Hammersmith, London, UK, 07/2014

Dissecting the molecular effects of (hormetic) stress response to counteract disease and aging processes. European Bioinformatics Institute, Hinxton/Cambridge, UK, 06/2014

Sequencing out the effects of metabolic stress response in single cells. 2nd Annual Food, Nutrition and Agriculture Congress, London, 04/2014

Proteomic approaches in nutrition research. 7th Central and Eastern European Proteomics Conference (CEEPC) on Proteomics Driven Discovery and Applications, Jena, Germany, 10/2013 (plenary lecture)

Refining our approach to research in genomics and drug discovery. World Research and Innovation Congress – Pioneers in Healthcare, Brussels, Belgium, 06/2013

Functional nutrigenomics, network biology and the prevention of metabolic diseases. The 5th Paris Workshop on Genomic Epidemiology, Paris, France, 05/2013

Teaching activities

Seminar Functional Genomics and Metabolism Research, Freie Universität Berlin, winter term 2012/13,

Seminar Bioinformatische Verfahren in der funktionellen Genomik und Proteomik, Freie Universität Berlin, summer term 2013

Praktical course Bioinformatische Verfahren in der funktionellen Genomik und Proteomik, Freie Universität Berlin, summer term 2013

Seminar Einführung in die Biotechnologie, Freie Universität Berlin, winter term 2013/14

Organisation of scientific events

ESGI Symposium on Functional Genomics and Metabolism Research, Berlin, 21st-22nd of March 2013
Otto Warburg Laboratory

Max Planck Research Group
Regulatory Networks in Stem Cells
Established: 01/2015

Scientist
Liat Ravid-Lustig (since 06/15)

PhD students
Oriana Genolet (since 02/15)

Students
Verena Mutzel (since 04/15)
Sungsoo Lim (02/15-07/15)

Scientific overview

During embryonic development, cells acquire distinct transcriptional and epigenetic states, which must then be stably maintained. Complex regulatory networks have evolved to govern these differentiation processes and to sustain a memory of the initial fate decision. The overall goal of our research is to elucidate the regulatory principles employed by such networks on the transcriptional and epigenetic level. Specifically, we study the developmental process of X-chromosome inactivation (XCI).
In mammals, X inactivation is an essential step in the early development of female embryos, which ensures X-chromosome dosage compensation between the sexes. Each cell in the embryo selects one randomly chosen X chromosome that will then undergo chromosome-wide gene silencing, mediated by the long non-coding RNA Xist, which is expressed only from the inactive X. A complex interplay of cis- and trans-acting regulatory mechanisms ensures that Xist is up-regulated only in cells with two X chromosomes (female-specific) and only from one of the two X’s in each cell (mono-allelic) (Schulz & Heard, Curr Opin Genet Dev 2013). To ensure correct developmental timing, the Xist regulatory network also integrates differentiation cues, and feeds back into the developmental program by blocking differentiation until X inactivation has occurred (Schulz et al., Cell Stem Cells 2014). However, the molecular mechanisms and the regulatory principles that are employed by the Xist regulatory network to ensure female-specific and mono-allelic expression and to coordinate X inactivation with the differentiation program, remain poorly understood.

For both, the initiation of X inactivation and its impact on differentiation, the underlying network must reliably sense an only two-fold difference in X-chromosome dosage and elicit a quantitative response. To understand such complex decision-making processes, quantitative experimental data must be interpreted with the use of mathematical models. For such a systems biology approach, it is essential to be able to quantify the network output together with central regulators in response to experimental perturbations (Figure 1). X inactivation is particularly amenable to this approach, because it is a developmental process with a clear quantitative output that can be recapitulated in an in vitro culture system, in differentiating...
mouse embryonic stem cells (Figure 2, top). With the rapid development of genome engineering techniques in recent years, it is now conceivable to perform a sufficient number of perturbations to be able to dissect this regulatory network. The resulting mutants can then be analyzed with quantitative multiplex-measurements of multiple regulators in a single cell with single-allele resolution. In addition, ever more high-throughput techniques become available to identify missing players in the network. Thus, X inactivation is an ideal model system for understanding the principles of quantitative information processing during mammalian development.

Although a series of hypotheses have been put forward over the years of how female-specific and mono-allelic expression of Xist might be ensured, few attempts have been made to rigorously study their potentials and limitations through mathematical modeling. To unambiguously test, which mechanisms would in principle be able to explain the expression pattern of Xist, we have compared a series of networks, modeled by ordinary differential equations, with respect to their stable steady states. This analysis showed that a negative feedback loop acting in trans, which has been proposed previously, is insufficient on its own and must be combined with a cis-acting positive feedback to allow for stable mono-allelic expression of Xist. Part of the proposed project aims at identifying the regulators and mechanisms that mediate these feedback loops. To achieve this, we study the interaction of known Xist regulators and try to identify additional new regulators.
Modeling the Xist regulatory network

The modeling analysis we have performed revealed that a negative feedback on its own is insufficient to recapitulate the Xist expression pattern, because expression cannot be maintained in the presence of a single activator dose. Therefore an additional mechanism must be in place to provide a memory of the Xist expression state. Since non-linear positive feedback loops can provide such a molecular memory, we propose the existence of such loop, acting \textit{in cis} on the allele level. In search for mechanisms that might mediate such a positive feedback loop, we are investigating a potential role of Xist’s antisense transcript Tsix. Tsix is a long non-coding RNA and is known to repress Xist expression \textit{in cis}, probably by transcription through the Xist promoter. Since Xist in turn also silences Tsix like other X-linked genes, Xist and Tsix mutually repress each other, thereby forming a double negative feedback loop, which can have similar properties as a positive feedback.

To understand, whether Tsix could indeed mediate the predicted positive feedback loop, we have developed a more detailed mathematical model by incorporating Tsix explicitly. In a series of stochastic simulations we could show that silencing of Tsix must be sufficiently fast to allow this feedback to provide the required expression memory. When the silencing rate was estimated experimentally, however, it turned out to be rather slow. An alternative hypothesis that we are currently testing is the possibility that Xist would interfere with Tsix function through direct transcriptional interference due to RNA polymerase (Pol II) collisions during antisense transcription.

In the context of his master’s thesis, Sungsoo Lim has developed and analyzed a mathematical model of Xist/Tsix antisense transcription. He could show that mutual inhibition due to Pol II collisions in combination with repression of the Xist promoter by Tsix transcription can indeed result in two stable states within realistic parameter ranges for initiation and elongation rates. In the next step, we will estimate the model parameters from experimental data to understand, whether they lie in the bistable region. To this end, Verena Mutzel, another master student in the group, is setting up a single molecule RNA-FISH quantification pipeline to perform absolute quantification of the Xist RNA, Xist nascent transcripts and Tsix transcripts. For these experiments, the institute has acquired a high-end, fully automated widefield fluorescence microscope. This data will then be used to estimate model parameters and will be the first step towards a quantitative understanding of Xist regulation. In the future, the quantification pipeline will be adapted to quantify Xist and its candidate regulators in the same cell in wildtype and mutant conditions to assess the quantitative relationship between Xist and its regulators. In an iterative process of mathematical modeling and collection of quantitative data the model of the Xist regulatory network will be refined and central model predictions will be tested experimentally.
Identification of unknown Xist regulators

In addition to a quantitative analysis of the interactions between known Xist regulators as described above, we will also attempt to identify new regulators of Xist and their mechanisms of action. To this end, we are setting up a screen to identify X-linked Xist activators, which are thought to play an important role in ensuring mono-allelic and female-specific regulation of Xist. Moreover, we are performing single-cell RNA sequencing to identify putative developmental regulators of Xist and we are using ATAC-Seq to identify cis-acting regulatory element of Xist.

Screen for X-linked activators

Since a double dose of the Xist activator is required for Xist up-regulation, its overexpression in male cells should induce ectopic Xist upregulation. Liat Ravid-Lustig, a postdoc in the group, will overexpress all X-linked genes in a male cell line using the CRISPRa system to identify genes that can induce Xist up-regulation. As a readout, we are developing an assay that allows identification and purification of Xist-expressing cells by flow cytometry. This will allow us to use a pooled lentiviral screening strategy based on FACS sorting of Xist positive cells and subsequent deep sequencing.

Developmental regulators

Xist is up-regulated upon exit from the pluripotent state. Several pluripotency factors have been suggested to function as Xist repressors, which would result in Xist up-regulation once these factors are down-regulated. On the single cell level, however, the expected negative correlation between Xist and these factors is not observed. To perform an unbiased search for putative developmental regulators of Xist, we are performing single cell RNA-sequencing of differentiating female ES cells together with the Sequencing Facility (Bernd Timmermann). This approach will allow us to identify genes with an expression pattern that is positively or negatively correlated with Xist. To further refine the analysis, we will quantify the “extend” of X-inactivation through analysis of gene silencing by allele-specific read-mapping in the hybrid cell line we are using. This allele-specific analysis will be performed in collaboration with Sarah Teichmann (Sanger Centre, Cambridge, UK). In this way, we hope to identify genes that are (anti)-correlated with initial Xist up-regulation. Their functional role will then be analyzed by induction or repression using CRISPRa or CRISPRi.
Cis-regulatory elements

To approach the identification of unknown Xist regulators and regulatory mechanisms from a complementary angle, we will identify cis-regulatory elements in the Xist locus by ATAC-Seq, which detects accessible regions in the genome. In collaboration with Edith Heard (Institut Curie, Paris, France) and Uwe Ohler (MDC, Berlin), we are performing ATAC-Seq in differentiation male and female ES cells at the time window, when Xist is being up-regulated. Through identification of regulatory regions only active in female cells and footprint analysis at these regions, we hope to identify candidate regulators. Moreover, we will use reporter assays to understand the temporal activity profile of these elements and thus narrow down their functional role in the Xist regulatory network.

New regulatory mechanisms that will be identified through these different approaches will then be included in the mathematical model described above. Data-based model analysis will help to further elucidate their functional role.

Linking the Xist network to stem cell differentiation

X inactivation is tightly coordinated with differentiation, since down-regulation of stem cell factors triggers Xist up-regulation. During my postdoc, however, I discovered that in turn also X inactivation modulates differentiation (Schulz et al., Cell Stem Cell 2014). Double X dosage modulates several signaling pathways such as MAPK, Akt and Gsk3 in XX ESCs, thereby blocking exit from the stem cell state. We are now attempting to uncover the underlying mechanism by performing a screen to identify the X-linked gene(s) responsible for the effect.

To identify the X-linked gene(s) that modulate signaling activity in ESCs, we will screen all X-linked genes using a reporter for MAPK pathway activity as a readout. Since ESCs with two X-chromosomes exhibit reduced expression of MAPK target genes (Schulz et al., Cell Stem Cell 2014), repression of the X-linked signaling modulator should increase expression of a fluorescent reporter for MAPK activity. A PhD student in the group, Oriana Genolet, is currently testing different MAPK reporter constructs. Once we have identified a sensitive readout, we will perform a pooled lentiviral screen using the CRISPRi technology to identify X-linked genes that when repressed in XX ESCs lead to increased MAPK activity. In this way we hope to gain a better understanding of the link between X-chromosome dosage, MAPK signaling and differentiation.
Planned developments

Our long-term goal is to gain a detailed mechanistic quantitative understanding of how female-specific mono-allelic expression of Xist is ensured. To this end, we will acquire quantitative single-cell data of previously known and newly identified regulators in wildtype conditions and in response to perturbations using the CRIPSR/Cas9 system.

This will hopefully allow us to identify the regulatory principles used to make such a quantitative reliable decision. This is also expected to yield new insight in the functional role of antisense transcription and in how exit from the stem cell state is regulated.

General information

Complete list of publications (2012-2015)

Group members are underlined.

2014


2013

Schulz EG & Heard E (2013). Role and control of X chromosome dosage in mammalian development. Current Opinion in Genetics and Development, 23(2), 109-115 (review)

2012


Student thesis

Otto Warburg Laboratory
Max Planck Research Group
Molecular Interaction Networks

Established: 06/2007 - 08/2015

Head
Prof. Dr. Ulrich Stelzl
Phone: +49 (30) 8413-1264
Fax: +49 (30) 8413-1960
Email: stelzl@molgen.mpg.de

Secretary of the OWL
Cordula Mancini
Phone: +49 (0)30 8413-1691
Fax: +49 (0)30 8413-1960
Email: mancini@molgen.mpg.de

Scientists
Jonathan Woodsmith (since 01/11)
Nouhad Benlasfer (since 05/10)
Victoria Casado (11/14-08/15)
Anna Hegele* (06/07-04/12)
Petra Birth (06/08-12/11)
Reynaldo López-Mirabal* (07/08-06/10)

PhD students
Stefanie Jehle (since 05/12)
Thomas Corwin (09/10-06/15)
Luise Apelt (07/10-06/14)
Mareike Weimann (12/07-02/12)
Atanas Kamburov* (01/10-12/11, jointly with R. Herwig, Dept. of Vertebrate Genomics)
Arndt Grossmann* (07/07-12/11)
Josephine Worseck* (09/07-11/11)

Undergraduate students
Hanna von Gries (03/14-07/14)
Marcel Michla (02/14-07/14)
Federico Apelt (11/10-04/12)
Ziya Özkan (08/09-10/12)
Franziska Wachsmuth (10/11-06/12)
Chrysovalantis Sourlis (10/09-06/10)
Sylvia Wowro (07/08-05/09)

* externally funded
Introduction

Protein interaction networks in cellular signaling and post transcriptional regulation

Comprehensive, high quality protein-protein interaction (PPI) networks represent a prerequisite for a better understanding of cellular processes and associated phenotypes in health and disease. In integrative approaches, interaction networks are very useful for studying complex genotype to phenotype relationships and result for example in a more accurate interpretation of the impact of genomic variation in cells. In addition to genetic variation that impacts gene expression, cells are subjected to a variety of other cellular and environmental cues. These signals are also mediated by interaction networks, with time-dependent, quantitative contributions of hundreds of proteins underlying the phenotypic outcome. In these networks, the fast response to internal and external cues is frequently mediated by the reversible covalent posttranslational modification (PTM) of present proteins, e.g. phosphorylation, ubiquitination, methylation or many others. Consequently, molecular interaction networks are conditional with respect to the signaling status of the cell. A crucial point for understanding the information flow in cellular signaling and regulation is to define how the spectrum of PTMs is coordinated and recognized by interacting molecules and how PTM-dependent alterations in protein interaction networks lead to signal propagation and cellular changes.

To comprehensively analyze human cellular networks governing the dynamics of molecular machines and signaling processes, we apply experimental functional genomics techniques, e.g. a high-throughput yeast two-hybrid screening approach in combination with mass spectrometry, biochemical, cell biological and computational methods. With this interdisciplinary approach our work aims at (i) high quality, large scale PPI data generation and analysis of human protein-protein interaction networks, (ii) integrative analyses of protein interaction dynamics to understand how information is processed in cellular networks and (iii) computational and experimental approaches to directly investigate patterns of protein modification that lead to conditional protein-protein interactions, such as phosphorylation-dependent interactions, that is PPIs requiring phosphorylation of one interaction partner.
Scientific methods and achievements

Generation and analyses of human protein-protein interaction networks

Protein interaction networks are essential in promoting our understanding of cellular phenotypes, however recording of a reliable, static representation of all possible human protein-protein interactions is work in progress: Maybe about 10-20% of the human interactome has been mapped today and large data biases exist. E.g., interactome data cover only a lim-

Figure 1: The protein methyltransferase interactome. We comprehensively screened proteins involved in arginine and lysine methylation and demethylation, i.e. protein methyltransferases (PMT) and demethylases (PDeM) such as AOF2/LSD1, for interacting partners and reported a major informational PPI network resource (523 PPIs involving 22 methyltransferases or demethylases). The annotated PPI network displays new candidate proteins with methyl transferase / de-methylase interaction partners in groups of proteins with shared interaction profiles (PMTs and PDeMs in green). Protein domains or functions that were enriched within the network are color coded to their respective prey proteins. Our data implicate protein methylation in specific cellular roles other than transcription and epigenetic regulation, i.e. splicing, Cullin-RING ubiquitin ligase-dependent protein degradation or the cytoskeleton (Weimann et al., Nature Methods 2013).
lected proportion of the human proteome with large regions of local interactome data paucity. In the last years, we developed strategies to address this data paucity applying our high quality Y2H PPI-screening techniques combined with computational approaches and functional analyses. We have contributed an interactome map for Alzheimer’s Disease, constructed a large directed PPI network to study signal flow, analyzed dynamic PPI patterns in pre-mRNA splicing and defined novel cellular functions for protein methylation (Figure 1). In the latter study we reported a novel Y2H-seq approach with significantly improved sensitivity through linking Y2H interaction selection with a second generation short-read Illumina sequencing readout. Interactome mapping requires both search space and interaction coverage: that is, scalability in the number of protein pairs that can be assayed and high sensitivity to actually detect the interactions in the search space. Increased sensitivity of Y2H-seq leads to higher interaction coverage in the screened interactome space.

Figure 2: Examples of multi-signal PTMi spot proteins. Examples of proteins annotated either as literature known, canonical PTMi spot containing proteins, proteins that integrate both ubiquitin and multiple phosphorylations in a PTMi spot (degron), or to one of five key regulatory cellular modules. Representative examples of PTM density distributions of a protein in each sub-group is linked to the panel (Woodsmith et al., PLoS Computat Biol 2013).
In addition, computational approaches for the assessment and functional prioritization of protein-protein interactions under various perspectives have been developed. This included PPI confidence assessment exploiting network topology or 3D structural constraints, integrative interaction prediction and the investigation of the interplay of distinct post-translational modifications in interaction networks. Because experiments that simultaneously study different modifications are very difficult, we investigated in a computational approach, whether different global post-translational modifications, i.e. phosphorylation (pY and pS/T), acetylation, and ubiquitination, are coordinated in human protein networks. Integrating these four globally measured PTMs (more than 3,000 modified proteins in human each) on protein complex datasets, we identified hundreds of human protein complexes that selectively accumulate PTMs. Also protein regions of very high PTM densities, termed PTMi spots, were characterized (Figure 2). Notably PTMi spots were exclusive of folded protein domains however show domain-like features e.g. equally high mutation rate in cancer. Systematic PTMi spot identification highlighted more than 300 candidate proteins for combinatorial PTM regulation.

This study suggest that analysis of PTMs coupled to protein interaction information will promote a better understanding of enzyme - substrate relationships, the dynamics of PTM-mediated signal flow and the consequences of PTM-mediated recognition events, i.e. the rewiring of molecular networks as a signaling response.

Differential / conditional networks

Though the generation of static interaction data at a proteome scale is important as it provides most valuable resources e.g. in the integrative analyses of genetic variation or PTMs, we aim at studying more densely mapped networks of specific functional modules at different conditions and cellular states. This is particular important as interactome networks are extensively re-wired during any cellular response. Principles of cellular regulation and signal processing can be learned from differential networks, while concomitantly getting mechanistic insight into the specific process. Two specific examples of differential/conditional network studies are briefly outlined below.

The human spliceosome assembles for each and every intron to be excised de novo. This assembly cycle likely involves more than 200 proteins, for which we provided a high quality interaction resource describing 632 interactions. A set of 76 affinity purification experiments collected from a series of proteomics studies was then used to infer PPI dynamics of the spliceosomal assembly cycle. We identified crucial sites of changing PPI patterns that contribute to the exceptional compositional dynamics - and
thus function - of this huge ribonucleoprotein machine (see also MPI/MG Research Report 2012). Our work continues on the definition of the role of higher eukaryote-specific proteins, e.g. in alternative splicing events, and aims at a better understanding of how the splicesomal assembly cycle connects to cellular signals.

We also develop direct experimental approaches to analyze conditional PPI patterns. Recently, we set up a modified Y2H system (employing e.g. human protein kinases) that is capable of detecting PTM-dependent interactions (Figure 3). In contrast to other methods for the identification of PTM-dependent protein interactions, our approach is examining the binary interactions of full-length human proteins at cellular concentrations in a cellular environment. The modified Y2H system is complementary to existing strategies such as peptide arrays or AP-MS/MS techniques and fully compatible with our large resources and unique proteome scale screening technology.

In a systematic, proteome wide approach, we have identified ~ 300 novel phospho-tyrosine(pY)-dependent PPIs that show high specificity with respect to human kinases and interacting proteins (Figure 3c). The data support the notion that many of the pY-interactions do not rely on known

![Image](image.png)
linear motifs encouraging the investigation of novel modes for phospho-tyrosine modification recognition. P-dependent interactions were further analyzed in mammalian cell culture systems using e.g. co-immunoprecipitation, protein complementation and functional reporter readouts. The more detailed characterization of selected P-dependent interactions demonstrated how PTM-dependent PPIs are dynamically and spatially constrained. Exemplary evidence was provided that two pY-dependent PPIs with the tetraspanin protein TSPAN2 dictate cellular cancer phenotypes.

**Conclusion**

Function clusters within protein interaction networks. Exploiting this feature, networks will be key to our understanding of complex phenotypes. The high quality interactome data sets generated are of significant biomedical relevance as they contribute directly to current attempts aiming to improve and individualize the practice of medicine on the basis of patient-derived genomic, proteomic and drug response information. Just as the interpretation of the large number of genetic variation between genomes is greatly aided by network information, evidence is piling up that protein-protein interaction networks will be successfully exploited in the analyses of the more than 100,000 cellular PTMs on 12,000 unique human proteins mapped across diverse conditions so far. In ongoing and future projects, we want to further elucidate global PTM patterns and investigate how PTMs together with differential splicing and genetic variation impact on cellular protein interaction networks. Network approaches are about to strongly shape our view on how specificity is achieved in cellular signal transduction and post-transcriptional regulation and may ultimately reveal the molecular changes in cellular processes that occur in human diseases such as cancer.
General information

Complete list of publications (2009-2015)

Group members are underlined.

2015


2014


2013


Stelzl U (2013). Molecular interaction networks in the analyseAs of sequence variation and proteomics data. PROTEOMICS - Clinical Applications, 7, 727-732


2012


**2011**


**2010**


**2009**


**Selected invited talks**

(Ulrich Stelzl)

- 12th [BC]2 - the Basel Computational Biology Conference, Congress Center Basel, Switzerland, 06/2015
- CSH Systems Biology: Networks, Cold Spring Harbor Laboratory, New York, USA, 03/2015
- HUPO 2014, 13th Human Proteome Organization World Congress, Madrid, Spain, 10/2014
- MRC National Institute for Medical Research, Mill Hill, London, UK, 06/2014
- Department of Biochemistry, University of Cambridge, UK, 06/2014
- Hybrigenics, Paris, France, 10/2013
- Universität Bern, Switzerland, 10/2013
- The International Conference on Systems Biology of Human Disease 2013 (SBHD 2013), Heidelberg, Germany, 06/2013
- CMBI Seminar Series: Center for Molecular Biosciences, Innsbruck, Austria, 06/2013
ESF-EMBO Symposium, Molecular Perspectives On Protein-Protein Interactions, Pultusk, Poland, 05/2013

Ian Lawson Van Toch Memorial Lecture, OCI Seminar in Computational Biology, Ontario Cancer Institute / University of Toronto, Canada, 09/2012

Integrative Network Biology 2012: Network Medicine, Helsingør, Denmark, 05/2012

Pharmacology and Molecular Sciences Wednesday Seminar Series, Johns Hopkins University School of Medicine, Baltimore, USA, 11/2011

MIT, Cambridge, Boston, USA, 11/2011

Max Planck Institute for Molecular Cell Biology and Genetics, Dresden, Germany, 11/2011

PPI Berlin: Current Trends in Network Biology (workshop), Max Delbrueck Communications Center Berlin-Buch, Germany, 10/2010

Joint Cold Spring Harbor Laboratory / Wellcome Trust Meeting on SYSTEMS BIOLOGY: NETWORKS, Hinxton, UK, 08/2010

Wellcome Trust 91st Advanced Course, Protein Interactions and Networks, Wellcome Trust Sanger Institute, Genome Campus Hinxton, Cambridge, UK, 12/2009

Green Seminar: Biotechnology Center TU Dresden, Germany, 10/2009

“What is a macromolecular Complex? Shades of Meaning Across Cellular, Systems, and Structural Biology (workshop), NKI, Amsterdam, Netherlands, 10/2009

EBI Seminars in Systems Biology: European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, UK, 06/2009

CSHL Meeting Systems Biology: Networks, Cold Spring Harbor Laboratory, New York, USA, 03/2009

PhD theses

Thomas Corwin: Deciphering human cytoplasmic tyrosine kinase phosphorylation specificity in yeast. Freie Universität Berlin, June 2015 (submitted)


Josephine Worseck: Characterization of phosphorylation-dependent interactions involving neurofibromin 2 (NF2, merlin) isoforms and the Parkinson protein 7 (PARK7, DJ1). Humboldt-Universität zu Berlin, 2012

Atanas Kamburov: More complete and more accurate interactomes for elucidating the mechanisms of complex diseases. Freie Universität Berlin, 2012
**Student thesis**


**Teaching activities**

Guest lecture „Networks“ in the context of the module Advanced Bioanalytik, Technische Universität Berlin, summer term 2015

Lecture at the module „Funktionelle Genomik“, Technische Universität Berlin winter term 2014/15, summer term 2014, winter term 2013/14

Guest lecture „Funktionelle Genomik“ in the context of the module Advanced Bioanalytik, Technische Universität Berlin, summer term 2013
Otto Warburg Laboratory
Research Group Gene Regulation and Systems Biology of Cancer
Established: 1996 (Dept. of Vertebrate Genomics), since 12/2014 Otto Warburg Laboratory

Head
Dr. Marie-Laure Yaspo (since 1996)
Phone: +49 (30) 8413-1356
Fax: +49 (30) 8413-1380
Email: yaspo@molgen.mpg.de

Scientists
Johanna Sandgren* (since 08/14, visiting guest scientist Karolinska Institute)
Hans-Jörg Warnatz* (since 05/09)
Catherine Sargent* (09/12-08/14)
Marc Sultan* (09/07-09/13)

Bioinformaticians
Christine Jandrasits* (since 12/13)
Thomas Risch* (since 02/13)
Vyacheslav Amstislavskiy* (since 08/11)
Meryem Avci* (01/09-12/12)
Karin Schlangen* (08/11-08/12)

PhD students
Nilofar Abdavi Azar (since 05/14)
Praneeth Reddy Devullapally*
(since 08/13)

Engineers and technicians
Carola Stochek (since 10/12)
Mathias Linser* (since 06/11)
Alexander Kovacsoviches* (since 01/11)
Simon Dökel* (since 04/10)
Daniela Balzereit* (since 02/02)
Aydah Sabah (12/12-12/14)
Agnes Wolf* (06/11-09/14)
Katharina Kim* (01/11-03/14)
Sabine Schrinner* (01/02-09/12)

* externally funded
Scientific overview

Our group focuses on cancer genomics and systems biology of cancer. Based on next generation sequencing (NGS) technologies, we aim at dissecting tumor molecular landscapes for investigating the interplay of pathways associated with malignant properties in individual tumors and the gene regulation networks operating in particular cancer groups. We are also interested in understanding the interactions of tumor cells in relation to their stromal niche, as well as in evaluating, to which extent patient-derived model systems such as cells or xenografts used for testing drug sensitivity recapitulate the original tumor features. Besides, we are developing NGS-based methods characterizing immune cell repertoires.

We made significant contributions within funded projects in the Department of Vertebrate Genomics (Hans Lehrach): Innovative Medicine Initiative (colon cancer, www.oncotrack.eu), International Cancer Genome Consortium with pediatric brain tumors (ICGC Pedbrain, www.pedbraintumor.org) and early onset prostate cancer (https://icgc.org/icgc/cgp/70/345/53039), pediatric acute lymphocytic leukemia (Ufoplan, German Federal Office for Radiation Protection), a pilot translational study on metastatic melanoma (Treat20-BMBF), FP7 EU consortia, the Blueprint epigenome (http://www.blueprint-epigenome.eu/) and TRIREME, a project tackling the cellular responses to ionizing radiation. Besides, under the umbrella of the Marie Curie Initial Training Network “VacTrain” (http://www.vactrain.eu/), and in “TREGeneration” EU Horizon 2020 Research and Innovation Program, we are developing NGS-based approaches for analyzing human immune cell repertoires, a promising step given their relevance in disease and cancer.

We have built powerful NGS-based pipelines integrating genome and transcriptome information, mutations, differential gene expression, long non-coding RNAs, alternative splicing, copy number alterations, and the accurate detection of gene fusions. Applying these methods to various cancer entities, we summarize below current progress and achievements.

Metastatic melanoma

Treat20 was a pilot translational project designed to predict the drug response for 20 metastatic melanomas based on their molecular characterization and modelling of drug action (Christoph Wierling’s group) We analyzed different melanoma types (cutaneous, uveal, iridal, mucosal, and acrolentiginous), and found novel molecular events in non-UV induced, non-BRAF mutated melanomas, such as a ROS1 gene fusion, focal amplifications or deletions of key cancer genes providing potential therapeutic targets.
This pilot phase prompted Treat20Plus, a new BMBF-funded follow up project, coordinated by Marie-Laure Yaspo, with the Charité Comprehensive Cancer Center (CCCC, Ulrich Keilholz) as clinical partner and the company Alacris Theranostics carrying out the tumor modeling for drug treatment prediction. Treat20Plus is a clinical trial for 35 metastatic melanomas refractory to treatment. Our group will carry out the NGS analysis and will establish melanoma-driven microtissues, which will not only provide a platform for drug testing and model validation, but also a tool for investigating the interactions between melanoma cells and their stromal environment, representing one of the future developments in the group.

**Integrated genomic and functional analyses of TCF3-HLF-positive ALL**

Pediatric acute lymphocytic leukemia (ALL) is characterized by chromosomal translocations that cause gene fusions involving master regulators of hematopoietic development. Our project aimed at comparing the molecular pathology between two different leukemia types both disrupting one allele of TCF3, which drives the B-cell differentiation program upstream of PAX5: t(1;19) fusing TCF3 to the DNA-binding domain of PBX1, associated with a good prognosis, and the dismal t(17;19)(q22;p13), resulting in the fusion TCF3-HLF. Their mutation profiles are summarized in Figure 1.
The two leukemia subtypes shared a gene expression signature of B-lymphoid cells, but also showed striking differences. We found that expression of TCF3-HLF leads to transcriptional reprogramming in B lymphoid progenitors in the context of PAX5 haploinsufficiency, towards an immature, hybrid hematopoietic state (Figure 2). TCF3 breakpoints occurred in hotspots indicative of terminal deoxynucleotide transferase (TdT) activity characteristic of an early B cell stage. Using *in silico* analysis, we predicted 39 potential HLF targets in TCF3-HLF samples, including SNAI2, GPC4, and BMP3 involved in stem cell proliferation. Profound cellular reprogramming occurred in TCF3-HLF towards a drug-resistant state reflected by stem cell, mesenchyme-derived and myeloid signatures. Drug response profiling in patient-derived xenografts, which had maintained the tumor’s genomic and transcriptome landscapes, identified resistance to most of the 98 tested drugs, but extreme sensitivity towards the BCL2-specific inhibitor ABT-199, indicating new therapeutic options for this fatal ALL subtype (Fisher et al., Nat Genet 2015).
Transcriptomics of pediatric brain tumors

In the context of the International Cancer Genome Consortium (ICGC) Pedbrain, we generated high coverage RNAseq data and analyzed the transcriptomes of 164 medulloblastomas, 90 pilocytic astrocytomas, and 50 glioblastomas.

Medulloblastoma (MB) is the most common malignant brain tumor in children, arising in the cerebellum or medulla/brain stem and showing biological and clinical heterogeneity. Previous studies identified four main groups of MB with distinct clinical, biological, and genetic profiles: WNT tumors, associated to a favorable prognosis and characterized by activated wingless pathway signaling; SHH tumors showing hedgehog pathway activation with intermediate prognosis; and group 3 and group 4 tumors that are less well characterized and clinically challenging.

Our group identified the first medulloblastoma-associated gene fusion involving the SHH gene (Jones et al., Nature 2012). Further, we could identify gene expression signatures for protein-coding as well as non-coding RNAs (Figure 3), specifying 9 different MB subgroups either reflecting their cells of origin or reprogramming of the tumor cells and corresponding to known and novel pathways operating in these groups. In group 3, we found one cluster enriched for rhodopsin and markers associated to visual perception, one myc-driven cluster, and one that is expressing a set of early developmental genes. We are particularly interested in the transcriptional

Figure 3: Hierarchical clustering of long non-coding RNAs specifying sub-clusters in group 3 and group 4 medulloblastomas
networks specifically active in the subgroups (Figure 4). By integrating to these findings in situ RNA expression data generated by us and others in the Eurexpress project (www.eurexpress.org) or from the Allen brain atlas (www.brain-map.org), we obtained a topological mapping for many group-specific factors, a work being carried out in collaboration with Gregor Eichele (Max Planck Institute for Biophysical Chemistry, Göttingen).

This work leads to an in-depth characterization of these tumor entities, which we will integrate with the genomic results of the consortium. We are establishing follow-up functional analysis to validate some of the newly identified gene regulation networks.

**Pilocytic astrocytoma (PA)** accounting for ~20% of all pediatric brain tumors, is typically associated with mitogen-activated protein kinase (MAPK) pathway alterations. Integration of genome and transcriptome sequencing of a cohort of 96 tumors demonstrated that PA is indeed a single pathway disease, featuring also novel gene fusions involving the RAS pathway, such as BRAF or the kinase domain of the NTRK2 oncogene (Jones et al., Nat Genet 2013).

Figure 4: Left: Nodes and edges of the different transcriptional networks specific to MB subgroups represented by the different colors. Right: in situ hybridization of the mouse orthologue of LMX1 and BARHI1
**Pediatric glioblastoma (pedGBM)** is one of the most common and aggressive malignant brain tumors, which presents complex genomic landscape and significant molecular heterogeneity. Integrative analysis of 52 pedGBMs led to the detection of genetic lesions activating the RTK-PI3K-MAPK signaling, but also to so far undescribed gene fusions involving the receptor tyrosine kinase MET affecting ~10% of the cases. The PTPRZ1-MET fusions lead to strikingly high expression of the truncated MET gene product encoding a constitutively active form of MET, driven by the strong PTPRZ1 gene promoter. The newly identified MET fusion chimeric proteins were characterized to be potent activators of MAPK signaling and were found in combination with genetic alterations in the tumor suppressors TP53 or CDKN2A/B. Our work allowed the characterization of a specific group of aggressive glial tumours, which could be successfully targeted using MET inhibitors.

**Analysis of a colorectal carcinoma cohort**

Colorectal carcinoma (CRC) is the third most common cancer worldwide and the second most common cancer in Europe, representing a major healthcare burden. KRAS mutations are used as predictive marker of therapeutic response to antibodies targeting EGFR, however, the resistance to therapeutics remains poorly understood. The IMI Oncotrack international project (www.oncotrack.com) aims at analyzing a prospective CRC cohort with “multi-omics” technologies for identifying novel biomarkers of disease and drug response. The Yaspo group is coordinating the genomic and transcriptomic analysis of the CRC cohort and matching patient-derived models. Together with the spin-off Alacris Theranostics, we have analyzed whole genome, exome, microRNA, and RNAseq data of more than 100 tumors, 60 xenografts and 40 spheroid cell models. Based on these data, our main research directions focus on

- identifying specific signaling pathways in CRC sub-groups;
- dissecting tumor heterogeneity and clonal evolution in tumor models;
- correlating drug sensitivity of the models with molecular markers.
We found that tumors carried between 50 and 1000 somatic alterations (mutations, indels, fusions, deletions, and focal amplifications), the highest range corresponding to microsatellite instable (MSI) cases. Besides hallmarks mutations in APC, Wnt signaling and RAF/RAS pathway, we identified novel recurrent changes in chromatin remodeling complex and epigenetic factors, which we are currently investigating. Besides, we identified three main tumor groups enriched for specific factors, EMT (e.g. SNAI1/2), stem cell determinants (e.g. ASCL2), enterocytes and goblet cell markers, respectively. Different biological processes such as Aurora kinase, Notch signaling, or high inflammation, correlate with these groups. We identified Cetuximab resistance patterns in PDX models correlating with particular sub-groups of tumors. However, we observed that models derived from EMT tumors showed different biological features and might be challenging in terms of clinical transferability. Further, expression of stem cell markers showed significant changes in the models, either due to epigenetic reprogramming and/or to clonal evolution. We are investigating this using a combination of multi-fluorescent labelling and \textit{in situ} hybridization (cooperation with Matts Nielson, Uppsala, Sweden).

We also investigate tumor molecular evolution in CRC synchronous tumors, providing an interesting biological context for identifying markers of metastatic spread.

\textbf{Technological developments: immune status}

The analysis of immune cell repertoires using NGS techniques is carried out by Hans-Jörg Warnatz.

We have developed a patent-pending technique for high-throughput paired amplification of antibody-encoding heavy and light immunoglobulin (Ig) chains from B cell populations ("PairedImmune" sequencing, international patent application PCT/EP2013/052328). The information on natively paired antibody-coding Ig gene pairs from populations of single B cells can be used for the generation of high-affinity antibodies useful as diagnostics and passive vaccines. We have successfully applied this technique to peripheral B cells from a healthy donor, and we are now analyzing B cell populations from a healthy donor before and after a tetanus vaccine boost.

Within the new European "TREGeneration" project that aims to develop within clinical trials new strategies for the treatment of graft versus host disease, a serious complication following bone marrow transplantation, our task is to characterize the T cell immune status of the patients and donors at key time points of the clinical trials.
Cooperation within the institute

We cooperate with the group of David Meierhofer for the proteomics analysis of melanoma and colon cancer, and worked with the bioinformatics group of Ralf Herwig and with the department of Martin Vingron in the development of methods for analyzing alternative splicing.

Planned development

The group plans to combine NGS data analysis with functional studies on ongoing projects addressing also long non-coding RNAs. The establishment of tumor microtissues for studying tumor-stromal interactions in melanoma will provide a model for interpreting molecular data and testing predicted drug responses. We also have a strong interest in high risk pediatric leukemias for linking the mutational pattern involving chromatin modifiers and components of the transcription initiation complex with specific transcriptional changes. Further, we plan single cell-based technologies for dissecting the heterogeneity of relapsed leukemias.
General information

Complete list of publications (2009-2015)

Group members are underlined.

2015


2014


2013

Otto Warburg Laboratory


2012


2011


2010


**Patent**


**PhD thesis**

Otto Warburg Laboratory

BMBF Research Group
Cell Signaling Dynamics

Established: 06/2014

Introduction

Cell signaling networks are complex and dynamic systems that enable living cells to sense and respond to changes in their immediate environment. One of the challenges in understanding cell signaling is to study how signaling proteins act together to determine cell fate decisions. Through mathematical modeling and experimental analyses, the long-term goal of the Cell Signaling Dynamics group is to study the mechanisms of cell signaling networks, by which extrinsic and intrinsic signals control cell

* externally funded
proliferation and differentiation at a systems level. Our current research focuses on the transforming growth factor-β (TGF-β) signaling, which is one of the most important signaling events as it regulates many cellular responses including cell proliferation, migration, and death. Abnormal activity of TGF-β has been connected to human diseases such as cancer, atherosclerosis, and fibrotic diseases of the kidney, lung, and liver.

TGF-β has different roles in the regulation of cell proliferation depending on the specific cellular context. In normal epithelial cells, TGF-β works as a tumor suppressor, as it induces sustained Smad signaling and inhibits cell proliferation. In contrast, TGF-β acts as a tumor promoter in cancer cell, as it triggers transient Smad signaling and stimulates tumor cell growth. This observation raises an important question: how can the same TGF-β signal induce very different signaling responses in different cellular contexts? Solving this question is crucial to understanding many dysfunctions of TGF-β signaling.

To address the aforementioned question, we work with three specific aims by investigating the spatiotemporal responses of TGF-β signaling across different cell lines and in single cells, which is summarized in Figure 1:

1. Identify the key feedback mechanisms that determine the duration of Smad signaling in different cell lines.
2. Study compartmental and temporal control of TGF-β signaling dynamics.
3. Investigate TGF-β signaling dynamics in single cells by quantitative single-cell measurements and stochastic modeling.

Figure 1: Systems biology approach to study TGF-β signaling

To address the aforementioned question, we work with three specific aims by investigating the spatiotemporal responses of TGF-β signaling across different cell lines and in single cells, which is summarized in Figure 1: (1) Identify the key feedback mechanisms that determine the duration of Smad signaling in different cell lines. (2) Study compartmental and tempo-
ral control of TGF-β signaling dynamics. (3) Investigate TGF-β signaling dynamics in single cells.

Scientific achievements and findings

Quantitative analysis of TGF-β signaling pathway

Mammalian cells can decode the concentration of extracellular TGF-β and translate this cue into appropriate cell fate decisions. We have studied, how variable TGF-β ligand doses quantitatively control intracellular TGF-β signaling dynamics, and how continuous ligand doses are translated into discontinuous cellular fate decisions on HaCaT cell proliferation. We developed data-driven mathematical models and analyzed the TGF-β signaling responses to different stimulation profiles of TGF-β. Our results indicated that TGF-β pathway displays different sensitivities to ligand doses at different time scales. While short-term TGF-β signaling responses are graded, long-term TGF-β signaling responses are ultrasensitive (Figure 2). These results suggested that long-term ultra-sensitive signaling responses in TGF-β pathway might be critical for cell fate decision making.

Figure 2: Model predictions of P-Smad2 signaling dynamics in response to different doses of TGF-β stimulations in HaCaT cells.
The development of model analysis tools

We have developed some software tools to perform model simulation, parameter estimation, sensitivity analysis and robustness analysis. In order to perform simulation and kinetic analysis of the models, one must know the parameter values for the modeled biological system. However, some parameter values cannot be experimentally measured in practice and their values are unknown. Therefore, parameter estimation is a common issue and very important for the mathematical modeling of biological systems. We developed a parallel implementation of a parameter estimation tool called SBML-PET-MPI. SBML-PET-MPI also parallelizes the algorithm of profile likelihood exploit for the identifiability analysis and confidence intervals analysis of the estimated parameters. The tool can run in the cluster or a server with multiple processors and provides good scalability with the number of processors. In addition, we have developed a tool to infer cellular regulatory networks based on Bayesian model averaging and linear regression methods.

Cooperation within the institute

Within the institute, the Cell Signaling Dynamics group is cooperating with David Meierhofer on absolute quantification of the phosphoproteins in TGF-β signaling. More internal cooperation is expected in the future.

Planned developments

Over the past decade, experimental studies have shown that most components of biological networks have substantial, unavoidable stochastic fluctuations in their levels and activities. As a result, even the same type of cells can react differently to the same environmental stimuli. Both experimental and modeling work of TGF-β signaling has been mainly conducted with the data from a population of cells in cell culture. However, averaging TGF-β signaling responses at cell population level can mask the variability of signaling dynamics within individual cells. Indeed, our preliminary analysis indicates that individual HaCaT cells have different nuclear Smad2 signaling in response to the same TGF-β stimulation. While the average responses from a population of cells show Smad2 nuclear accumulation in response to TGF-β, individual cells display large cell-to-cell variability. This observation raises questions about the source and mechanisms of cell-to-cell variability in TGF-β signaling responses. In addition, new emergent behaviors such as bimodal (all-or-none) and oscillation responses can exist at single cells, while they might be hidden by averaging from a population of cells. Therefore, it is important to quantitatively measure and investigate TGF-β signaling dynamics in single cells.
In the future, we plan to study cell-to-cell variability in TGF-β signaling with live-cell imaging, \textit{in situ} proximity ligation assay and optogenetics approaches. The temporal dynamics of TGF-β-regulated transcriptional outputs in single cells will be investigated with single molecule FISH and/or Fluidigm-based single-cell gene expression analysis. We expect to gain a predictive understanding of TGF-β signaling in single cells by integrating stochastic mathematical models with quantitative experimental data.

Besides our own research on TGF-β signaling dynamics, we are collaborating with several external labs on mathematical modeling of signaling networks in different systems. In cooperation with Xuedong Liu’s group at University of Colorado, Boulder, we will develop mathematical models in two projects: to investigate mitophagy signaling dynamics in response to distinct mitochondrial damage signals and to study the effect of mechanical force on the dynamic properties of wound responses by TGF-β and EGF signal. In addition, we will work with Vasso Episkopou’s group at Imperial College London to understand how targeted genes are specifically induced by different doses and timing of TGF-β signaling. Another collaboration is planned with Mei Li’s group at the Institute of Genetics and Molecular and Cellular Biology in Strasbourg, which will study the dynamic interactions of the cytokines induced by thymic stromal lymphopoietin (TSLP).

**General information**

**Complete list of publications (2012-2015)**

Group members are underlined.

**2014**


**2012**


Otto Warburg Laboratory
Minerva Group Neurodegenerative disorders
Established: 09/2008 – 02/2015

Introduction

Human life expectancy is steadily rising in industrialized western countries and as a result fatal late-onset neurodegenerative disorders, such as Alzheimer’s disease, Parkinson’s disease, or the polyglutamine-related diseases, are among the leading causes of disability and death, representing one of the major challenges of today’s modern medicine. Millions of people worldwide suffer from these devastating disorders or are at risk, and a marked rise in the economic and social burden caused by these disorders will be noticed over the upcoming decades. Even though these diseases are quite common, the mechanisms responsible for their pathologies are in most cases still poorly understood and effective preventative therapies are

* externally funded
currently not at hand. For the heritable forms of these neurodegenerative disorders, linkage studies have led to the discovery of the causative genes. Current knowledge of the underlying molecular mechanisms accountable for the observed neurodegenerative processes was gained mainly from studying inherited disease variants, and these resulted in the identification of genetic and metabolic factors modulating disease onset and progression. Of note, similarities in the clinical and neuropathological features have suggested that neurodegenerative diseases may share similar mechanisms of pathogenesis related to abnormal protein folding, aggregation, cellular dysfunction and cell death. In consequence, a comprehensive characterization of the molecular mechanisms implicated in the clinical heterogeneity of specific neurodegenerative disorders should help in defining the complete picture of potential pathomechanisms.

The main research interest of my group was to elucidate molecular mechanisms contributing to neurodegenerative processes in the polyglutamine disorder spinocerebellar ataxia type 2 (SCA2) and whether and how these pathways can be correlated to other polyglutamine or neurodegenerative disorders, such as spinocerebellar ataxia type 1 (SCA1) and amyotrophic lateral sclerosis (ALS), on different cellular levels by combining yeast genetics, humanized yeast models, and functional genomic approaches. Moreover, we were interested in studying the biology of cell stress responses as well as stress granules (SGs) and P-bodies, central self-assembling structures regulating mRNA metabolism, and their relevance in age-related human disorders including cancer.

**Ataxin-2 proteins: role in disease and stress granule biology**

**Splicing of ataxin-2 transcripts**

To gain insight into the cellular function of ataxin-2 (ATXN2), the disease-causing protein in SCA2, we have performed comprehensive protein-protein interaction studies using the yeast-2-hybrid system. These studies revealed that ATXN2 is found in association with a number of proteins implicated in the cellular mRNA metabolism, amongst others. Further analyses showed that ATXN2 and a number of its protein interaction partners are components of SGs, cellular sites assembling in mammalian cells as response to specific cellular stresses that are central for regulating and controlling mRNA degradation and translation. In particular, we were interested in investigating the interaction between ATXN2 and the splicing factor FOX-2. The rationale for this was based on the finding that FOX-2 is part of a main protein interaction hub in a network related to human inherited ataxias. In addition, we discovered that the SCA2 gene bears two putative FOX-binding sites ~ 30-100 nucleotides downstream of exon 18 in the
ATXN2 transcript, suggesting that FOX-2 could potentially be involved in ATXN2 pre-mRNA splicing. RNAi experiments revealed that this splicing event indeed depends on FOX-2 activity, since reduction of FOX-2 levels led to an increased skipping of exon 18 in ATXN2 transcripts. In this line, we also discovered that the localization and splicing activity of FOX-2 is affected in the presence of nuclear ataxin-1 inclusions, a pathological hallmark in SCA1. Most striking, we observed that splicing of ATXN2 transcripts is affected in the presence of these nuclear ataxin-1 inclusions as well. Since ATXN2 has been shown to modulate SCA1 pathogenesis, it is quite tempting to speculate that alterations in ATXN2 transcripts and their cellular consequences could affect SCA1 pathogenesis. Moreover, we observed recently that this ATXN2 splicing event is regulated by specific stress conditions. Therefore, further insight into the cellular significance of this transcript variant and its regulation and whether and how this relates to mechanisms underlying disease would be an interesting aspect in the future.

**Humanized yeast and synthetic lethality analyses**

Much of our knowledge on protein misfolding, aggregation, and toxicity in human neurodegenerative disorders was derived from genetic and functional approaches exploiting model organisms. The baker’s yeast *Saccharomyces cerevisiae* represents a well-established model system in this research field, since it combines a high level of conservation between its cellular processes and those of mammalian cells and represents a powerful genetic system with a wealth of tools for genome-wide analyses. Over the years, global genetic screens in yeast led to the unbiased identification of cellular processes implicated in a number of neurodegenerative disorders, e.g. Alzheimer’s disease, Parkinson’s disease, polyglutamine diseases, and amyotrophic lateral sclerosis, as well as modifiers of aggregation of the respective disease proteins. Most importantly, these findings have been translated to human cell lines and transgenic animal models, demonstrating the value of these genetic yeast approaches. Therefore, we decided to analyze expression of human ATXN2 in yeast in more detail. For this, we generated respective constructs with inducible expression of full-length normal ataxin-2 as well as N- and C-terminal fragments thereof. The rationale behind analyzing ataxin-2 fragments was the following. As reported by other colleagues, a 42 kD N-terminal ATXN2 cleavage product comprising the polyglutamine domain was detectable in brain material of SCA2 patients, which was elevated in the primarily degenerating Purkinje cells in comparison to control material. In this light, a C-terminal ATXN2 cleavage product with a molecular mass of 70 kD has also been detected in SCA2 brain material as well as in cell lines, in which additional C-terminal regions with different molecular masses were also detectable with poten-
tial impact in disease (Huynh et al., J Clin Neurophysiol 1999). Finally, and most importantly, recent results show that ATXN2 intermediate-length polyglutamine expansions cause activation of caspase 3 generating N- and C-terminal ATXN2 cleavage products under stress conditions (Hart & Gittler, J Neurosci 2012), suggesting that these cleavage products might also have an impact on ALS pathogenesis.

In order to further gain insight into the physiological function of ATXN2 in RNA-processing pathways and the molecular mechanisms of its modifier effect, we performed directed unbiased genetic synthetic lethality approaches based on respective humanized yeast model systems for SCA1, SCA2, and ALS. Since ATXN2 function modulates SG and P-body presence in mammalian cells, and deficiency of SGs and P-bodies might add to pathogenesis in some neurodegenerative disorders, we set out to express full-length ATXN2 protein with normal polyQ domain in a panel of 23 yeast strains deficient for SG or P-body components from the Euroscarf deletion library. We observed that expression of full-length ATXN2 protein with a normal polyQ domain was slightly toxic in five yeast deletion strains deficient for SG or P-body formation. We also performed this approach with ATXN1, the disease-causing protein in SCA1; however, we did not observe any growth defects due to expression of ATXN1 in WT and the 23 yeast deletion strains tested. Consequently, we performed such synthetic lethality experiments with TDP-43. Interestingly, we observed that TDP-43-induced toxicity was reproducibly slightly stronger in two yeast deletion strains, in which expression of ATXN2 results in growth defects as well. This finding encouraged us to perform a more defined analysis, in which we also included the respective N- and C-terminal cleavage products to have a more complete picture of ATXN2 function. Interestingly, expression of the C-terminal ATXN2 caspase-3 cleavage product showed a strong reduction in growth of these two yeast strains, whereas expression of the N-terminal ATXN2 caspase-3 cleavage product had no significant effect on growth. Consequently, we performed further validation experiments in yeast including complementation analyses with yeast proteins as well as with the human homologs. Moreover, we used the self-made yeast strains for further independent validation, and obtained similar results. In parallel, we have performed the unbiased genetic approaches with mutant ATXN2 as well as the N-terminal cleavage product of ATXN2 with an expanded polyQ domain, and did obtain same results compared to normal full-length proteins or the N-terminal cleavage product of ATXN2 with normal polyQ domain, indicating surprisingly no effect of the expanded polyQ domain within ATXN2 in this experimental setting.

In addition, we exploited the yeast system to investigate the effect of human ATXN2 overexpression on stress granule assembly in yeast. For this, constitutively expressed Redstar-labeled ATXN2 N- and C-terminal cleavage products were expressed in a chromosomally Pbp1-GFP tagged strain.
and the effect on Pbp1 localization under normal and stress conditions was investigated. Treatment of cells with either NaN3 or heat resulted in normal SG-formation in control and N-terminal ATXN2 expressing cells. Notably, expression of the C-terminal ATXN2 domain causes Pbp1-GFP to co-localize with C-terminal ATXN2-positive foci rather than smaller SGs under both conditions, thereby interfering with SG-formation. To test whether the C-terminal ATXN2 foci represent SGs, we also conducted cycloheximide experiments. Those showed that SG formation is indeed inhibited by cycloheximide in control cells as well as cells expressing the N-terminal ATXN2 domain. Notably, C-terminal ATXN2-positive foci still assemble under these conditions, and most importantly, Pbp1-GFP remains co-localized. The same set of experiments was performed in Dhh1-GFP-tagged strain and gave same results under the chosen stress conditions. These findings indicate that expression of the C-terminal ATXN2 domain interferes with SG assembly in yeast.

Ataxin-2-like: role in stress granule/P-body biology

Since valuable information about the molecular mechanism contributing to disease pathogenesis in neurodegenerative disorders have been gained through studying the cellular function of paralogs of neurodegenerative proteins, we also included the ATNX2 paralog, termed ataxin-2-like (ATXN2L), in our studies. We discovered that ATXN2L and ATXN2 functionally overlap regarding their function in the cellular mRNA metabolism. We demonstrated that ATXN2L associates with known interaction partners of ATXN2, the RNA-helicase DDX6 and the poly(A) binding protein, and with ATXN2 itself. Moreover, we showed that ATXN2L is important for the proper formation of two key cellular structures, SGs and P-bodies, which are important for regulating and controlling mRNA translation, degradation and stability. Sole ATXN2L overexpression induces SGs, while a reduction of SGs was detected in case of a low ATXN2L level. On the other hand, we observed that ATXN2L overexpression as well as its reduction has an impact on the presence of microscopically visible P-bodies indicating a role for ATXN2L in the regulation of SGs and P-bodies.

Over the last few years evidence has been provided that posttranslational modifications of several SG components that are involved in the primary nucleation step play a significant role in the regulation of SG assembly / disassembly and function. Arginine methylation is a common posttranslational modification of RNA-binding proteins taking place in the arginine-glycine-rich (RGG) domains of these proteins. Since recent proteomic mass spectrometry approaches identified ATXN2L as being arginine methylated, we set out to investigate whether methylation of ATXN2L might be important for its localization. We demonstrated that ATXN2L is asymmetrically dimethylated in vivo exploiting specific dimethyl arginine
antibodies. Of note, the nuclear localization of ATXN2L to splicing speckles is altered under methylation inhibiting cellular conditions as well as under a reduced PRMT1 level. Furthermore, we demonstrated that ATXN2L associates with PRMT1. However, arginine methylation within the RGG motifs of ATXN2L is not essential for its localization to SGs, and we observed that ATXN2L is also part of SGs under methylation inhibiting cellular conditions as well as under PRMT1 reduction. These results indicate that methylation of ATXN2L seems not mandatory for its SG localization.

Relevance of stress granules and P-bodies in cancer

In cancer therapy clinical applications of chemotherapeutics are often limited due to drug resistance, for which the underlying mechanisms are not completely understood. Recently, stress granules have been linked to apoptotic processes and to cellular pathways contributing to chemotherapy resistance in cancer treatment. To dissect the underlying mechanisms in more detail, we have performed yeast genetic approaches and functional yeast studies. In particular, we have performed global yeast screens utilizing the yeast deletion strain library to identify strains sensitive for particular chemotherapeutics. Of note, these unbiased yeast screens identified candidate genes implicated in ribosomal function, tRNA modification, transport, and processing of mRNAs, amongst others. Moreover, some gene products are either components of stress granules or P-bodies per se or are involved in mediating interplay between stress-activated pathways and apoptosis.

In addition, we investigated in a direct approach, whether 5-fluorouracil treatment may possibly be related to SG assembly. This widely used therapeutic activates protein kinase PKR and leads to phosphorylation of eIF2, which is a criterion for SG assembly. We demonstrated that 5-fluorouracil treatment of cells indeed results in de novo assembly of stress granules. Moreover, 5-fluorouracil affects SG assembly under stress conditions as well as disassembly. Remarkably, RACK1, a protein mediating cell survival and apoptosis is sequestered into 5-fluorouracil-induced SGs. Interestingly we discovered that 5-fluorouracil metabolites incorporating into RNA, not DNA, caused SG assembly. Moreover, we demonstrated that other RNA incorporating drugs also cause SG assembly. Accordingly, incorporation of chemotherapeutics into RNA may result in SG assembly with potential significance in chemoresistance and cancer biology.
Cooperation within the institute

Within the institute, the Neurodegenerative Disorders group closely cooperated with the following people and their groups: Zoltán Konthur, Michal Schweiger, Lars Bertram, Holger Klein/Martin Vingron, Hans Lehrach, and David Meierhofer.

Cooperation outside the institute

Outside the MPIMG, we cooperated with the following labs:

- Matteo Barberis/Edda Klipp, Humboldt University, Berlin, Germany
- Georg Auburger, Dept. of Neurology, Goethe University Medical School, Frankfurt, Germany
- Jörg Isensee/Tim Hucho, University of Cologne, Germany

General information

Complete list of publications (2009-2015)

Group members are underlined.

2015

2014

2013


2012


2011


2010


2009
Habilitation/State doctorate


PhD theses

Christian Kähler: Die Rolle zellulärer Stressantworten bei der Chemoresistenzen gegenüber 5-Fluorouracil. Freie Universität Berlin, 2014


Student theses


Marcel Schulze: Untersuchungen zur Interaktion von proteolytischen APP-Produkten und dem Transkripti-


**Teaching activities**

Practical course: *Physikalische Übungen für Pharmazeuten*, Universität Hamburg, winter term 2009/10

Single lecture in the series „Gene und Genome: die Zukunft der Biologie“; Freie Universität Berlin, winter term 2009/10; winter term 2010/11; winter term 2011/12; winter term 2012/13

In addition, several students have been supervised during their internships in the group.
Max Planck Fellow Group
Efficient Algorithms for Omics Data
Established: 07/2014

Head
Prof. Dr.-Ing. Knut Reinert
Freie Universität Berlin
Institut für Informatik
Takustrasse 9, 14195 Berlin
Phone: +49 (0)30 8387 5222
Fax: +49 (0)30 8387 5218
Email: knut.reinert@fu-berlin.de

PhD students
Stefan Budach (since 09/15, IMPRS-CBSC, jointly with Annalisa Már-sico)
Christopher Pockrandt (since 09/15, IMPRS-CBSC)
Jongkyu Kim (since 01/15, IMPRS-CBSC)
Hannes Hauswedell (since 08/14)
Xiao Liang (since 10/13, IMPRS-CBSC)
Leon Kuchenbecker (since 01/12, IMPRS-CBSC, jointly with Peter N. Robinson)
Temesgen Dadi (since 10/13, IMPRS-CBSC)
Dr. Sandro Andreotti (10/08-09/14, IMPRS-CBSC)
Dr. Enrico Siragusa (08/10-08/14, IMPRS-CBSC)
Dr. Anne-Katrin Emde (05/08-04/12, IMPRS-CBSC jointly with Stefan Haas)
Dr. Birte Kehr (05/08-04/12, IMPRS-CBSC)
Dr. Chris Bielow (10/07-07/11, IMPRS-CBSC)
Introduction

Knut Reinert is professor of Algorithmic Bioinformatics at the Freie Universität (FU) Berlin. His cooperation with the MPIMG started already in 2002 when he accepted an appointment at FU Berlin. At the beginning, Reinert and his partners at the MPIMG concentrated on the analysis of mass spec data with scientists from the Dept. of Vertebrate Genomics (Hans Lehrach), as well as read mapping and detection of genomic deletions and copy number variations with the Dept. of Computational Molecular Biology (Martin Vingron). Since 2004, he has been one of the key faculty from the university side for the establishment of the International Max Planck Research School for Computational Biology and Scientific Computing (IMPRS-CBSC). In this context he jointly supervises many PhD students together with Martin Vingron and many scientists from his department of Computational Molecular Biology. In addition, Knut Reinert’s expertise on algorithmics is extremely helpful for the MPIMG for the analysis and interpretation of next generation sequencing (NGS) data.

In order to further strengthen the close ties between the Reinert lab and the MPIMG, he has been appointed as a Fellow of the Max Planck Institute for Molecular Genetics in 2014 (limited to five years initially), where he heads the Efficient Algorithms for Omics Data group, additionally to his group on Algorithmic Bioinformatics at the Freie Universität Berlin (FUB). At the same time the FU Berlin further strengthened the tie to the MPIMG by appointing Annalisa Marsico as a Junior Professor for High-throughput genomics, a step supported by Knut Reinert as a mentor and collaborator in new projects. Although the fellow group was established only mid-2014, this report starts in 2012 to give a better overview.

Scientific overview

The long-term goal of the Reinert lab is to bridge the existing gap between advanced results in algorithmic research and their practical application as bioinformatics tools for real world data. The group is achieving this by working on solving well-posed computational problems arising in the context of -omics data analysis, mostly for next generation sequencing (NGS) data as well as proteomics data produced by HPLC/MS methods.

The applications of these new data are manifold. In NGS analysis they range from DNA sequencing for reference-guided assembly or for ChIP-Seq, over sequencing of RNA transcripts (RNA-Seq) to sequencing and mapping bisulfite-treated reads. In proteomics various protocols for producing data for qualitative and quantitative analysis are constantly being developed and need adequate algorithms for processing.
The aim of our groups is not solely to develop algorithms and tools for those applications, but rather to dissect them into well-defined algorithmic components. We then generalize those components as much as possible and implement them efficiently and robustly in software libraries. The two most mature projects in the Reinert lab are SeqAn (http://www.seqan.de) and OpenMS (http://www.openms.de), jointly with the University of Tübingen and the ETH Zurich. Both libraries, SeqAn and OpenMS, have been used for solving problems at the MPIMG, and have been further extended by many students, in particular Marcel Schulz, Jonathan Göke, Ann-Katrin Emde and Alexandra Zerck.

Being a Max Planck Fellow will benefit the Reinert lab not only by being able to hire more PhD students. It rather deepens the existing collaborations with groups at the MPIMG and will allow the lab access to experimental data generated in the MPIMG wet labs.

### Read mapping and variant detection

[Anne-Katrin Emde, Enrico Siragusa, Sandro Andreotti, Leon Kuchenbecker, Jongkyu Kim]

Read mapping, a seemingly easy problem, was initially solved by heuristically accelerating standard q-gram-based methods due to the demands on speed for large NGS data sets. We addressed the problem of optimally solving read mapping in several publications. Our last two implementations, Masai (Siragusa et al, Nucleic Acids Res 2013) and Yara, are 2–5 times faster and more accurate than Bowtie 2 and BWA. The novelties of our read mappers are the use of filtration with approximate seeds and a method for multiple backtracking (Masai) as well a mapping by strata using an adaptive filtration scheme (Yara). In the future, we will address the problem of mapping reads to pan-genomes (see below) and work on even faster filtration and seeding schemes using the bi-directional Burrows-Wheeler transform and adaptive, gapped seeds.

![Figure 1: Duplication and translocation alignment patterns](image)

On the left side, we show alignment patterns of a reference g (upper sequence) with a donor genome d where sequence Block 2 is either duplicated (upper figure) or translocated within the donor genome. In a read-to-reference alignment, read r1 indicates the duplication or translocation event (dup_impr(v, z)) through the different order of the read parts within the reference. We also observe pseudodeletions for both variants (highlighted on right side) through reads r2 and r3: an upstream duplication of Block 2 creates a pseudodeletion del(w,z), an upstream translocation pseudodeletions del(w,z) and del(v,w). From observing both dup_impr(v,z) and del(w,z), we infer the duplication dup(v, w, z). If we also observe del(v,w), we infer the translocation tra(v, w, z).
The landscape of structural variation (SV) including complex duplication and translocation patterns is far from resolved. SV detection tools usually exhibit low agreement, are often geared toward certain types or size ranges of variation and struggle to correctly classify the type and exact size of SVs. We were successful in implementing two algorithms in tools named Gustaf (Emde et al., Bioinformatics 2014) and Basil/Anise (Holtgrewe et al., Bioinformatics 2015). Gustaf (Generic mUlti-SpliT Alignment Finder) is a sound generic multi-split SV detection tool that detects and classifies deletions, inversions, dispersed duplications and translocations of >30 bp. Our approach is based on a generic multi-split alignment strategy that can identify SV breakpoints with base pair resolution. We show that Gustaf correctly identifies SVs, especially in the range from > 30 to 100 bp, which we call the next-generation sequencing (NGS) twilight zone of SVs, as well as larger SVs (> 500 bp). Furthermore, we are able to correctly identify size and location of dispersed duplications and translocations, which otherwise might be wrongly classified, for example, as large deletions (see also Figure 1). These methods have also been applied by Stefan Haas in the context of genetics and cancer genomics (work of Mike Love and Ruping Sun).

For large insertions, we developed approaches for detecting insertion breakpoints and targeted assembly of the insertions from HTS paired data. After detecting breakpoints with the tool Basil, we conduct a targeted, local assembly with the tool Anise. One major point of Anise is that we employ a repeat resolution step on near identity repeats that are hard for assemblers. This results in far better reconstructions than obtained by the compared methods like MindTheGap.

In addition to our work in SV detection, we also considered the problem of correctly identifying different isoforms of expressed genes obtained from mixture data from RNA-seq experiments. The Cidane (Canzar et al., RECOMB 2015) framework for genome-based transcript reconstruction and quantification from RNA-seq reads assembles transcripts with significantly higher sensitivity and precision than existing tools, while competing in speed with the fastest methods. In addition to reconstructing transcripts ab initio, the algorithm also allows to make use of the growing annotation of known splice sites, transcription start and end sites, or full-length transcripts, which are available for most model organisms.

The tool Imseq (Kuchenbecker et al., Bioinformatics 2015) implements a method to derive clonotype repertoires from next generation sequencing data with sophisticated routines for handling errors stemming from PCR and sequencing artefacts. The application can handle different kinds of input data originating from single- or paired-end sequencing in different configurations and is generic regarding the species and gene of interest.
(Pan)-Genome comparison and metagenomics

[Birte Kehr, Sandro Andreotti, Temesgen Dadi, Hannes Hauswedell, Christopher Pockrandt]

The plummeting costs of sequencing technologies have made it possible to investigate data sets stemming from a mixture of different genomes as a whole (a so called metagenome) and to sequence large populations of individuals from a species or clade (a so called pangenome). For those applications we worked on basic algorithms like mapping a set of metagenomic reads to protein databases, and on identifying graph-based data structures to represent similar sequences succinctly, while allowing fast computations on the representation.

Our tool Lambda (Hauswedell et al., Bioinformatics 2014) is an alternative for BLAST in the context of sequence classification. Lambda often outperforms the current best tools at reproducing BLAST’s results and is the fastest compared with the current state of the art at comparable levels of sensitivity.

For representing a set of similar strings (i.e. a pangenome), we provide a data type called Journaled String Tree (JST) (Rahn et al., Bioinformatics 2014) (see also Figure 2) that exploits data parallelism inherent in the set by analyzing shared regions only once. In real-world experiments, we can show that algorithms, that otherwise would scan each reference sequentially, can be sped up by a factor of 115.

Concerning the representation of multiple genomes we surveyed several graph-based data structures (Kehr et al., BMC Bioinformatics 2015) for genome-sized alignment and compared the structures of those graphs in terms of their abilities to represent alignment information (e.g. see Figure 3). We show that crucial pieces of alignment information, associated with inversions and duplications, are not visible in the structure of all
graphs. Based on these findings, we outline a conceptual framework for graph-based genome alignment that can assist in the development of future genome alignment tools.

We also investigated the structure of real and computed genome-sized alignments induced by gene gain, loss, duplication, chromosome fusion, fission, and rearrangement. When gene gain and loss occurs in addition to other types of rearrangement, breakpoints of rearrangement can exist that are only detectable by comparison of three or more genomes. A very large number of these “hidden” breakpoints can exist among genomes that exhibit no rearrangements in pairwise comparisons. Developing an extension of the multi-chromosomal breakpoint median problem to genomes that have undergone gene gain and loss, we demonstrate that the median distance among three genomes can be used to calculate a lower bound on the number of hidden breakpoints. We applied our approach to measure the abundance of hidden breakpoints in simulated data sets under a wide range of evolutionary scenarios and could demonstrate that hidden breakpoint counts depend strongly on relative rates of inversion and gene gain/loss. Applying current multiple genome aligners to the simulated genomes we show that all aligners introduce a high degree of error in hidden breakpoint counts, and that this error grows with evolutionary distance in the simulation, which suggests that hidden breakpoint error may be pervasive in genome alignments (see Kehr et al., WABI 2012).
The reconstruction of the history of evolutionary genome-wide events among a set of related organisms is of great biological interest since it can help to reveal the genomic basis of phenotypes. However, a high sequence similarity often does not allow one to distinguish between orthologs and paralogs. We show how to infer the order of genes of (a set of) families for ancestral genomes by considering the order of these genes on sequenced genomes in the evolutionary model of duplications and loss. Our branch-and-cut algorithm solves the two species small phylogeny problem about 200 times faster than the previously published method (see Andreotti et al., J Comput Biol 2013).

In the upcoming years we will work on analysis of 16S RNA metagenomics samples and expand our repertoire for RNA analysis.

**Genomic RNA analysis**

[Stefan Budach, Christopher Pockrandt]

So far, SeqAn does not offer direct support for the analysis of RNA sequences with associated structure of base pair probabilities. However, it is well known that secondary structure conservation is a crucial element for functional conservation in RNAs. Especially IncRNAs show little evidence for evolutionary conservation, while there is evidence of structural conservation. Hence we plan, based on previous work, to infer structural motifs of IncRNAs, cluster them into functional groups and devise methods to identify such structural motifs fast in novel genomic sequence. This work will be conducted in collaboration with the group of Annalisa Marsico (FU Berlin and MPIMG) and a joint PhD student (Stefan Budach). Here it will be especially helpful to have access to the experimental resources at the MPIMG, for example to determine RNA-protein interactions.

**Proteome analysis with HPLC-MS**

[Xiao Liang, Chris Bielow]

In the context of HPLC-MS we worked during the report period in several applied projects on pipelines for quantitative HPLC-MS analysis. In the next years, we will work on more basic research questions, namely on the problem of protein identification resp. protein inference from a set of HPLC-MS measurements. The traditional algorithmic approach for protein identification disregards the inherent information wealth available in a set of mass spectra. By blindly splitting the identification of fragment spectra into several parts, it is not possible that peptide identifications re-enforce each other and evidence from one spectrum supports the identification of a similar one. Further, the MS1 spectra carry more information that is of
relevance for identification than just the parent mass of the fragment ion. This information such as further supporting mass positions or retention time information regularly remains either unobserved or is only used out of context and thereby lost for further steps.

**General information about the whole group**

**Complete list of publications (2012-2015)**

Research group members are underlined.

**2015**


**2014**


2013


2012


**Selected invited talk (Knut Reinert)**

*Bridging the gap: Enabling top research in translational research*. Australasian Genomics and Technologies Association (AMATA) 2013, Surfers Paradise, Australia, 10/2013

**PhD theses**

MPIIMG students are underlined.

Sandro Andreotti: *Optimization algorithms for computational proteomics and next generation sequencing*. Freie Universität Berlin, 2015

Manuel Holtgrewe: *Engineering of algorithms for personal genome pipelines*. Freie Universität Berlin, 2015


Birte Kehr: *Efficient algorithms for genome comparison*. Freie Universität Berlin, 2013

Alexandra Zerck: *Optimal precursor selection for LC-MS/MS based proteomics*. Freie Universität Berlin, 2013


Anne-Katrin Emde: *Next generation sequencing algorithms: from read mapping to variant detection*. Freie Universität Berlin, 2012


**Teaching activities**

Since 2003, I have been teaching various lectures on Bioinformatics, NGS data analysis, and proteomics as part of the BSc and MSc programs at the Freie Universität Berlin. In addition, I have been supervising many MSc and BSc students, both in German and English. Furthermore, I am the liaison officer of a group of students of the “Studienstiftung des deutschen Volkes”.

Lecture titles include:
- Discrete Mathematics
- Algorithms
- Algorithms and data structures for bioinformaticians
- Proteomics
- Sequence analysis
- Algorithmic Bioinformatics
Department of Vertebrate Genomics

Now: Emeritus Group Vertebrate Genomics

Head
Prof. Dr. Hans Lehrach, director emeritus since 12/14
Phone: +49 (0)30 8413-1220
Fax: +49 (0)30 8413-1380
Email: lehrach@molgen.mpg.de

Scientific Management
Ingrid Stark* (OncoTrack, since 12/14 half time Treat20plus)
Phone: +49 (0)30 8413-1682
Fax: +49 (0)30 8413-1380
Email: stark@molgen.mpg.de

Wilfried Nietfeld* (until 11/14)
Phone: +49 (0)30 8413-1682
Fax: +49 (0)30 8413-1380
Email: nietfeld@molgen.mpg.de

Heads of project groups
Silke Rickert-Sperling*
(11/99-02/11, guest 03/11-10/17)
Lars Bertram*
(10/08-11/14, guest 12/14-11/16)
Christoph Wierling*
(01/09-11/14, guest 12/14-11/15)

Michal Ruth Schweiger*
(01/07-11/14, guest 12/14-12/15)
Heinz Himmelbauer*
(03/02-09/14, guest 10/14-12/15)
Albert Poustka*
(01/02-11/12, guest 01/13-12/15
Bodo Lange*
(05/03-12/11, guest 01/12-12/15)
Margret Hoche (01/02-11/14)
Ralf Herwig (03/01-11/14)
Marie-Laure Yaspo (10/95-11/14)

Scientists and postdocs
Ute Postel (since 04/15)
Philipp Peters (since 01/15)
Juliane Dohm (since 10/08)
Martin Kerick (since 08/08)
Hannah Müller (since 06/08)
Günter Zehetner (since 01/95)
Christina Grimm*
(12/02-05/11, guest 06/11-07/16)
Melanie Isau*
(04/14-02/15, guest 03/15-12/15)
Moritz Schütte*
(11/12-02/13, guest 03/13-12/15)
Tatiana Borodina* (06/08-12/11, guest 01/12-11/15)
Thomas Keßler*
(07/07-11/14, guest 12/14-11/15)
Johanna Sandgren*
(guest 08/14-11/14)
Moritz Beber (02/14-11/14)
Christine Jandrasits (12/13-11/14)
Thomas Risch (08/12-11/14)
Christopher Hardt (03/12-11/14)
Vyacheslav Amstislavskiy
(12/11-11/14)

PhD students
Vera Rykalina*
(11/10-04/13, guest 05/13-10/16)
Sabrina Grasse* (guest 07/15-07/16)
Matthias Mark Megges*
(09/10-02/14, guest 03/14-12/15)
Sophia Schade* (10/10-11/14, guest 12/14-12/15)

* externally funded
Introduction

Structure and organisation of the department

Since the last evaluation in October 2012, the Department of Vertebrate Genomics reduced in size and finally closed in November 2014. The research group of Marie-Laure Yaspo continues as an independent research group of the Otto Warburg Laboratory (see report Yaspo). Ralf Herwig’s group joined the Department of Computational Molecular Biology (Martin Vingron) in December 2014, and Margret Hohe continues her work on the haplotype/diplotype architecture of human genomes, also within Martin Vingron’s department. Christoph Wierling and Albert Poustka have joined Alacris Theranostics GmbH, the spinout company developed from the department. Michal Schweiger and Lars Bertram have accepted Professorships at the University of Cologne and the University of Lübeck, respectively.

I continue to maintain an Emeritus laboratory at the institute. In addition, I serve as CEO and scientific director of the Dahlem Centre for Genome Research and Medical Systems Biology (DCGMSB), a fully grant-funded institution, and act as chairman of the Scientific Advisory Board (Beirat) of Alacris Theranostics.

Research concept

Over the last decade the focus of the core groups of the department has been on the development of computer models of individual cancer patients, and their use to identify the optimal therapy. In parallel, the same concepts can be used to virtualize steps in drug development, which in itself could have an enormous impact on this area. In taking such an approach, we are following the example of essentially all other areas in which we face complex problems with dangerous and/or expensive consequences. We do not build skyscrapers to watch them fall over during the first autumn storms; instead we first test computer models of all possible designs against hur-
ricane scale winds. We no longer build cars to find out in real accidents that they are “unsafe at any speed”, but carry out extensive “virtual crash tests” in silico, making unavoidable mistakes safely, cheaply and quickly on computer models.

This is, however, currently not the case in medicine and drug development. Here, therapies are selected based on the response of groups of patients, with no guarantee that the individual will benefit from the therapy he/she gets. Drugs are primarily tested in large, costly and, in a sense, also dangerous, clinical trials, leading to high failure rates and very high costs for the few drugs that are finally approved. Every patient is different. Every tumor and potentially every tumor cell can be different and react differently to the drugs the patient receives. Many patients therefore do not respond or are even harmed by the drugs they take. Many clinical trials continue to fail, at high cost, and sometimes with negative consequences for participating patients.

To change this, we have increasingly focused our work on what we consider to be the only realistic chance to really improve this situation: the establishment of “virtual patient” models based on a detailed molecular characterization of every patient, allowing the testing of drug therapies on the individual patient in silico, rather than potentially harming real patients. To support this goal, the group of Marie-Laure Yaspo has carried out detailed analyses of tumors and patients for a number of projects and has established a bioinformatics infrastructure (see also separate report). Christoph Wierling (now carrying on this work at Alacris Theranostics) has continued to develop PyBios3, the object-oriented modeling engine used to establish the tumor/patient models. Related work has also been a major focus of the groups of Ralf Herwig, concerned with the development of new algorithms and a functional genomics database (ConsensusPath database, CPDB), but also contributing to the analysis of specific tumors; and Michal Schweiger, who is working on the molecular characterization and the epigenetic regulatory mechanisms in different cancer types. Other relevant work has been carried out by the groups of Margret Hoehe, focused on the analysis of the haplotype/diploidtype architecture of human genomes and its implications for genome biology and individualized medicine; Lars Bertram, interested in the mapping and characterization of complex disease genes, predominantly in the field of neuropsychiatric diseases; and Albert Poustka, focusing on the evolution of multicellular organisms, as well as on the disease mechanisms of autism disorders.

After the termination of my department, major parts of the core project had to be moved to other institutions (Alacris, DCGMSB), since I have not been able to apply for the follow-up grants needed to continue the work from this institute, and key group leaders required for the project were unable to continue their work here. A large component of this work, i.e.
Figure 1: Development of Virtual Patient Models.
Publicly available resources representing the sum of knowledge on cancer, cell signaling and drug action (e.g. dissociation constants and molecular targets) are used to construct a large-scale mechanistic model of cellular signaling. This generic large scale signaling network is personalized with omics data (e.g. transcriptome/exome/proteome) from individual patient tumors / cell lines / experimental tissues. The effects of identified molecular alterations on pathway function and cross-talk can then be simulated using a mechanistic modeling approach. Response to molecularly targeted drugs (single or in combination) can be predicted, through establishment of a molecular readout (as a proxy for phenotypic effects, e.g. cell proliferation, senescence and migration, apoptosis induction), allowing identification of the optimal treatment. Applications range from personalized medicine in the clinic to virtual clinical trial scenarios, enabling *in silico* testing of drug effects (single or combination) and potential side effects on individual or large patient (or preclinical model) cohorts.
the detailed molecular analysis of patient samples and data analysis, does however continue at the institute (mostly in collaboration with the group of Marie-Laure Yaspo).

**Scientific methods and achievements**

Since the biological processes relevant to a disease, but often also the mechanisms of action of the potential therapies are exceedingly complex, with many factors influencing both effects and side effects of any given therapy on the individual patient, patients are treated with drugs, which help in some cases, but in others may cause more damage than good. This element of “experimentation” is likely to be unavoidable in such complex situations. In building predictive in silico models of the patient, we “virtualize” this “experimentation”, exploring effects and side effects of hundreds (or potentially thousands) of different drugs or drug combinations on the patient in silico, and only exposing the real patient to the therapy optimal for him or her (see Figure 1).

Every patient is different and can react differently to a specific therapy. Every patient will therefore have to be characterized in sufficient detail to be able to ultimately predict this different response. In a number of projects (e.g. OncoTrack and Treat20plus) we have been taking such an approach. Sequencing of the exome/genome and transcriptome (and in some cases proteomes) of tumors, and the exome or genome of the patients is conducted. As outlined in the report of Marie-Laure Yaspo, sequences are then interpreted to derive information on biologically relevant parameters, including mutations, deletions/insertions, copy number variation, regions of LOH (loss-of-heterozygosity), abundance changes in coding/non-coding sequences and microRNAs, splice variants, allele specific expression patterns, abundances of proteins and protein modification forms. This information is then used to individualize models of the normal biological networks in the different cell types and organs involved in the disease process. These objects then automatically generate systems of differential equations, which can then be solved numerically. Although some (but not all) of the starting concentrations in these equations can be set, based on the results of the -omics characterization, the parameters are generally unknown. For smaller models we have therefore used a Monte Carlo strategy, carrying out multiple modeling runs of all relevant conditions using parameters drawn from probability distributions ideally reflecting any previous knowledge. As the models increase in size, this is being complemented by parameter fitting strategies, based on comparing predictions to experimental results (e.g. the drug response of xenografts with known exome and transcriptome data), in collaboration with Alacris Theranostics (especially Alexey Shadrin und Christoph Wierling), and the groups of Marie-Laure Yaspo and Jan Hasenauer (Helmholtz Centre, Munich). To allow this type of analysis on increasingly large models, with rapidly increasing datasets,
major effort is focused on optimizing algorithms to increase the numerical performance of many of the components of this analysis, such as numerical integration of very large systems of differential equations and efficient parameter optimization in high dimensional spaces.

Project groups within the department

Gene Regulation and Systems Biology of Cancer (Yaspo group, 1995 - 2014)

In developing the sequencing and sequence analysis infrastructure within the department, and carrying the responsibility for essentially all sequencing-based projects as well as some of the technology development, the group of Marie-Laure Yaspo has been carrying out much of the core work of the department. This work will continue on a collaborative basis. The group has carried the main responsibility for work conducted within numerous departmental projects, including the 1000 Genomes project, Blueprint, ESGI, PREDICT, PedBrain Tumor, Treat20, OncoTrack, Treat20plus etc. Additional projects have addressed the genetics of gene regulation (e.g. through analysis of consomic strains) and, together with Alacris Theranostics, on the development of new techniques to deeply characterize the status of the immune system in individuals.

Marie-Laure Yaspo’s group continues as an independent research group of the Otto Warburg Laboratory.

Bioinformatics (Herwig group, 2001 - 2014)

The group of Ralf Herwig has contributed to (i) the development of computational methods for the analysis of molecular data, in particular high-throughput sequencing (HTS) data, derived from complex phenotypes and (ii) the integration and interpretation of these data in the context of biological networks. It has, for example, contributed to the informatics side of the work of the department in the 1000 Genomes project, has developed and maintains the ConsensusPath database (CPDB), one of the inputs into the tumor models, has been involved in applying our modeling techniques in a number of additional projects (PREDICT, APO-SYS, CarcinoGenomics, diXa etc.), and has collaborated on specific components of our sequence analysis pipeline. The group has developed computational methods for HTS applications, in particular for exome sequencing, RNA-seq and MeDIP-seq, and works on the integrative analysis of these data in order to elucidate the interplay of methylation, gene expression and genome structure that are operative in human (disease) processes, for ex-
ample related to cancer progression, drug toxicity, renal dysfunction and stem cell development. During the reporting period the group published 69 scientific publications.

Ralf Herwig’s group joined the Department of Computational Molecular Biology in December 2014.

**Diploid Genomics Research (Hoehe group, 2002 - 2014)**

The group of Margret Hoehe has focused on the analysis of haplotypes and diploidy as a fundamental feature of the human genome. Work includes (i) the development of novel molecular genetics and bioinformatics approaches as well as methods to haplotype-resolve whole genomes, and their application to data production; (ii) the analysis and annotation of haplotype-resolved genomes at the individual and population level; (iii) the establishment of public resources to advance integration of phase information at the gene and genome level and prepare the ground for “phase-sensitive” personal genomics and individualized medicine.

Margret Hoehe will now continue her research on the haplotype/diplotype architecture of human genomes and its implications for genome biology and medicine within the Department of Computational Molecular Biology.

**Evolution & Development (Poustka group, 2003 - 2014)**

The group of Albert Poustka has continued its research on phylogenomics of animal evolution. Major scientific contributions of the group comprise the identification of a significantly more complex immune system than ever expected in a major lophotrochozoan clade, the bivalves, i.e. the mussel *Mytilus edulis*. In addition, the group has contributed to work on the genetics of Autism.

Albert Poustka joined Alacris Theranostics in 2015.

**Cancer Biology (Schweiger group, 2007 – 2014)**

Complementing the main focus of the department on Systems Medicine of Cancer, the group of Michal Schweiger has, in particular, focused on tumor epigenetics. Major scientific contributions of the group comprise the identification of a so-called DNA methylator phenotype for TMPRSS2:ERG translocation-negative patients. This implies the possibility to specifically treat these patients with epigenetic therapies. Furthermore, members of the group have found that the bromodomain protein BRD4 regulates the oxidative stress response and that this might partly explain sensitivities of prostate cancers against bromodomain inhibitors.

Michal-Ruth Schweiger was appointed Professor at the University of Cologne in April 2014.
Neuropsychiatric Genetics (Bertram group, 2008 – 2014)

From the work on the cloning of the Chorea Huntington gene in the late eighties/early nineties, the department has had a continuing interest in neurodegenerative diseases, continued by a number of research groups within the department, and, more recently, also ageing, with work initiated by the groups of James Adjaye and Lars Bertram. The group of Lars Bertram has, in particular, focused on the mapping and characterization of complex disease genes, predominantly in the field of neuropsychiatric diseases. In this function, they led (and continue to lead) the genetics / genomics work packages on a number of national (e.g. Berlin Aging Study II [BASE-II], funded by the BMBF) and international (e.g. European Medical Information Framework Alzheimer Disease [EMIF-AD] project, funded by EU FP7) research projects. In addition to the laboratory work, Lars Bertram and his team have spearheaded the development of several bioinformatics / biostatistics tools facilitating the analysis and interpretation of large-scale genomics data (e.g. the AlzGene and PDGene databases), and to model the effects of disease-associated DNA sequence variants on micro-RNA function. During their six years at MPIMG they co-authored a total of 70 peer-reviewed publications.

Lars Bertram was appointed Professor (W2) of Genome Analytics at the University of Lübeck, Germany, in December 2014.

Systems Biology (Wierling group, 2009 - 2014)

The group of Christoph Wierling focused on further developing its systems biology approach to mathematical modeling of cellular processes with respect to complex diseases, such as cancer, expanding and optimizing the large-scale cancer signaling model, based on PyBioS. In particular, the group has recently finalized the third generation (PyBios3) of the modeling platform and has contributed to solving a number of the very significant computational bottlenecks in modeling and parameter optimization in high dimensional parameter spaces. The group has also contributed to a number of projects, in which the modeling platform has been used to integrate and interpret the molecular data generated in a range of diseases.

Christoph Wierling joined Alacris Theranostics in December 2014 as head of Modeling and Bioinformatics.
Planned developments

In the future, I plan to continue the collaboration with the group of Marie-Laure Yaspo on current projects (e.g. OncoTrack, Treat20plus), with a particular focus on data analysis and data interpretation (e.g. development and use of statistical procedures to identify and validate biomarkers predicting responders/non-responders in the data sets).

A second major focus of current and future work (in collaboration with Alacris Theranostics) is the further improvement of the “virtual patient” model, especially in its capability to “learn” from experimental systems e.g. preclinical and clinical data from our own projects, including PDX and 3D cell line models from OncoTrack, cell line and patient data from Treat20plus, involving very significant computational and theoretical challenges. This includes the further development of computationally efficient parameter optimization strategies in high dimensional spaces. Our largest tumor model has ~35,000 parameters, and still continues to grow.

A third focus has been and will continue to be the further development of techniques required to characterize each patient in sufficient detail (in collaboration with Marie-Laure Yaspo and Alacris Theranostics). This will include the further development of techniques to characterize the deep immune status, (including, if resources can be identified, development of models of the patient’s immune system as part of the virtual patient model), the development of spatially resolved transcriptome and proteome analysis techniques to be able to address tumor heterogeneity in the models, and the continuing development of new techniques for the robust assembly of long sequences from short read data.

In this, I will continue to collaborate with former members of my department on topics of common interest: In particular, with the groups of Marie-Laure Yaspo, continuing the department’s core work; Margret Hoehe, on the analysis of the haplotype structure of the human genome; and Ralf Herwig, on ongoing projects relevant to the “virtual patient” system.
General information about the whole department

For Ralf Herwig, Margret Hoehe, and Marie-Laure Yaspo, only the publications and other general information from 2009-2014 are listed here. The publications and information since 2015 are listed in the reports of the Dept. of Computational Molecular Biology (R. Herwig, M. Hoehe) and of the Research group Gene Regulation and Systems Biology of Cancer, OWL (M.-L. Yaspo).

Complete list of publications (2009 – 2015)

Department members are underlined.

2015


Genome sequencing identifies a novel mutation in ATP1A3 in a family with dystonia in females only. *Journal of Neurology*, 262(1), 187-193

**2014**


crop plant sugar beet (Beta vulgaris). *Nature, 505(7484), 546-549*


memory performance. *Translational Psychiatry, 4*, e454


ing in RNA-seq data. *Nucleic Acids Research*, 42(14), e110


Wierling C (2014). Bridging the gap between metabolic liver processes and


2013


**2012**

MW, Amstislavskiy VS, Borodina TA, Lienhard M, Mertes F, Sultan M, Herwig R, Yaspo ML.


Lill CM, Roehr JT, McQueen MB, Kavvoura FK, Bagade S, Schjieide BM, Schjieide LM, Meissner E, Zauf


*International Journal of Developmental Biology*, 56(10-12), 853-858

2011


ity of tumor cells. *PLoS Computational Biology*, 7(3), e1001113


mouse embryonic stem cells identifies conserved enhancers active in early embryonic development. PLoS Computational Biology, 7(12), e1002304


HV (2011). ITFoM - The IT Future of Medicine. Procedia Computer Science, 7, 26-29


Pfeiffer T, Bertram L, & Ioannidis JPA (2011). Quantifying selective reporting and the Proteus phenomenon for multiple datasets with similar bias. PLoS one, 6(3), e18362


Aggresome formation. Traffic, 12(3), 330–348


2010


as quantitative phenotypes: Genetics core aims, progress, and plans. *Alzheimer’s & Dementia*, 6(3), 265-273


2009


Heid IM, Huth C, Loos RJF, Kronenberg F, Adamkova V, Anand SS, Ardlie K, Biebermann H, Bjerregaard P, Boe-
do not drive gene duplicate retention. *Genome Research*, 19(11), 2036-2051


Parkhomchuk D, Amstislavskiy V, Soldatov A & Ogryzko V (2009) Use of high throughput sequencing to observe genome dynamics at a single cell level. *Proceedings of the National Academy of Sciences of the USA*, 106(49), 20830-20835


Zuccotti M, Merico V, Sacchi L, Bellone M, Brink TC, Stefanelli M, Redi

Scientific honours

Ralf Herwig: PerMediCon Award (3rd place) for the project EPITREAT Personalized lung cancer treatment based on epigenetic biomarkers, PerMediCon 2014, Cologne, Germany, 2014

Yuliya Georgieva: 3rd prize, 5th Speed Lecture Award, 10th BIONALE, BioTOP Berlin-Brandenburg and vfa bio, 2012

Nana-Maria Grüning: Nachwuchs:wissenschaftlerinnen-Preis, Forschungsverbund Berlin, 2012

Ralf Herwig: 31st Animal Protection Research Prize, Federal Ministry of Food and Agriculture (BMEL), Berlin, Germany, 2012

Markus Ralser: European Molecular Biology organisation (EMBO) Young Investigator Award, 2012

Alexander Kühn: Winner „Gesundheit 2030 – Deine Ideen für die Zukunft der Gesundheitsforschung“, Bundesministerium für Bildung und Forschung (BMBF) and WELT-Gruppe, 2011

Markus Ralser: Wellcome-Beit Prize, awarded to top four basic biomedical or clinical intermediate fellows of the Wellcome Trust in the UK, 2011

Markus Ralser: Wellcome Trust Research Career development fellowship, Wellcome Trust, UK, 2011

Markus Ralser: ERC starting grant, European Research Council, 2011

Lars Bertram: Special-Award of the Hans-und-Ilse-Breuer Foundation for Research in Alzheimer’s, 2010

Stephan Klatt: Posterprize, Peptalk – 9th Annual Protein Science Week, Cambridge-Healthtech Institute, 2010

Silke Sperling: W3 Heisenberg professorship for Cardiovascular Genetics, DFG, 2010

Lars Bertram: Recipient of the 2009 Independent Investigator Award, NARSAD, 2009

Thore Brink: PhD prize, Berliner Wissenschaftliche Gesellschaft e.V. and TSB Technologie Stiftung Berlin, 2009

Selected invited talks (Hans Lehrach)

The virtual patient, the key to a truly personalised medicine and prevention. The 2nd Precision Medicine Congress London, London, UK, 09/2015

From NGS to a truly personalised therapy. VIBT - Vienna Institute of BioTechnology, Universität für Bodenkultur Wien, Vienna, Austria, 05/2015

Next generation sequencing: a key factor in the personalization of medicine. The Royal Swedish Academy of Sciences Jubilee symposium, Gothenburg, Sweden, 04/2015

From NGS to a truly personalised therapy (keynote). Molecular Diagnostics Europe Conference, Cancer Sequencing, Lisbon, Portugal, 04/2015

DNA sequencing methods in human genetics and disease research (keynote). EMBO Workshop: Modern DNA concepts and tools for safe gene transfer and modification, Evry, France, 03/2015
Personalized medicine and virtualized drug development. Biocenter Oulu Day 2015, Personalised Medicine: Hype or Hope, Oulu, Finland, 03/2015


Virtualised drug development for an individualized medicine. 2014 International Symposium on Clinical and Translational Medicine, China, 05/2014

Big Data Medicine. Advances in NGS&Big Data, Barcelona, Spain, 05/2014

Virtualised drug development for an individualized medicine. COST Conference, 5th Session: Computational Epigenetics and Constellation Thinking, Athens, Greece, 05/2014

Virtualised drug development for (truly) personalised medicine. World CDx meeting, Frankfurt, Germany, 04/2014

Systems Biology of Cancer. 5th Paris Workshop on Genomic Epidemiology, Paris, France, 05/2013

Blueprint for integration of genomic techniques into CDx development. World CDx Frankfurt, Germany, 03/2013

The future of medicine (and health). 2nd World Congress of Cutaneous Lymphoma, Germany, 02/2013

ITFoM – IT Future of Medicine. The International Conference on Genomics in Europe (ICG-Europe), BGI, Denmark, 05/2012

ITFoM – IT Future of Medicine. Intel Leadership Conference on HPC for Life Sciences, Hungary, 05/2012

Systems Genomics – the future of Medicine. Robert B Church Lecture in Biotechnology, University of Calgary, Canada, 03/2012

The Big Six-Spotlight on EU Flagship: Medicine in the 21st Century: IT as a “Magic Bullet”? / Talk: Integrating data on genetic information, Metabolic processes, behavior and Environment to build up the virtual patient. Schweiz. Akademie der Wissenschaften, Swiss National Science Foundation (SNSF), Switzerland, 03/2012

Genetics, Genomics & Systems Biology. 2nd Heidelberg Forum for Young Scientists, Germany, 02/2012

Presentation ITFoM. National Research Foundation (NRF), Dubai, Vereinigte Arabische Emirate, 02/2012

The IT Future of Medicine. Spanish National Cancer Research Centre, CNIO, Spain, 01/2012

Environment and Genetics: what can we conclude, what should go on? EPH Environment and Public Health in modern society; Focus: reports from basic research; mechanisms of disease origin & interdisciplinary research, Potsdam, Germany, 11/2011

Forward Look on personalized Medicine. ESF Workshop, Disease Summit, Netherlands, 10/2011

Systems Patientomics: The Future of Medicine. Molecular Diagnostics Summit Europe, Germany, 10/2011

The Future of Medicine. European Health Forum Gastein, Austria, 10/2011
The IT Future of Medicine - a flagship initiative to revolutionize our health care system. NGFN Jahrestagung, Symposium III, From Genomics to Application, Germany, 09/2011

Patient Stratification based on individual full genome sequencing and disease modelling. Mühldorfer Conference Science to Market 2011, Germany, 09/2011

An in silico reference model of humans as goal for personalised medicine. European Science Foundation (ESF) Technology Workshop, UK, 09/2011

Der Virtuelle Patient-Systembiologie als Chance für eine individualisierte Medizin. 18. Dresdner Palais-Gespräch, Germany, 09/2011

From Biobanks, clinical and molecular phenotypes towards systems biology (IT Future of Medicine - initiative). European Congress of Pathology, Session: Biobanks - the power of many, Finnland, 08/2011

Genetics, Patient Modelling and IT. CARS 2011, Computer Assisted Radiology and surgery, 25th International Congress and Exhibition, Germany, 06/2011

Other Childhood Cancers - Current Molecular Approaches to address Environmental Risk Factors. International Conference on Non-Ionizing Radiation and Children’s Health, Slovenien, 05/2011


Genetics, Genomics & Systems Biology. VIB-Future and history of Genomics, Belgium, 04/2011

The 1010$ model for all and then? FEBSX-SysBio2011: From Molecules to Function, Austria, 02/2011

Latest developments in genomics research. Medizinische Universität Graz, Opening Event, Institute of Pathology, Austria, 02/2011

Krankheiten verstehen und vorbeugen - Neue Ansätze der Systembiologie (Keynote Lecture). 6. Wissenschaft trifft Wirtschaft (WtW), Uni Konstanz, Germany, 12/2010

Genetics, Genomics & Systems Biology, 6th Genetic Workshop, Germany, 11/2010


Next Generation Sequencing. DGHO Jahrestagung 2010, Deutsche Gesellschaft für Hämatologie und Onkologie, ICC Berlin, Germany, 10/2010

Genetics, Genomics and Systems Biology: the Path to an Individualized Treatment of Cancer Patients. International Society of Oncology and BioMarkers - ISOBM 2010, Germany, 09/2010

Genomforschung und die Zukunft der Krebsmedizin. Alpbacher Technologiegespräche 2010, Österreich08/2010

Genetics and individualizing therapy: learning from cancer and gambling. Heart Failure Congress 2010, ICC Berlin, Germany, 05/2010

Genetics, Genomics & Systems Biology. Symposium in Honour of Professors Sir Kenneth and Noreen Murray, UK, 05/2010

Systembiologie. Heinrich-Warner-Symposium, Translationale Forschung beim
Prostatakarzinom, Hamburg, Germany, 02/2010


Genetics, Genomics & Systems Biology: Baltic Summer School 2009 - Genetic Basis of Medicine, Kiel, Germany, 09/2009


Appointments of former members of the department

Lars Bertram: Professorship (W2) for Genome Analytics, University of Lübeck, 2014

Christoph Wierling: Group leader, Alacris Theranostics GmbH, Berlin, 2014

Ruth-Michal Schweiger: Lichtenberg Professorship, University of Cologne, 2014

James Adjaye: Professorship (W3) and Director of the Institute for Transplantations Diagnostics and Cell Therapies (ITZ), Heinrich-Heine-University of Düsseldorf, 2012

Bodo Lange: Managing Director, Alacris Theranostics GmbH, Berlin, 2012


Harald Seitz: Group leader, Fraunhofer Institute for Biomedical Engineering, Potsdam, Germany, 2012

Silke Sperling: Professorship for Cardiovascular Genetics (Heisenberg professorship), Experimental Clinical Research Center (ECRC), joint institute between Charité - Universitätsmedizin Berlin and Max Delbrück Center for Molecular Medicine), 2011

Andreas Dahl: Head of Deep Sequencing Group, Technische Universität Dresden, Germany, 2010

Georgia Panopoulou: Lecturer, Institute of Biochemistry and Biology, University of Potsdam, Germany, 2010

Eckhard Nordhoff: Group Leader, Medizinische Proteom-Center, University Bochum, Germany, 2009

Heinz Himmelbauer: Head of the Ultrasequencing Unit, Centre for Genomic Regulation (CRG) Barcelona, Spain, 2008. Since 2012, Head of Genomics Unit, CRG Barcelona, Spain. Since 2015, Professor of Bioinformatics, Universität für Bodenkultur (BOKU), Vienna, Austria

Habilitationen/State doctorate


Silke Sperling: Discovering the transcriptional networks for cardiac development, function and disease with a systems biology approach. Charité-Universitätsmedizin Berlin, 2009

### PhD theses


Christina Röhr: Analyses of regulatory mechanisms and identification of microRNAs as biomarkers in colorectal cancer. Freie Universität Berlin, 2015

Julia Schröder (Dr. rer. med.): Funktionelle Charakterisierung genomweiter Assoziationssignale der Berliner Altersstudie II. Charité – Universitätsmedizin Berlin, 2015


Antje Krüger: The Influence of Metabolism on Gene and Protein Expression during the Oxidative Stress Response. Freie Universität Berlin, 2014

Björn Lichtner: The impact of distinct BMPs on self-renewal and differentiation of human embryonic stem and induced pluripotent stem cells. Freie Universität Berlin, 2014


Vikash Kumar Pandey: Metabolic modeling of different degrees of steatohepatitis in mice. Freie Universität Berlin, 2014

Stephan Klatt: Evaluation of the protozoan parasite L. tarentolae as a eukaryotic expression host in biomedical research. Freie Universität Berlin, 2013

André E. Minoche: The genome of sugar beet (Beta vulgaris) - Assembly, annotation and interpretation of a complex plant genome. University of Bielefeld, 2013


Andrea Wunderlich: Direct regulation of the interferon signaling pathway by the bromodomain containing protein 4. Freie Universität Berlin, 2013

Alexandra Zerck: Optimal precursor ion selection for LC-MS/MS based proteomics. Freie Universität Berlin, 2013


Katharina Drews: Generation and characterization of induced pluripotent stem cells from human amniotic fluid cells. Freie Universität Berlin, 2012

Nana-Maria Grüning: Pyruvate kinase triggers a metabolic feedback
loop that controls redox metabolism in respiring cells. Freie Universität Berlin, 2012

Atanas Kamburov: More complete and more accurate interactomes for elucidating the mechanisms of complex diseases. Freie Universität Berlin, 2012


Christian Linke: Serial regulation of yeast cell cycle by Forkhead transcription factors Fkh1 and Fkh2. Freie Universität Berlin, 2012


Markus Schüler: Bioinformatics Analysis of Cardiac Transcription Networks. Freie Universität Berlin, 2012

Tatjana Schütze: SELEX Technologieentwicklung im Hochdurchsatz. Freie Universität Berlin, 2012


Jenny Schlesinger: Regulation of Cardiac Gene Expression by Transcriptional and Epigenetik Mechanisms and Identification of a Novel Chromatin Remodeling Factor. Freie Universität Berlin, 2011


Anja Berger: Molecular analysis of the Oryzias latipes (Medaka) transcriptome. Freie Universität Berlin, 2010

Lukas Chavez: Multivariate statistical analysis of epigenetic regulation with application to the analysis of human embryonic stem cells. Freie Universität Berlin, 2010


Daniela Köster: Rolling Circle Amplification auf Biochips. Freie Universität Berlin, 2010

Cornelia Lange: Generation and application of genomic tools as important prerequisites for sugar beet genome analyses. Freie Universität Berlin, 2010


Axel Rasche: Information Theoretical Prediction of Alternative Splicing with Application to Type-2 Diabetes mellitus. Freie Universität Berlin, 2010


Gina Ziegler: Funktionelle Charakterisierung von differenziell exprimierten Genen in einem Mausmodell für
die zerebrale Ischämie. Technische Universität Berlin, 2010

Young-Sook Baek: Gene Expression Analysis of Differentiating U937 Cells. Freie Universität Berlin, 2009

Thore Brink: Transcriptional and signaling analysis as a means of investigating the complexity of aging processes in human and mouse. Freie Universität Berlin, 2009

Hendrik Hache: Computational Analysis of Gene Regulatory Networks. Freie Universität Berlin, 2009

Justyna Jozefczuk: Analysis of dynamic regulatory events during human embryonic stem cell differentiation into hepatocytes: new insights on downstream targets. Freie Universität Berlin, 2009


Theam Soon Lim: Parameters affecting phage display library design for improved generation of human antibodies. Freie Universität Berlin, 2009

Martje Tönjes: Transcription networks in heart development and disease with detailed analysis of TBX20 and DPF3. Freie Universität Berlin, 2009


**Student theses**


Sunniva Förster: *Optimizing cDNA phage display for high throughput biomarker discovery.* Humboldt University of Berlin, Diploma thesis, 2011


Matthias Linser: *Biotin-basiertes de-enrichment von low-copy Vektorsequenzen in humanen genomischen Fosmidbanken für die Next Generati-
on Sequenzierung. Humboldt Universität zu Berlin, Bachelor thesis, 2011


Sandra Schmökel: Establishment of a purification method for the parathyroid hormone from human tissue extracts and phage display analysis of diagnostically important targets. Universität Potsdam, Master thesis, 2010


Thomas Bergmann: Development of a Multi-Wavelength Fluorescence


Oliver Herrmann: Funktionelle Analyse der konservierten, nicht kodierenden Elemente des sonic hedgehog Gens in Fugu (Takifugu rubripes), Stickelback (Gasterosteus aculeatus) und Medaka (Oryzias latipes). Fachhochschule Zittau/Görlitz, Bachelor thesis, 2009


Teaching activities

From 1998 - 2013, the department held the lecture “From functional genomics to systems biology” at Freie Universität Berlin (each winter term, 2hrs/week).

Inventions, patents and licences


Strand-specific cDNA sequencing. Aleksey Soldatov, Tatiana Borodina, Hans Lehrach. EP 09008808.9, 2009


Computer implemented model of biological networks. Ralf Herwig, Christoph Wierling, Hans Lehrach. WO2010025961

Spin-offs

Global Action 4 Health Institute Limited, Dublin 2, Ireland, 2012

Dahlem Centre for Genome Research and Medical Systems Biology gGmbH, Berlin, Germany, 2010

Organization of scientific events

DataFest, OncoTrack Meeting, Berlin, May 20-21, 2014


Organisation of two parallel forums: session on “The future of medicine – developing an infrastructure for personalized medicine” at the European Health Forum Gastein, Bad Hofgastein, Austria, October 5-8, 2011


2nd Annual NGFN Meeting, Berlin, November 26-28, 2009 (co-organizer: H. Lehrach, W. Nietfeld)

Genomic variations underlying common neuropsychiatric diseases and cognitive traits in different human populations., EU-project meeting, Berlin, October 5-6, 2009

EMBO Practical Course: Next generation sequencing: ChIP-seq and RNA-seq, Berlin, February 1-13, 2009
Department of Human Molecular Genetics

Now:
Emeritus Group Human Molecular Genetics

Heads of associated groups
Harry Scherthan (01/04 – 10/14)
Wei Chen (01/09-12/11)
Susan Schweiger (04/05 – 09/11)

Scientists and postdocs
Melanie Hambrock (06/14-10/14)
Annibale Garza Caraval (08/11-01/12)
Masoud Garshasbi (05/08-10/11)
Nils Paulmann (04/08-08/11)
Lars Riff Jensen (04/02-03/11)
Sybille Krauss* (05/05-2010)
Chandan Goswami (07/08-06/09)

PhD students
Friederike Hennig (04/14-10/14)
Grit Ebert (10/09-10/14)
Melanie Hambrock (12/09-05/14)
Lucia Püttmann (06/08-12/13)
Daniel Mehnert (03/11-03/13)
Anne Steininger (07/08-02/13)
Roxana Kariminejad*
(at times at the MPIMG until 11/12)
Hyung-Goo Kim*
(at times at the MPIMG until 09/12)
Silke Stahlberg (07/08-08/12)
Agnes Zecha (07/08-08/12)
Vivian Boldt (03/08-02/12)
René Buschow (09/09-01/12)
Juliane Schreier* (01/09-01/12)
Christine Andres (07/08-10/11)
Paul Hammer (01/07-10/11)
Jakob Wovinckel (07/06-10/11)
Lia Abbasi Moheb (10/05-09/11)
Julia Hoffer (09/10-08/11)
Julia Kuhn* (10/06-08/11)
Robert Weißmann (09/09-04/11)
Sahar Esmæeeli-Nieh* (11/06-10/10)
Na Li (09/07-09/10)
Hui Kang (09/07-09/10)
Eva Kickstein* (03/06-03/10)

Senior scientific staff
Thomas F. Wienker (05/11-12/15)
Luciana Musante* (01/11-07/15)
Hao Couger Hu* (05/09-07/15)
Reinhard Ullmann (02/04-10/14)
Vera Kalscheuer (07/05-10/14)
Diego Walther (02/03-07/12)
Tim Hucho (09/05-01/12)
Andreas Tschach (03/05-06/11)
Andreas Kuss (06/05-10/10)

* externally funded
For 20 years, from its establishment in 1994 to its closure in 2014, the research of the Department of Human Molecular Genetics revolved around monogenic or “Mendelian” disorders. Monogenic disorders had also been an early target of the Human Genome Project. In the early 1990ies, however, when the – now refuted – “Common Disease-Common Variant” (CDCV) hypothesis and genome-wide association studies (GWAS) gained popularity and predictions about the imminent eradication of common disorders became commonplace, leading researchers reoriented this project on common diseases to ensure continued funding by the American Congress. In spite of largely futile attempts to identify clinically relevant genetic markers for common diseases, cogent explanations for the failure of the HapMap (and GWAS) paradigm and strong pleas for monogenic disorders as a more promising target for genome research, it was only after the introduction of Next Generation Sequencing (NGS) when the most prominent advocates of GWAS gave in and genome research re-focused on monogenic disorders.

Before joining the MPIMG, H.-H. Ropers and his group had been among the first employing positional cloning to elucidate monogenic forms of blindness, deafness as well as cognitive and/or behavioral disorders. In 1995, he and his coworkers embarked on two large transnational projects aiming to identify disease-causing gene defects in a systematic way (“Breakpoint cloning in patients with disease-associated balanced chromosome rearrangements”, collaboration with N. Tommerup, Copenhagen, DK; “European X-linked Mental Retardation Consortium”, collaboration with four other European groups). Later on, through collaboration with Pieter de Jong (Oakland), they were also among the few groups worldwide to develop arrays of overlapping BAC clones spanning the entire human genome. These and other high-resolution microarrays were instrumental in the identification of disease-associated microdeletions and duplications.
(Copy Number Variants, CNVs), particularly in individuals with intellectual disability (ID) and related disorders\textsuperscript{7}, where CNVs account for 10 to 15\% of the patients\textsuperscript{8}.

In the late 1990ies, ID became the central research topic of the Department of Human Molecular Genetics, because (i) ID is the diagnosis with the highest socio-economic costs of all disorders listed by the International Classification of Disease (ICD10), (ii) it is by far the most common reason for referral to genetic services, and (iii) at the outset, very little was known about the molecular causes of ID. Our work and that of our collaborators has contributed significantly to the growing popularity of research into the genetic causes of cognitive impairment, but even today, most genetic defects that give rise to ID are still unknown (see\textsuperscript{9} and below).

During our ground-breaking research into X-linked ID, initially an almost exclusively European domain, we realized that X-linked forms of ID are far less common than previously assumed\textsuperscript{10} and that in turn, autosomal forms had to be more frequent than expected. Theoretical considerations as well as empirical data indicated that most functionally relevant gene defects are inherited as recessive traits (discussed in 3), but in Western countries, their investigation was hampered by small family sizes. In Iran, autosomal recessive forms of ID (ARID) are common because 40\% of the children have consanguineous parents, and families are usually large. Therefore, in 2004, we joined forces with a strong Iranian group to study ARID families in a systematic fashion.

At that time, no more than three ARID genes were known. By serial autozygosity mapping in consanguineous families with 2 or more intellectually disabled children we identified numerous additional loci for autosomal recessive ID (ARID) and showed that this condition is an extremely heterogeneous genetic disorder. Subsequent Sanger sequencing of positional and functional candidate genes revealed causative mutations in numerous of these families, and this approach led to the identification of a dozen novel ARID genes\textsuperscript{9}.

In 2008, the MPIMG had become the first (Continental) European customer of Solexa-Illumina, now the leading manufacturer of high throughput – low cost sequencing systems worldwide. Two years later, Next Generation Sequencing techniques had entirely replaced Sanger-based mutation screening at the MPIMG. This significantly accelerated our research into the genetic causes of ID and related disorders, as documented by an article describing 50 additional candidate genes for autosomal recessive ID\textsuperscript{11} and a recent study encompassing more than 400 families with X-linked ID\textsuperscript{12} (for details, see Report Kalscheuer).

In October 2011, the head of the Department reached his official retirement age, but was re-installed as Acting Director for a final period of 3 years. At the same time, support for this research from the Max Planck Innovation
Funds and the Federal Ministry of Education and Research expired. In keeping with a deconstruction plan, several scientists left the department, and the structural budget was progressively reduced until the department was closed at the end of October 2014. During this period, an EU grant partially compensated the diminishing structural support and enabled us to continue and round off our research into X-linked and autosomal recessive forms of ID until and even beyond April 30th, 2015.

On July 31st, 2015, the last full-time scientists of our group, Hao Hu and Luciana Musante, have left the institute, but both still participate in the ongoing analysis of unpublished data generated by our team since 2012. These efforts also involve Vera Kalscheuer, who is continuing our XLID research after having joined the research group of Stefan Mundlos in November 2014, and Thomas Wienker, whose part-time appointment as Biostatistician and Clinical Geneticist will terminate in December 2015.

After having declined an offer to serve as founding director of an Institute of Genomic Medicine at the Berlin Institute of Health (BIH) and failing attempts to establish the BIH as front-runner in the field of Medical Genome Sequencing (see below), H.-H. Ropers has accepted a part-time appointment as Board-Certified Medical Geneticist and adviser at the Human Genetics Institute of the University of Mainz. This institute (head: Susann Schweiger), a Center for Rare Diseases with a focus on ID and neurodevelopmental disorders, will continue the collaboration with our Iranian colleague and his team, and it will serve as safe haven for sequencing data and other resources that are the shared property of his group and ours.

Scientific activities and results, 2012-2015

X-linked ID (for details, see Report Kalscheuer)

In view of the diminishing return of our long-standing search for novel X-linked ID genes; we had envisaged that this research would reach its natural end by the time our department would be closed, and that the same might soon apply to our research into disease-associated balanced chromosome rearrangements. However, more than nine months after our department was closed and Vera Kalscheuer joined the group of Stefan Mundlos, both projects are still internationally competitive, and there is no reason for discontinuing them any time soon. Apart from shedding more light on the pathogenesis of specific forms of ID and the identification of more than a dozen novel XLID genes, these studies and recent observations of a French group eventually confirmed our previous finding that in males, mutations inactivating of the MAOA gene may cause low frustration tolerance and aggression.
For more than 20 years, this observation had been considered as highly controversial, since it contradicted the common belief that due to its complexity, human behavior cannot be influenced significantly by single genes – and that it should not because otherwise, this would undermine “morale, jurisdiction and religion”\(^\text{16}\). Recent studies have shown that MAOA deficiency is not the only gene defect predisposing to aggression and violence\(^\text{17}\). While being a lesser concern for morale and religion, it remains to be seen whether and how the legislation and jurisdiction will react to these findings.

**Hunting genes for autosomal recessive ID**


Because of the remarkable locus heterogeneity of ARID\(^\text{18}\) we could not expect to identify more than a fraction of the underlying gene defects before our department would be dissolved. Still, in order to make the most of the productive collaboration with our Iranian partner, we decided to pursue this research until all collected ARID families had been analyzed – or until the very end of the GENCODYS project. Indeed, since 2012, we managed to investigate the entire remaining cohort of more than 400 multiplex consanguineous families by targeted exon enrichment and sequencing, whole exome sequencing and whole genome sequencing, respectively. Employing MERAP, a novel integrated sequence analysis pipeline providing a one-stop solution for identifying disease-causing mutations\(^\text{19}\), we have detected apparently causative mutations co-segregating with ID in more than half of these families and collected detailed clinical information for fine-grained genotype-phenotype comparisons. About 40% of these mutations involve hitherto unknown (candidate) genes for ID. To our knowledge, this is the largest study of its kind ever conducted and it is three times the size of our previous one which quadruplicated the number of known ARID genes\(^\text{11}\).

Many of the consanguineous ID families recruited through the genetic service of our Iranian partner had only a single affected child and were therefore not included in this study. In outbred Western populations, where families are usually small and most ID patients are sporadic cases, dominant de novo mutations have been identified as the predominant cause of ID\(^\text{20}\). In populations with frequent parental consanguinity, the incidence of ID is 2-3 times higher, and inherited recessive forms should be more common than autosomal dominant ones. However, there are no empirical data yet to support this assumption. Therefore, as our final collaborative project, we and our Iranian partner have set out to address this problem by performing WES in 100 isolated patients with ID and their healthy consanguine-
uous parents. Whole exome enrichment and sequencing as well as quality checks could be completed before the EU-funded GENCODYS project expired, and sequence alignment, variant calling, filtering and Sanger confirmation are underway. Assuming that the incidence of de novo mutations does not differ much between populations, we expect that in Iran, the proportion of sporadic ID patients carrying dominant de novo mutations will be much lower than in Western countries. These studies should shed more light on the global importance of recessive gene defects in the etiology of ID and other severe childhood disorders.

**Diagnostic and translational activities**

[Thomas F. Wienker, Hao Hu, H.-Hilger Ropers, Tomasz Zemojtel (since 2014 Dept. Medical Genetics, Charité)]

By serially elucidating the molecular basis of single gene disorders, we have provided the basis for early molecular diagnosis, prevention and improved disease management in patients and families. From the start, this has been an important motivation and driving force of our research. Since the advent of NGS, the translation of these methods into the clinic to improve the quality of genetic health care has lagged behind. As the first user of the Solexa-Illumina technology in Germany and Continental Europe, we joined forces with Stephen Kingsmore, NCGR Santa Fe and Children’s Mercy Hospital, Kansas City, to generate a clinical entry test for children with severe ID and/or unexplained developmental delay. This NGS-based panel test, called MPIMG-1, comprised 1220 genes and allowed the molecular diagnosis of >550 severe recessive childhood disorders as well as the vast majority of the monogenic causes of ID identified until 2012.

After encouraging pilot studies in Berlin (collaboration with Chen Wei, Max Delbrück Center, and the Pediatric Department of the Charité), this test and our MERAP pipeline for the identification of disease-causing DNA variants\(^\text{17}\), were adopted by several Human Genetics Institutes and research groups in Germany and beyond. In some of these, including Mainz and Innsbruck, it is still being used for selected patients. When Thomas Zemojtel left us to join the department of Medical Genetics at the Charité Berlin (head: Stefan Mundlos), the MPIMG-1 panel became integral part of an even larger panel test with a diagnostic yield of almost 30% in a clinical setting\(^\text{21}\).

Since 2011, we have offered WES and/or WGS to selected patients and families with undiagnosed and presumably novel genetic conditions, in order to end the mostly long diagnostic odyssey of these families. Another aim of this project was to demonstrate the superior power of these
approaches for diagnosing hitherto unknown genetic defects and to overcome the irrational, but wide-spread reluctance to adopt these novel tools in genetic health care.

Contrary to our initial intentions, we decided to publish some results of this initiative to enhance their acceptance by the media. Indeed, our report on the very first family studied received an Editorial as well as a commentary in Nature. Moreover, the novel syndrome described in our article has been named Bainbridge-Ropers syndrome (OMIM #615485) after the first and last authors of this publication.

Approximately one third of the gene defects identified in patients and families with unclear diagnoses had not been described before and would not have been detected with panel tests. Recent publications have confirmed our conclusion that WES and WGS are better suited for the diagnosis of unclear genetic disorders than gene panels, particularly for common, but heterogeneous disorders like ID and ASD where most of the underlying gene defects are still unknown.

However, there is an even more compelling argument for the introduction of comprehensive diagnostic tests covering all genes (i.e., WES) or even the entire human genome (i.e., WGS): given the rarity of most individual gene defects, existing initiatives to sequence the genomes of 100,000 volunteers or patients are much too small to identify more than a fraction of the genetic defects causing rare Mendelian disorders. A solution for this problem is the implementation of medical genome sequencing as universal diagnostic test at large centers for Rare Diseases and the storage of all disease-associated sequence variants in a central database, in line with our previous recommendations.

However, despite considerable efforts, it proved impossible to overcome the opposition of various stakeholders against these proposals; and when novel, ultrahigh-throughput sequencing “factories” became available in Spring 2014 and whole genome sequencing costs fell below the 1000€/genome barrier, the Berlin Institute of Health (BIH) missed even this unique opportunity to establish itself as the unrivalled German leader and a global player in this emerging field. Shortly afterwards, H.-H. Ropers accepted a position as board-certified Human Geneticist and scientific adviser at the University of Mainz.

**ID genes: genetic determinants of the normal IQ distribution?**

In his “Theory of X-linkage of major intellectual traits”, Robert Lehrke had speculated that variants of X-linked ID genes might also be responsible for the normal variation of the IQ. Today, more than 40 years later, the
search for the genetic factors underlying the heritability of the IQ is still ongoing, but it is no longer entirely focused on the X chromosome since it is now established that X-linked genes do not play a predominant role in ID. However, even very large GWAS have not identified genetic variants robustly associated with intelligence, and a more recent endeavor employing WGS to look for genetic variants enhancing the IQ in 1600 British and Chinese “geniuses” was equally unsuccessful. Therefore we wondered whether the rationale for these studies might be wrong, and that high IQ may be due to the absence of mildly deleterious genetic variants rather than being caused by IQ-enhancing factors.

In a pilot study conducted together with a Dutch group, we have performed targeted exon enrichment and NGS to look for rare genetic variants in 168 selected ID genes in individuals sampled from the high and low ends of the IQ distribution. No convincing evidence was obtained for a significant inverse relationship between the IQ of the probands and the number of subtle mutations in these genes. However, given the small size of our cohort and the limited number of ID genes screened for genetic variants, more comprehensive studies of this kind will be required before ruling out the hypothesis that it is the absence of negative genetic factors, not the presence of IQ-stimulating ones, that predisposes to high IQ.

**Epilogue**

During the past 25 years, the Human Genome Project and the development of high-throughput/low cost sequencing techniques have revolutionized the identification of genetic defects underlying monogenic disorders. Since single gene disorders were rediscovered as important targets of genome research, their elucidation and analogous studies of knockout animal models have already provided more insights into the role of genes in health and disease than 20 years of heavily funded, but largely futile association studies. Apart from the significant spinoff of this research for genetic health care, these studies have provided the basis for in-depth research into the function and interaction of human genes and eventually, the function of the entire human genome. Thus, studies into the function, interaction and regulation of the many dozen ID genes identified by our group alone will continue to shed light on the function of the human brain for many years to come.

According to recent estimates and observations in mouse models, mutations in at least two thirds of all genes will give rise to disease, which means that the majority of disease-associated single gene defects still wait to be discovered. Systematic studies into genetic factors modifying the expression of these genes and the penetrance of genetic disorders are about to add an additional dimension to this research, with far-reaching implica-
tions for genetic health care and for understanding genome function. Thus, the era of human genetics as bridge between genome and clinical research is not approaching its end – instead, it is just beginning, and the best is yet to come. Against this background, the intention of the Max Planck Society to discontinue its engagement in this pivotal area of life sciences is therefore questionable and may deserve reconsideration.

Acknowledgements

I thank the Max Planck Society for 20 years of generous support for a discipline that has been and still is undervalued by some of its members. Thanks are also due to the German Federal Ministry of Education and Research and to the EU for several research grants, and I am grateful to many colleagues in Europe and world-wide for long-standing, productive interactions, which have enriched my life.

References

2 e.g., see Ropers HH, *Frankfurter Allgemeine Zeitung*, 21.01.2001
5 see Ropers HH, *Dialogues Clin Neurosci* 12:95-102, 2010
8 see *Scientific Report* 2012
9 see Ropers HH & Hamel BCJ, *Nat Rev Genet* 6:46-56, 2005
12 Genetic and Epigenetic Networks in COgnitive DISorders [GENCODYS]FP/ref 241995, 05/2010 - 04/2015
13 see Outlook, *Scientific Report* 2012
18 Hu H et al., *Hum Mutat* 35:1427-35, 2014


Bainbridge et al., *Genome Med* 5:11, 2013


e.g., see Gilissen et al., *Nature* 511:344-347, 2014


Personal Genome Project, http://www.personalgenomes.org/

Genomics England, http://www.genomicsengland.co.uk/

Ropers HH et al., Neue Sequenzierungstechniken und Krankenversorgung, Berlin-Brandenburgische Akademie der Wissenschaften (BBAW), 2013

see Ropers HH, 3. Gentechnologiebericht der BBAW, 2014


see Yong E, Chinese project probes the genetics of genius, *Nature* 497:297-299, 2013

Franić S et al., *Eur J Hum Genet*, 2015, doi:10.1038/ejhg.2015.3 [Epub ahead of print]

Franić S et al., *Intelligence* 49:10-22, 2015

Cooper et al., *Hum Mutat* 31:631–655, 2010
General information about the whole department

For Vera Kalscheuer, only the publications and other general information from 2009-2014 are listed here. The publications and information since 2015 are listed in the report of the Research group Development & Disease (S. Mundlos).

Complete list of publications (2009 – 2015)

Department members are underlined.

2015


2014


De Brouwer AP, Nabuurs SB, Verhaart IE, Oudakker AR, Hordijk R, Yntema HG, Hordijk-Hos JM, Vo-


Cleft palate, digital and genital abnormalities define novel Xq25q26 duplication syndrome. Human Genetics, 133(5):625-638


2013

cating mutations in ASXL3 are associated with a novel clinical phenotype with similarities to Bohring-Opitz syndrome. *Genome Medicine*, 5(2), 11


JNK3 truncation mutations associated with cognitive disorders in two unrelated patients. *Human Genetics*, 132(4), 461-471


Rademacher N, Kunde SA, Kalscheuer VM & Shoichet SA (2013). Synaptic MAGUK multimer formation is mediated by PDZ domains and promoted by ligand binding. *Chemistry & Biology*, 20(8), 1044-1054


2012


Cingoz S, Bache I, Bjerglund L, Ropers HH, Tommerup N, Jensen H, Bron- 
dum-Nielsen K & Tumer Z (2011). Interstitial deletion of 14q24.3-q32.2 
in a male patient with plagiocephaly, BPES features, developmental delay, 
and congenital heart defects. *American Journal of Medical Genetics A*, 
155A, 203

Otto C (2011). Impaired left-ventricular cardiac function in male GPR30-
deficient mice. *Molecular Medicine Reports*, 4, 37

Fullston T, Gabb B, Callen D, Ullmann R, Woollatt E, Bain S, Ropers 
HH, Cooper M, Chandler D, Carter K, Jablensky A, Kalaydjieva L & Gezz J 
(2011). Inherited balanced translocation t(9;17)(q33.2;q25.3) concomitant 
with a 16p13.1 duplication in a patient with schizophrenia. *American Journal 
of Medical Genetics Part B: Neuropsychiatric Genetics*, 156, 204

Garshasbi M, Kahrizi K, Hosseini M, Nouri Vahid L, Falah M, Hemmati 
new nonsense mutation in TUSC3 is responsible for non-syndromic au-
tosomal recessive mental retardation in a consanguineous Iranian family. 
*American Journal of Medical Genetics A*, 155A, 1976

Gilling M, Lind-Thomsen A, Mang Y, Bak M, Moller M, Ullmann R, 
Kristoffersen U, Kalscheuer VM, Henriksen KF, Bugge M, Tumer Z & 
Tommerup N (2011). Biparental inheritance of chromosomal abnormali-
ties in male twins with non-syndromic mental retardation. *European Journal 
of Medical Genetics*, 54, e383

Goswami C, Kuhn J, Dina OA, Fernandez-Ballester G, Levine JD, 
through an ion-conductivity-independent TRPV1 pathway. *Journal of Neu-
rochemistry*, 117, 995

Gregor A, Albrecht B, Bader I, Bijlsma EK, Ekici AB, Engels H, Hack-
mann K, Horn D, Hoyer J, Klapecki J, Kohlhase J, Maystadt I, Nagl S, Prott 
Zweier C (2011). Expanding the clinical spectrum associated with defects 
in CNTNAP2 and NRXN1. *BMC Medical Genetics*, 12, 106

Hu H, Eggers K, Chen W, Garshasbi M, Motazacker MM, Wrogemann K, 
Kahrizi K, Tzschach A, Hosseini M, Bahman I, Hucho T, Muhlenhoff M, 
Gerardy-Schahn R, Najmabadi H, Ropers HH & Kuss AW (2011). ST-
3GAL3 mutations impair the development of higher cognitive functions. 
*American Journal of Human Genetics*, 89, 407

Jakobsen LP, Bugge M, Ullmann R, Schjerling CK, Borup R, Hansen L, 
Eiberg H & Tommerup N (2011). 500K SNP array analyses in blood and 
saliva showed no differences in a pair of monozygotic twins discordant for 
cleft lip. *American Journal of Medical Genetics A*, 155A, 652

Jensen LR, Chen W, Moser B, Lipkowitz B, Schroeder C, Musante L, 
Tzschach A, Kalscheuer VM, Meloni I, Raynaud M, Van Esch H, Chelly J, 
De Brouwer AP, Hackett A, Van Der Haar S, Henn W, Gezz J, Riess O, 
resequencing of 17 X-linked intellectual disability genes in 135 patients 
of Human Genetics*, 19, 717

Kahrizi K, Hu CH, Garshasbi M, Abedini SS, Ghadami S, Kariminejad 
R, Ullmann R, Chen W, Ropers HH, Kuss AW, Najmabadi H & Tzschach 
A (2011). Next generation sequencing in a family with autosomal recessive 
Kahrizi syndrome (OMIM 612713) reveals a homozygous frameshift mu-


neural function in Drosophila and humans. *Proceedings of the National Academy of Sciences of the USA*, 108, 12390


2010


hibernator, the Arctic ground squirrel. *Physiological Genomics*, 42A, 39


2009


tions cause a novel syndrome characterized by ataxia and mild mental retardation with predisposition to quadrupedal gait. *PLoS Genetics*, 5, e1000487


### Scientific honours

Hans-Hilger Ropers: *EURORDIS Scientific Award 2014*, European Organization for Rare Diseases, 2014

Hans-Hilger Ropers: *Honorary medal and honorary membership*, German Society for Human Genetics, 2009

### Selected invited talks

**(H.-Hilger Ropers)**


- *Genome Sequencing meets the clinics*. 1st International and 13th National Conference on Genetics, Tehran, Iran, 11/2014
Was uns das Genom erzählt. 23. Lessing-Gespräche, Jork (Altes Land), 11/2014

Predictive Genetic Diagnosis: Time for a Paradigm Shift. Exploratory Round Table Conference 2014 on Personalized Medicine: From Risk Factors to Disease Predispositions – 40th Anniversary of the MPG-CAS Cooperation. Shanghai, China, 05/2014

Beyond GWAS: New Sequencing techniques and their clinical utility. Microsomes and Drug Oxidations, Stuttgart 05/2014

Geschlechtschromosomen und Krankheiten. Joint Symposium Deutsche Akademie der Naturforscher Leopoldina/Österreichische Akademie der Wissenschaften (ÖAW), Geschlechtsabhängige Vererbung – mehr als Gender und Sex, Veterinärmedizinische Universität Wien, Austria, 03/2014

New Sequencing Techniques: Why they are indispensable for Health Care, and what we can do to speed up their clinical implementation (keynote). Clinical Exome Sequencing Conference, Lisboa, Portugal, 12/2013

Elucidation, diagnosis and prevention of intellectual disability: meeting the challenge. Golden Helix Conference, Dubai, 12/2013

On elucidating the genetic basis of intellectual disability (keynote). Frontiers in Biomedical Research, HKU 2013, Hong Kong, China, 12/2013

Disease gene discovery through NGS and data exchange. MEDICA, Düsseldorf, 11/2013

Diagnose ohne Therapie? Was können wir von der molekularen Genetik erwarten? Seminar of the Konrad Adenauer Foundation, Cadenabbia, Italy, 09/2013

On NGS in intellectual disability, research and diagnosis. MRC/IGMM Human Genetics Unit, University of Edinburgh, 02/2013

Intellectual disability: Genetic dissection of a complex disorder and implications for health care. Center for Clinical and Translational Studies, Rockefeller University New York, USA, 11/2012

Predictive diagnosis of cognitive impairment: Where are we heading? Molecular Neurobiology Today and Tomorrow; Russian-German Symposium of the Russian Academy of Sciences and the Berlin-Brandenburgische Akademie der Wissenschaften, Moscow, Russia, 04/2012


High-Throughput sequencing in Intellectual Disability Research and Health-Care: The Future is now (keynote). 16th Annual Meeting of the German Society of Neurogenetics, Erlangen, Germany, 10/2011

Molecular dissection of intellectual disability: of chromosomes aberrations, linkage studies and high-throughput sequencing (keynote). Mental retardation: from genes to synapses, functions and dysfunctions. CNRS-Conferences Jacques Monod, Roscoff, France, 10/2010
Appointments of former members of the department

Hao Hu: Professor of Genetics, Zhongshan School of Medicine, Sun Yat-sen University, and Director of the Department of Molecular Diagnostics, Guangzhou Women and Children’s Medical Center, Guangzhou, China, 2015

Reinhard Ullmann: Head of Research Group, Bundeswehr Institute of Radiobiology affiliated to the University of Ulm, Munich, 2014

Tim Hucho: Professorship (W2) on Anaesthesiology and Pain Research, University Hospital of Cologne, 2012

Andreas Walter Kuss: Professorship (W2) on Molecular Human Genetics, Ernst-Moritz-Arndt Universität, Greifswald, 2010

Sybille Krauss: Head of Research Group, Deutsches Zentrum für Neurodegenerative Erkrankungen (DZNE), Bonn, 2010

Habilitationen / State doctorates

Reinhard Ullmann: Chromosomenveränderungen in ihrer Beziehung zu humanpathologischen Erkrankungen. State doctorate (Habilitation) on Genetics and Cell Biology, University of Salzburg, Austria, 2011

PhD theses

Melanie Hambrock: Untersuchungen zur Funktion der Proteinkinase CDKL5 und des CDKL5-Proteinkomplexes. Freie Universität Berlin, 2015

Lucia Püttmann: Identification and characterization of gene defects underlying autosomal recessive intellec-

Jakob Vowinckel: Die Protein-Monoaminylierung als regulatorischer Mechanismus in der Signaltransduktion. Freie Universität Berlin, 2010

Masoud Garshasbi: Identification of 31 genomic loci for autosomal recessive mental retardation and molecular genetic characterization of novel causative mutations in four genes, Freie Universität Berlin, 2009

Stella Amrei Kunde: Untersuchung zur Funktion des PQBP1-Komplexes, Freie Universität Berlin, 2009

Artur Muradyan: SPON2 and its implication in epithelial-mesenchymal transition, Freie Universität Berlin, 2009


Franziska Sotzny: Der Einfluss der Serotonylierung auf die Proliferation von Karzinomzellen. Universität Bayreuth, Master thesis, 2010

Grit Ebert: Veränderungen der Genexpressionsmuster im Verlauf der Re-

Student theses


Fabian Roske: Die Veränderung der Aufnahme und Inkorporation von biogenen Monoaminen durch Adipogene-

se in 3T3-L1-Zellen. Universität Bayreuth, Bachelor thesis, 2009


Caroline Schwarzer: Etablierung eines immunologischen Nachweissystems zur Erfassung des Aktivierungszustandes der kleinen GTPase Cdc42. Universität Bayreuth, Bachelor thesis, 2009

Christine Technau: Monoaminylierung von Transkriptionsfaktoren am Beispiel des CREB. Freie Universität Berlin, Bachelor thesis, 2009

Teaching activities


Diego Walther, Technische Fachhochschule Wildau, Freie Universität Berlin

Organization of scientific events


Scientific services

Animal facility

Established: 2003

Head
Dr. Ludger Hartmann
Phone: +49 (0) 8413-1189
Fax: +49 (0) 8413-1197
Email: hartmann@molgen.mpg.de

Technician
Judith Fiedler (since 07/11, also transgenic unit)

Animal care takers
Nadine Lehmann (since 12/14, part time)
Niclas Engemann (since 08/14)
Heike Schlenger (since 07/14)
Larissa Schmidtke (since 08/13)
Anne Heß (since 03/13)
Katharina Hansen-Kant (since 04/09)
Christin Franke (since 09/07)
Dijana Wrembel (since 09/07)
Eileen Jungnickel (since 09/06)
Sonja Banko (since 04/05)
Mirjam Peetz (since 01/05)
Carolin Willke (since 09/02)
Julia Wiesner (since 05/02)
Katja Zill (since 06/99)
Ulf Schroeder (since 09/96, master)
Nadine Lehmann (08/11-10/14)
Janina Hoppe (09/08-08/12)

Apprentices
Vanessa Nolting (since 09/15)
Robin Walbrodt (since 09/15)
Celine Hilgardt (since 09/14)
Lisa Rieger (since 09/14)
Corinna Schwichtenberg (since 09/13)
Ceszendra Kaufmann (since 09/12)
Nathalie Michaelis (since 09/12)
Niclas Engemann (09/12-08/14)
Larissa Schmidtke (09/10-08/13)
Mareike Wegmann (09/11-03/13)
Laura Kühn (09/10-01/13)
David Brandenburg (09/09-08/12)
Sarah Hackforth (09/09-08/12)
Nadine Lehmann (09/09-08/11)

Service
2.5 persons from a service company
(cage washing etc.)
Overview

The animal facility of the institute provides an optimal research environment in the field of laboratory animal science, which includes the basic animal breeding and maintenance service for approximately 330 genetically modified and 30 wildtype mouse strains and technical services with a highly motivated staff. The mouse strains are kept under specified pathogen free (SPF) conditions in areas with restricted access. By using several physical barriers and standard operation procedures, we have been strongly committed ourselves to keep our rodent colony free of rodent pathogens. All strains are housed in individually ventilated caging systems (approximately 6,500 cages) and are handled under sterile conditions (with changing hood).

The animal facility provides high standard services which include

- animal husbandry
- colony management
- assistance in experimental design and techniques
- experimental work
- tissue biopsies
- blood and organ collection
- health monitoring (FELASA)
- cryopreservation of mouse embryos and sperm freezing (approximately 380 strains are cryopreserved)
- *in vitro* fertilisation (IVF)
- sterile embryo transfer
- training for researchers, care takers, and trainees
- import & export of animals
- quarantine
- rederivation

For the management of these mouse strains and the offered services, a mouse-colony management software program (PyRAT®) enables scientists to see and modify all their research data online.

The zebrafish facility of our institute is set up to raise and keep up to 15,000 zebrafish (*Danio rerio*). The aquaria system is located in the animal house and consists of approximately 150 single tanks that are used for breeding and maintenance of zebrafish lines, as well as for providing eggs, embryos and larvae to the researchers of the institute. For zebrafish embryo manipulation, the facility offers a DNA/RNA microinjection setup.
Scientific services

Transgenic unit

Established: 2004

Head
Dr. Lars Wittler (since 08/10, Dept. of Developmental Genetics)
Phone: +49 (0) 8413-1192/-1435
Fax: +49 (0) 8413-1197
Email: wittler@molgen.mpg.de

Technician
Judith Fiedler (since 07/11, also animal facility)

Overview
The mouse serves as the major animal model for investigating gene function in a mammal. Its genome is well characterized, defined genetic backgrounds are available and, most of all, the systematic and directed manipulation of the murine genome is easy to accomplish.

The Transgenic unit provides a centralized resource and standardized platform to utilize the technologies that are necessary to generate genetically modified mouse models for the institute. It consists of Lars Wittler (Dept. of Developmental Genetics, B. Herrmann), Judith Fiedler (Transgenic unit/Animal facility), Karol Macura (Dept. of Developmental Genetics, B. Herrmann) and Mirjam Peetz (Animal facility) with additional support from other members of the animal facility. Our team has built up a robust routine employing the morula aggregation technique. Morula aggregation can be used for both, the conventional generation of chimeras and for tetraploid complementation experiments, a method for generating embryos
that are almost exclusively made up by the donor embryonal stem (ES) cells. In the recent years, we optimized this platform in such way that it enables us to process up to 13 different ES cell lines per week. Since 2012, we perform approximately 300 experiments per year, about half of them as tetraploid complementation assays. For the majority of these experiments, F0 embryos are directly taken from the foster mother to analyze the genetic and molecular basis of embryonic development. However, we also establish about 60 novel genetically altered mouse lines per year.

Although the main methodological emphasis of the transgenic lab relies on ES cell-based technologies, with the rise of novel genome editing tools like the CRISPR/Cas9 technology, we also acquired a new microinjection workstation in 2014. Currently we are working out the various possibilities for CRISPR/Cas9-based applications by direct manipulation of zygotes through microinjection.

Besides the described routine generation of genetically modified mouse models, the transgenic lab also offers practical training and assistance in the generation of transgenic murine ES cell lines and ES cell culture, gives advice in planning the optimal strategy for the generation of genetically modified mouse models and contributes to the education of students and scientist of the institute in transgenic technologies.
General information

Complete list of publications (2009-2015)

Group members are underlined.

2015


2013

2012

Scientific services
Scientific services

Sequencing core facility

Established: 12/2007

Overview

During the last years DNA sequencing has undergone a revolution, which has changed the shape of biological research. By next generation sequencing techniques we are now able to get complete genomic sequences of model organisms, to resequence whole genomes or target regions, and to analyze methylation, expression and splicing patterns with very high throughput at relatively low cost.

To implement and accelerate the use of these new techniques within the Max Planck Institute for Molecular Genetics (MPIMG), the Next Generation Sequencing (NGS) Group was founded at the end of 2007. The group is a central service unit, which offers its service to all research groups of the institute. Since 2010 the acquired expertise is also provided to scien-
tists outside the MPIMG and the group attained the status of a Sequencing Core Facility for all institutes of the biological-medical section (BMS) of the Max Planck Society. Today our facility operates several next generation sequencers and maintains a fully equipped lab and staff able to perform a variety of sequencing applications - from sample preparation to data analysis. Currently we are providing expertise for two different technical platforms: Illumina Hiseq 2500/NextSeq 500 and Roche/454 FLX+ systems. The high throughput of our Illumina systems in terms of gigabases and number of reads produced per run offers a real advantage for many applications. Expression profiling (RNA-Seq), methylation analysis (MeDIP-Seq and Bisulphite-Seq), copy number analysis as well as the identification of protein binding sides (ChIP-Seq) and the analysis of whole exomes or genomes profit to a great extent from the large output of this system.

In our function as Max Planck Sequencing Core Facility our group was evaluated by the Max Planck Society at the end of 2013. A team of internal and external reviewers lead by the Vice President of the BMS examined our group for two days. An integral part of this evaluation was an extensive user survey. 90% of our collaboration partners attested us an excellent communication and a very short processing time. As overall result the committee rated our project management as outstanding.

As core facility our group is involved in a broad range of successful projects with other institutes, especially in the field of de novo sequencing of new model organisms.

**De novo sequencing of model organisms**

Many genome-related findings in science lead to questions, which can only be addressed in model organisms. For a decade large scale vertebrate or plant genome projects were only possible for extensively funded research groups or international sequencing consortia. Especially data production for gigabase-sized genomes required access to sequencing centers, which employed several hundred people to run Sanger sequencing. With the emergence of NGS technologies this picture completely changed. Today it is not challenging anymore to produce a high sequencing coverage of a genome using NGS technologies, the question is, how much data do you really need from which DNA library types and in particular how do you assemble it to get the best results possible? Furthermore, classical genome annotation strategies (ab initio or evidence-based gene prediction by protein homology) nowadays can be combined with NGS-derived RNAseq data to result in highly valid gene models including information on alternative splicing and expression levels, which in combination with a web-based visualization strategy are a key prerequisite for our collaborators to extract biological meaning from the assembled genomes.
During the evaluation period we have assembled and annotated a huge number of different genomes, like the canary (genome size about 1.2 Gb), the rusty-margined flycatcher (1.2 Gb), the blue tit (1.2 Gb), the purple-spined sea urchin (800 MB), the Iberian mole (2.5 Gb), as well as highly repetitive plant genomes, like coyote tobacco (2.4 Gb) and desert tobacco (800 Mb). Comparison of assembly statistics with related species that were already sequenced showed that our assemblies are of superior short and long range continuity (see Canary assembly below). All genome assemblies took profit from a mixture of different sequencing libraries and involved several assembly strategies. The genomes have been annotated by different approaches including a vast amount of RNAseq data of different tissues and are made available to our collaborators through a web-based version of the UCSC genome browser.

The canary is an excellent model for behavioral neuroscience, because it facilitates the study of speech learning, memory formation, hormone-driven sexual dimorphisms and behavioral (song) plasticity involving neural phenomena, such as large-scale brain restructuring and adult neurogenesis. The project started with the de novo sequencing of the canary genome and was then extended by complex transcriptome analysis of different brain regions of the canary and other singing and non-singing bird species. As first step we calculated a genome assembly of the canary generated by combining long-read and short-read next generation sequencing technologies as well as genome collinearity (Figure 2). The result was a high quality assembly and annotation of a female 1.2-Gbp canary genome. Whole genome alignments between the canary and 13 genomes throughout the bird taxa show a much-conserved synteny, whereas at the single-base resolution, there are considerable species differences.

Figure 1: Singing of the canary is regulated by hormones: The Canary Genome Project (collaboration with Manfred Gahr and Carolina Frankl, MPI for Ornithology, Seewiesen)
These differences impact small sequence motifs like transcription factor binding sites such as estrogen response elements and androgen response elements. To relate these species-specific response elements to the hormone sensitivity of the canary singing behavior, we identify seasonal testosterone-sensitive transcriptomes of major song-related brain regions, HVC and RA, and find the seasonal gene networks related to neuronal differentiation only in the HVC. Testosterone-sensitive up-regulated gene networks of HVC of singing males concerned neuronal differentiation. Among the testosterone-regulated genes of canary HVC, 20% lack estrogen response elements and 4 to 8% lack androgen response elements in orthologous promoters in the zebra finch. The canary genome sequence and complementary expression analysis reveal intra-regional evolutionary changes in a multi-regional neural circuit controlling seasonal singing behavior and identify gene evolution related to the hormone sensitivity of this seasonal singing behavior. Such genes that are testosterone- and estrogen-sensitive specifically in the canary and that are involved in rewiring of neurons might be crucial for seasonal re-differentiation of HVC underlying seasonal song patterning (Frankl-Vilches et al., Genome Biol 2015). This project demonstrates the need for high-quality genome assembly to detect the evolution of genes in comparative studies.
Cooperation within the institute

Within the institute the Sequencing core facility cooperates intensively with all departments and with the majority of the Otto Warburg groups (selected projects below):

- Bernhard Herrmann and Hermann Bauer (sequencing and analysis of the mouse t-haplotype)
- Martin Vingron (sequencing of the Sweet Potato)
- Stefan Mundlos and Dario Lupianez (sequencing of the Spanish mole genome)
- Hans Lehrach (OncoTrack project)
- Hilger Ropers and Thomas Wienker (whole exome sequencing of patients with mental retardation)
- Ho-Ryun Chung (DEEP project)
- Ralf Herwig (HeCaTos project)
- Thorsten Mielke (visualization of ultrastructure of cancer cells)
- Ulf Ørom (identification of ncRNAs)
- Sascha Sauer (ESGI project)
- Edda Schulz (single cell transcriptome analysis)
- Michal-Ruth Schweiger (transcriptome, whole exome and Medip analysis of prostate cancer patients)
- Ulrich Stelzl (yeast two-hybrid interaction screening approach involving next-generation sequencing)

Cooperation outside the institute

In our function as sequencing core facility of the BMS, our group was involved in a broad range of successful projects with other institutes, like the de novo sequencing of the canary genome (collaboration with Manfred Gahr, MPI for Ornithology), the coyote tobacco genome project (collaboration with Ian T. Baldwin, MPI for Chemical Ecology), the blue tit genome (collaboration with Jakob Müller, Dept. Kempenaers, MPI for Ornithology), the sea urchin genome (collaboration with Ulrich Benjamin Kaupp and Timo Struenker, Caesar Institute) or the analysis of the mouse B cell repertoire (collaboration with Hedda Wardemann, MPI for Infection Biology).

Outside of the Max Planck Society we are involved as data production site in large international sequencing projects, like the 1000 Genomes Project (www.1000genomes.org), the Oncotrack project (www.oncotrack.eu) the ESGI project (www.esgi-infrastructure.eu) or the HeCaTos project (http://www.hecatos.eu/). In the area of Berlin we have different collaborations running, like the analysis of T cell splicing (Florian Heyd, Freie Universität Berlin), the analysis of the basic principle of allergic contact dermatitis by sequencing T cell transcriptomes (Katherina Siewert and Hermann-Josef...
Thierse, Federal Institute of Risk Assessment) or whole exome analysis of patients with lymphoblastic leukemia (Renate Kirschner-Schwabe, Department of Pediatrics, Division of Oncology and Hematology, Charite). To optimize the workflow and to balance the sequencing load at peak times we established strong relationships to other core facilities. Especially with Andreas Dahl (Deep Sequencing Group, Biotechnology Center Dresden) and Jochen Hecht (Genomics Unit, Centre for Genomic Regulation, Barcelona, Spain) we are in a permanent process of exchange and collaboration.

Outlook

During the last years, the sequencing core facility accomplished a large number of different projects and in most cases these collaborations generated follow up projects, like the burrowing owl project with the MPI for Ornithology or the desert tobacco project with the MPI for Chemical Ecology. Currently we consult with more than a dozen Max Planck Institutes and other external sites that are interested in future collaborations.

In a joined effort with the Ralser Group at the University of Cambridge we just started the Alpine marmot (Marmota marmota) genome project. The idea that started this project is to get insights in the biological phenomenon of hibernation. Within the frame of this project, we would like to establish a high quality genome sequence and combine it with RNAseq to support gene annotation.

As key assignment of our site we are part of the evolution of sequencing technologies. We permanently evaluate and improve our protocols. As beta test site we are now taking part in a project with Oxford Nanopore to test the new MinIon technology. This technology delivers reads up to a length of 100 kb. Applications for this long read technology include de novo assembly of complex genomes, like repeat rich plant genomes, human genome phasing, and cancer sequencing. The aim of this collaboration is to establish the new technology and to make it available as a service for collaboration partners. With a similar motivation we are negotiating to become an alpha test site for the new sequencing system that Roche is developing together with PacBio. A first alpha system of this technique will be available at the beginning of 2016 and promises to combine long read length with a high accuracy. This system would be helpful for a broad range of different applications like genome sequencing or also for full length transcriptome analysis.

Besides the evaluation of new sequencing technologies we are currently establishing new workflows for sample preparation. Especially single-cell RNA-seq provides new opportunities for developmental biology or cancer research. Out of this reason the institute acquired a Fluidigm C1 system at the beginning of the year and we have now established different protocols, like single-cell transcriptome profiling.
General information

Complete list of publications (2009-2015)

Group members are underlined.

2015


Genome Biology, 16(1), 19 -


2014


Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. *Nature Communications*, 5, 3934

*among the list of collaborators: Timmermann B


*shared first author


*among the list of collaborators: Timmermann B

2013


2012


2011


Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, Clark AG, Yu F, Gibbs RA, *1000 Genomes Project* & Bustamante CD. Demographic history and rare allele sharing among human populations. *Proceedings of the National Academy of Sciences of the USA*, 108(29), 11983–11988 *among the list of collaborators: Timmermann B*


*among the list of collaborators: Timmermann B*


*among the list of collaborators: Timmermann B*


*among the list of collaborators: Timmermann B*
stem cells harbor homoplasmic and heteroplasmic mitochondrial DNA mutations while maintaining human embryonic stem cell-like metabolic reprogramming. *Stem Cells*, 29(9), 1338–1348


2010


*among the list of collaborators: Timmermann B


Kotschote S, Wagner C, Marschall C, Mayer K, Hirv K, Kerick M, Timmer-


*among the list of collaborators: Timmermann B*


*shared first author*


*equal contribution

2009


---

**Selected invited talks (Bernd Timmermann)**


Clusterkonferenz Gesundheitswirtschaft Berlin-Brandenburg 2014, Berlin, Germany, 10/2014

Transfusionsmedizinische Gespräche, Hannover, Germany, 02/2014

20. Arbeitstagung DGPF, Micromethods in Protein Chemistry, Bochum, Germany, 06/2013

NGS Pharma 2013, Frankfurt, Germany, 01/2013

MipTec 2012, Basel, Switzerland, 09/2012

Next Generation Sequencing in Research and Diagnostics, Basel, Switzerland, 09/2012

Sequencing Meeting Taiwan, Taipei, China, 09/2012

Next Generation Sequencing and Next Generation Molecular Diagnostics conferences, Prague, Czech Republic, 06/2012

IPFA/PEI Workshop, Budapest, Hungary, 05/2012

Analytica 2012, Genomic Sequencing Workshop, Munich, Germany, 04/2012

German Ethics Council, Public Hearing about Multiplex Technologies in Diagnostics, Berlin, Germany, 03/2012

44th annual congress of the DGTI, Hannover, Germany, 09/2011

Monash Institute of Medical Re-
search, Melbourne, Australia, 06/2011
Centenary Institute, Sydney, Australia, 06/2011
National Cancer Center, Seoul, South Korea, 06/2011
Chinese University of Hong Kong, Prince of Wales Hospital, Hong Kong, China, 05/2011
3rd Annual Meeting of NGFN-Plus and NGSN Transfer, Berlin, Germany, 11/2010
3rd Benelux Next-Generation Sequencing Users Meeting, Nijmegen, Netherlands, 11/2010
Jornadas Salud Investiga, Cadiz, Spain, 10/2010
Hematology Focus Group Meeting, Munich, Germany, 05/2010

21st Annual Meeting of the German Society for Human Genetics, Hamburg, Germany, 03/2010
29th German Cancer Congress, Berlin, Germany, 02/2010
Pharma IQ Meeting, Next-Generation DNA Sequencing, London, UK, 01/2010
Seminario Científico Biomedicina y Genómica, Barcelona, Spain, 12/2009
Molecular Biology Days Belgium, Brussels, Belgium, 10/2009
NGS 2009 Conference, Barcelona, Spain, 10/2009
Next Generation Sequencing Symposium, Dresden, Germany, 03/2009
Scientific services

Mass spectrometry facility

Established: 03/2012

Head
Dr. David Meierhofer (sind 03/12)
Phone: +49 (0) 8413-1121/-1567
Fax: +49 (0) 8413-1380
Email: meierhof@molgen.mpg.de

PhD students
Yang Ni (since 04/15)
Ina Gielisch (since 04/12)

Technician
Beata Lukaszewska-McGreal
(since 03/12)

Overview

Structure of the MS facility
The mass spectrometry facility provides proteomics and metabolomics support for the entire institute. It also conducts research in the field of mitochondrial pathologies.

Research concept
Metabolomics and proteomics explore the complexity and dynamics of metabolites and proteins in biological systems to understand their functions in the cell and in the organism. Mass spectrometry has evolved and matured to a level, where it is able to assess the complexity of the human proteome and metabolome.
In our laboratory, we established several mass spectrometry-based methodologies for identification and quantification of metabolites and proteins. We tuned and optimized pure metabolites and created a reference library for multiple reaction monitoring (MRM)-based LC-MS approaches. For high confidence in identification and relative quantification of metabolites, we are additionally calculating MRM ion ratios of the transitions. These are molecule-specific fragments from parent ions, generated in a collision cell of a mass spectrometer. Data sets are further analyzed and integrated by bioinformatic tools for protein-protein network and pathway analyses to elucidate the overall impact, e.g. of pathogenic mutations or treatments.

Our research focuses on the field of mitochondrial pathologies, a heterogeneous group of metabolic disorders that are frequently characterized by anomalies of oxidative phosphorylation, especially in the respiratory chain. Respiratory chain diseases (RCD) represent a large subset of mitochondrial disorders and are biochemically characterized by defective oxidative phosphorylation, leading predominantly to neurological and muscular degeneration. They occur at an estimated prevalence of 1 in 5,000 live births and are collectively the most common inborn error of metabolism. These pathologies show a wide spectrum of clinical manifestations and variation in the mode of onset, course and progression with disease.

We are studying pathogenic mutations, which cause a respiratory chain disease, especially of complex I (CI, NADH-ubiquinone oxidoreductase) genes, encoded by either the nuclear or mitochondrial DNA of human-derived fibroblast cells. A CI deficiency can cause Leber Hereditary Optic Neuropathy (LHON), Mitochondrial Encephalopathy, Lactic Acidosis, Stroke-like episodes (MELAS) and Leigh Syndrome or Leigh-like Syndrome (LS). We use fibroblasts, derived from patients carrying CI mutations or mutations leading to a CI deficiency, to explore altered protein expression and metabolic pathways in order to discover the pathomechanisms of a respiratory chain disease. Additionally, a CI deficiency was chemically mimicked by rotenone as a model, to elucidate the biochemical consequences of a defect respiratory chain.

Elimination of the respiratory electron chain by depleting the entire mitochondrial DNA (mtDNA, $\rho^0$ cells), which encodes 13 proteins of the respiratory chain as well as rRNAs and tRNAs, has therefore one of the most severe impacts on the mitochondrial energy metabolism in eukaryotic cells. We used $\rho^0$ cells as a model to study the mitochondrial energy metabolism.
Scientific achievements/findings

We reported a rapid LC-MS-based method for relative quantification for targeted metabolome profiling, with an additional layer of confidence by applying multiple reaction monitoring (MRM) ion ratios for further identity confirmation and robustness.

Pathway and protein-protein interaction (PPI) network analyses of a CI deficiency, mimicked by rotenone, revealed that the respiratory chain was highly deregulated. Metabolites such as FMN, FAD, NAD+ and ADP, direct players of the respiratory chain, and metabolites of the TCA cycle decreased up to 100-fold. Synthesis of functional iron-sulfur clusters, which are of central importance for the electron transfer chain, and degradation products such as bilirubin were also significantly reduced. Glutathione metabolism on the pathway level, as well as individual metabolite components such as NADPH, glutathione (GSH), and oxidized glutathione (GSSG) was down regulated.

The entire depletion of the mtDNA in a human osteosarcoma cell line resulted in a non-uniform downregulation of 55S mitochondrial ribosomal proteins, the respiratory electron chain, the tricarboxylic acid (TCA) cycle and the pyruvate metabolism in ρ0 cells (Figure 1).

Surprisingly, some metabolites of the TCA cycle were reduced in ρ0 cells, such as citric acid and aconitic acid (both 6-fold), others instead increased, such as lactic acid (2-fold), succinic acid (5-fold) and oxalacetic acid (2-fold). Exceptionally, the mitochondrial retrograde response, a pathway of communication from mitochondria to the nucleus, was up-regulated in ρ0 cells, such as signaling by GPCR, EGFR, G12/13 alpha cAMP and RhoGTPase. This was supported by our phosphoproteome data: The most significant up- as well as downregulated pathways were GTPase signaling and cytoskeleton organization. Furthermore, we observed a striking
de-ubiquitination, for example, 80S ribosomal proteins were 3-fold and SLC amino acid transporters 7-fold de-ubiquitinated in ρ⁰ cells. The latter might cause the observed significant increase of amino acids levels in ρ⁰ cells. We conclude that mtDNA depletion not only leads to an uneven downregulation of mitochondrial energy pathways, but also to a strong retrograde response.

In a cooperation project with the research group Development & Disease (Stefan Mundlos), we analyzed patients with a novel mutation in ALDH-H18A1 encoding pyrroline-5-carboxylate synthase (P5CS), causing a progeroid form of autosomal-dominant cutis laxa disease by a pulsed chase LC-MS experiment. Therefore, we quantified the flux through the glutamate-proline pathway to determine, whether an isotopically labeled substitution of glutamic acid interferes with the catalytic function of the enzyme. We provided ¹³C⁵¹⁵N glutamic acid to the cells and after 6 and 12 hr we determined the amount of ¹³C⁵¹⁵N-labeled proline by a targeted mass spectrometry approach. In the controls we found similar amounts of labeled proline at the indicated time points, whereas in the P5CS-deficient cell line and the fibroblast lines bearing the heterozygous p.Arg138Trp substitution a reduced proline accumulation was detectable (Figure 2).

Figure 2: Reduced proline accumulation in fibroblasts from affected individuals harboring the P5CS-p.Arg138Trp substitution. ¹³C⁵¹⁵N proline levels relative to Ctrl 1 6 hr values. The resulting ratios are given with a log₂ scale. Note clearly reduced efficiency of ¹³C⁵¹⁵N proline accumulation in cells harboring biallelic or monoallelic ALDH-H18A1 mutations.
Cooperation within the institute

- Hermann Bauer, Dept. of Developmental Genetics
- Sebastiaan Meijsing, Dept. of Computational Molecular Biology
- Hans Lehrach, Emeritus group Vertebrate Genomics
- Vera Kalscheuer; Uwe Kornak, Research Group Development & Disease
- Thorsten Mielke, Microscopy & Cryo Electron Microscopy Service Group
- Ulf Ørom, Otto Warburg Laboratory
- Sascha Sauer, Otto Warburg Laboratory
- Ulrich Stelzl, Otto Warburg Laboratory
- Marie-Laure Yaspo, Otto Warburg Laboratory

External cooperation

External cooperation’s are established with

- Alessandro Prigione, Max Delbrück Center for Molecular Medicine (MDC) Berlin-Buch
- Johannes Mayr, Wolfgang Sperl, Paracelsus Medizinische Privatuniversität, Salzburg, Austria
- Ilka Wittig, University of Frankfurt
- Georg Auburger, University of Frankfurt
- Volker Zickermann, University of Frankfurt
- Sascha Al Dahouk, Bundesinstitut für Risikobewertung (BfR), Berlin
- Holger Prokisch, Helmholtz Zentrum München
- Oliver Kann, University of Heidelberg
- Markus Schülke, Charité-Universitätsmedizin Berlin
- Georg Seifert, Charité-Universitätsmedizin Berlin
- Hudert Christian, Charité-Universitätsmedizin Berlin
- Hermann-Georg Holzhüter, Charité-Universitätsmedizin Berlin
- Silke Sperling, Charité-Universitätsmedizin Berlin
- Claus-Eric Ott, Charité-Universitätsmedizin Berlin

Special facilities / equipment

The following mass analyzers are available in our MS facility

QTrap 6500 from AB/Sciex

The AB SCIEX Triple Quad™ 6500 LC/MS/MS system merges triple quadrupole and linear ion trap capabilities and is one of the world’s most sensitive and selective triple quadrupoles. This mass spectrometer is used for targeted approaches (Multiple Reaction Monitoring, MRM), for both, metabolomics and proteomics approaches. For each metabolite or peptide
of interest, an individual method has to be created and optimized on the instrument. The QTrap can be online attached to an Agilent 1290 Infinity ultra-high pressure (1200 bar) liquid chromatography unit (routinely used for metabolite identification and quantification) or alternatively to a Dionex Ultimate 3000 nanoHPLC (routinely used for peptide identification and quantification).

For metabolome profiling, we tuned 400 pure metabolites in order to establish a reference set of MRM methods. Our metabolite library contains key cellular components such as amino acids, acyl-carnitines, ceramides, sphingomyelins, bile acids, carbohydrates, vitamins, hormones, nucleotides and biogenic amines. Metabolites are extracted by a water/methanol/chloroform or methyl-tert-butyl ester (MTBE) method including internal standards and are analyzed by an online coupled targeted LC-MS/MS approach. As metabolites have very diverse chemical features, they cannot be separated by only one LC condition. Hence, we use following three different columns for an adequate separation of metabolites: Reprosil-Pur C18-AQ, a zicHILIC, and an ACQUITY UPLC BEH Shield RP18 column.

Three transitions (molecule specific fragments from parent ions, generated in a collision cell of a mass spectrometer) are measured for each targeted metabolite via multiple reaction monitoring (MRM) in positive or negative electro spray ionization (ESI) mode. Peak areas from the total ion current for each metabolite transition are integrated using Multi-Quant software (sMRM; AB/Sciex). For a comprehensive data identification and relative quantification, we additionally determine the MRM ion ratios and compare to our tuned metabolites. Only metabolites, which show the identical retention time and MRM ion ratios are considered as identified. Pathway analysis of differentially expressed metabolites is further explored with bioinformatics tools.

**Q Exactive Plus from Thermo Scientific**

The Exactive Plus Orbitrap MS is a benchtop LC-MS system for high-performance, high-throughput screening, compound identification, and quantitative analysis. The Exactive Plus Orbitrap delivers high-resolution, accurate-mass full-scan MS for fast, precise and reproducible results. Our Q Exactive Plus Orbitrap is equipped with a Dionex Ultimate 3000 nanoHPLC.

For proteome profiling, we apply fractionation by strong cation exchange chromatography of digested proteins prior to LC-MS analysis. For post-translational modifications, e.g. phosphoproteome or ubiquitylome, we use specific enrichments like TiO$_2$ beads for phosphorylated peptides or specific antibodies for ubiquitinylated peptides.
NanoHPLC-MS/MS runs are quantitatively analyzed by MaxQuant software, running on a 8 and 16 core Dell server, by either using stable isotope labeling by amino acids in cell culture (SILAC) or the label-free quantification algorithm. Subsequent statistical analyses are performed using Perseus, a post data acquisition package of MaxQuant. Pathway analyses of differentially expressed proteins are further explored with bioinformatic tools such as Gene Set Enrichment Analysis (GSEA). Protein-protein interaction network analyses and identification of hub proteins are done by programs such as String and Cytoscape.

Planned developments

We will continue to establish valuable LC-MS methods to support the needs of the institute as well as our own research interests, ranging from new instrument configurations to secondary data analyses tools.

General information

Complete list of publications (2012 – 2015)

Group members are underlined.

2015


2014

Cui H, Schlesinger J, Bansal V, Dunkel I, Meierhofer D & Rickert-Sper-
ling S (2014). Regulation of myogenesis via kinase driven activation of DPF3a, a BAF complex member and its interaction with transcription repressor HEY1. *Cardiovascular Research*, 103, S1


2013


2012


**Selected invited talks (David Meierhofer)**

*Omics profiling of ρ⁰ cells (without mtDNA) triggers the retrograde response and de-ubiquitination of solute carrier amino acid transporters*. 2nd International Symposium on Profiling, ISPROF, Carapica-Lisbon, Portugal, 09/2015

*Integrated metabolome- and proteome profiling of complex I deficient cells, applying internal multiple reaction monitoring (iMRM) ratios for improved metabolite identification*. 7th OpenMS User Meeting “High-performance software for high throughput proteomics and metabolomics”, Berlin, 09/2014

*ρ⁰ cells, a Metabolome and Proteome survey of cells lacking mitochondrial DNA (mtDNA)*. International Workshop on MS-Based Proteomics Bioinformatics and Health Informatics, Izmir, Turkey, 05/2014

*Protein sets define disease states and predict in vivo effects of drug treatment*. Leadnet meeting at Castle Waldthausen, Mainz, 05/2013
Scientific services

Microscopy & Cryo-Electron Microscopy Group

Established: 1978 (microscopy); 2004 (cryo-electron microscopy)

Overview

From imaging service to structure determination of macromolecular complexes

The microscopy and cryo-electron microscopy service group headed by Thorsten Mielke combines the former microscopy group headed by Rudi Lurz from 1978 until his retirement in 2012 and the cryo-electron microscopy group which was installed in 2004 within the framework of Berlin-Brandenburg-wide research consortia “UltraStructure Network” (USN) and “Anwenderzentrum” (AWZ). Our group provides a broad range of imaging techniques for all departments and OWL research groups of the institute. The lab has established a wide range of transmission electron mi-

* externally funded
Scientific services

croscopy (TEM) methods such as ultra-thin sectioning of plastic-embedded samples, immune-labelling of sections or isolated structures as well as visualization of nucleic acids and nucleic acid-protein complexes using metal-shadowing technics.

We further provide a technology platform for cryo-electron microscopy including sample screening, semi-automated sample vitrification, data acquisition as well as computing resources for image processing with a strong focus on structure determination of macromolecular protein complexes using single particle methods. Single particle cryo-EM has emerged as a key technology in structural biology, and ground-breaking technological progress has been made in recent years to reach near-atomic resolution. After establishing our cryo-EM facility within the “UltraStructure Network”, our successful cryo-EM activities are now well-embedded within the Berlin research landscape contributing e.g. the central service project Z1 to the collaborative research centre SFB740 and joined activities within the NeuroCure cluster of excellence and the SFB958.

Due to the continually increasing demands on light-microscopic techniques at the institute, the former microscopy group took over the responsibility for technical service, maintenance and training of an increasing number of users and light microscopes (mainly fluorescence microscopes), which are operated as shared equipment and which are accessible for all members of the institute. The joined microscopy and cryo-electron microscopy group now continues these service activities and supports all users according to their specific biological questions. Besides routine service, we also aim at implementing new technologies and applications to serve the increasing demands on imaging techniques within the institute for visualization of biological structures. In the last years, we hereby focused on automated data collection strategies for cell-biological TEM applications and correlative microscopy techniques in order to bridge light and electron microscopy.

Special facilities / equipment

The group currently operates three transmission electron microscopes, which have been moved into newly built EM-facility rooms in spring 2015. Two LaB6 instruments (a 100 kV Philips CM100 equipped with a TVIPS Fastscan CCD camera and a 120 kV FEI Tecnai T12 Spirit equipped with a TVIPS 4k F416 CMOS camera) are used for routine sample screening, imaging of ultrathin-sections and initial cryo-data collection. The core instrument of our facility is a 300 kV Tecnai G2 Polara cryo-electron microscope (FEI), which has been upgraded with k2summit direct electron detector (Gatan) in May 2015, jointly funded by the NeuroCure cluster and the collaborative research centres SFB740 and SFB958. Additionally we provide the infrastructure for TEM sample preparation including carbon
evaporators, ultra-microtomes and a semi-automated Vitrobot cryo-plunger (FEI). In April 2015, we additionally installed a new Leica UC7 ultra-microtome equipped with a FC7 cryo-chamber suitable for Tokuyasu cryo-sectioning and sectioning of vitrified samples (CEMOVIS).

In light-microscopy, we support a broad range of epifluorescence and confocal microscopes including an Axio-Imager Z1, a LSM510meta, a LSM 700 (all instruments from Zeiss), and a Cellomics high-content array scanner (Thermo-Fischer Scientific). In particularly to support the work of the new OWL groups of Edda Schulz and Zhike Zi, the institute acquired a new inverse fluorescence microscope (Axio-Observer Z1, Zeiss) equipped with a life-cell-imaging chamber in December 2014. Furthermore, we upgraded our Zeiss LSM710NLO two-photon laser-scanning microscope with additional laser lines (440 nm and 405 nm) in June 2015. This instrument was especially configured for two-photon imaging and can now be accessed by a broader range of users. Since imaging of large scale biological objects such as mouse, zebrafish or chicken embryos becomes increasingly important for many groups studying developmental or disease processes, the institute decided to acquire a Zeiss lightsheet.Z1 microscope within 2015. Light sheet technology allows for imaging of fixed and living samples with up to 5 mm thickness (10 mm, when imaging from two sides) at dramatically reduced light intensities. In 2016, several of our light microscopes have to be moved due to the refurbishment of tower 1. Therefore, and to cope with the increasing imaging demands, the institute plans to centralize shared light microscopes in joint facility rooms in direct neighborhood to the EM lab.

Activities of the microscopy group

Besides light microscopy support, most of our projects within the institute are on ultra-thin sectioning of tissues, cells and cell components. We hereby collaborate with all departments and research groups having a wet lab. This involves individual projects to answer specific scientific questions as well as long-lasting collaborations. To give examples for individual projects, we visualized *Leishmania tarentolae* cells (collaboration with the Konthur group; Klatt et al., J. Proteome Res 2013), studied stress granules in Hela cells upon treatment with Fluorouracil and sodium arsenide (collaboration with the Krobitsch lab; doctoral thesis C. Kähler, FU Berlin, 2013) and analyzed the location of proteins e.g. within mouse sperm cells (Herrmann lab; doctoral thesis S. Schindler, FU Berlin 2012). Examples for long-term collaborations are TEM studies related to stem cell differentiation (together with the Adjaye lab), cancer (Schweiger lab) and other disease processes such as segmental progeroid disorders (Mundlos lab, Björn Fischer-Zirnsak).
Furthermore, the microscopy group collaborated for many years successfully with a large number of national and international research groups e.g. on degenerative brain diseases (AG Wanker, MDC Berlin; AG Multhaup, FU Berlin). Classical mica-adsorption techniques have been applied to analyze protein-DNA interactions involved in bacterial and yeast replication (AG Speck, Imperial College, London, UK; AG Zawilak-Pawlik, Institute of Immunology and Experimental Therapy, Wroclaw, Poland; AG Bravo, CIB, Madrid, Spain). Another focus lay on the structure, assembly and infection-mechanism of bacterial phages such as SPP1 infecting \textit{B. subtilis} (collaborations with P. Tavares, CNRS, Gif-sur-Yvette, France; P. Boulanger, Université de Paris-Sud, Orsay, France; C. Breyton, CNRS, Grenoble, France, and M. Loessner, ETH Zürich, Switzerland, amongst others). Originally introduced by the former director Thomas Trautner, SPP1 phages until today serve as a model system for eukaryotic dsDNA viruses. Some of these collaborations ended with the retirement of Rudi Lurz, others are continued e.g. the analysis of DnaA binding to the oriC region in \textit{Helicobacter pylori} (Donczew et al., J Mol Biol 2014) and cryo-EM studies on phage assembly intermediates (together with P. Tavares and E. Orlova, Birkbeck College, London, UK).

Within the collaborative research centre SFB740 (project Z1) our group aims at implementing new technologies to overcome three main hurdles in single particle cryo-EM, namely sample heterogeneity, the large amount of data required and the poor contrast and high noise-level of cryo-EM data. In single particle cryo-EM, 3D information is derived by averaging over thousands if not hundred thousands of projection images of individual molecules (referred to as particle images), assuming that they are all identical. In practice, however, preparations of protein complexes rather have to be considered as heterogeneous samples due to variable complex assembly and conformational flexibility. Moreover, it is now commonly accepted that these conformational states often represent intermediate states within a metastable energy landscape; and tuning this energy landscape e.g. by ligand binding and/or hydrolysis of nucleotides is the fundamental basis for the biological function of these complexes (Munro et al., Trends Biochem Sci 2009).

Together with C. Spahn (IMPB, Charité Berlin), we implemented and tested multiparticle refinement strategies developed in his lab to analyze sample heterogeneity \textit{in silico}. Introducing 3D variance analysis and variability mapping (Loerke et al., Methods Enzymol 2010) we could split heterogeneous data sets of bacterial as well as eukaryotic ribosomal complexes into sub-populations representing different conformational and/or functional states. This allowed us to identify new functionally important conformational modes occurring during the elongation cycle of protein biosynthesis such as a novel intra-subunit pe/E hybrid state showing a partly translocated tRNA (Ratje et al., Nature 2010), the large swivel movement of the
30S head (Ramrath et al., Nature 2012), new translocational intermediate states (Ramrath et al., PNAS 2013), spontaneous movements of tRNAs on the eukaryotic PRE 80S ribosome (Budkevich et al., Mol. Cell 2011) and subunit rolling between the mammalian PRE and POST complexes (Budkevich et al., Cell 2014). Furthermore, we could determine several structures of ribosomal complexes involved in IRES-mediated initiation (Yamamoto et al., Nat Struct Mol Biol 2014) and termination (Muhs et al., Mol Cell 2015).

Multiparticle refinement, however, requires even larger data sets comprising up to a million of particles and more. We therefore implemented the Leginon system (Suloway et al., J Struct Biol 2005) for automated data collection on our Spirit and Polara microscopes, which allows us to acquire up to 20,000 digital micrographs per week at routine level. Leginon is also used for automatic acquisition of tilt-pairs required for random conical tilt analysis as well as tomographic tilt series. Another major breakthrough in cryo-EM was the recent introduction of direct electron detector devices (DEDs). In contrast to scintillator-based camera systems, DEDs directly register primary electrons and thus provide a much higher sensitivity and dramatically improved signal-to-noise ratio. We had the great chance to test the prototype of a Falcon II DED on a Titan Krios microscope with FEI in Eindhoven and a Gatan k2summit DED on our Polara. Combining multiparticle refinement, automatic data collection and the potential of new DED detectors, we could identify more than 10 different functional states from ex vivo derived human polysomes and refine the highest-occupied state, the 80S POST-complex, to near-atomic resolution (3.4 Å using the 0.143 criterion, Behrmann et al., Cell 2015, Figure 1). In May 2015, we installed a k2summit DED on our Polara, which will
now allow us to obtain high-resolution structures also from smaller protein complexes such as RNA-polymerase (collaboration with M. Wahl, FU Berlin), TFII-complexes (collaboration with J. Kuper, Virchow Center for Biomedical Research, Würzburg) or the DPE2-complex from plants (DGF project Mi 940/1-1), amongst others.

In collaboration with Stephan Sigrist (FU Berlin), we applied electron tomography to study electron dense specializations (“T-bars”) at neuromuscular terminals of Drosophila larvae, comparing the effect of different isoforms of the “Bruchpilot” protein on proper T-bar formation (Matkovic et al., J Cell Biol 2013). T-bars represent a protein scaffold, which structures the release of synaptic vesicles at the presynaptic active zone. To proof the hypothesis, that age induced memory impairment (AMI), which is also observed in flies and which can be supressed by restoring juvenile polyamine levels (Gupta et al., Nature Neuroscience 2013), is related to structural rearrangements at the active zone, we implemented Leginon for automated data collection on ultrathin-sections, which enables us to image regions up to 100 µm$^2$ at a resolution of ~10-20 nm (Figure 2). The acquired micrographs (> 1000 per region of interest) are then stitched to one large image using FIJI. Imaging the calyx-region of Drosophila brains from animals at different age and feeding conditions, we could show statistically that the number of active zones and the density of synaptic vesicles per synaptic bouton decrease with age. Spermidine treatment did not restore these changes, instead, it suppressed an increase of both, the average size of T-bars and the average number of Bruchpilot molecules within this scaffold, indicating significant restructuring at the active zone with increasing age. We further applied automated TEM to study the differentiation of induced pluripotent stem cells (iPSCs). In collaboration with J. Adjaye, we found that bile canaliculi, a unique ultrastructural feature of hepatocytes,

Figure 2: (A) Montage of 1,080 TEM images showing the entire mushroom body calyx region from the Drosophila brain. (B) Zoom into the region marked in (a) Several bouton regions are visible, one of which showing a synaptic active zone is highlighted in (C).
start to form within the hepatic endoderm (HE) state. Similarly, we could show that iPSCs derived from patient fibroblasts, which potentially can be used as a model to study rare mtDNA-linked genetic diseases, develop a typical triangular shape upon differentiation towards neuron progenitor cells (collaboration with A. Prigione, MDC Berlin).

These findings clearly demonstrate the potential of automated TEM to visualize the ultrastructure of entire cells and excerpts from tissues. In order to combine ultrastructural information provided by TEM and the power of fluorescence microscopy, which however is limited to sub-micrometer resolution according to Abbe’s law, we now aim at implementing correlative microscopy approaches. In preliminary experiments, we could localize the stem cell specific marker protein TRA1-60 using both, confocal light microscopy and automated TEM-imaging on immuno-labelled ultrathin sections (Figure 3). In collaboration with M. Schweiger (now University of Cologne) and G. Schönfelder (Bundesinstitut für Risikobewertung, Berlin) we are currently testing correlative approaches including Tokuyasu cryo-sectioning techniques to study cultured cancer cells and Xenocraft tumor models, also aiming at potentially linking information at ultrastructural level (e.g. the occurrence of autophagosomes) to sequencing data obtained by our in-house sequencing facility. Similarly, a statistical analysis of sub-cellular structures such as lysosomes and autophagosomes related to defects in the enzyme Pyrroline-5-carboxylate reductase 1 (PYCR1) in fibroblasts from patients suffering from a not yet described syndrome with severe progeroid features (collaboration with the Mundlos lab) may be linked to proteomics data obtained by our mass spectrometry facility.
General information

Complete list of publications (2009-2015)

Group members are underlined.

2015


**2014**


2012


2011


**2010**


2009


Scientific services

IT Group

Established: 1995

Heads
Donald Buczek (IT group since 12/96; head since 01/11)
Phone: +49 (0)30 8413-1433
Fax: +49 (0)30 8413-1432
Email: buczek@molgen.mpg.de

Peter Marquardt (IT group since 08/95; head since 01/11)
Phone: +49 (0)30 8413-1430
Fax: +49 (0)30 8413-1432
Email: marquardt_p@molgen.mpg.de

IT staff
David Schrader (since 09/15)
Marius Tolzmann (since 08/07)
Sven Püstow (since 10/96)
Frank Rippel (since 01/95)
Alfred Beck (since 05/93)
Lena Zander (06/14-06/15)
Henriette Labsch (10/14-04/15)
Matthias Rüster (08/12-09/13)
Tobias Dreyer (09/11-08/13)

Apprentices
Robin Nielas Hofmann (since 09/15)
Tobias Gomes Danzeglocke (since 09/13)
Lena Zander (08/11-06/14)
Maximiliano Orsini (09/12-04/13)
Matthias Rüster (08/09-08/12)
Tobias Dreyer (09/08-08/11)

Students
Lena Zander (since 06/15)
Merlin Buczek (02/14-11/14)
Henriette Labsch (09/09-09/14)
Elmar Frerichs (11/12-11/13)
Overview

In January 2011, the responsibility for the Information Technology (IT) group has been taken over by Donald Buczek and Peter Marquardt. The IT group is in charge of the operation and development of the IT-infrastructure of the institute. This includes workstation and server systems, storage, archives, wire-based and wireless LAN, Internet access, Internet services and remote access. Examples of our current activities are listed below.

Windows desktop and servers

Most scientists use a Windows-based desktop for their research documentation. Over the years, we inherited a wide range of different versions of Windows installations and a huge variety of hardware. In 2012, we started a complete rollout and exchange program and replaced all used desktop workstations with new hardware running Windows 7 to get rid of Windows 98/Vista and XP systems. This also opened the possibility to centralize our software deployment. Today, we are able to roll out desktop machines, on which programs like Internet Explorer, Outlook and Adobe Acrobat Reader are deinstalled and replaced by more reliable and secure alternatives. This adds another layer of security to our system, so that virus and Trojan infections are quite rare. We still have a couple of scientific devices running with ancient versions of Windows, but they have been put in a separate network without direct connection to our network. We also establish virtual Windows installations accessible via a remote desktop for scientists, who need special applications for a limited time.

Storage

In recent years, the requests of NGS technology have dramatically pushed the requirement for cost effective storage. To accomplish this, we developed a storage concept based on generic hardware, almost independent of any hardware manufacturer or operating system release, and with constant monitoring and exchange. The concept involves “cheap” hard disks with five years manufacturer warranty.

The warranty is the lowest level of the data persistence concept. If a disk fails within the first four years, we replace it based on warranty. After four years, we observe that the replacement disks have more errors than the ones, we originally sent in. Since 2012, when we had 3,000 disks online, we had to replace up to 50 disks per week at peak times.

The next level of data persistence includes homogenous storage infrastructure like disk enclosures and raid controllers. We use only one specific type of hardware RAID controller to be able to swap, exchange, or replace the
underlying disks anywhere in our storage pools. All disks have the same physical dimensions, run in the same type of enclosures, and are interconnected with the same cables to all the same kind of controllers. In addition, the RAID configuration is always the same for all units. Due to RAID level 6 we are able to keep our data consistent, even when two out of sixteen disks fail.

The third level targets error monitoring. Starting in November 2009, we developed a daily logging mechanism, which keeps track on all relevant disk error statistics. For each disk, we log its SMART data (Self-Monitoring, Analysis and Reporting Technology) every 24 hours. Depending on the daily output, we decide with anticipatory obedience, which disk is going to die and replace it as soon as possible. Over the last six years, we processed more than 10,000 hard disks this way, so a lot of long term experience flows into the interpretation of individual hard disk errors or influence of infrastructural impacts like physical moving, power cuts and temperature changes.

Level four of our persistence concept targets the filesystem. NGS data requires a filesystem capable of handling billions of files and terabytes of data. As an internal decision, we work with an open source filesystem, which dictates the use of XFS; a system also supported by almost all Unix-like operating systems (OS).

Level five defines distinct identification of the units in our network. All units have individual RAID and file system labels, which will never change. They are accessible inside our network independent of their content. Even every single disk is labelled, when moved physically.
Level six is based on mirroring. Due to increasing hard disk size, the loss of complete units requires extremely long time to restore the lost data from backup. To prevent this, we provide a second unit, where the complete unit is synced on overnight. So in the worst case, we only lose changed files of a couple of hours.

Level seven is called backup. We backup everything in our Linux server farm; individual project data folders are only excluded on demand or when they fill up beyond one terabyte. This policy keeps the size of our backup structure clearly laid out and makes it possible to preserve any version of a file for 120 days. For all excluded project data folders, we rely on persistence concept level 1 to 5 or 6.

**Backup**

Backup is an essential service we offer. As long as files are stored on our served disk units, they have backup, except the files are larger than one terabyte or excluded on demand. The backup server itself has the same hardware setup as all the other file servers, but with disk units marked “confidential” attached. To gain even more security, we are running two backup servers, which are separated by open ground. Each backup server is saving the files from all servers in the opponent server room, with the two rooms located in different towers of the institute building. In case of damage to one room, we still have all the data available in the other room. In addition, we also participate in the Joint Network Center (see below) to improve the security level of our archives. Even in case of major structural damages affecting the whole institute building, our long-term archives are safely stored on a remote archive server in the GNZ.

Technically, our backup concept has two approaches. The first and most important is to keep the system open source. We don’t allow any software and hardware maintenance licenses or any form of manufacturer dependency. This lesson has been learned years ago, when our old DEC/OSF Digital Unix backup servers went out of service. Based upon our well defined hardware storage concept, we keep track of individual project data folders in central distributed maps. These data folder locations, including operating system, home directories or encapsulated software installations, are handled by either one of the backup servers and the complete folders get synced on a differential base to a defined confidential file system. Our sync method stores only changed or new files in comparison to the last backup run. This reduces the occupied space on the backup server to a minimum.

The second approach is an extremely simple way to handle the data. By prefixing a date code to the folder names, we have entire copies of every folder at that backup cycle including permissions, ownership and mod-
ification dates. Upon user request to restore a file, we just walk into the filesystem with standard Unix command line tools and copy the requested files back to their original location. Keeping everything online allows us to be able to search through contents or create difference log files of any file. The only condition is that the files had to be online at least once during the last four months. In these cases, restoration can be done instantly without waiting for tapes or time slots.

All software to accomplish these approaches is written in Perl and developed in our IT group. The system runs on any machine in our network without limits; we also are able to set up additional servers, or add disk space or network bandwidth within a couple of minutes.

Archives

Long-time archiving of data is an extremely important topic. According to the Max Planck Society Rules of Good Scientific Practice, we have to keep primary data for at least ten years and, of course, have to ensure that the data remains readable for this period of time. To accomplish this task, we classify our data in homogeneous raw data and heterogeneous generic data.

**Homogeneous raw data** consists of well-defined and structured output from sequencing machines, microscopes, or other devices generating scientific data. The amount of data varies over time, when newer technologies get established or the scientific background changes; but all devices still generate terabytes of data. As archiving these on tape robots is extremely expensive, we developed a simple concept for big data archives built on our storage concept. We copy the data directly to our well-managed storage units and remove them physically, when they are full. Storing a mirror unit in a different location adds a second level of persistence. To fulfill the requirement of keeping the archives at least ten years, we migrate units to freshly installed units after five years, what is exactly the duration of hard disk warranty. Due to hard disk size evolution, the new units for archiving are usually about four times larger than the old ones, so that we can copy four warranty-expired units on it. The old disks get physically destroyed and trashed, and the data location map gets adjusted.

**Heterogeneous generic data** like home directories, mail folder, project directories, and unstructured data are managed regardless of their contents. They are considered as a folder with a specific path to a given time. This metainformation is converted to a folder name on our archive server. When archiving, the original data get mirrored into this folder. The archive server then splits this folder into small archive files that get encrypted, compressed, and transferred to the GNZ, where it is stored on a tape robot.
Compute cluster software

Especially due to the development of new sequencing techniques, the computational and storage needs for data processing and analysis have dramatically increased. To fulfill the computational demands, we developed a job queuing system, which fits absolutely firm in our server setup concept. The idea is to have a generic compute job submitting system, which allows the computational power to be used as effective as possible. The queuing system consists of a MySQL database and clients without any master server. We start clients on compute nodes and are able to split a single machine into different nodes for distinct purposes, with CPU cores and memory divided. The users submit their jobs via simple commands. The compute nodes query the database for jobs, which fit to the node limits; the jobs get started; and job status information is fed into the database. The system runs productive on 20 multicore compute servers supplying about 1,200 CPU cores, and sums up to 12 TB RAM, and we are still working on developing it further.

GitHub Enterprise for MPG.DE

Our in-house software development, either from the bioinformaticians of the different department and groups, or from our IT service group, requires a distributed revision control system, which as state of the art is ‘git’. For many years, we were used to put our work on public servers like github.com, but recently, the demand for private repositories came up, to keep scientific work in-house until it is published. In 2015, we decided to invest in a locally installed GitHub Enterprise Instance, a commercial product that allows us to create and manage private repositories running on our own hardware. After successful tests, we went productive and are now able to offer accounts to every other member of the Max Planck Society and their collaborative partners. Up to now, we have 126 registered users from 19 MPIs and a few cooperation partners in Europe, China, and the US, as well as more than 250 repositories.

GNZ cooperation

For many years, the Max Planck Institutes in the Berlin-Brandenburg region join forces to run a Joint Network Center (Gemeinsame Netzwerkzentrum, GNZ) located at the Fritz Haber Institute of the Max Planck Society in Berlin. The GNZ offers several IT services to its members. All heads of the IT groups of the Berlin-Brandenburg MPIs meet regularly to discuss, plan, or present issues, ideas, and strategies concerning MPG-related IT. These meetings usually concentrate on local matters and are very productive. Documented protocols since 2003 are online available for reference.
**Education**

The IT group is very active in the training and education of young technicians, students, trainees, and apprentices. Our IT apprentices participate at the Bundeswettbewerb Informatik regularly and already entered the competitions round two out of three successfully.

**Material resources, equipment and spatial arrangements**

The online storage capacity of the MPIMG file servers exceeds 7 PB spread over approximately 3,200 hard disks. The active disk-based backup capacity sums up to about 420 TB, whereas the tape and disk-based offline archived data currently comprises about 2 PB. Presently the group serves about 250 Windows-based PCs and 220 Linux/Unix systems with a variety of hard- and software components and about 80 OSX systems. A variety of web servers are protected by a fire-wall installation, about 50 web servers are active and maintained. The active hard- and software development of the group serves the scientific departments as well as the service and administration groups.

The IT group installed a storage capacity of more than 7000 TB. Currently, we are running about 4,300 CPU cores with 26,000 GB RAM spread over about 220 Linux systems ranging from single core systems with 256 MB RAM up to 80 multicore servers with 2 TB RAM including our dedicated compute cluster.

Our internal network backbone is based on 10 GbE technology and is currently fed by approx. 50 interconnected network interfaces, from Isilon storage systems via multicore compute servers up to huge file- and archive servers. The in-house LAN is segmented by about 150 manageable switches giving us the flexibility to control each segment and, if necessary, to configure each switch port individually.

To supply a stable and reliable infrastructure for our IT equipment, we planned and implemented two physically separated server rooms. The storage and archive server room located in tower 4 is capable of supplying 180 kW cooling capacity and houses 20 server racks. The room has been reconstructed life without service interrupts from a laboratory equipment room with free flow air cooling to a closed cold aisle containment system. In tower 3, a second server room comprising a warm aisle containment system was implemented. It is capable of cooling down 450 kW with full redundancy and contains 30 racks.
General information

Complete list of publications (2009-2015)

Group members are underlined.

2012

2011


2010


Scientific services

Library

Librarians
Praxedis Leitner, M.A.
Dipl.-Inf. Sylvia Elliger

Phone: +49 (0)30 8413-1314
Fax: +49 (0)30 8413-1309
Email: library@molgen.mpg.de

Overview

The library is more than a room; it connects people with information by providing diverse resources and services to the academic community. The main responsibility is to offer the scientific literature as comprehensively as possible, primarily to the institute’s staff. The library collection is focused on the fields of research covered by the departments and independent research groups of the institute: Computational Molecular Biology, Developmental Genetics, Human Molecular Genetics, Vertebrate Genomics, Development & Disease, Neurosciences, Mass Spectrometry, and Microscopy.
- Holding of about 75,000 volumes (monographs, journals, series, proceedings etc.)
- e-journals licenses: 44 packages of scientific publishing houses
- e-book collections: more than 500,000 titles
- Databases and aggregators: 23 (ever-increasing)
- Managing the institute’s own publications via PuRe (Publication Repository of the MPG, formerly PubMan)
- Migrating from the actual web opac Allegro C to the Open Source discovery system VuFind

The library is responsible for acquiring, collecting, processing, storing and distributing all printed and electronic scientific literature needed by the institute’s scientists, students, project staff. As a scientific service unit, the library strives to create a dynamic e-information environment that fosters research, learning and innovation. The library’s goals are

- to develop and maintain high quality collections with an emphasis on digital content management for timely access to information resources;
- to provide services and tools that support research and education.

Support for reference management systems is provided to the faculty, staff and guests of the MPIMG. All offers and services like online catalogue, access to databases and ejournals, ebooks, internet services and intranet are permanently checked and developed according to the needs of our scientists.

As a modern research library, we take on transformative ventures, many of which will change the very nature of libraries. These projects, for example PuRe, involve innovative partnerships with peer libraries and companies worldwide and range from augmenting our existing strengths to improve search and academic collaboration with new technologies. We organized the XXXVII. Library Meeting of the Max Planck Society (12th to 14th May 2014) with 140 participants, which took place in the new building of our institute. At current time, the librarians of the MPG create a linked data environment for discovery and navigation among the rapidly, expanding array of academic information resources.

**Memberships**

- Permanent guest of the “Friedrich-Althoff Konsortium”, Berlin
- Member of the “Arbeitsgemeinschaft der Spezialbibliotheken”
- Member of the OpenAccess Initiative within the Max-Planck-Society
Research Support

Administration

Head
Dr. Christoph Krukenkamp  
(since 10/12)  
Phone: +49 (0)30 8413-1360  
Fax: +49 (0)30 8413-1394  
Email: krukenkamp@vw.molgen.mpg.de

Secretaries
Tamara Safari (since 09/10, part-time)  
Jeannine Dilßner (since 02/96, part-time)  
Sebastian Klein (06/05-09/10)  
Sara Aziz (03/09-04/10, part-time)  
Phone: +49 (0)30 8413-1399/1364  
Fax: +49 (0)30 8413-1394  
Email: sek-verwalt@molgen.mpg.de

Human resources
Gisela Rimpler (since 07/11)  
Kathleen Linnhoff (since 11/07, part-time)  
Jeanette Brylla (since 04/96)  
Jeannette Bertone (since 03/96, part-time)  
Udo Boettcher (head, 01/15-03/15)  
Heiko Figge-Boettger (head, 07/14-10/14)  

Accounting
Julia Zlotowitz (head; 07/12-05/14)  
Stephanie Cuber (11/11-12/12)  
Ruth Schäfer (head; 05/75-07/12)  
Margrit Pomerenke (07/08-03/12)  
Hilke Wegwerth (01/92-03/11)

External project funding
Joachim Gerlach (since 06/97)  
Anke Badrow (since 02/96)

Guest houses, apartments
Tamara Safari (since 09/10, part-time)  
Marion Radloff (since 09/10)  
Marianne Hartwig (since 10/07)  
Sara Aziz (03/09-09/10, part-time)  
Eleonora Volcik (08/09-02/10)
Overview

The administration of the Max Planck Institute for Molecular Genetics (MPIMG) secures smooth operations and stable infrastructures for the institute. Besides the core administrative tasks, personnel administration, and accounting, the administration takes care of purchasing and of all financial aspects of national and international grants. Scientists receive support in legal questions pertaining to technology transfer and patenting. This, like many other issues, is dealt with in close cooperation with the respective departments of Max Planck Headquarters in Munich.

Subsequently to a retirement-related change in the management of all areas (purchasing, personnel administration, accounting and external funding) as well as in the overall administration itself between 2012 and 2014, the administration plans to set up a new management structure for the coming years. A new head of the personnel administration (currently vacant) is currently being sought.

Due to the retirements of H.-Hilger Ropers and Hans Lehrach in 2014, the budget of the MPIMG has already been gradually reduced since 2012. This counts for the institutional budget from the Max Planck Society as well as for the external funding, especially from the German Ministry for Education and Research (BMBF) and the European Commission. In addition, the number of employees decreased from about 450 people in 2009 to about 250 people in 2015. This downsizing was achieved only by the expiration of temporary contracts in the departments Lehrach and Ropers and, of course, required diverse interim solutions to allow as smooth changes of group leaders and scientists to other research institutions as possible as well as the completion of dissertations for all PhD students of the two departments.
The ongoing refurbishment of the old buildings required a general audit of all existing equipment, as the lab space has been reduced by one complete tower and it was not possible to continue operating all equipment. On the other hand, the resources especially of the sequencing facility and the microscopy group had to be enhanced with e.g. new sequencers (NextSeq, HiSeq 2500), a Fluidigm system for single-cell genomics and a light sheet microscope to meet the ongoing demands of the remaining departments and research groups.

In 2015, the Max Planck Society revised its funding structures for PhD students and postdocs. Starting from July 1st, 2015, all PhD students and postdocs starting at the MPIMG will only be funded by a MPG funding contract. This, of course, does not count for students or postdocs who are funded by third parties like Humboldt Foundation, Volkswagen Foundation or others. In addition to the regular funding by contracts, the MPIMG developed a guest program to award a limited number of fellowships to excellent postdocs from abroad. Candidates have to be nominated by a scientific member or a head of an (independent) research group of the Institute. The selection is based on the scientific excellence of the candidates as well as on the originality of the research project; a grant committee decides upon all grants.

Special attention has been paid on topics of work-life balance, especially in assisting (female) scientists with small children. The MPIMG has contracts with three nearby day-care facilities including a German-English and a German-French nursery to allow the admission of children on short notice. Additionally, an in-house childcare service is available that can be requested for seminars or events taking place at the institute.
Research Support

Technical management & workshops

**Head**
Dipl.-Ing. (FH) Ulf Bornemann  
(since 04/01)  
Phone: +49 (0)30 8413-1424  
Fax: +49 (0)30 8413-1394  
Email: bornemann@molgen.mpg.de

**Building services**
Reinhardt Strüver (head) (since 07/02)  
Reinhard Kluge (since 12/10)  
Bernd Roehl (since 08/97)  
Frank Kalaß (since 01/88)  
Thomas Oster (08/05-03/11)

**Electromechanics**
Karsten Beyer (since 07/07)  
Florian Zill (since 08/95)  
Carsten Arol (since 05/92)

**Glass instruments construction**
Peter Ostendorf (since 01/02, part-time)

**Technical supply service**
Dirk Grönboldt-Santana (since 01/98)
Overview

The technical management of the MPIMG is responsible for the operation and maintenance of the whole institute, including the buildings, the building services like electric energy supply, cooling and heating systems, steam generator, water conditioning and so on, and the outdoor facilities. To secure operational reliability of all relevant systems beyond working hours, the workshop staff maintains an emergency service for 24 hours a day including weekends and holidays.

In addition to the building services, the technical management also assists the scientific departments and groups at lab moves, lab reorganizations or with the installation of new technical equipment.

The main challenge during the last years has been (and still is) the new construction of tower 3 and the complete structural and electrical renovation of the nearly 50 years old buildings tower 1 and 2 to meet the requirements of today’s fire safety, occupational safety, and energy efficiency. Right from the beginning, the head of the technical management has been involved in all planning phases, as his team will have to operate the buildings after of the refurbishment. In 2013, tower 3 has been finished and handed over to the institute. Amongst many others, this included a new sprinkler and gas extinguishing system, which were installed for the first time at the MPIMG, so that the staff had to be trained to operate it.

After tower 3 came in operation, tower 2 had to be cleared out completely, which meant that all departments and groups had to be distributed to the remaining towers 1, 3, and 4. In October 2013, the structural and technical refurbishment of tower 2 commenced, starting with the insulation of the new ventilation system for the seminar rooms in tower 3 after the construction of the new building.
the building shell, i.e., the façade, including the replacement of windows and doors. All floor plans have been adapted to new laboratory utilization concepts. By the end of the project, the complete technical media for electrical energy, heating, cooling, water, ultra-pure water, steam, gas, and compressed air supplies, as well as for ventilation and air conditioning, fire detection technology, communication and building automation will be newly installed.

The entire construction work is carried out during full research operations in the adjoining buildings, which leads to considerable restrictions for the research groups, e.g., repeated interruptions in the media supply, as temporary equipment has to be connected, or old supply systems have to be switched over to the new systems.

Tower 2 is expected to be completed and handed over to the MPIMG in spring 2016. Subsequently, tower 1 will be completely refurbished, too; its completion is scheduled for 2018. Until now, the staff of the institute’s technical management has mastered the challenges arising from the new technical installations, particularly the central cooling unit, the fire detection technology and the extensive building automation quite well, despite all difficulties during the construction and initial start-up phases.