5th Bioinformatics and Stem Cells Satellite Workshop
Monday, April 22, 2013
University Hospital Cologne, Lecture hall 2 (first floor), Joseph-Stelzmann-Str. 9, 50931 Köln (Cologne)

http://www.ibima.med.uni-rostock.de/stemcellbioinf_2013/

9:20 Welcome
9:30-10:00 Carsten Marr: Models of blood stem cell differentiation based on single-cell time-lapse microscopy
10:00-10:30 Antony Gaspar: Genes controlling Stemness and lineage specification of Pluripotent Stem cells- learnt from Transcriptomics
10:30-11:00 Ivan G. Costa, Eduardo Gade Gusmão: Unravelling Differentiation Processes with Statistical Analysis of Genome Wide Data

Coffee
11:15-11:40 Marcos J. Araúzo-Bravo: Hacking the Human Genome to Discover Combinatorial Regulatory Motifs Reveals a Transcriptional Recursive Arrangement
11:40-12:05 Loi Luu Phuc: DNA methylation somatic memory switches distal regulation in iPS

Lunch break
Lightning talks (based on late submissions)
13:20-13:35 Vlad Cojocaru: DNA recognition by the transcription factors defining stem cell pluripotency: insights from computer simulation
13:35-13:50 Mohamed Hamed: Co-expressed gene set involving imprinted genes plays key role in blood stem cell differentiation
13:50-14:05 Andreas Heider: Merging raw data from different microarray platforms using the R/Bioconductor package virtualArray
14:05-14:20 Davood Sabour: Identification of genes specific to mouse primordial germ cells through dynamic global gene expression
14:20-14:25 Discussion
14:25-14:50 Leila Taher: Modeling approaches to understand transcriptional changes of stemness genes

Coffee
15:05-15:50 Bruce Conklin: Patient-specific Stem Cells and Cardiovascular Genetics
15:50-16:35 Marius Wernig: Direct Reprogramming towards the Neural Lineage
16:35-17:00 Mathias Ernst: Gene expression data of pluripotent stem cells – A comparative analysis using a network-guided approach

19:00-22:00 Dinner (planned, self-pay), venue tba.

Supported by the DFG Priority Program SPP 1356 "Pluripotency and Cellular Reprogramming", by the Institut für Neurophysiologie Cologne, by the Stem Cell Network NRW and by the IBIMA Rostock.

Abstracts

Hacking the Human Genome to Discover Combinatorial Regulatory Motifs Reveals a Transcriptional Recursive Arrangement

Marcos J. Araúzo-Bravo

Computational Biology and Bioinformatics Group, Max Planck Institute for Molecular Biomedicine, Roentgenstraße 20, 48149 Muenster, NRW, Germany

To know the map between transcription factors (TFs) and their binding sites it is essential to reverse engineer the regulation process. Only about 10%-20% of the transcription factor binding motifs (TFBMs) have been reported. This lack of data hinders understanding gene regulation. To address this drawback, we propose a computational method that exploits never used TF properties to discover the missing TFBMs and their sites in all human gene promoters. The method starts by predicting a dictionary of regulatory "DNA words". From this dictionary, it distills 4098 novel predictions. To understand the crosstalk between motifs, an additional algorithm extracts TF factor combinatorial binding patterns creating a collection of TF regulatory syntactic rules. Using these rules, we narrowed down a list of 504 novel motifs that appear frequently in syntax patterns. We tested the predictions against 509 known motifs confirming that our system can reliably predict de novo motifs with an accuracy of 81% - far higher than previous approaches. We found that on average, 90% of the discovered combinatorial binding patterns target at least 10 genes, suggesting that to control in an independent manner smaller gene set, supplementary regulatory mechanisms are required. Additionally, we discovered that the new TFBMs and their combinatorial patterns convey biological meaning, targeting TFs and genes related to developmental functions. This reveals a recursive arrangement in the transcription regulation process where among all the possible available targets in the genome, the TFs tend to regulate other TFs and genes involved in developmental functions. We provide a comprehensive resource for regulation analysis that includes a dictionary of DNA words, newly predicted motifs and their corresponding combinatorial patterns. Combinatorial patterns are a useful filter to discover TFBMs that play a major role in orchestrating other factors and thus, are likely to lock/unlock cellular functional clusters.

Patient-specific Stem Cells and Cardiovascular Genetics

Bruce Conklin

Gladstone Institute of Cardiovascular Disease Gladstone Scientific Officer, Research Technology & Innovation UCSF Division of Genomic Medicine, 1650 Owens Street, Mission Bay,San Francisco CA 94158

We use induced pluripotent stem (iPS) cells to model human disease. Our major focus is on genes that cause "sudden death" due to abnormal heart rhythm, and heart failure from cardiomyopathy. We use iPS cells from patients and also engineering iPS cells to have specific mutations. Engineering iPS cells allows direct comparison of cells with the exact same (isogenic) background. Recent genetic studies provide gene variant associations that are largely untested.

Comparing iPS cells with discrete mutations in an isogenic background provides an experimental system to directly test these genetic associations. Personalized medicine can benefit from experimental testing of gene variants to prove (or disprove) hypothetical genetic associations.

Gene expression data of pluripotent stem cells – Comparative analysis using a network-guided approach

Mathias Ernst, Leila Taher & Georg Fuellen

Institute for Biostatistics and Informatics in Medicine and Ageing Research, Medical Faculty, University of Rostock, Ernst-Heydemann-Str. 8, 18055 Rostock, Germany

Pluripotency refers to a cell being able to give rise to all three germ layers. It is now well understood that pluripotent cells can be subdivided into two distinct states, namely the naïve or ground state and the primed state, which has advanced further on the path to differentiation. From mice, cells resembling both states can be retrieved ex-vivo and via reprogramming of somatic cells; furthermore, both states are readily interchangeable under appropriate conditions. In human, however, only stem cells that resemble the murine primed state are generically accessible. Transforming such human primed stem cells into cells that resemble the naïve state of pluripotency has been attempted and indeed claimed, yet such studies have met with criticism. We have recently developed ExprEssence, a tool for the analysis of high-throughput data within the context of gene/large protein-protein interaction (PPI)and regulation networks. Here we employ this tool to reanalyse two published microarray transcriptomics data sets that deal with primed-to-naive reprogramming, one in mouse and the other one in human. Based on a large all-purpose PPI network derived from STRING 9.0 we identify the network interactions that experience the most prominent changes, i.e. startups or shutdowns, in performing a combinatorial, pairwise cross-state and cross-species comparison of the transcriptomics data. We perform a comprehensive functional analysis of the gene sets that underlie the selected interactions by identifying over- and underrepresented, respectively, terms from the biological process division of the GeneOntology (GO). Aggregating the obtained evidence for each gene set into larger, biologically meaningful metagroups and subjecting those to principal components analysis (PCA) allows us to obtain insights as to the similarities and differences between the distinct states of pluripotency in human and mice.

Genes controlling Stemness and lineage specification of Pluripotent Stem cells- learnt from Transcriptomics

John Antonydas Gaspar[1], Michael Xavier Doss[2], Herbert Schulz[3], Shiva Potta[1], Johannes Winkler[4], Jürgen Hescheler[1], Agapios Sachinidis[1]

[1]Center of Physiology and Pathophysiology, Institute of Neurophysiology, and Center of Molecular Medicine, University of Cologne (CMMC), Robert-Koch Str. 39, 50931 Cologne, Germany

[2]Stem Cell Center, Masonic Medical Research Laboratory, 2150 Bleecker Street, Utica, NY 13501 USA

[3]Max-Delbrueck-Center for Molecular Medicine - MDC, Robert-Rössle Str. 10, 13092 Berlin, Germany

[4] Institute for Genetics, University of Cologne, Zülpicher Str. 47a, 50674 Cologne, Germany

Investigating the molecular mechanisms controlling the post embryogenesis developmental program *in vivo* is challenging and time-consuming. Similar to the totipotent cells of the inner cell mass, gene expression and morphological changes in cultured PSCs (Pluripotent Stem cells) occur hierarchically during their differentiation, with epiblast cells developing first, followed by germ layers and finally somatic cells. An attempt is made to understand differentiation networks controlling embryogenesis *in vivo* using a time kinetic in a combination of high throughput -omics

technologies with murine embryonic stem cells, by identifying molecules defining fundamental biological processes in the pluripotent state as well as in early and later differentiation stages.
To gain a better understanding of the complex biological processes and functions involved in stemness as well as in differentiation processes, the candidate molecules significantly expressed in the stemness state and gene clusters participating in early development during multilineage specification are identified. In a specific manner, the transcriptomic signatures, focusing on the central genes and signaling pathways of mesodermal cells and cardiovascular cell derivatives are defined. Furthermore, we have identified decisive genes and functional annotations characterizing inducible pluripotent stem cell (iPSC)-derived $Acta2^{+}$ cardiomyocytes that could provide a point of reference for following an early development of heart tissues as well as the molecular mechanisms which could be manipulated to advance the embryonic phenotype of iPSC-derived cardiomyocytes to an adult cardiac phenotype.

Unravelling Differentiation Processes with Statistical Analysis of Genome Wide Data

Eduardo G. Gusmão[1,2], Ivan G. Costa[1]

[1]Interdisciplinary Centre for Clinical Research (IZKF) & Institute for Biomedical Engineering, RWTH University Medical School, Aachen, Germany

[2]Center of Informatics (CIn), Federal University of Pernambuco (UFPE), Recife, Brazil

Cell differentiation is controlled by a circuit of transcription factors and chromatin modifications that determine cell fate by activating cell type specific expression programs. The identification of cis-acting elements that dictate expression of genes is crucial for the understanding of such regulatory networks. Advances in sequencing technologies allow the measurement of cis-acting elements of parts of such regulatory circuitries on a genome-wide scale. However, the integrative analysis of such large, unstructured and noise prone data poses great challenges in computational biology methods. It will be described in this talk two computational approaches related to these problems; (1) the use of multivariate hidden Markov models for the integrative analysis of histone modification and DNA sequence of transcription factor binding sites and (2) the use of mixture of sparse regression models for the detection of transcription factor and histones regulatory networks in blood cell development.

Co-expressed gene set involving imprinted genes plays key role in blood stem cell differentiation

Mohamed Hamed

Center for Bioinformatics, 66041 Saarbrücken, Germany

Maintenance of pluripotency, cell differentiation and reprogramming in mouse are regulated by a complex gene-regulatory network termed PluriNetwork. We found that during hematopoiesis, 272 genes that were earlier found to be involved in this network show similar expression changes as 86 well-known imprinted genes suggesting a functional connection of both gene sets. When enriching the imprinting network by genes with known regulatory effects on imprinted genes, we found that 20 genes are shared between the PluriNetwork and the imprinting network. Among these genes are the Yamanaka factors Oct4 and c-Myc. Subsets of the imprinted gene set are found to be markedly co-expressed with the subsets of plurigenes along six separate hematopoietic cell lines. Both gene

sets show pronounced functional similarities. This hints at an important regulatory role of imprinted genes during cell differentiation.

Merging raw data from different microarray platforms using the R/Bioconductor package virtualArray

Andreas Heider[1], Rüdiger Alt[2]

1, Translational Centre for Regenerative Medicine (TRM), Phillip-Rosenthal-Strasse 55, 04103 Leipzig

2, Vita 34 AG, Perlickstrasse 5, 04103 Leipzig

One of the original goals of public microarray databases such as NCBI GEO or EBI ArrayExpress was to make data collected by one research group available to other groups to analyze and possibly integrate with their own data. However, a growing number of different types and generations of microarray chips exists. This diversity prohibits the integration of raw data. Current software implementations for the analysis of multiple data sets mainly deal with meta-analysis. While the merging of data generated on the same chip platform is possible even here there are batch effects to be taken care of. A method to deal with multiple chip types and batch effects at the same has been missing.

The software "virtualArray" presented here was designed to solve both of these problems in a straightforward and user friendly way and was written in the R programming language extending the Bioconductor bioinformatics packages. The package was designed to be able to combine raw data from almost any chip type based on current annotations from NCBI GEO or Bioconductor [1].

The process performed by virtualArray involves the establishment of a consistent data set incorporating raw data of all input data sets. virtualArray can then directly employ one of six implemented methods to adjust for batch effects resulting from different chip types used. Both steps can be tuned by the user. The generated data set is presented as a conventional Bioconductor "ExpressionSet" object allowing further downstream analysis using other Bioconductor packages.

The software was designed in a modular fashion permitting an easy extension with other functions. Using this software package, researchers can easily integrate their own microarray data with data from public repositories or other sources that are based on different microarray chip types. The default approach applies a robust and up-to-date batch effect correction technique to the data.

Heider, A., 2011. virtualArray [WWW Document]. URL http://www.bioconductor.org/packages/2.10/bioc /html/virtualArray.html

DNA methylation somatic memory switches distal regulation in iPS

Loi Phuc Luu, Marcos J. Araúzo-Bravo

Computational Biology and Bioinformatics Laboratory, Max Planck Institute for Molecular Biomedicine, Roentgenstraße 20, 48149 Muenster, NRW, Germany

Many studies have focused on the search for the remaining fingerprint of somatic memory of the iPS cells reprogrammed by transcription factors. Such studies have been performed based on high throughput data analysis of transcriptomics or methylomics data, using microarrays or next generation sequencing (NGS) technologies. The conclusions of these studies are controversial. The transcriptomics based studies from somatic, iPS and ES cells usually identify very few somatic memory and concerns exist of whether the detected remaining fingerprint of somatic origin at transcriptomics level is due to poor quality of iPS cells at low cell passages. Several works point out

that an enough number of cells passages erase the transcriptomics fingerprint of the original somatic cells. However, on methylomics level, there are studies disclosing differences between iPS and ES cells, thus, indicating that some fingerprints of such memory remain. Such studies do not take into consideration the methylomics state of the original somatic population. We have developed algorithms for searching differentially methylated sites from NGS data, and applied them not only for comparing methylomics data from iPS and ES cells as previous studies, but we also integrated in the analysis the methylomics data from the original somatic cell lines from which reprogrammed cells were derived. Using human iPS cells derived from several fibroblast lines, we have found more than 30000 sites identified as somatic memory loci consistent across the DNA methylomes profiles of all the iPS cell lines. Surprisingly, more than 70% of these differently methylated sites are in the distal regulation regions and the leftovers are in promoters.

Models of blood stem cell differentiation based on single-cell time-lapse microscopy

Carsten Marr

Helmholtz Zentrum München Deutsches Forschungszentrum für Gesundheit und Umwelt (GmbH), 85764 Neuherberg, Germany

Hematopoiesis is often pictured as a hierarchy of branching decisions, giving rise to all mature blood cell types from the stepwise differentiation of a single cell, the hematopietic stem cell. Different aspects of this process have been modeled with different experimental and theoretical techniques, and on a spectrum of scales.

Here, we integrate three different scales to study the possibility of inference of mechanistic knowledge from the differentiation process. We focus on a submodule of hematopoiesis, the differentiation of granulocyte-monocyte progenitors (GMPs) to granulocytes or monocytes. Within a branching process model, we infer the differentiation probability of GMPs from experimentally quantified heterogeneity of colonies in assays under permissive conditions, where both granulocytes and monocytes can emerge. We compare the predictions with the differentiation probability in genealogies determined from single-cell time-lapse microscopy. Contrary to the branching process model, we find the differentiation probability to rise with the generation of the cell within the genealogy. We study this effect with a differentiation model that takes the tree structure, and a possible time delay between decision and marker onset into account. To study the differentiation probability from a molecular perspective, we set up a stochastic toggle switch model, where we execute the intrinsic lineage decision with two antagonistic transcription factors. We find parameter regimes that allow for both time dependent and time independent differentiation probabilities. Finally, we infer parameters for which the model matches experimentally observed differentiation probabilities via Approximate Bayesian Computing. These parameters suggest a separation of timescales in the dynamics of granulocyte and monocyte differentiation.

In silico molecular analysis reveals unique properties of the transcription factors defining the stem cell pluripotency

Felipe Merino, Vlad Cojocaru

Department of Cell and Developmental Biology, Max Planck Instiute for Molecular Biomedicine

Röntgenstr. 20, 48149 Münster

Oct4 and Sox2 are in the core of the transcription regulation network that is essential for the

maintenance and induction of pluripotency. Oct4 cannot be replaced by any other member of its family in the process of reprogramming to pluripotency, whereas only a few other Sox factors are able to substitute for Sox2 in this process. Oct4 is a member of the POU family of transcription factors that have a bipartite DNA-binding domain (POU domain) composed of two subdomains: the POU specific domain (POUS) and the POU homeodomain (POUHD), both binding in the major groove of the DNA. Sox2 has one DNA-binding domain (HMG domain) that binds the minor groove and induces a sharp kink in the DNA. It has been shown that Oct4 cooperates with Sox2 to bind regulatory regions that are responsible for the regulation of pluripotency-defining genes. The DNA elements bound by Oct4 and Sox2 differ in the number of the base pairs that separate the binding sites of the POUs and HMG domains. Furthermore, one DNA element has been identified on which Oct4 cooperates with Sox17 and not Sox2. A mechanism in which Oct4 changes partners between Sox2 and Sox17 to interpret a different enhancer code during early differentiation into primitive endoderm has been proposed.

What are the unique properties of Oct4 and Sox2 that confer them the functional specificity towards regulating stem cell pluripotency? We use computational structural biology methods to reveal these properties. First, we will present a comparative analysis of the electrostatic potential of Oct4 orthologues that explains the ability or inability of these orthologues to replace the mouse Oct4 in the induction of pluripotency in mouse fibroblasts and predicts potential protein-protein interaction interfaces on the surface of Oct4. Then, we will show how we are using molecular dynamics simulations to reveal the potential differences in the DNA recognition mechanisms between Oct4 and Oct1 (another member of the POU family). Finally, we will show how molecular modeling and simulations successfully predict the amino acids that are responsible for the different functions of the Oct4/Sox2/DNA and Oct4/Sox17/DNA complexes. These predictions lead to the design of a functional switch between Sox2 and Sox17 that was validated experimentally.

Identification of genes specific to mouse primordial germ cells through dynamic global gene expression

Davood Sabour, PhD

Department of Cell & Developmental Biology; Max Planck Institute for Molecular Biomedicine, Röntgen Str.20, D-48149, Münster-Germany

*Current Address: Institute of Neurophysiology, Medicine Faculty, University Hospital Köln, Robert-Koch Str.39, D-50931, Köln-Germany

Molecular mechanisms underlying the commitment of cells to the germ cell lineage during mammalian embryogenesis remain poorly understood due to the limited availability of cellular materials to conduct in vitro analyses. Although primordial germ cells (PGCs)--precursors to germ cells--have been generated from embryonic stem cells (ESCs)-pluripotent stem cells derived from the inner cell mass of the blastocyst of the early embryo in vitro-the simultaneous expression of cell surface receptors and transcription factors complicates the detection of PGCs. To date, only a few genes that mark the onset of germ cell commitment in the epiblast-the outer layer of cells of the embryo--including tissue non-specific alkaline phosphatase (TNAP), Blimp1, Stella and Fragilis--have been used with some success to detect PGC formation in in vitro model systems. Here, we identified 11 genes (three of which are novel) that are specifically expressed in male and

female fetal germ cells, both *in vivo* and *in vitro*, but are not expressed in ESCs. Expression of these genes allows us to distinguish committed germ cells from undifferentiated pluripotent cell populations, a prerequisite for the successful derivation of germ cells and gametes in vitro.

Modeling approaches to understand transcriptional changes of stemness genes

Leila Taher

Institute for Biostatistics and Informatics in Medicine and Ageing Research, Medical Faculty,
University of Rostock, Ernst-Heydemann-Str. 8, 18055 Rostock, Germany

Understanding the process of stem cells differentiation is not only of mechanistic interest, but also holds potential for regenerative medicine. We present a computational method for analyzing regulatory interactions between the genes involved in embryonic stem cell (ESC) differentiation based on single-cell time resolved gene expression profiles. In particular, we address the following questions: a) how the network-based analysis relates to the traditional gene-based differential expression analysis; and b) how network connectivity relates to gene co-expression. Our approach is widely applicable to other biological systems.

Direct lineage reprogramming towards the neural lineage

Marius Wernig

Institute for Stem Cell Biology and Regenerative Medicine, 265 Campus Drive, Rm. G3141, Stanford, CA 94305

Cellular differentiation and lineage commitment are considered robust and irreversible processes during development. Challenging this view, we found that expression of only three neural lineage-specific transcription factors could directly convert fibroblasts into functional neurons in vitro. These induced neuronal (iN) cells expressed multiple neuron-specific proteins, generated action potentials, and formed functional synapses. This demonstrated that even very distantly related cell types can be directly induced from a given cell type. We next wondered whether a defined non-ectodermal cell can be converted into iN cells given the heterogeneity of fibroblast cultures. We therefore tested whether (endoderm-derived) hepatocytes can be reprogrammed to iN cells. Surprisingly, using the same 3 transcription factors primary mouse hepatocytes could be converted very efficiently into fully functional iN cells. Moreover, gene expression studies on the global and single cell level confirmed not only the induction of a neuronal transcriptional program but also the efficient silencing of the hepatocyte-specific expression pattern. We therefore conclude that iN cells are not a hybrid cell with equal identities of the starting cell and neurons, but cells with a predominant neuronal identity with an epigenetic or transcriptional memory of the starting cell population.

Finally, we recently extended the direct lineage reprogramming approach and successfully generated induced tripotent and self-renewing neural precursor cells (iNPCs) and induced oligodendrocyte progenitor cells (iOPCs) from rodent fibroblasts. We believe iOPCs are particularly interesting as these cells could be used for transplantation-based strategies to treat de- and dysmyelinating diseases such as Pelizaeus-Merzbacher's disease and other leukodystrophies.