Abstracts of the 3rd Satellite Workshop on Bioinformatics in Stem Cell Research July 15 2010, Dresden

A data integration approach to mapping OCT4 gene regulatory networks required for sustaining self-renewal and pluripotency in embryonic stem cells.

James Adjaye
Max-Planck Institute for Molecular Genetics, Department of Vertebrate Genomics (Molecular Embryology and Aging) Berlin

Background: Deciphering the transcriptional networks operative in human embryonic stem cells (hES), induced pluripotent stem cells (hiPS) and embryonal cancinoma cells (hEC), is essential for enhancing our understanding of self-renewal and pluripotency. The transcription factor OCT4, is a master regulator of the transcriptional networks required for inducing and maintaining pluripotency. Therefore, employing a systems biology approach whereby correlating gene expression resulting from the ablation of OCT4 function in hES and hEC cells with potential OCT4-binding sites within the promoters of target genes allows a higher predictability of motif-specific driven expression modules important for inducing pluripotency and maintaining self-renewal.
Methology/Principal Findings: We have conducted ChIP-on-Chip and ChIP-Seq experiments using OCT4 antibodies to obtain a defined dataset related to OCT4-bound regions up to 50 kb upstream of the transcription start sites of putative target genes. To achieve this, we compared several peak finding analysis programs to arrive at a refined list of OCT4 targets in hEC cells and compared this data to hES specific OCT4-binding and expression. We identified and verified a highly enriched POU/OCT4 - motif by employing a de novo approach, this enabled us to unviel six distinct OCT4-binding modules which are evolutionary conserved. Of these are for instance, the classic OCT4-SOX2 motif present within the NANOG proximal promoter. Other target genes such as USP44 and GADD45G have the POU-motif but not the classical HMG/SOX2 motif within their proximal promoter region. Additionally, we observed preferred distances for the HMG and the POU motif, thus further evidence for additional binding modules other than the classical HMG-POU consensus sequence.
In undifferentiated hEC and hES cells, USP44 and GADD45G are positively and negatively regulated by OCT4 respectively. Furthermore, over-expression of GADD45G in hEC cells resulted in an enrichment of up-regulated genes related to differentiation pathways. Due to the large nature of available datasets pertinent to embryonic stem cell biology, we have integrated our and already published datasets and developed an interactive embryonic stem cell database.
Conclusion/Significance: Employing a systems biology approach, we have uncovered new OCT4-binding modules and regulated targets, and highlighted their importance in the hEC/hES self-renewal circuitry. In this era of high-throughput functional genomics, which results in large datasets, our database allows rapid and convenient assess and comparisons between published datasets related to embryonic stem cell biology.

Study of stem cell reprogramming using profiles of gene expression

Miguel Andrade, Nancy Mah
Computational Biology and Data Mining group, Berlin

Somatic cells can be reverted to an embryonic-like state simply by over-expressing a combination of four transcription factors (OCT4/POU5F1, SOX2, KLF4 and MYC or OCT4/POU5F1, SOX2, NANOG and LIN28). These so-called induced pluripotent stem (iPS) cells have enormous potential for applications in regenerative medicine, for example, cell replacement therapy in neurodegenerative diseases. The use of iPS cells, instead of embryonic stem (ES) cells, is advantageous because generation of iPS cells does not require destruction of the embryo, and transplant rejection can be avoided by using the patient's own cells (e.g. fibroblasts) to generate iPS cells for therapeutic purposes. However, the reprogramming protocol is currently an inefficient and lengthy process. Additionally, the induction of oncogenes (MYC, KLF4) should be avoided, since they may cause tumours. Before iPS cells can be safely used in patients, we require a better understanding of the processes involved in reprogramming. Using publicly available genome-wide microarray expression datasets from reprogramming experiments, we identify sets of genes that define signatures characteristic of donor cell type (human fibroblasts), self-renewal (iPS, ES) and partially

reprogrammed cells (partially induced pluripotent cells (PiPSC)). We find that pluripotency is favoured by the inhibition of epithelial mesenchymal transition (EMT), leading to mesenchymal epithelial transition (MET).

A comprehensive model of the spatio-temporal stem cell and tissue organisation in the intestinal crypt

Jörg Galle 1, Peter Buske 1, Nicholas Barker 2 , Hans Clevers 2, Markus Loeffler 3
1 Interdisciplinary Centre of Bioinformatics, University Leipzig; 2 Hubrecht Institute, Utrecht
3 Institute for Medical Informatics, Statistics and Epidemiology, University Leipzig

We introduce a novel three-dimensional model of stem cell and tissue organisation in murine intestinal crypts. Integrating the molecular, cellular and tissue level of description this model links a broad spectrum of experimental observations encompassing spatially confined cell proliferation, directed cell migration, multiple cell lineage decisions and clonal conversion.
Using computational simulations we demonstrate that the model is capable of quantitatively describing and predicting the dynamic behaviour of the tissue during steady state as well as after cell damage and following selective gain or loss of gene function manipulations affecting Wnt- and Notch-signalling. Our simulation results suggest that reversibility and flexibility of cellular decisions are key elements of robust tissue organisation of the intestine. We predict that the tissue should be able to fully recover after complete elimination of isolated cellular subpopulations including that of actual functional stem cells. This challenges current views of tissue stem cell organisation.

Molecular decision making in embryonic and adult stem cells

Ingmar Glauche, Maria Herberg, Tilo Buschmann, Ingo Roeder
Institute for Medical Informatics and Biometry, Dresden

Molecular interactions between transcription factors are considered to be major control mechanisms for stem cell fate decisions. By translating these interactions into an appropriate mathematical state space formulation it is possible to investigate the dynamics of cellular development on the molecular level and establish a conceptual understanding of stem cell fate decisions. In particular, we have established different mathematical models for the description of molecular switches in embryonic and haematopoietic stem cells.
It has been recently demonstrated that individual embryonic stem (ES) cells reversibly change expression level of the crucial transcription factor Nanog and that cells with a low Nanog level are more likely to undergo differentiation. We show that alternations between high and low protein concentrations can be explained by different assumptions yielding similar experimental characteristics. Based on the model results we argue that Nanog variability is a potential “gate-keeper” mechanism to the control of ES cell differentiation.
For the haematopoietic system we discuss a particular molecular switch in myeloid progenitor cells involving the antagonistic transcription factors PU.1 and Gata-1. Using an ODE approach we study the influence of different sources of noise on the transition dynamics changing the system from the co-expression state towards the dominance of one factor in each individual cell. Based on these results we analyzes the emerging heterogeneity on the population level with respect to the overall dynamics of the molecular switch.

Elucidating the regulatory program of adult neurogenesis

Jacob J. Michaelson, Andreas Beyer
Biotechnology Center, TU Dresden

Even though most parts of the adult mammalian brain are precluded from neurogenesis, the olfactory bulb and the hippocampus are capable of developing new neurons after birth. Whereas external factors affecting neurogensis are known (such as physical activity or stimulating environments), molecular and genetic causes of neurogeneis are much less understood, largely due to the complex interaction of these factors.

In this work, we used a variety of data sources and novel computational techniques to investigate two questions: first, what is the logic ordering of genes implicated in neurogenesis; second, how do these genes interact when regulating expression of relevant genes. As a first step in this analysis, we employed text-mining to assemble a set of high-confidence neurogenesis-associated genes. To further explore the regulatory context of these genes, we exploited systems genetics data from a recombinant inbred panel of mice (BXD) that vary widely in their neurological phenotypes including adult neurogenesis. Specifically, we examined three types of regulatory relationships to give context to these genes: upstream regulators, downstream targets, and lateral (epistatic) co-regulators. Examining these complex relationships was made possible by a novel extension of the Random Forests machine learning method that utilizes latent information in the forest structure. The insight gained through this work demonstrates the value of a systems approach when deconstructing the genetic factors controlling adult neurogenesis, and the presented framework also has wider applicability to complex traits in general.

Efficiently Finding Homologous Pluripotent Proteins by Using Similarity Search

Arnoldo J. Muller-Molina and Marcos Araúzo-Bravo
Max Planck Institute for Molecular Biomedicine, Münster

Given the large amounts of biological data available, it is necessary to employ similarity search indexes because the brute force approach does not scale up. In this talk we introduce general similarity search index techniques. We also give an overview of popular data structures and pointers to different open source implementations. As an example of an application we will have a simple use case where we find pluripotent proteins while using a very expensive distance function.
Given the large amounts of biological data available, it is necessary to employ similarity search indexes because the brute force approach does not scale up. In this talk we introduce general similarity search index techniques. We also give an overview of popular data structures and pointers to different open source implementations. As an example of an application we will have a small tutorial that explains how to use an open source similarity search engine to find pluripotent proteins while using a very expensive distance function.

Continuous single cell data as the basis for stem cell systems biology

Timm Schroeder
Institute of Stem Cell Research, Helmholtz Zentrum Muenchen - German Research Center for Environmental Health (GmbH) Munich / Neuherberg

Many long-standing questions in stem cell research remain unsolved. Stem cell driven regenerative systems are highly complex and dynamic, consisting of large numbers of different cells expressing many molecules controlling their fates. Therefore, mathematical models are necessary - both to aid the interpretation of experimental data, and to simulate the behavior of stem cell systems based on hypothetical assumptions about their complex cellular or molecular composition. However, the generation of models is hampered by the lack of precise experimental data. Even moderate numbers of unknown variables quickly lead to uncertainties within the models which can render them largely useless for solving biological questions. In particular, it is a major problem that stem cell systems are usually followed by analyzing populations of cells - rather than individual cells - at very few time points of an experiment, and without knowing individual cell identities. Continuous real-time tracking of individual cells would be an important prerequisite to fully understand the developmental complexity of stem cell driven systems. We have therefore developed culture and imaging systems to follow the fate of individual cells over long periods of time. Our approaches also allow the continuous long term quantification of protein expression levels in living stem cells. This novel kind of quantitative data of single cell behavior and molecule expression is used as the basis for the improved generation and falsification of models describing stem cell systems. I will discuss how the role of dynamic extracellular signaling and cell intrinsic transcription factor networks in controlling stem cell fates can be addressed.

THE EUROPEAN HUMAN EMBRYONIC STEM CELL REGISTRY, HESCREG - A CELL SEARCH AND CHARACTERIZATION ENGINE FOR PLURIPOTENT CELLS AND SYSTEMS BIOLOGY

Stachelscheid, Harald1, Damaschun, Alexander1, Aran, Begoña2, Elstner, Anja1, Veiga, Anna2, Stacey, Glyn3, Borstlap, Joeri1, Kurtz, Andreas1

1Berlin-Brandenburg Center for Regenerative Therapies, Charité - University Medicine Berlin, Berlin 2Barcelona Stem Cell Bank, Centre of Regenerative Medicine in Barcelona, 3UK Stem Cell Bank, UK Health Protection Agency, NIBSC, South Mimms

The European human embryonic stem cell registry - hESCreg - was set up in 2007 as a global registry for human embryonic stem cells (hESC). Its goal is to provide comprehensive information on available hESC lines including their derivation, culture, genetics, potency and procurement/ethical provenance. There are 74 hESC providers from 23 countries worldwide who cooperate with the hESCreg platform to provide and validate the data of 618 hES cell lines registered in hESCreg (January 2010). About one third of all registered hESC lines, i.e. 226 lines, are available for further research. 72 hESC lines in the registry carry a genetic modification. The vast majority of these possess inherent genetic defects that cause common diseases such as Cystic Fibrosis or Hemophilia. These cell lines represent prime candidates for disease-specific lines of research. Lines with induced modifications carry reporter genes such as the fluorescence marker GFP. Whilst European providers make about three-quarters of their lines available, non-European providers grant access to about one-eighth of their stocks. Completeness of the provided information for each cell line varies and is filterable for each category and symbolized by an indicator bar system. hESC lines are linked to publications in which their application in research is described, to EU research projects with hESC as well as to a cell characterization tool which enables comprehensive linking of hESCreg data with information from other databases.

The cell characterization tool CellFinder functions are aimed at developing hESCreg into a stem cell navigation tool facilitating the linkage of individual cell lines or groups of cells to genetic or functional characteristics from sources outside of hESCreg, e.g. expression profiles for differentiated or pluripotent cells. This tool allows comparison and analysis of cells within hESCreg on multiple layers, but also expansion of data in the registry which can ultimately be used to design research projects of the user. To further develop its utility, hESCreg has started to register human induced pluripotent stem cell lines (hiPSC). Currently 20 hiPSC from 3 providers in 3 countries are listed. The integration of the registry into a systems biology platform for regenerative medicine will be outlined.

Data analysis and modeling of Pancreatic cell reprogramming

Joseph Xu Zhou

Systems Biology Group,Centre for Information Services and High Performance Computing (ZIH,) Technische Universität Dresden

Cell fate reprogramming, such as the generation of insulin-producing beta cells from other pancreas cells can be achieved by external modulation of key transcription factors. However, the known gene regulatory interactions which form a complex network with multiple feedback loops makes it increasingly difficult to view regulatory pathways as schemes of causal influences that serve as basis for predicting the effect of transcriptional perturbation. Furthermore, not sufficient information on regulatory networks is available for detailed data analysis. Here we demonstrate that by using the qualitatively described gene interaction as basis for a coarse-grained dynamical ODE(ordinary differential equation) model it is possible to recapitulate the differentiation process of the exocrine and beta, alpha, delta endocrine cells types and to predict which gene perturbation can result in desired lineage reprogramming. Our model indicates that the constraints imposed by the incompletely elucidated regulatory gene network architecture suffice to erect a predictive model for making informed decisions in choosing the set of transcription factors that need to be modulated for fate reprogramming.