Realia et naturalia

DISSERTATIONES
BIOLOGICAE
UNIVERSITATIS
TARTUENSIS
276

HEDI PETERSON

Exploiting high-throughput data for establishing relationships between genes

TARTU ÜLIKOOL
1632
UNIVERSITAS TARTUENSIS

DISSERTATIONES BIOLOGICAE UNIVERSITATIS TARTUENSIS
276

DISSERTATIONES BIOLOGICAE UNIVERSITATIS TARTUENSIS
276

HEDI PETERSON

Exploiting high-throughput data for establishing relationships between genes

UNIVERSITY OF TARTU
PRESS
1632

Institute of Molecular and Cellular Biology, University of Tartu, Estonia

Dissertation was accepted for the commencement of the degree of Doctor of Philosophy in bioinformatics at the University of Tartu on 28th of April by the Council of the Institute of Molecular and Cellular Biology, University of Tartu.

Supervisors:	Prof. Jaak Vilo, PhD University of Tartu Tartu, Estonia
	Prof. Juhan Sedman, PhD University of Tartu Tartu, Estonia
Opponent:	Prof. Boris Lenhard, PhD Imperial College London London, United Kingdom
Commencement:	Room No 105, 23B Riia St, Tartu, on June 17th, 2015, at 10.15

The publication of this dissertation was financed by the Institute of Computer Science, University of Tartu

ISSN 1024-6479
ISBN 978-9949-32-828-4 (print)
ISBN 978-9949-32-829-1 (pdf)

Copyright: Hedi Peterson, 2015

University of Tartu Press
www.tyk.ee

"It’s just that simple and that hard!"

Richard Hamming

Contents

LIST OF ORIGINAL PUBLICATIONS

The current dissertation is based on the following publications which will be referred to in the thesis by Roman numerals:

I. Adler P[1], **Peterson H**[1], Agius P, Reimand J, Vilo J. Ranking genes by their co-expression to subsets of pathway members. Annals of the New York Academy of Sciences 2009, 1158:1-13

II. Krushevskaya D, **Peterson H**, Reimand J, Kull M, Vilo J. VisHiC - hierarchical functional enrichment analysis of microarray data. Nucleic Acids Research 2009, 37:W587-W592

III. Billon N, Kolde R[1], Reimand J[1], Monteiro MC, Kull M, **Peterson H**, Tretyakov K, Adler P, Wdziekonski, Vilo J, Dani C. Comprehensive transcriptome analysis of mouse embryonic stem cell adipogenesis unravels new process of adipocyte development. Genome Biology 2010, 11(8):R80

IV. Jung M[1], **Peterson H**[1], Chavez L, Kahlem P, Lehrach H, Vilo J, Adjaye J. A data integration approach to mapping OCT4 gene regulatory networks operative in embryonic stem cells and embryonal carcinoma cells. PLOS ONE 2010, 5:e10709

V. **Peterson H**[1], Abu Dawud R[1], Garg A, Wang Y, Vilo J, Xenarios I, Adjaye J. Qualitative modeling identifies IL-11 as a novel regulator in maintaining self-renewal in human pluripotent stem cells. Frontiers in Physiology 2013, Oct 28;4:303.

The articles listed above have been printed with the permission of the copyright owners.

My contribution to these articles are the following:

Ref. I - I designed and co-conducted the study, formatted the input data, drafted algorithms and analysed the results. I also drafted the first manuscript.

Ref. II - I provided biological insight, interpreted and analysed biological case studies. I defined some of the visualisation solutions and analysis steps, contributed to the VisHiC web page. I participated in writing the manuscript.

[1]Equal contribution

Ref. III - I contributed with promoter analysis expertise, designed and performed integration of motif matching and conservation data.

Ref. IV - I led the chromatin immunoprecipitation signal peak analysis, motif discovery and matching experiments. I identified the binding modules of OCT4. I gathered and formatted all publicly available data for the Embryonic Stem Cell Database. I designed and implemented the database and web site. I co-wrote the paper.

Ref. V - I collected all the data except for the initial literature network. I created all the networks, performed data filtering and conducted all *in silico* perturbation experiments. I implemented the novel model evaluation approach and chose the models for experimental validation. I drafted the initial manuscript and co-wrote the final paper.

LIST OF ABBREVIATIONS

ANOVA	analysis of variance
bp	base pair
cDNA	complementary DNA
ChIP	chromatin immunoprecipitation
chip	microarray
Cy3	cyanine 3
Cy5	cyanine 5
DNA	deoxyribonucelic acid
ECC	embryonic carcinoma cell
ESC	embryonic stem cell
ESCDb	Embryonic Stem Cell Database
FC	fold change
GEO	Gene Expression Omnibus
GO	Gene Ontology
hESC	human embryonic stem cell
HPRD	Human Protein Reference Database
IUPAC	The International Union of Pure and Applied Chemistry
Kb	kilobase, 1000 base pairs
KEGG	Kyoto Encyclopedia of Genes and Genomes
mESC	mouse embryonic stem cell
mMSC	mouse mesenchymal stem cells
mRNA	messenger ribonucleic acid
PCA	principal component analysis
PCR	polymerase chain reaction
PWM	position weight matrix
Q-RT-PCR	quantitative real time polymerase chain reaction
RMA	robust multi-array average
RNA	ribonucleic acid
RNAi	RNA interference
RNA-seq	RNA sequencing
TCA	tricarboxylic acid
TF	transcription factor
TFBS	transcription factor binding site
TSS	transcription start site

INTRODUCTION

All organisms operate through activating and repressing genes in a coordinated and timely manner between biological conditions. No gene or protein can work in isolation. Understanding how interactions are formed and activated at specific time points or conditions is a central part of understanding how specific cells work.

With the increase of high-throughput methods it is possible to obtain large amount of data to study a cellular system from complementary aspects more easily. The vast quantity of data produced also enables data re-use and provokes new methods for integration of heterogeneous data.

Despite all the research performed and data produced, only about one third of human genes have been annotated to biological pathways. Hence, for the majority of genes, interactions and their activation conditions have yet to be identified.

Production of induced pluripotent stem cells has lead to the importance of understanding transcription factors and signalling molecules that are capable of determining cellular states. In order to direct stem cells into desired differentiation paths, the underlying regulatory networks have to be identified. Computational methods can help to identify new regulators for such networks, and cell behaviour in different conditions can be estimated using *in silico* modelling.

In the first part of the thesis, I provide a short overview of high-throughput data origin and analysing methods used for identifying relationships between genes. I also cover ways of representing the large amount of interaction data.

In the research section of the thesis, I demonstrate how different types of high-throughput datasets can be exploited to find new connections between genes. First, we investigated the potential of data re-usability by measuring the strength of co-expression for extracting functionally related genes using known biological pathways. Next, I introduce an integration of an enrichment and clustering method. This method enables automatic identification of gene sets that are not only co-expressed but also share known functional annotations.

I also demonstrate utility of data integration by combining evolutionary sequence conservation data together with *in silico* predictions of regulatory binding events. Additionally, we investigated embryonic stem cell regulation by adding value to experimental results with *in silico* regulatory motif detection methods. We refined richness of public data by making embryonic stem cell specific perturbation and DNA-binding data available for public use in a concise manner. I conclude the thesis by demonstrating how data integration and filtering methods followed by *in silico* modelling helps to prioritise a list of novel candidate genes for embryonic stem cell regulation.

CHAPTER 1

REVIEW OF LITERATURE

1.1 Measuring gene activity

Genes encoded in the DNA define how to produce RNA and proteins. However, the amount and timing of RNA production and protein coding is affected by many factors. In order to describe how an organism functions we study its gene activity levels.

1.1.1 Experimental analysis

Gene activity is studied by measuring mRNA levels in the cell as only transcribed genes can produce proteins and there are no high-throughput protein expression methods available. Extracted mRNA from the cell has to be reverse-transcribed to complementary DNA (cDNA). cDNA, unlike mRNA, is stable. Also, signal amplification using PCR works only on DNA.

Low-throughput analysis

The most sensitive way of measuring gene activity is by low-throughput quantitative real-time primer chain reaction (Q-RT-PCR) (Gibson, Heid, & Williams, 1996). There are two quantification methods – relative and absolute. The widely used relative quantification method amplifies the complementary DNA (cDNA), generated by reverse transcription of mRNA, and measures the difference in expression levels compared to reference genes (Bustin, 2002). Usually, housekeeping genes are considered as reference genes, as their expression should be stable throughout different experimental conditions (Vandesompele *et al.*, 2002). As the expression values from Q-RT-PCR are relative to the reference genes, it makes different Q-RT-PCR results comparable. Absolute quantification allows to measure the absolute amount of a specific mRNA sequence in a sample.

The Q-RT-PCR method however is very time-consuming and cannot be executed in a high-throughput manner. Thus, methods that allow to measure activity of a large number of genes simultaneously, like microarrays and RNA-seq, are used to obtain a general overview of gene activity in biological conditions.

High-throughput analysis

Microarrays are glass or silica slides to which tens of thousands of different short oligonucleotides, called probes, are attached to (Lockhart *et al.*, 1996; Schena *et al.*, 1995). The oligonucleotides are designed to bind with unique regions in the transcriptome and are complementary to the cDNA sequences derived from mRNAs. In each spot on the slide there are millions of copies of the same short oligonucleotide sequence.

There are two main types of widely used microarrays divided by the number of samples studied simultaneously – two-channel and single-channel chips.

In two-channel microarrays, two samples are, after labelling with different fluorophores (e.g. Cy3 and Cy5), mixed and hybridised to the same chip. The ratio between the intensities of the fluorescence are used in further analysis. Alternatively, in single-channel microarrays (e.g. Illumina Bead-arrays, Affymetrix GeneChip), only one fluorophore (usually Cy3 or biotin) is used. The fluorophore labels emit light when excited by laser at certain wavelengths (usually read for Cy3 from 532 to 570nm and for Cy5 from 635 to 670nm). The intensity of the signal reflects the quantity of the respective transcript in the original sample and is represented with a real number in the resulting data matrix.

The limitation of microarrays lies in the fact that one can only measure mRNAs that have a complementary probe defined and spotted to the slide. It may be complicated to design unique probes for all the alternative transcripts of the genes, distinguish pseudogenes from protein coding genes and, thus, the specificity of the readings can be below the expected accuracy and coverage level (Draghici *et al.*, 2006; Johnson *et al.*, 2005). Also, microarrays have problems with background noise and the dynamic range of detection is limited (Wang *et al.*, 2006).

In order to overcome such limitations, RNA sequencing (RNA-seq) is used (Wang, Gerstein, & Snyder, 2009). The RNA-seq method in principle is not limited to predefined complementary probe sets but allows to theoretically measure all the messenger RNA in the cells. Compared to microarrays, it also has a higher dynamic range for the quantification of gene expression levels and for distinguishing alternative and rare transcripts (Marioni *et al.*, 2008). However, microarrays still detect low-expressed genes better, as most RNA-seq measurement power is spent on the top 7% of the known transcriptome (Łabaj *et al.*, 2011). Also, current sequencing methods are still error-prone and sequence read lengths are often not long enough (Metzker, 2010). Thus, there are often no matches to the known transcriptome due to sequencing errors or multiple matches in the known transcriptome when reads are not long enough to be matched uniquely (Trapnell & Salzberg, 2009).

The limitation of both microarrays and RNA-seq is that these methods measure average cell population. As the concentration of mRNA from a single cell is too low and even when a cell population is used, the signal has to be amplified (Levsky & Singer, 2003).

Significant results obtained from high-throughput approaches are then validated mostly using the low-throughput Q-RT-PCR method. Q-RT-PCR and alternative gene expression validation methods are further reviewed in Chuaqui *et al.* (2002).

1.1.2 Computational analysis

Outcome of both microarray and RNA-seq experiments is typically a numerical data matrix where rows describe probe sets, transcripts or genes, and columns represent the samples. This data matrix is the starting point for the following computational analysis.

Data preprocessing

The first step in the analysis workflow is quality control (Zhang, Szustakowski, & Schinke, 2009). The aim of quality control is to check if all samples are suitable for further analysis.

Dye distribution needs to be examined when analysing microarrays in order to validate that the signal is distributed equally across the array and there are no blobs, smears or thumb prints on the chip, as these produce artefacts in the gene activity values (Reimers & Weinstein, 2005). If original scanner files are not available, then the images can be reproduced based on the data matrix for easier artefact identification by human eye. Microarray samples where obvious signal distribution artefacts are detected should be removed from the downstream analysis.

Other quality control steps involve checking correlation values between biological and technical replicates if available (Wilkes, Laux, & Foy, 2007). The latter should produce high correlation coefficients and explain the technical variation in the experiment.

Usually three biological replicates are used when detecting differences between two conditions which is less than the suggested minimum of 5 replicates (reviewed in Allison *et al.* (2006)). Ideally, all the samples are analysed in the same manner and at the same time. However, this is often impossible due to sample collection capabilities or technical limitations. In these cases, a batch effect might originate from the set-up of the experiments. For example, differences between the experiments conducted on different dates, the person conducting the experiments, the origin of the samples, and other factors need to be taken into account when performing downstream analysis. Batch removal algorithms such as Combat (Johnson, Li, & Rabinovic, 2007) might be needed to remove the experimental set-up factor and to measure the true biological differences between conditions (batch removal algorithms are reviewed in Chen *et al.* (2011)).

The next step in the analysis workflow is normalisation. Normalisation aims at making data from different samples or analyses more comparable to each other by eliminating systematic biases or artefacts. Most common methods for expression data normalisation are quantile and model based normalisation (Figure 1).

In quantile normalisation, the empirical distributions of gene expressions are made to be the same across every array in the experiment (Bolstad *et al.*, 2003). The most widely used methods for microarrays that apply quantile normalisation, are the robust multi-array average (RMA) algorithm (Irizarry *et al.*, 2003) and FARMS (Hochreiter, Clevert, & Obermayer, 2006). For RNA-seq, further developments have been proposed, such as conditional quantile normalisation (Hansen, Irizarry, & Zhijin, 2012).

Sometimes in microarray analysis, probe sets that have detection p-values above the threshold (e.g. set to 0.05), are removed. The aim of this filtering step is to remove probes that are not expressed in any of the arrays. Each probe set on the array has its own detection p-value that explains the probability that the gene measured by this probe set is actually expressed. It is calculated using the signal intensity of the perfect match and mismatch probes in a probe set for Affymetrix, and the background signal from negative probes for Illumina arrays. Removal of non-expressed probes from further analysis improves the power for detecting differentially expressed genes.

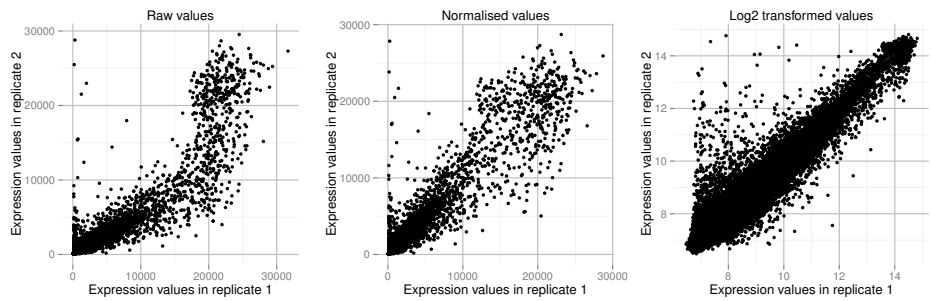


Figure 1: Illustration of expression value and correlation changes of two biological replicates during standard bioinformatics analysis steps. Both the *x*-axis and the *y*-axis represent expression values. The left plot shows that sample 1 has higher raw values than sample 2 for respective probes (the Pearson correlation coefficient 0.9242). The middle plot shows that after normalisation this bias is reduced (the Pearson correlation coefficient 0.9503). On the right plot, after log transformation, the overall correlation is high but some probes have higher values in replicate 2 compared to replicate 1 (the Pearson correlation coefficient 0.9643).

Expression matrix values representing the intensities from image analysis usually vary from zero to several thousands. In order to describe better the magnitude of expression changes, expression values are log transformed, usually on base 2 ($\log_2$) (Figure 1) (Cui, Kerr, & Churchill, 2003; Irizarry *et al.*, 2003).

Given the expression matrix, two major questions can be answered – which genes are differentially expressed between conditions and which genes are co-expressed across the conditions.

Differentially expressed genes

Differential gene expression analysis is used to identify genes that are significantly changed between different samples. Most common methods are the t-test, linear models (Smyth, 2004) and rank products (Breitling *et al.*, 2004). In addition, fold change (FC) is calculated between the mean expression values from the biological replicates under investigation and can be used to filter out genes with the greatest changes between conditions. As discussed before, usually the expression matrix is $\log_2$ transformed and log fold change values are used to express changes in expression levels. Commonly a combination of thresholds to define significantly changed genes is used, e.g. the following thresholds are applied: 2 or 1.5 fold change (log 2 is 1 or 0.585) and 0.05 for multiple testing corrected p-value (Patterson *et al.*, 2006; Raouf *et al.*, 2008).

After applying FC and statistical significance thresholds, we get lists of up- and down-regulated genes. These lists can be studied further using previously collected information compendia by applying functional characterisation approaches (described below). Log fold change values or significance p-values can also be used for prioritising genes in the gene set.

Some of the functional characterisation tools are capable of taking the order of the gene list into account while finding the most significant annotation terms (Reimand, Arak, & Vilo, 2011). This is beneficial when genes with the highest fold change belong to small and specific functional groups that would be diluted while querying within a larger gene set (e.g. the top five genes out of a hundred differentially expressed genes belong to *endodermal cell fate specification* with the total size of nine genes).

Two genes are co-expressed when their expression profiles are correlated, meaning that if one gene's expression levels go up then the same happens with the other gene. Co-expression can be measured with several distance measures (e.g. Pearson, Euclidean and Spearman's rank correlation coefficients). The Pearson correlation coefficient is among the top two most widely used correlation measures for gene expression studies (Piasecka, Robinson-Rechavi, & Bergmann, 2012).

The Pearson correlation coefficient uses real expression values and expresses linear dependence between the variables. For the Pearson correlation coefficient, data is expected to have normal distribution. The Pearson correlation coefficient can be viewed as a measure of similarity between the shape of two expression profiles. When the directionality of expression change is the same, the correlation coefficient is positive, and in case of the opposite expression patterns, the coefficient is negative. The disadvantage of the Pearson correlation coefficient is that it is affected by outliers (Langfelder & Horvath, 2012). This problem is reduced if log-transformation is applied first. The Absolute Pearson correlation coefficient is the absolute value of the Pearson correlation coefficient and it expresses only the change of values but not directionality.

In comparison, for the Spearman's rank correlation coefficient, real expression values are converted to ranks by first sorting them. The Spearman's rank

correlation coefficient expresses the squared differences of ranks. Rank correlations, such as the Spearman's rank correlation coefficient, are less sensitive to non-normal distribution.

Co-expressed gene pairs are often used as input for finding potentially functionally related gene sets. The "guilt by association" paradigm states that if genes behave similarly in at least one biological condition then they are potentially carrying out similar biological functions (Wolfe, Kohane, & Butte, 2005). This paradigm is used widely in different functional enrichment tools where a gene without a known function is linked to the annotations of the other genes in the gene set.

1.1.3 Clustering genes

When analysing large biological data matrices using clustering methods, the goal is to identify genes that behave similarly and could be co-regulated in similar biological conditions, or share common functions.

Clustering is a method that groups data points based on the distances between the points. For example, genes are clustered together by similarities of their expression values. Usually the correlation coefficients between the individual genes are calculated and coefficients are converted to distance metrics (e.g the Pearson's distance = 1 - the Pearson correlation coefficient).

Hierarchical clustering

In hierarchical clustering clusters are obtained based on the distance between values where the two points with the smallest distance are put into one cluster. Then the next closest pair is found until all the points are clustered. This is known as full agglomerative hierarchical clustering. Another option is divisive clustering where, starting from one cluster containing all the observations, the sets are split recursively. Full hierarchical clustering has a quadratic complexity which makes this approach too slow for large datasets that need to be analysed fast.

Another alternative is to optimise the number of computations needed and use a set of pivot points in the data. This allows to limit the number of computations in an order of magnitude. For example, the approximate hierarchical clustering algorithm HappieClust (Kull & Vilo, 2008) for an average size microarray matrix requires only 2-3% of the calculations that are needed for full agglomerative hierarchical clustering to produce comparable results (Kull & Vilo, 2008).

In order to obtain individual gene clusters, the full hierarchical clustering tree needs to be cut at a certain distance. The cutting distance is usually arbitrarily selected, and has to be chosen by the user. Visualising the clustering results with dendrograms (described below) might help to choose the optimal cluster cutting threshold.

K-means clustering

An alternative way is to pre-define a number of clusters so that all the genes will be covered. For example, if we have only two conditions, we can expect to see three major clusters of genes (genes that go up, genes that go down and genes that do not change significantly). In this case we can set k, denoting the number of clusters, to 3. However, there are always genes that have expression profiles somewhere between the expected clusters. Because of that, we get a lot of noise in the clusters unless we are capable of pre-defining the optimal k.

The rule of thumb would set $k \approx \sqrt{\frac{n}{2}}$ which for an average of 50000 probe sets would be close to 160. The suitability of a given k, however, depends a lot on the studied biological conditions. Additionally, there are other methods like the Elbow method where the number of clusters is chosen at the point where the percentage of variance stops adding extra information (Tibshirani, Walther, & Hastie, 2001). For a given matrix, the Elbow method helps to find the optimal k which is dependent on the specific dataset, and needs to be calculated again for the next experiment. The clustering outcome is also dependent on the initial cluster centres, and can vary when using different initiation points on the same data matrix.

1.1.4 Visualising expression data

Large numerical data matrices contain information about the measured genes but for comprehensible presentation of similarities between data points, additional steps need to be taken. Although programmatically it is easy to pick genes by specific behaviour or similarity measure, different visualisation approaches help to get a more comprehensive overview of the data for a larger audience. The most popular approaches to visualising gene expression data are dendrograms, heat maps, line plots and scatterplots (Gehlenborg *et al.*, 2010).

As discussed earlier, clusters identify genes that behave similarly. In addition to having gene lists for each cluster, it is beneficial to represent relationships between clusters. Therefore, clusters are often visualised by dendrograms – trees of clusters linked to each other by hierarchical clustering. For each dendrogram the distance measure is represented on the *y*-axis and nodes are described on the *x*-axis.

Heat maps are graphical representation of data where numerical matrix values are visualised as coloured rectangles. Usually a gradient from one colour to another through a mid point (e.g. blue to white to yellow or green to black to red) is used to demonstrate the levels of expression (e.g. red for highly expressed genes, green for lowly expressed genes). The intensity of colour in each matrix cell in the heat map corresponds often to the mRNA expression level of a given gene in a specific tissue or treatment.

Blocks of cells that have similar colours represent genes and conditions where expression values are similar. It is easy to distinguish such coloured blocks from each other by human eye and make a fast decision about the outcome of the experiment.

Heat maps and dendrograms can also be combined. In this case expression values of individual genes or cluster averages can be represented by heat maps while the relationships between genes (or clusters) and conditions can be described by attached dendrograms. Original visualisation approaches by (Eisen *et al.*, 1998) have been developed further, for example, in the pheatmap R package developed by Raivo Kolde.

Gene expression profiles are also often represented as line plots where the x-axis contains conditions and the y-axis denotes expression levels. A coloured line then represents expression changes across conditions.

High-throughput methods create very large datasets with tens of thousands of rows and hundreds of columns. These data matrices are easily manageable by computers but very hard to grasp by human brains if kept numerical. In order to understand the patterns and trends in the data, usually the dimensionality of the huge matrices needs to be reduced and visualisation methods applied (Slonim, 2002).

When analysing biological data, the main questions are often about determining sets of similarly behaving genes and pointing out significant differences between experimental conditions. Various standard analysis and visualisation steps are applied to large data matrices, such as clustering, principal component analysis, dendrograms and heat maps (Gehlenborg *et al.*, 2010).

Principal component analysis (PCA) is used for reducing the dimensionality of large datasets and finding the variables (contributing factors) that separate samples the most. The principal components represent the directions where data points are most spread out and are themselves linear combinations of the original variables.

An eigenvector is a direction in the data with the largest variance, and the eigenvalue explains the variance of the data in that direction. The eigenvector with the highest eigenvalue is the first principal component. The eigenvector with the next highest eigenvalue is the second principal component and so on.

Each principal component is described by contributing factors and their loadings that explain the magnitude of the contribution for a specific component. The loadings can be used to explain for example the gene's contribution to the components and represented in a biplot. On the plot, the genes with the largest weights extend towards the respective component directions (Landgrebe, Wurst, & Welzl, 2002). The usage of PCA for gene expression analysis is further explained in (Ringnér, 2008).

Functional characterisation

Given a list of co-expressed genes, the next natural question is to find the potential functionality the gene set could carry out and characterise the genes without a known function.

Knowledge about genes and proteins is gathered to large structured databases. These databases can hold either specific information (e.g. human protein-protein interactions) or describe broader terms and relationships. The Gene Ontology (GO) (Ashburner *et al.*, 2000) is a bioinformatics initiative that gathers information about genes and gene products across all species. Terms used in GO are species-neutral. Each gene is described with terms from a controlled vocabulary. The vocabulary is composed of three main domains: *molecular function (MF)* describing activities at the molecular level; *biological process (BP)* describing operations or molecular events with a defined beginning and end; and *cellular component (CC)* describing cellular and extracellular parts.

The Gene Ontology is described with a directed acyclic graph where each term has defined relationships to one or more terms inside the domain (in rare occasions to other domains). All the genes associated with a term also belong to the parent term (e.g. *stem cell differentiation* (GO:0048863) is a child term of *cell differentiation* (GO:0030154)).

Gene Ontology annotations are widely used to search for overrepresented terms among co-expressed gene groups. There exists a variety of software applications that enable the users to search for enriched GO annotations for the provided gene lists (Bauer, Gagneur, & Robinson, 2010; Huang *et al.*, 2007; Reimand, Arak, & Vilo, 2011).

Functional characterisation is not limited to finding overrepresented GO annotations, but can be applied to any collection of previous knowledge or even to predicted data. For example, g:Profiler uses TRANSFAC (Matys *et al.*, 2006) regulatory motif enrichment information; miRNA target prediction data from miRBase (Griffiths-Jones *et al.*, 2006); KEGG (Kanehisa *et al.*, 2014) and REACTOME (Croft *et al.*, 2014) pathways; protein complexes from CORUM (Ruepp *et al.*, 2010); and Human Phenotype Ontology (Köhler *et al.*, 2014) (Reimand, Arak, & Vilo, 2011).

1.1.5 Evaluating experiments

Every time an experiment is set up, a pair of hypotheses (H_0, H_1) is formed to be tested. The aim of any statistical test is to limit detecting the effect that is not present, or missing the effect that actually is present in the data. Typically, H_0 states that there is no difference between two conditions and the alternative hypothesis H_1 states that there is a difference. A type I error describes the situation when a true null hypothesis is incorrectly rejected (false positive) while a type II error describes the situation when a test fails to reject a false null hypothesis (false negative).

In expression analysis, the main task is to look for genes that are significantly changed between the conditions. In this case, a comparison of two groups can be done using the t-test that assumes that the sample sizes are equal and variance among the samples is same in both case and control groups.

When more than two groups are compared at the same time, it is beneficial to use analysis of variance (ANOVA) statistical models (Cui, Kerr, & Churchill, 2003). With ANOVA one can test whether or not the means of several groups are equal while controlling type I errors so the error rate threshold remains at 5%.

Multiple testing correction

With high-throughput data produced to study different conditions, and the availability of large annotation datasets, multiple testing problems come into focus (Dudoit, Shaffer, & Boldrick, 2003). When we test the same group of genes against different annotation groups, we have to take into account the total number of tests we perform in order to limit the false positive results. The most common ways to adjust significance thresholds are Bonferroni correction and false discovery rate (Benjamini & Yekutieli, 2001).

Bonferroni correction is the most conservative method to control the probability of making one or more false discoveries among all the hypotheses (family-wise error rate). When a test involves multiple comparisons the comparison-wise error has to be corrected by the number of comparisons done, so that the test experiment-wise error would be at an expected level. For example, when performing 50 pair-wise comparisons the Bonferroni corrected threshold for each individual pair-wise test (comparison-wise error) should be 0.05/50=0.001 so the experiment wise error would be 0.05.

False discovery rate (FDR) is one of the most common multiple testing approaches used in high-throughput biology. FDR controls the proportion of type I errors made among all significant results (Benjamini & Hochberg, 2000). FDR is less conservative than Bonferroni correction and leads to loss of fewer true positive results than Bonferroni correction.

For n alternative hypotheses $H_1, .., H_n$ tested, p-values $P_1, ..P_n$ are calculated. FDR ranks the p-values in increasing order and, for a given significance threshold α, finds the largest k that satisfies

$$P(k) \geq \frac{k}{n} \cdot \alpha \ .$$

All alternative hypothesis $H_{(1)}, ..., H_{(k)}$ are then accepted as significant.

1.2 Gene functions

Section 1.1 of measuring gene activity was represented in a simplified manner assuming that there are linear steps from gene to mRNA to protein levels. In a real biological system, these steps are not linear as there are many levels that control the amount protein present in a cell. With our current understandings and

technological abilities, the final level of protein activities has to be approximated using data from mRNA experiments.

In the context of this thesis we concentrate on subsets of genes that are capable of regulating the expression of other genes or mediate the signal in stem cells and during differentiation.

1.2.1 Embryonic stem cells

Embryonic stem cells (ESCs) are pluripotent cells that are capable of differentiating to any cell type in the organism. The first human embryonic stem cells (hESCs) were derived from the inner cell mass of a blastocyst in 1998 by Thomson *et al.* (1998). Human embryonic stem cells are the basis of three primary germ layers: ectoderm, endoderm and mesoderm. These layers develop further to more than 200 different cell types in the human body (Figure 2).

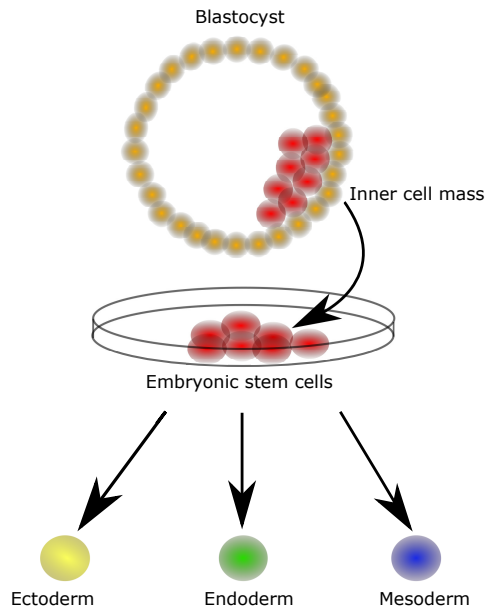


Figure 2: Pluripotent embryonic stem cells are derived from the inner cell mass of a blastocyst. These cells are capable of differentiation first to three primary germ layers (ectoderm, endoderm, mesoderm) and finally to over 200 distinct cell types in the human body.

The pluripotent stem cells at each cell division give birth to another pluripotent stem cell and to one cell that starts to differentiate towards one of the three main germ layers. This self-renewal property keeps stem cells in an undifferentiated state.

Human embryonic stem cells are kept in the pluripotent state by expressing three transcription factors – OCT4, SOX2 and NANOG (Boyer *et al.*, 2005). These three transcription factors regulate each other and a variety of target genes

(Boyer *et al.*, 2005; Macarthur, Ma'ayan, & Lemischka, 2009). In order to start differentiation, the activity of these three genes needs to be decreased and the activity of the differentiation driver genes has to increase (Yeo & Ng, 2013).

Exit from pluripotency state can be achieved by knocking down the core regulators of OCT4, SOX2 and NANOG (Greber, Lehrach, & Adjaye, 2007; Matin *et al.*, 2004). Alternatively, overexpressing cell type specific genes individually or in combinations leads also towards differentiation, e.g. overexpression of HEX for hepatoblasts (Inamura *et al.*, 2011); APP for neural differentiation (Freude *et al.*, 2011); GATA4 and TBX5 for cardiac specification (Dixon *et al.*, 2011).

The commonly used term pluripotency covers at least two sub-states - naïve and primed. The naïve or ground state describes embryonic stem cells that have no differentiation bias. These cells are capable of forming blastocyst chimeras when additional cells are added to preimplantation embryos; have two active *X* chromosomes (if female karyotype); express Rex1, NrOb1 and FGF4 markers. Primed pluripotent embryonic stem cells, however, cannot form blastocyst chimeras any more, they have an inactivated chromosome *X* and also have a differentiation bias. (Nichols & Smith, 2009)

According to these properties, mouse embryonic stem cells are considered to be in the naïve state, while the majority of human embryonic stem cells represent the latter, primed, state. However, it is possible for hESCs to adopt the naïve phenotype, at least for limited number of passages (Hanna *et al.*, 2010). Also, primed mouse embryonic stem cells resembling hESCs can be produced when isolated from the post-implantation epiblast (Tesar *et al.*, 2007). The differences between mESCs and hESCs, attributed for a long time to species differences, may be due to differences of cell states, instead. It has been shown recently that when human embryonic stem cells are treated with 5i/L/A (MEK inhibitor, GSK3 inhibitor, ROCK inhibitor, BRAF inhibitor, SRC inhibitor, Activin, hLIF in N2B27 medium) then primed hESCs are converted to the naïve state (Theunissen *et al.*, 2014).

Although the mechanisms of cell state transition from naïve to primed are not clear yet, it can be expected to take place through gradual changes of the activity levels of key factors. For example, it has been shown that an alternative OCT4 enhancer is used only in naïve ESCs and it is inactivated in primed state (Theunissen *et al.*, 2014). Also, gradual changes can be encoded in proteins changing their dimerisation partners and, thus, the regulated genes. For example, OCT4 changes its known dimerisation partner SOX2 to SOX17 during the first differentiation steps (Kallas *et al.*, 2014).

Alternative cellular systems used for studying gene regulation in pluripotent stem cells are embryonic carcinoma cells. These cells are derived from germ cell tumours, like the testicular tumour. Embryonic carcinoma cells share cell characteristics with hESCs, and core regulatory proteins are the same as in hESCs (Clark, 2007). Embryonic carcinoma cell lines are easier to culture and were available as a model system for studying pluripotency before the first hESC lines were established (Andrews *et al.*, 2005). Embryonic carcinoma cell lines like

NCCIT and NTERA2 have been used for studying self-renewal and pluripotency regulatory networks (Andrews *et al.*, 2005; Greber, Lehrach, & Adjaye, 2007; Josephson *et al.*, 2007).

1.2.2 Adipocytes

Adipocytes are cells of mesenchymal origin and their main function is storing energy as fat. Mesenchymal stem cells originate from the mesoderm and differentiate into bone, muscle, connective tissue and adipocytes. There are three main types of adipose tissue – brown, brite (beige) and white (Giralt & Villarroya, 2013; Harms & Seale, 2013).

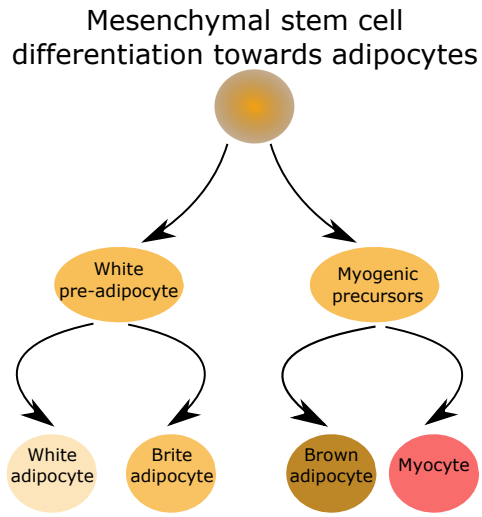


Figure 3: Differentiation of adipocyte lineages from mesenchymal stem cells. White and brite adipocytes are differentiated from white pre-adipocytes while brown adipocytes, like monocytes, are derived from myogenic precursors.

During adipocyte differentiation from the mesenchymal stem cells, first precursory cells are formed. These cells cannot differentiate any more towards alternative lineages such as muscle or bone. In the final step, terminal differentiation into mature and functional adipocytes takes place (Figure 3).

Mouse embryonic stem cells can be differentiated directly into adipocytes when treated with retinoic acid (Dani *et al.*, 1997). Also PPARγ is a key regulator of adipocyte development (Rosen *et al.*, 1999) and is capable of converting non-adipose cells to adipocyte-like cells together with C/EBPα (Wu *et al.*, 1999). Although some regulators that drive differentiation of adipocytes are known, the detail regulatory steps and participating genes are still poorly described.

1.3 Gene regulation

Measuring gene activities across conditions and various time periods provides us with a set of snapshots of cell populations. Based on the data of active and inactive genes, the goal is to identify regulatory mechanisms and switches that define when and how much certain genes are activated. Below we address mainly the transcriptional regulation of genes.

1.3.1 Transcription factors

Transcription factors are proteins that control transcription rates of other genes by binding to the regulatory regions. These regions are mainly promoters located in the proximity of the gene transcription start sites (TSS) or enhancer regions further away from the TSS. Transcription factors work as single proteins, form homodimers or heterodimers with other transcription factors. Some transcription factors are broadly expressed and bind to a large set of genes (e.g. TBX1, CLOCK), while others are more specific (e.g. ZBTB32 with testis specific expression) (Vaquerizas *et al.*, 2009). There are approximately 1500 transcription factors in humans that can be grouped into 347 protein domains and families (Vaquerizas *et al.*, 2009).

Chromatin immunoprecipitation

The most common high-throughput method for identifying potential transcription factor binding sites across the genome is chromatin immunoprecipitation (ChIP). First, the transcription factor is cross-linked with DNA, cells are lysed, for example, with sonification, and TF-DNA complexes are pulled down using a transcription factor specific antibody. After TF-DNA complex cross-linking is reversed by heating, the DNA strands are purified and amplified, and identified by various methods. Usually, the resulting DNA sequences have a length of up to 1000 base pairs. A technical overview has been provided by Nelson, Denisenko, & Bomsztyk (2006).

Previously, sequence identification was done using specialised promoter tiling arrays where short overlapping oligonucleotides complementary to promoter regions were present (ChIP-chip method). The DNA sequences bound by the transcription factor were identified by a fluorescent signal like in microarray analysis.

Alternatively paired-end-tag libraries are used where unique regions from the genome are attached with short unique sequences for further identification (ChIP-PET method).

Lately, the most common method is to sequence the DNA regions pulled down in the immunoprecipitation step (ChIP-seq method). The advantage of ChIP-seq is better coverage of the sequences, as it is not limited to pre-defined promoter sequences. However, if the sequences are short, mapping them uniquely back to the genome can be difficult. Longer methodology review can be found in (Ho *et al.*, 2011).

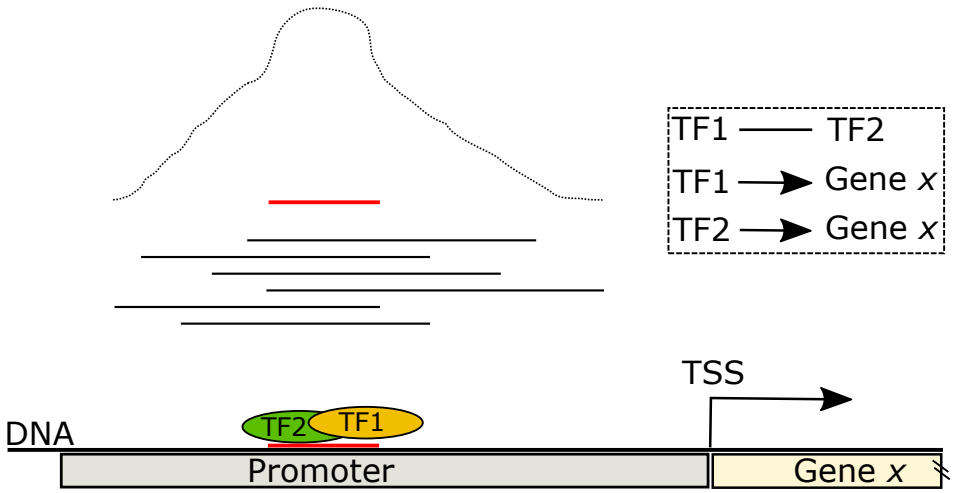


Figure 4: Chromatin immunoprecipitation followed by high-throughput readout leads to a set of DNA sequences. When aligning these we can define a peak region and map that to the promoter of gene x. The real transcription factor binding to DNA usually happens under the highest peak region (red). Often transcription factors work as dimers and bind together to DNA (TF1, TF2). From this example we can extract the following relationships – protein-protein interaction between TF1 and TF2 and transcriptional regulation of gene x by both TF1 and TF2 (dashed rectangle).

Whichever method was used to identify the sequences bound by the transcription factor, the following analysis step is peak detection (Figure 4). Most of the transcription factors bind to short DNA regions, but from the ChIP experiments we obtain regions that are several hundred base pairs long. The sequences are cut at random positions around the area where the transcription factor is bound to DNA (Nelson, Denisenko, & Bomsztyk, 2006). Therefore, by aligning the sequences, the overlap between sequences should identify transcription factor binding sites.

There are several alternative algorithms for peak detection that are reviewed by Barrett, Cho, & Palsson (2011).

Transcription factor binding sites

After the peak regions are identified, usually transcription factor binding sites are searched to confirm the binding. Transcription factors recognise short, (4-12 base pairs), DNA motifs called transcription factor binding sites and bind only to the corresponding site either alone or in complex. Each transcription factor can recognise usually more than one unique motif.

The set of transcription factor binding sites is usually represented as a consensus motif (Table 1) with alternative options at some positions, position count matrices (Table 2) and position weight matrices (PWM) (Table 3) where each nucleotide and position has some weight. Ideally, the sequence score described by a

PWM should represent the binding energy for the sequence (Stormo, 2000). Position weight matrices are often represented with sequence logos to illustrate the tolerance of substitutions at every position (Figure 5) (Schneider *et al.*, 1986).

M_1	C	T	T	G	C	A	A	A	T	G
M_2	A	A	T	G	C	A	A	A	T	A
M_3	A	A	T	G	C	A	A	A	T	G
M_4	T	A	T	G	C	A	A	A	T	A
M_5	A	C	T	G	C	A	A	A	T	A
				.	.	.				
M_{n-1}	A	A	T	G	C	A	T	A	T	A
M_n	C	A	T	G	C	A	G	A	T	C
$Consensus$	W	A	T	G	C	A	A	A	T	R

Table 1: List of n=1582 potential OCT4 binding sites. A consensus base is found for each position and represented under the line. IUPAC nucleotide code W stands for A or T, R stands for A or G.

A	465	1228	0	0	0	1582	1364	1582	243	472
C	356	158	0	0	1492	0	0	0	40	210
G	261	8	0	1582	90	0	97	0	241	548
T	500	188	1582	0	0	0	121	0	1058	352

Table 2: Count matrix of the OCT4 binding sites listed in Table 1.

The count matrix is obtained by first aligning the initial list of binding sites. At each position across a set of sequences, all possible nucleotides are summed up. In the resulting matrix, sums of columns have to be equal.

Finally, to represent the binding energy, the position count matrix is converted to a weight matrix according to

$$M_{k,j} = log(M_{k,j}/b_k)$$

where b is a background model describing the expected frequency of individual nucleotides in the dataset (for the example in Table 3 G and C had frequency of 0.2, A and T 0.3) (Stormo, 2000).

A	0.083	0.099	−0.038	−0.038	−0.038	0.103	0.101	0.103	0.072	0.083
C	0.085	0.072	−0.038	−0.038	0.109	−0.038	−0.038	−0.038	0.049	0.076
G	0.080	0.023	−0.038	0.110	0.062	−0.038	0.064	−0.038	0.079	0.092
T	0.084	0.068	0.103	−0.038	−0.038	−0.038	0.061	−0.038	0.096	0.078

Table 3: Position weight matrix of the OCT4 binding sites listed in Table 1.

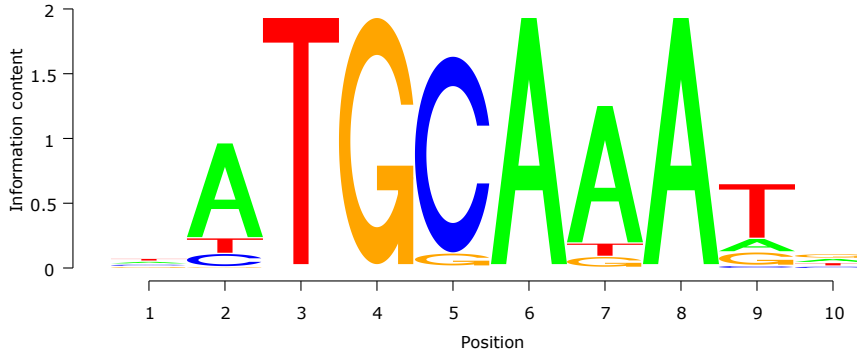


Figure 5: Sequence logo of OCT4 binding sites based on 1582 motifs from Table 1. The height of each letter describes information content that represents the tolerance for substitutions in that position (Schneider *et al.*, 1986).

A matrix of binding sites is often represented with a sequence logo (Figure 5). The height of each letter in the logo describes information content that represents the tolerance for substitutions in that position (Schneider *et al.*, 1986). Positions with a low tolerance for substitutions have high information content and are represented with larger letters. Positions illustrated with smaller letters represent positions with low information content and high tolerance for substitutions.

Transcription factor binding data is collected to databases like TRANSFAC (Matys *et al.*, 2006) and JASPAR (Sandelin *et al.*, 2004). These databases keep both individual experimentally validated binding sites and position weight matrices built on these sequences. It is common to use third party tools like STORM (Smith *et al.*, 2006) and STAMP (Mahony & Benos, 2007) for matching the position weight matrices obtained from those two databases to promoter sequences.

1.3.2 Cell perturbations

Another way to study the effect of a gene for cell behaviour is to perturb its expression. Positive perturbations are usually done by over-expressing a gene by introducing corresponding cDNA to the cells. Negative perturbations are performed by knocking down mRNA expression by introducing complementary interfering RNA (RNAi) to the cells (Figure 6). The short regulatory RNA binds to already transcribed mRNAs and inhibits the gene activity by initiating double RNA degrading mechanisms.

Perturbation experiments are widely used to obtain regulatory edges for gene regulatory networks (Ideker, Thorsson, & Karp, 2000; Molinelli *et al.*, 2013; Peterson *et al.*, 2012; Wagner, 2002). Using only perturbation data does not allow to distinguish if the regulatory relationship is direct or indirect between the perturbed gene and its target genes. But combining the data with chromatin immunoprecipitation (Kubosaki *et al.*, 2010) or proteomics (Sopko *et al.*, 2014) data could help to differentiate these two groups.

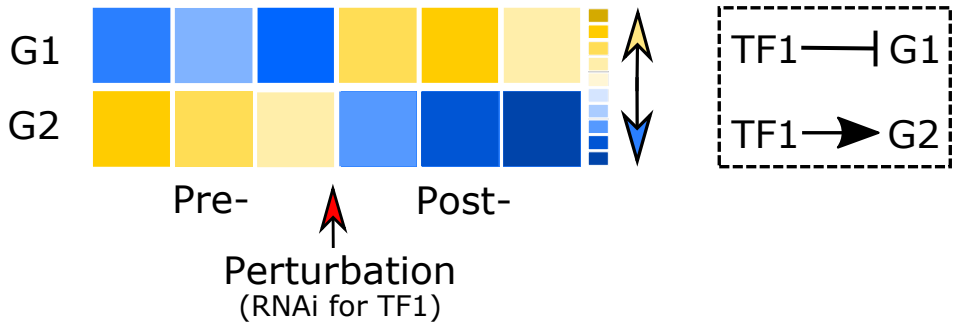


Figure 6: In perturbation experiments, such as negative perturbation of transcription factor TF1, genes regulated by TF1 would change their expression. Genes with high expression levels (yellow) are down-regulated (blue) and vice versa. From perturbation experiments we can conclude both negative (TF inhibits G1) and positive (TF1 activates G2) regulatory edges (in the dashed rectangle).

1.4 Biological networks and pathways

Networks and pathways can be considered or modelled as graphs describing connections between biological entities carrying out some functions in a specific biological setting.

1.4.1 Networks

Knowledge about relationships between genes are represented with networks. The networks can be built manually by annotating each edge with experimental evidence and literature data. Alternatively, biological networks are constructed using high-throughput data such as mRNA expression, yeast two-hybrid experiments or chromatin immunoprecipitation. In addition, automatically mined literature co-mentions in abstracts or DNA conservation is used.

Functional relationships between genes are usually described for a pair of genes or proteins. Each such relation can have a different experimental origin, confidence value and direction. In order to handle such pairs in a more comprehensive way, the pairs are linked together forming a network. In such networks, genes or proteins are represented by nodes, and relationships between them are represented by edges.

Edges can either be directed or undirected. Undirected edges describe connections between two genes or proteins where only the association is known but not its directionality. Such edges could be co-mentions or protein-protein interactions.

In gene regulation networks, directed edges explain the direction of regulation between the nodes. For example, a node representing a transcription factor will have an outgoing edge to the target gene. The same edge will be an incoming edge for the same target gene (Figure 7). A node with only incoming edges is a target node or leaf. Sometimes edges are also weighted to represent the trustworthiness

of the given interaction. The edge weight could illustrate, for example, correlation in a co-expression network, peak score from an immunoprecipitation study or co-mentions in a literature network.

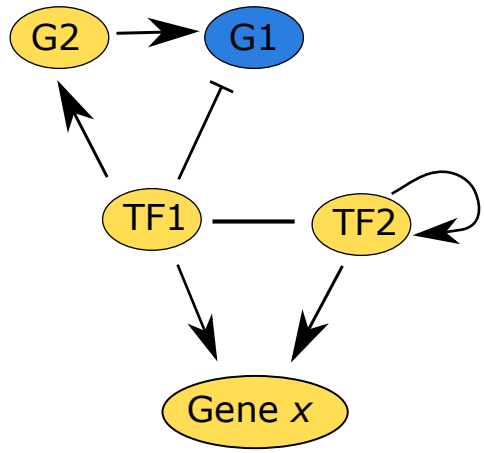


Figure 7: Minimalistic regulatory network. When TF1 is active all other genes will be active (yellow) except G1 that is in an inactive (blue) state (in network modelling, inhibition is considered stronger than activation, therefore G1 stays inactive in this example). TF1 forms a feed-forward loop through G2 to G1 and can itself be considered a hub. TF1 and TF2 have undirected interaction that could represent protein-protein interaction or literature co-mentioning, for example. TF2 is an autoregulative transcription factor.

The directed edges in the network can be either positive or negative. While the majority of directed edges represent positive regulation (e.g. transcription factor up-regulating a target gene), there are also negative regulation events that describe inhibitory relations between the objects. For example, a repressor protein binding to a promoter element and, thus, blocking transcription. Alternatively, genes that show over-expression after negative perturbation of a gene will have a negative regulatory edge from the perturbed gene.

Each biological network of representative size includes the most common network motifs – connected components, hubs, cliques, feed-forward and feed-back loops (Figure 8) (Alon, 2007). Transcription factors are usually hubs, having outgoing edges to many other genes while incoming edges from few other nodes. Cliques are sets of nodes that are all connected to each other. Protein complexes where each gene or protein has interactions with many other complex partners, usually form cliques in networks. Also, tight regulatory subunits where transcription factors regulate their own regulators such as OCT4, SOX2 and NANOG core network in embryonic stem cells, form a clique (Boyer *et al.*, 2005).

In a biological organism, functions are often coded in at least two alternative ways to ensure robustness of the signal, as the systems must be robust against genetic and environmental perturbations to be evolvable (Kitano, 2004). For a system to have periodic behaviour, at least one negative feed-back loop is needed,

while for the existence of multiple steady states, at least one positive feed-back loop is needed (Gouzé, 1998; Snoussi, 1998).

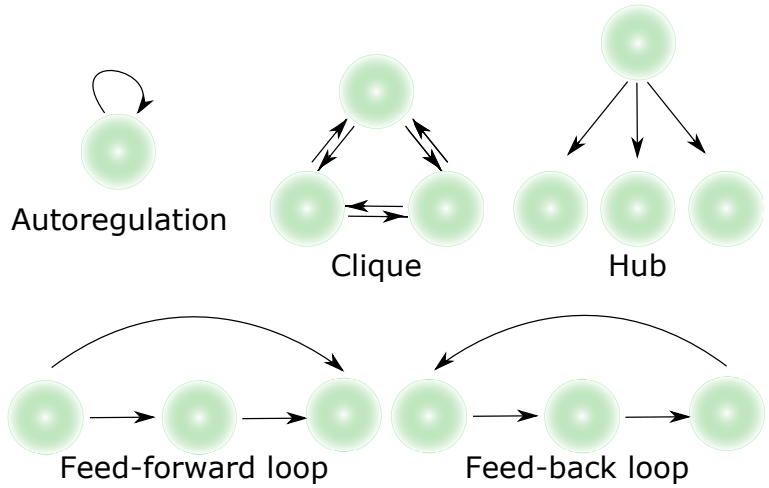


Figure 8: Network motifs: autoregulation – when a gene is a transcription factor and regulates itself; clique – transcription factors that have cross-regulation (like OCT4, SOX2, NANOG); hub – transcription factor regulating many other genes; feed-forward loop – e.g. two transcription factors regulating a common target gene (right circle) while one transcription factor (left circle) regulates the other (middle circle) as well; feed-back loop – e.g transcription factor (left circle) regulating a transcription factor (middle circle) and being regulated by its own indirect target (right circle).

The most common network motif is the feed-forward loop that is composed of a transcription factor TF1 that regulates a gene G2, and both TF1 and G2 regulate gene G1 while TF1 regulates G1 directly as well (Figure 7) (Mangan & Alon, 2003). The OCT4, SOX2 and NANOG triplet form both self-regulatory and feed-forward loops in hESCs regulatory networks (Boyer *et al.*, 2005).

The feed-forward loop enables different regulatory behaviour based on the type of regulatory edges. For example, when both regulatory edges are positive, the feed-forward loop creates stable activation levels that are rather insensitive to temporary changes of the input (Mangan, Zaslaver, & Alon, 2003). When the edges are of opposite type (one positive, one negative) then the loop provides a switch that creates a delayed oscillating response by inactivating the initial regulator (Mangan & Alon, 2003; Mangan, Zaslaver, & Alon, 2003). A network without various types of feed-back loops will only reach a unique fixed state regardless of the initial conditions (Pigolotti, Krishna, & Jensen, 2007).

Although regulatory networks of embryonic stem cells have been composed for years (Boyer *et al.*, 2005; Kim *et al.*, 2008; Wang, Levasseur, & Orkin, 2008), there is still a lack of knowledge about what the full network of pluripotency regulation looks like. It is known so far that there is a core triplet of transcription factors, OCT4, SOX2 and NANOG, that regulate themselves and many other genes to keep the hESCs in pluripotent state (Boyer *et al.*, 2005). Also, it has

been recently shown that ERK signalling plays an important role in keeping the pluripotency (Göke *et al.*, 2013). Alternatively to gene expression data, ENCODE DNAse I footprinting and ChIP data has been used to come up with an alternative ES regulatory network (Neph *et al.*, 2012).

1.4.2 Pathways

Pathways characterise certain biological processes that are described as series of reactions and interactions, and are often represented as graphs. Unlike in networks, where edges are often undirected, in pathways edges mostly represent reactions and have direction. Reactions are usually operated by enzymes and represent conversion of one substance to another.

While regulatory networks are commonly derived from high-throughput experiments or through literature mining, pathways are described one reaction at a time by human curators. Annotating certain biological functions is, therefore, time consuming. Of course, some information can be extracted using computational approaches but the value lies behind the trustfulness of nodes and edges in the pathway.

Biological pathways are collected and stored to major databases from where they can be queried and visualised further. Reactome (Milacic *et al.*, 2012) is a hand-curated database that covers various diseases, metabolic and signalling pathways (Figure 9). All the information is curated and externally reviewed. Latest version 51 covers 7760 human genes and 1428 small molecules that are organised into 1597 pathways annotated by 17939 literature references (Croft *et al.*, 2014). By using orthologous proteins the information can be projected to other model organisms.

The KEGG pathway database also collects manually curated knowledge and automatically maps knowledge from one organism to another when there is enough support from orthology mapping (Kanehisa *et al.*, 2014). The current KEGG database covers 470 pathway maps for 3660 organisms supported by 356 006 literature references (March 2015).

Although accurate, manual data curation is time-consuming and, thus, there exists a need for automatic approaches. Pathways can be constructed using literature mining. For example, mining the vast amount of published scientific literature helps to point out genes that are mentioned together (Wu, Schwartz, & Nenadic, 2013). More advanced methods, like natural language processing approaches, can also detect relation types between proteins and small molecules or diseases, for example (Yuryev *et al.*, 2006).

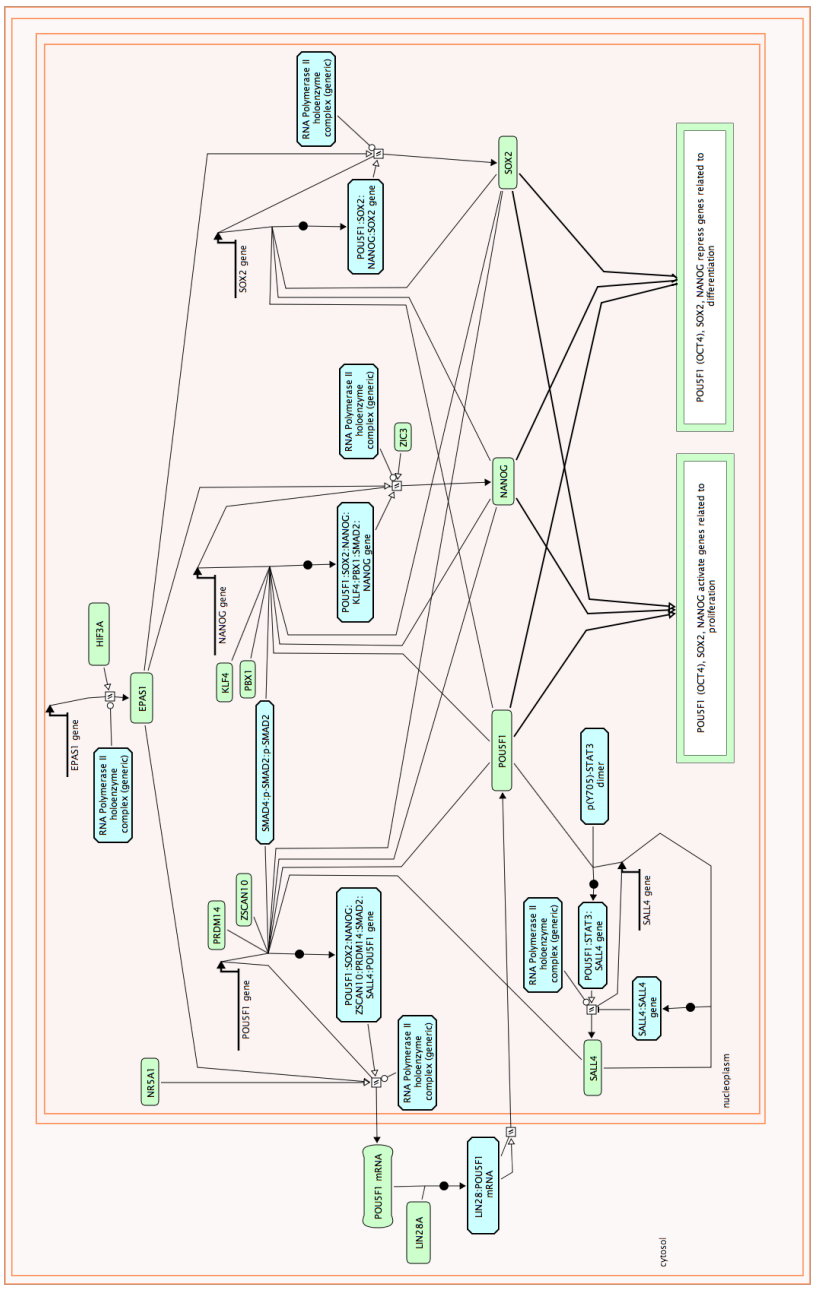


Figure 9: Reactome pathway number 263856 of *Transcriptional regulation of pluripotent stem cells*. Light blue rounded rectangles represent complexes, light green rounded rectangles represent proteins. Edges represent various reaction types. Double pink lines mark the membrane (nuclear and cell membrane). Figure adapted from www.reactome.org (Croft *et al.*, 2014)

1.4.3 *In silico* modelling

Collecting data into networks helps to get a visual overview of the relationships between genes, identify network modules and determine core genes. However, networks are static as they only describe interactions but not their potential activation states across time or conditions.

In silico modelling helps to further study biological functions described by networks. With modelling, it is possible to find all activity states, called steady states, that the given network is capable of describing. These steady states can represent, for example, different stages of cell differentiation, disease states or cell cycle phases.

Modelling can also be used to test network conformity with the biological functions it should describe. For example, if a network should describe differentiation towards specific cell lineages, but only some of the final states are reachable from the network by simulations, then probably some regulatory edges are missing from or are incorrect in the current network.

There are several network modelling approaches, each with different data needs and outcomes. Roughly, the modelling methods can be separated into qualitative and quantitative modelling.

Qualitative modelling is less stringent data-wise and a network with undirected edges is sufficient for such a modelling approach. Qualitative modelling is also referred to as Boolean modelling due to the binary logic of ON and OFF states of each node. At any given time point, each node can either be active or inactive, thus ON or OFF. With qualitative modelling, we can measure the movement of a signal through the network, study network activation states that are stable and need perturbations to get out of the state – steady states.

Steady states represent stable cell types or differentiation stages where cells will stay if there is no external stimulus. For example, in embryonic stem cells we can distinguish two steady states. One is the pluripotency state where the cells can differentiate to any cell type and where certain key transcription factors, such as OCT4, SOX2 and NANOG, are active. The other is an early differentiation state towards exiting pluripotency where these factors are inactive and differentiation markers are active. In order to transition from one steady state to the other, key regulators need to be affected.

Qualitative modelling is widely used for regulatory networks where the link between genes or proteins might be known, but not necessarily the size of the effect, nor even if the regulation is direct or indirect. Qualitative modelling is also applicable to larger networks. Networks suitable for qualitative modelling are easier to compose from high-throughput data, unlike quantitative modelling approaches that need quantitative data with high quality (Schlitt & Brazma, 2007).

Quantitative modelling is more data-demanding and is usually used for smaller networks. For example, ordinary differential equation models need many kinetic parameters which limits the network, as it is difficult to produce such data in large amounts. Thus, this type of modelling is less used for larger gene regulatory networks established from noisy high-throughput data.

1.5 Combining heterogeneous data

Each individual interaction measurement method allows to dissect only one part of the regulatory complexity of cells. The datasets become more valuable if they can be combined to tackle the complex regulatory mechanisms active in specific cell types and conditions. For example, when studying a transcription factor's direct targets in a specific cell type, perturbation experiments can be used to get the downstream regulated genes; DNAse I hypersensitivity sites give open chromatin regions of potential active promoters; chromatin immunoprecipitation experiments add binding events of the transcription factor (Neph *et al.*, 2012).

Following up with the same types of experiments in other cell types or other organisms, the regulatory event can be confirmed either to be common with more than one cell type or even evolutionarily conserved. The latter often means that the functional relationship is significant for the lineage of organisms where it is conserved (e.g. primates, mammals, vertebrates).

Reconciling results from different experimental data types is not straightforward. Experiments are analysed by various platforms, results are provided in different formats, and genes are described with a variety of identifier types. This leads to identifier mapping problems where namespaces do not have one-to-one matches, platform identifier intersections are far from perfect and, thus, data loss is inevitable.

The most widely used microarray producers, Illumina and Affymetrix, provide mapping files where the identifiers are mapped to commonly known gene identifiers, such as Ensembl IDs. This allows us to match gene names with different platform or database specific identifiers and, thus, combine the datasets. There are several tools such as g:Convert (Reimand *et al.*, 2007), Syngerizer (Berriz & Roth, 2008) or BridgeDb (van Iersel *et al.*, 2010) that provide such mapping services.

CHAPTER 2

AIMS OF THE PRESENT STUDY

The aim of the present thesis is to apply different methodologies and data types for studying relationships between genes with systematic approaches. The specific aims of the thesis are the following:

1. To study the power of gene expression usage for biological pathway reconstruction;
2. To develop a methodology for automatic highlighting of significant gene clusters based on their co-expression and functional similarities;
3. To identify new regulatory events of adipocyte differentiation in mouse by combining different data sources;
4. To identify alternative regulatory modules of human embryonic stem cells;
5. To prioritise genes responsible for pluripotency using qualitative modelling;

CHAPTER 3

RESULTS AND DISCUSSION

3.1 Power of gene expression similarities for studying pathway relationships (Ref. I)

One of the challenges in systems biology is to understand how previously published data can be exploited for knowledge expansion. This involves understanding the possibilities of different data types and might lead to detecting limitations as well. Large quantities of mRNA expression measurements across a variety of conditions and tissues have been and continue to be produced and gathered into databases like ArrayExpress (Parkinson *et al.*, 2007) or GEO (Barrett & Edgar, 2006).

Obtaining pathway relationships from gene expression data

These large-scale publicly available gene expression data collections can be reused for various bioinformatic experiments to confirm or extract new knowledge. While there are plenty of experimental datasets available, the majority of human genes are still not annotated in any of the major pathway databases. At the time of this study (Adler *et al.*, 2009) there were 4220 proteins described in KEGG database (Kanehisa & Goto, 2000) and 1804 annotated in Reactome database (version 23, 28.01.2008)(Västrik *et al.*, 2007). Since the publication, the numbers have grown to 6869 proteins annotated in KEGG database (16.02.2015) (Kanehisa *et al.*, 2014) and 7757 in the Reactome database (version 51, 20.02.2015) (Croft *et al.*, 2014). Although the numbers have increased tremendously, the majority of human proteins are still not described in any pathway. Therefore, the question of how to exploit all the public data to expand our knowledge about the functions and interactions among the proteins remains substantial.

In this paper we studied the predictive power of using only gene expression data to identify members of known biological pathways. We wanted to understand if genes belonging to a pathway are co-expressed and to what extent. In addition, we aimed to find out if there are subsets of genes that have more similar expression profiles and could, therefore, be co-regulated and used for further candidate gene searches.

We used 35 curated biological pathways from the Reactome database version 23 (Västrik *et al.*, 2007). These pathways could be further divided to signalling pathways, metabolic processes and large protein complexes. The gene expression dataset that we used was kindly provided to us by Margus Lukk from the European Bioinformatics Institute in pre-publication state (Lukk *et al.*, 2010). This expression compendium covers 6108 manually curated and quality controlled samples across 369 different human cell and tissue types, cell lines and disease states all measured on the AffyMetrix HG_U133A microarray platform. The expression data was already normalised using the RMA algorithm (Irizarry *et al.*, 2003) and covered 22283 probe sets that converted to 12579 Ensembl gene identifiers at the time of the analysis.

Limitations of gene expression data

Our first aim was to study the predictive power of gene co-expression to derive members of known pathways had we not known them in advance. For this, we performed a leave-one-out computational experiment where we excluded one gene at a time from the pathway under study. For the remaining pathway members we calculated correlations with all the remaining probe sets available in the large data compendium. The resulting correlation values were converted to ranks with the highest correlation being ranked 1 and the lowest correlation being the total number of correlations calculated. In this study we used two main distance measures for calculating the correlation scores. First, we started with the Pearson correlation coefficient that is widely used for gene expression similarity analysis as it represents the relationship between two numerical vectors (in our case expression values for two probe sets at a time). The Pearson correlation coefficient shows the relatedness and strength of such a relationship. After initial analysis we also performed the same steps with the Absolute Pearson correlation coefficient that takes into account the absolute correlation and not its direction. The latter is useful when expecting to see correlations between genes that have opposite effects, e.g. suppressor-target gene pairs.

Many genes present on the HG_U133A platform have more than one probe set. In this analysis we always picked the most favourable probe set for each gene to maximise the correlation score for any given gene pair.

After calculating all the individual correlation scores between pathway genes and other genes on the platform, we converted these to rank values. Finally, we calculated the average rank across the entire pathway. If genes among the pathway were strongly co-expressed, we would get a small rank for the genes removed during the leave-one-out experiments. However, this was not true for the majority of the pathways under study. The highest proportion of pathway genes, 6%, that were ranked as first in the leave-one-out experiments were from the *Translation* pathway.

When looking at the top ten genes, the results are improved with up to 43% of pathway genes having rank 10 or less when using the Pearson correlation coefficient. The top 10 genes represent less than 0.1% of all the possible genes on the platform, therefore, the results can be considered significant.

Using only a subset of pathway genes

Overall, the results were not satisfactory when using all the pathway genes. This can be due to the fact that biological pathways describe complicated biological processes that are carried out by many genes and are inter-related over a period of time. Therefore, all the pathway members do not need to be expressed simultaneously but, rather, in a more complicated and orchestrated manner. Also, the pathway can contain subsystems or protein complexes that carry out specific functions that cannot be detected by mRNA expression alone. To benefit from this, we decided to look for an optimal subset of pathway genes that would produce the smallest rank possible for the pathway gene left out. In this analysis we did not take into account any topological information about the pathways as we wanted to test the opportunities and limitations of using only gene expression values for pathway reconstruction.

We performed an exhaustive search using the Pearson correlation coefficient and the Absolute Pearson correlation coefficient, and converted correlation values to distance that varies from 0 to 1. Our aim was to find the optimal threshold that would produce the highest average rank for each of the leave-one-out genes for a given pathway. We considered thresholds from the set of values that varies from 0 to 1 with a step of 0.02.

This approach resulted in lower ranks for all the pathways as seen in Table 1 and Figure 1 in Paper I. For example, for the *Translation* pathway, the results improve from 6% to 28% of pathway genes ranked 1st and from 43% to 53% of pathway left out genes ranked in the top 10. Similar improvements were seen when we used the Absolute Pearson correlation coefficient as a similarity measure. The most promising results were achieved for *Translation*, *Electron transport chain* and *Pyruvate metabolism and TCA cycle* pathways, while *Signaling by NGF* had the worst outcomes in the leave-one-out experiments.

This can be explained due to the nature of genes that form the given pathways. For example, the *Translation* pathway consists mostly of ribosomal genes that need to be expressed simultaneously to form the large protein complex. At the same time signalling pathways, like *Signalling by NGF*, probably have low ranks for the known pathway genes due to the fact signalling pathways do not function through large protein complexes or transcriptional control but have a more dynamic regulation where many factors are still to be discovered. Such pathways often utilise post-translational modifications and short-living ligand-receptor bindings that carry out the pathway functions (Fortini, 2009).

Potential new candidate genes

There were many genes from the platform that did not belong to a given pathway but produced smaller similarity ranks to the pathway than the known left out pathway genes. We considered as potentially interesting candidates all genes that had an average rank up to 100 when using a subset of pathway genes and the Absolute Pearson correlation coefficient. Using Gene Ontology (Ashburner *et al.*, 2000) information extracted from g:Profiler (Reimand *et al.*, 2007) and protein-protein interaction data from Intact (Kerrien *et al.*, 2007), HPRD (Peri *et al.*, 2003) and ID_SERVE (Ramani *et al.*, 2005) through GraphWeb software (Reimand *et al.*, 2008), we estimated how many of the candidate genes having small ranks could actually be linked to the pathways.

For the majority of 35 pathways we had more than 100 genes (Table 2 in Paper I) that could be either pathway members or linked to the pathway based on their high co-expression to characterised pathway members. For some pathways, like *Cell cycle checkpoints* and *Cell cycle, mitotic*, more than half of the candidate genes actually had known protein-protein interactions with the pathway members. For some other pathways, like the *Electron transport chain*, the known protein-protein interactions were limited or even missing. The full results are given in Table 2 in Paper I.

Already during the article publication process we learned that, for example, the TNFRSF10C candidate gene that had been shown to have protein-protein interactions with four *Apoptosis pathway* genes, was annotated in the next Reactome version to the *Apoptosis pathway*.

Summary

The main outcome of our work is the detected limitation of using only gene expression data to describe pathway membership in different biological pathway types. We also show that it makes more sense to use only a subset of strongly correlated pathway members for such tasks. Many of the candidate genes we found can be related to the pathways they were proposed from. Our results can also be used as a baseline for benchmarking other, more complicated pathway reconstruction methods.

3.2 Compact visualisation of microarray data based on functional enrichments (Ref. II)

In the previous section we addressed the importance and widespread distribution of microarray experiments available to the public. These datasets can be used in a combined manner to perform systematic analysis with bioinformatics methods to find functional connections between genes. However, it is also important to study each individual dataset separately. This approach is needed for experimentalists

to interpret the results, make relevant conclusions and compare the outcomes of similarly executed experiments.

Current microarray platforms cover tens of thousands of probe sets at a time. Even with minimalistic visualisation approaches this leads to high dimensional heat maps or dendrogram images that are hard to fit onto a screen and grasp by the human eye. The large images covering all the information in the data might also present data points that are not relevant for the current experimental set-up and, therefore, exploit much needed space in the output.

Clustering expression data

Experimentalists use microarrays mostly to find genes with similar expression profiles across the studied conditions. Genes that behave similarly are often co-regulated and take part in common processes. This forms the basis of *guilt by association* reasoning (Wolfe, Kohane, & Butte, 2005). Therefore, clustering genes based on their expression patterns can help in creating hypothesis about the potential molecular function or biological process that a gene without a known function is carrying out.

The most common approach to get an overview of a large-scale expression experiment is to perform full hierarchical clustering and then dissect the large dendrogram into smaller clusters. This is usually done either by using some fixed distance measure to cut the tree or by predefining the number of expected clusters when using the K-means approach. Both of these strategies require user provided input that is difficult to know beforehand as each dataset might have unique cluster forming parameters that produce the best results.

Once the clusters are identified, each of the clusters is characterised using Gene Ontology terms (Ashburner *et al.*, 2000), common pathways or over-represented regulatory motifs. When there are hundreds or thousands of clusters with potential impact, this step can be exhausting for a human to analyse.

Automatically finding functionally related gene sets

To overcome the problems stated above, we came up with a web based tool that helps to find biologically significant clusters in the microarray data. In order to do this, we combine fast approximate hierarchical clustering, functional enrichment analysis and the smart cluster cutting algorithm to create a compact heat map and a dendrogram.

In large microarray datasets there are close to 50 000 probe sets that would lead to almost 1 250 000 000 values representing all against all distances. Even with currently available computational resources, this leads to a vast number of computations, especially if computed for many datasets on the fly at a web server. With the help of the HappieClust algorithm (Kull & Vilo, 2008) we are able to reduce the computation times an order of magnitude compared the standard approaches. This is achieved by first computing a subset of pairwise distances and then computing all pairwise distances between similarly expressed genes, as well

as distances to a subset of more distant pivot genes chosen by random. Based on this, HappieClust is able to approximate the full hierarchical clustering that has been shown to produce biologically comparable results to standard agglomerative hierarchical clustering (Kull & Vilo, 2008). These hierarchical clustering results are then functionally characterised. For all clusters with size ranging from 5 to 1000 genes, we find all statistically significant Gene Ontology (Ashburner *et al.*, 2000) terms, Reactome (Västrik *et al.*, 2007) and KEGG (Kanehisa & Goto, 2000) pathways and regulative motifs from Transfac (Matys *et al.*, 2006) and miRBase (Griffiths-Jones *et al.*, 2006) using g:Gost (Reimand *et al.*, 2007). The overrepresentation significance is calculated using a hypergeometric test that is used to assess if a function is enriched among the sample and is represented by:

$$P_\alpha = \sum_{k}^{min(n,K)} \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

where for a functional annotation α, the probability P_α is calculated based on k genes from a set of n genes having annotation α when out of the N genes in the genome, K genes have annotation α. These probabilities are then corrected for multiple testing.

The analysed clusters are hierarchically related while differing in size, therefore we calculate the cluster size weighted score for each cluster. We normalise the summarised individual hypergeometric p-values with the cluster size n and score each cluster with score q:

$$q = \frac{1}{n}\sum_{\alpha}(-log_e(p_\alpha)) \ .$$

Compact visualisation of functional clusters

The enrichment analysis helps to identify clusters that are composed of genes that share functions or could be regulated in similar a manner. Our aim is to highlight clusters that are most significantly annotated in the full dendrogram.

We achieve this by first looking for clusters that have a high annotation score q. We start from the highest scoring cluster and consider that as an *enriched cluster*. At the same time we exclude all of its child and parent clusters from further analysis and do not show these in the output.

Second, in order to minimise the output, we look for clusters that have annotation scores below the significance threshold or inappropriate size. Starting from the root of the dendrogram we pass it recursively. All the clusters that do not have an *enriched cluster* as a child cluster are compressed and hidden from the final output.

By carrying out these two steps we achieve the compression of the original heat map of approximately 50 000 pixels in height if a one pixel row is used per gene to few hundreds of clusters on one image with less than 5000 pixels in

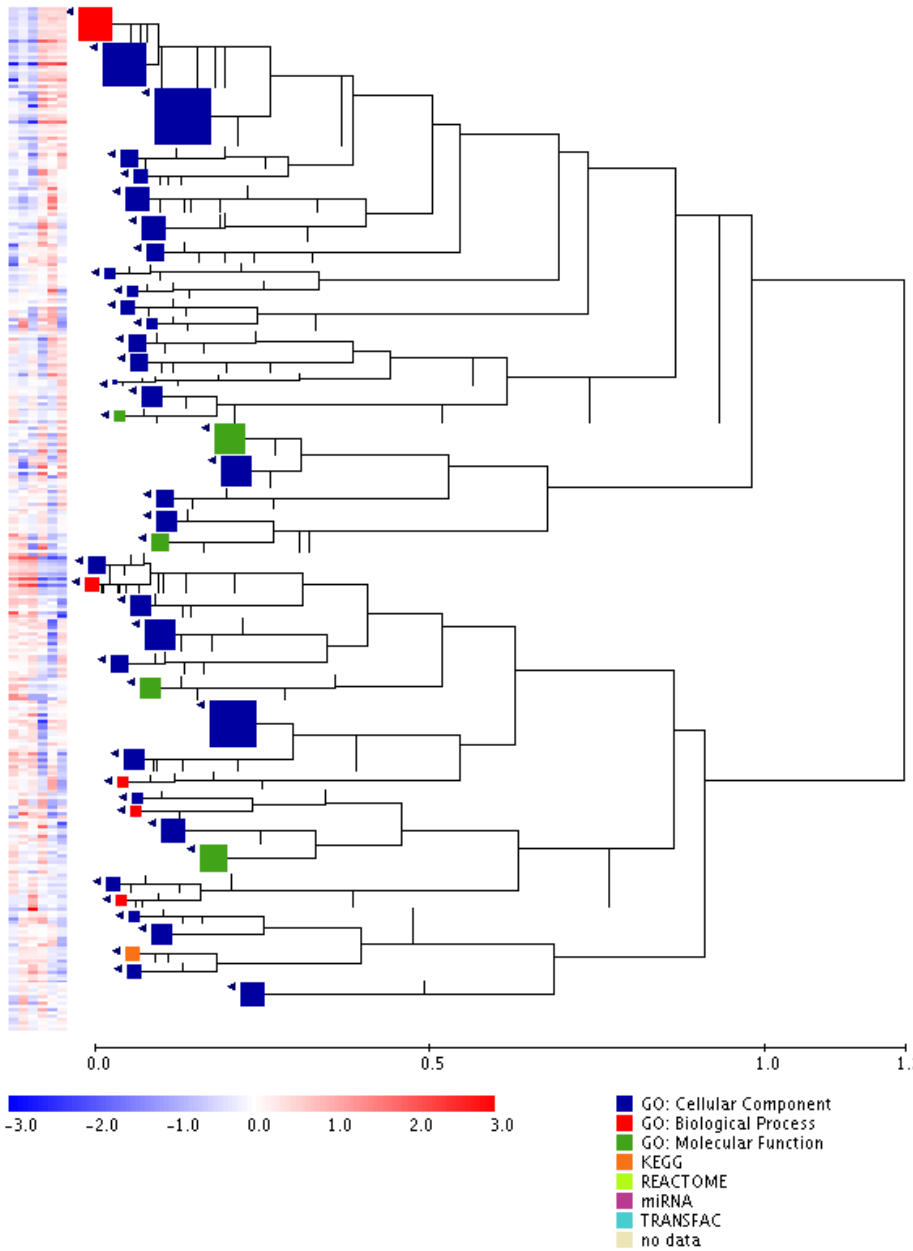


Figure 10: Overview of VisHiC output based on microarray experiment GSE2248 (mesenchymal precursor cells derived from embryonic stem cells) (Barberi *et al.*, 2005) from GEO database (Edgar, Domrachev, & Lash, 2002).

height with the same row parameters. The size of the latter images can be fit to the screen unlike the image that shows all the expression profiles of individual genes. In order to make the final output carry as much information as possible, we combine the heat map that depicts gene activation and repression profiles together with the dendrogram that features the functional enrichments of clusters (Figure 10). We use colour-coded squares to signify *enriched clusters*. The size of the square illustrates the size of the cluster. The colour of the square represents the functional domain that had the lowest p-value for that specific cluster, e.g. red denotes the *biological process* domain from GO.

Cluster id: 42762
Cluster score: 1.79E1
Size: 30
Contains 2 unique annotations.
Domains Top Annotations

Domain	Name	Description	p-value
CC	**GO:0044428**	**nuclear part**	**1.81E-23**
BP	GO:0006139	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	3.58E-23
MF	GO:0003723	RNA binding	1.09E-20
KE	KEGG:04110	Cell cycle	4.72E-5
RE	REAC:72107	Formation of the Spliceosomal E complex	1.02E-8
MI	MI:hsa-miR-489	MI:hsa-miR-489	2.06E-5
TF	TF:M00428_4	NKTSSCGC:4	1.01E-8

Figure 11: Enriched annotation groups of VisHiC cluster 42762 based on microarray experiment GSE2248 (mesenchymal precursor cells derived from embryonic stem cells) (Barberi *et al.*, 2005) from GEO database (Edgar, Domrachev, & Lash, 2002).

For each of the clusters we provide best annotations for every functional domain, mark functional annotations that are uniquely related to the given cluster and provide hyperlinks for further detailed analysis in g:Profiler (Figure 11).

Summary

VisHiC is a web tool that allows to create an easily interpretable overview of expression datasets. It highlights clusters of genes that not only are expressed in the same manner but also share functional annotations. The complicated decisions of choosing at what distance to cut an expression dendrogram or how many clusters to look for with the K-means algorithm, are removed from the user side. Automatic functional enrichment analysis and the following cluster filtering saves time in getting an initial overview of affected processes and helps to find gene sets to concentrate on in the downstream investigation. VisHiC can be the second step of every mRNA level study using microarray or next generation sequencing after quality control and before the differential expression analysis steps. With VisHiC one can identify sets of genes that otherwise could go unnoticed as they are not part of the initial hypothesis.

3.3 Integration of data layers for revealing new regulatory processes (Ref. III, IV)

Biological objects, especially multicellular organisms can be viewed as complex molecular machineries carrying out complicated processes. The aim of experimental biologists together with bioinformaticians is to understand the processes by measuring the system in every possible way. This leads to a large variety of experiments for the biological system under study and, thus, creates a challenge of how to combine all the results in a meaningful way. Usually there are major experimental data types collected by a laboratory, e.g. microarrays for measuring mRNA abundance, or chromatin immunoprecipitation followed by chip or sequencing for identification of protein binding locations. Initially, the primary data is analysed individually, but often this does not lead to major breakthroughs. Instead, new questions arise that the data in hand is not able to answer. Combining the results with other complementary experiments can reveal additional findings.

3.3.1 Adipocyte differentiation regulation

With the progress of studying a variety of different organisms, from simple bacteria to more complex mammals or plants, it is possible to carry over knowledge from one organism to another. This promotes more complicated studies in the model organism with the goal of transferring knowledge to the human. Also, the more organisms are sequenced and studied, the greater our understanding of the evolution and emerging functionalities becomes. General understanding is that functional DNA regions evolve more slowly and fewer mutations are tolerated compared to "junk" DNA. These functionally active regions, such as protein coding genes or to some extent transcription factor binding sites, are highly conserved across species. Comparison of rodent-human sequences has been used to identify regulatory regions in humans (Wasserman *et al.*, 2000). In our work we use DNA conservation as an additional layer of evidence when identifying functional transcription factor binding sites in the mouse genome.

Adipocyte regulation

In this study our goal was to understand the gene regulation mechanisms active during adipocyte development from mouse embryonic stem cells. Adipocytes are cells derived from mesenchymal stem cells (MSCs) that form adipose tissue. Although the general two-step process of adipogenesis, preadipocyte formation from MSCs and terminal differentiation into mature adipocytes is known, the exact regulatory events of early events of adipocyte development are not fully clear yet.

In this study we contributed to data analysis of adipocyte differentiation. Retinoic acid receptor β agonist CD2314 was used to stimulate mESCs differentiation towards adipocytes. In parallel adipocyte differentiation was repressed through addition of glycogen synthase kinase 3 (GSK3) inhibitor BIO. For comparison,

combination of the agonist and inhibitor and control media were used. RNA was extracted on day 3 before any compounds were added; just after stimulation on day 6; and for effects on later differentiation stages on day 11.

Clustering genes by expression profiles

RNA levels were measured with Affymetrix GeneChip Mouse Genome 430 2.0 microarrays (available as E-TABM-668 in ArrayExpress (Rustici *et al.*, 2013)). Differentially expressed genes were identified using the Limma package from Bioconductor (Smyth, 2004). Genes were clustered by their activation timings. Five main clusters were identified:

- Cluster 1 with up-regulated genes by CD2314 on day 6;
- Cluster 2 with down-regulated genes by CD2314 on day 6;
- Cluster 3 with up-regulated genes by CD2314 on day 11;
- Cluster 4 with down-regulated genes by CD2314 on day 11;
- Cluster 5 with up-regulated genes by CD2314 on both day 6 and day 11.

Initial microarray results were confirmed by measuring gene expression using quantitative real-time PCR. Out of the 30 tested genes 29 showed comparable expression patterns to microarrays. Clusters were further characterised using g:Profiler (Reimand *et al.*, 2007) and annotated with known protein-protein interactions from the Human Protein Reference Database (Peri *et al.*, 2003).

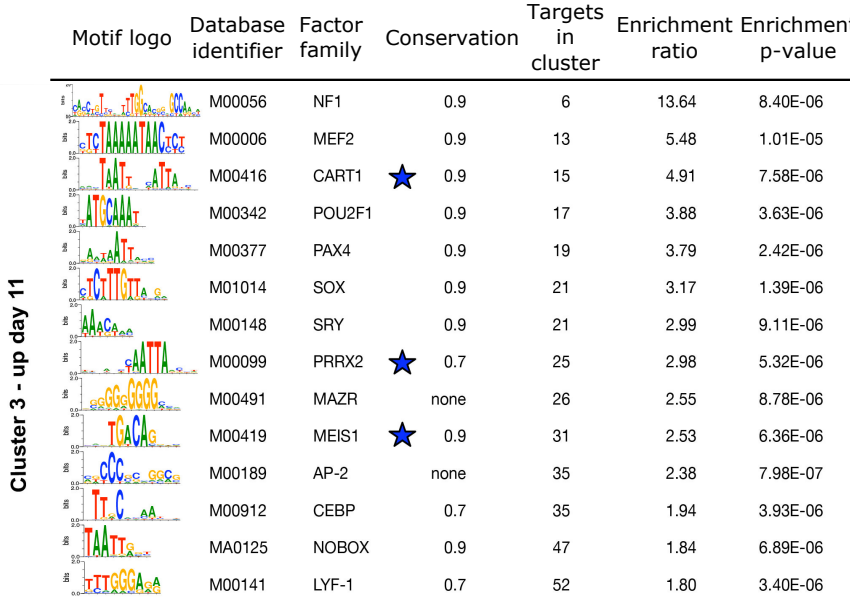
In Cluster 1 we found that early adipocyte differentiation is related to blood vessel formation in mESCs; in both Cluster 1 and 5 we saw enrichment of neural development related genes in early adipocyte development; in Cluster 1 and 3 we observed WNT pathway genes being up-regulated during adipocyte development (Figure 2 in paper III).

Significant regulatory motifs

In addition to identifying differentially expressed transcription factors, we looked for overrepresented regulatory motifs of known transcription factors. We extracted 3000 bp long promoter regions for every gene using UCSC genome database (mm8 release) (Karolchik *et al.*, 2008). These regions consisted of 1000 bp downstream of the transcription start site (TSS) and 2000 bp upstream of the TSS.

Promoter regions were scanned with position weight matrices describing regulatory motifs of transcription factors. Matrices were obtained from TRANSFAC (version 11.4) (Matys *et al.*, 2006) and JASPAR (Sandelin *et al.*, 2004) databases and matched with STORM software (Smith *et al.*, 2006).

Many of the known adipocyte development regulators from homeo box (HOX), forkhead box (FOX family, e.g FOXC2 (Gerin *et al.*, 2009)) and nuclear receptor (Nr family e.g. NR2F2 (Xu *et al.*, 2008)) gene families were found. We also saw down-regulation of known adipogenesis inhibitors, such as MSX2. Transcription factors expressed during adipocyte development but without a known relation to adipogenesis were characterised using literature. Additionally, their expression



Motif logo	Database identifier	Factor family		Conservation	Targets in cluster	Enrichment ratio	Enrichment p-value
	M00056	NF1		0.9	6	13.64	8.40E-06
	M00006	MEF2		0.9	13	5.48	1.01E-05
	M00416	CART1	★	0.9	15	4.91	7.58E-06
	M00342	POU2F1		0.9	17	3.88	3.63E-06
	M00377	PAX4		0.9	19	3.79	2.42E-06
	M01014	SOX		0.9	21	3.17	1.39E-06
	M00148	SRY		0.9	21	2.99	9.11E-06
	M00099	PRRX2	★	0.7	25	2.98	5.32E-06
	M00491	MAZR		none	26	2.55	8.78E-06
	M00419	MEIS1	★	0.9	31	2.53	6.36E-06
	M00189	AP-2		none	35	2.38	7.98E-07
	M00912	CEBP		0.7	35	1.94	3.93E-06
	MA0125	NOBOX		0.9	47	1.84	6.89E-06
	M00141	LYF-1		0.7	52	1.80	3.40E-06

Figure 12: Statistically over-represented TFBSs in CD2314-regulated gene promoters on day 11. Differentially expressed TFs during mESC adipogenesis are marked with stars. Adapted from Figure 7 in Ref. III.

data were explored for mesenchymal differentiation patterns. Majority of these transcription factors were found to be expressed during mouse embryogenesis in mesenchymal areas or/and in the adipose progenitor cell compartments of mouse adipose tissue.

It is known that some of the adipocyte regulators, e.g. C/EBP are conserved (Rosen *et al.*, 2000) on the DNA level across species. Therefore, we incorporated DNA conservation data as an additional layer of information to understand the adipocyte regulation. We downloaded the euarchontoglires DNA conservation track from UCSC and used it to filter motif matches. Every transcription factor binding site match was evaluated using a conservation threshold from the set of thresholds that varies from 0.7, to 1.0 with a step of 0.1. The value represents the strength of conservation where 1.0 denotes perfect conservation across the included species. For some of the promoters, none of the thresholds led to significant results and, therefore, we also performed an analysis using no conservation data. In this case we took the average of the best three hits in the promoter, while in the initial analysis we only took the highest scoring hit per promoter.

We detected 16 significantly enriched motifs in Cluster 1 and 14 motifs in Cluster 3 (Figure 12). Some of the transcription factors related to the enriched motifs had been shown before to be active in adipocyte differentiation (e.g. Leukemia-related factor (LRF) (Laudes *et al.*, 2004) or Early growth response protein 2 (EGR2) (Chen *et al.*, 2005)). Additionally, we found enriched motifs related to the Activator protein 2 transcription factor family known to be expressed highly in

neural crest cells (Mitchell *et al.*, 1991). It has been previously shown that neural crest cells can differentiate into adipocytes (Billon *et al.*, 2007), therefore, we hypothesised that AP-2 could regulate adipocyte generation through the neural crest developmental pathway.

Finally, we combined expression profiles, protein-protein interaction data and transcription factor binding site enrichment analysis. We noticed that three of the enriched motifs on day 11 after CD2314 stimulation belong to transcription factors up-regulated on day 6 just after the compound treatment. These are pair related homeobox 2 (PRRX2), myeloid ecotropical viral integration site 1 (MEIS1) and cartilage homeo protein 1 (CART1) transcription factors (marked with $\star$ on Figure 12). We hypothesised that MEIS1 could be regulating adipocyte commitment through protein complex with PBX1 and HOXB4 proteins. After Paper III was published our co-authors showed that PBX1 is an adipocyte development regulator that promotes adipocyte generation from neural crest cells during embryogenesis (Monteiro *et al.*, 2011).

Summary

We studied adipocyte differentiation by treating mouse embryonic stem cells with differentiation promoting compound CD2314 and inhibitor BIO. We clustered genes by their expression patterns. For each cluster we searched overrepresented and evolutionarily conserved regulatory motifs. Motifs found in Cluster 3, CD2314 treatment measured on day 11, were matched to transcription factors being differentially expressed already on day 6. Based on a combination of complementary data sources we proposed that three transcription factors (MEIS1, CART1, PRRX2) could be new potential regulators of early adipocyte differentiation from mouse mesenchymal stem cells.

3.3.2 Alternative regulation of human embryonic stem cells

Embryonic stem cells are pluripotent cells that are capable of differentiating into any cell type in the organism. Human ESCs are kept in the pluripotent state by expressing core regulators SOX2, NANOG and OCT4 (Boyer *et al.*, 2005). Although the core genes are known, the full process of keeping the pluripotency state and even more importantly, the mechanisms controlling early differentiation steps are still unclear. With this study our aim was to investigate the downstream targets of OCT4.

Finding peaks in ChIP-enriched regions

We performed in the embryonic carcinoma cell line NCCIT, OCT4 chromatin immunoprecipitation followed by hybridisation on NimbleGen two-array set of human promoter tiling arrays. Three biological replicates of OCT4 binding to the DNA were analysed. After quality control and quantile normalisation we had to exclude one array out of six due to uneven dye distribution in the chip. This left

us with 3 replicates for array 1, and 2 replicates for array 2.

We applied three independent peak finding algorithms to the normalised data MA2C (Song *et al.*, 2007), TAMALPAIS (Bieda *et al.*, 2006), brute-force algorithm developed in-house at Max Planck Institute (Chavez *et al.*, 2009). A large number of peaks were detected only by one algorithm out of three. This could be due to the set-up of the algorithms, e.g. TAMALPAIS assumes less than 5% of the probes are actual binding sites. It has also been shown previously that there is a larger variation in results between the algorithms applied on the same array than same algorithm applied on different array platforms (Johnson *et al.*, 2008).

Results obtained from each of the three algorithms were considered equal. Regions identified by one algorithm in three replicates creates an equal score to a peak identified by three separate programs. If a region was identified in all the replicates by all three programs then we gave it the peak score 1.0. If only one program identified a peak in two replicates out of three then the score was 0.222. With this approach we found 497 genes with peak scores of at least 0.5.

For experimental validation, we randomly picked 4 peaks that were identified by all three programs, and 2 peaks identified uniquely by each of the programs, 10 regions in total. All ten peak regions were successfully validated by real-time PCR (Figure 2D in Paper IV).

When we compared the identified target genes with similar experiments done with another embryonic carcinoma cell line NTERA2 (Goto *et al.*, 1999) and embryonic stem cell line H9 (Boyer *et al.*, 2005), we found an overlap of 31 genes (including core regulators OCT4, SOX2 and NANOG). When comparing with only the H9 data, the overlap was 46 genes.

Regulatory modules of OCT4 target genes

OCT4 is known to bind to the ATGCAAAT consensus sequence (Chew *et al.*, 2005; Loh *et al.*, 2006). In all peak regions we searched for motifs with Levenshtein distance up to two from the consensus sequence. We aligned the matching sequences and created a new position weight matrix out of this set of sequences (Figure 4C in Paper IV).

We defined a PWM threshold score for downstream analysis by taking the median motif score from peaks with score 0.5 or above. The resulting PWM matching threshold 7.3 was used for downstream motif analysis. OCT4 is known to form a heterodimer with SOX2 (Remenyi *et al.*, 2003; Williams, Cai, & Clore, 2004). Hence, we did the same analysis with the SOX2 transcription factor and the SOX2 consensus sequence CATTGTT (Chambers & Smith, 2004; Pesce & Scholer, 2001).

We identified 372 target genes that had both the OCT4 and SOX2 motif with a score above the threshold. We got an overlap of 293 genes when comparing to 332 OCT4 and SOX2 target genes from (Boyer *et al.*, 2005).

We identified 6 distinct potential modules of OCT4 transcriptional regulation by its binding to DNA (Figure 13). The first module covers 39 genes that have

both SOX2 and OCT4 motifs in the peak regions. From this group we validated OCT4 binding in the NANOG promoter with a band shift assay. The second module consists of 122 genes that had OCT4 motif but no SOX2 motif in the peak regions. We validated OCT4 binding to an evolutionary conserved region in USP44 promoter. The third module covers 65 promoters with only the SOX2 motif and no strong OCT4 motif.

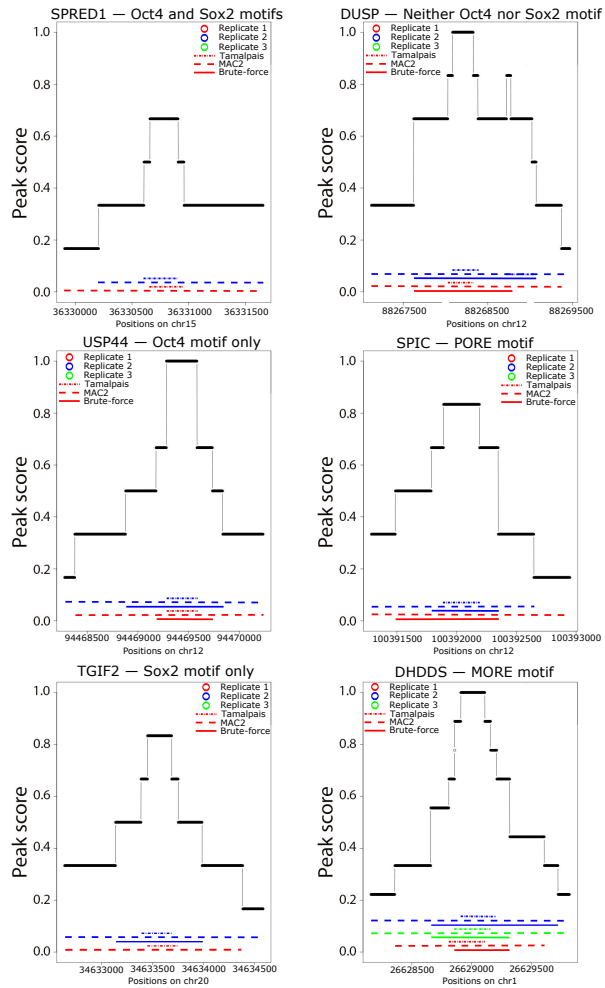


Figure 13: Six distinct OCT4 transcriptional regulation modules. DNA positions illustrated in *x*-axis. *y*-axis represents peak score defined by overlap of individual peak regions. Peaks were identified by TAMALPAIS (continuous line), MAC2 (dashed line) and the brute force algorithm (dotted line). Colours of the line represent replicates (red – replicate 1, blue – replicate 2, green – replicate 3).

For more than half of the peak regions we were not able to find neither the OCT4 nor the SOX2 motif. These promoters of 271 genes might be targeted by OCT4 binding together with unknown protein-protein interaction partners that bind to DNA; OCT4 might bind in some unknown configuration to DNA and thus recognizing motif that we did not know to look for; our motif matching threshold was too stringent to find all the true binding occurrences.

For these promoter regions we applied *de novo* motif discovery algorithms and looked for overrepresented motifs. Promoter regions from the defined peaks were analysed with the TAMO package (Gordon *et al.*, 2005) that combines MEME (Bailey & Elkan, 1994), MDScan (Liu, Brutlag, & Liu, 2002) and AlignACE (Hughes *et al.*, 2000) methods. The identified motifs were then matched against the known transcription factor binding site data in the TRANSFAC (version 11.3) (Matys *et al.*, 2006) and JASPAR (version 3) (Sandelin *et al.*, 2004) databases using the STAMP tool (Mahony & Benos, 2007). In total we found 4 distinct motifs that could be related to 8 potential transcription factors (Figure 10 in Paper IV).

Previously it has also been shown that OCT4 is able to co-operatively bind to the Palindromic Oct factor Recognition Element (PORE) ATTTGAAATG-CAAAT by two OCT4 molecules (Botquin *et al.*, 1998) and in a similar manner to the MORE element (more PORE, ATGCATATGCAT) (Tomilin *et al.*, 2000). We were able to identify 4 genes with PORE and 10 genes with MORE element in their peak regions.

We were also aware that our experiments might miss true positive targets of OCT4. So we looked for genes that had a positive direct regulation in other experiments, such as in (Boyer *et al.*, 2005), but no significant binding in our data. Out of these genes the most promising gene was GADD45G that, based on the OCT4 motif in its promoter, would belong to module 2. We could confirm the binding of OCT4 to the predicted motif by a band shift assay and ChIP-real-time-PCR. GADD45G activity increases during early differentiation from ESCs, thus, indicating that GADD45G could be involved in inducing loss of self-renewal. Also our microarray results with over-expression of GADD45G for two days suggest that increased levels of GADD45G lead to up-expression of genes belonging to cell cycle and differentiation processes (e.g. differentiation markers EOMES, BMP4, MSX1), while not significantly affecting the core regulatory complex of OCT4, SOX2 and NANOG. Since the publication of Paper IV it has been shown that GADD45G is required for early embryonic cells to exit pluripotency and enter differentiation in *Xenopus* (Kaufmann & Niehrs, 2011) and its expression across species is conserved from *Xenopus* to medaka fish to mouse (Kaufmann, Gierl, & Niehrs, 2011).

Creating the Embryonic Stem Cells Database

Embryonic stem cells came into focus in 1998 when Thomson and colleagues were able to isolate human ESCs from blastocysts (Thomson *et al.*, 1998). Since

then many research groups around the world have investigated these cells in a variety of ways. This has led to similar and complementary experiments produced in different labs. Data richness allows to study similarities and differences across the experiments, makes it possible to pinpoint the uniqueness of cell lines or organisms; identify core targets of the regulators identified in different organisms using a variety of chip types or microarray producers. In order to benefit from this variety of data we need to collect it in one place and make it comparable.

We collected a number of experiments performed on the most widely used embryonic stem cell lines for human and mouse. In addition to embryonic stem cell lines, we also included embryonic carcinoma cell lines that are easier to maintain but are biologically similar to the embryonic stem cells (Andrews *et al.*, 2005; Damjanov, Horvat, & Gibas, 1993). We searched for already published articles studying the regulation of embryonic stem cells and looked for high-throughput experiments in the supplementary files or in uploaded datasets in the public domain. We were able to collect 31 datasets from 13 different publications. For mouse we found 9 perturbation datasets (Ivanova *et al.*, 2006; Loh *et al.*, 2006; Masui *et al.*, 2007; Sharov *et al.*, 2008; Walker *et al.*, 2007) and 18 TF-DNA interaction datasets (Chen *et al.*, 2008; Loh *et al.*, 2006; Mathur *et al.*, 2008). For human we were able to find 7 perturbation datasets (Babaie *et al.*, 2007; Greber, Lehrach, & Adjaye, 2007, 2008) in addition to the GADD45G experiment produced in this study (Jung *et al.*, 2010) and 5 public TF-DNA interaction datasets (Boyer *et al.*, 2005; Jin *et al.*, 2007; Lister *et al.*, 2009) in addition to the OCT4 immunoprecipitation experiment carried out in this study (Jung *et al.*, 2010).

All the original datasets were checked for missing data, converted to common format and imported to a MySQL database. The computational complexity to merge all these experiments to one database comes from the inclusion of two organisms and use of different experimental platforms leading to different identifier types and a varying number of present genes. In order to make the datasets comparable and enable searching across the datasets, we converted all the original identifiers to Ensembl gene IDs (Ensembl version 53) using the g:Convert tool (Reimand *et al.*, 2007). In cases when we were not able to convert all the probe IDs to Ensembl identifiers, only the original identifiers are kept in the database. We used a central identifier namespace across all the experiments so we could match different datasets and provide all the results in one table. Ensembl identifiers between mouse and human species are mapped using orthology data from the Ensembl database through the g:Convert tool from g:Profiler (Reimand *et al.*, 2007).

All perturbation datasets included in the database are microarray experiments. We used the already analysed form of the data provided by the authors. We applied the standard statistical significance threshold of 0.05 and considered genes to be differentially expressed only if they had log2 fold change above 1.5 between the control and perturbed conditions. In the output we illustrate gene behaviour with coloured table cells (Figure 14). Strong green and red colours describe significant fold changes while lighter tones show events that are significant but the

fold change value is below the threshold we set. When no information is present in a particular dataset for a given gene, we mark it with a gray cell. White cells represent no detectable changes in the datasets.

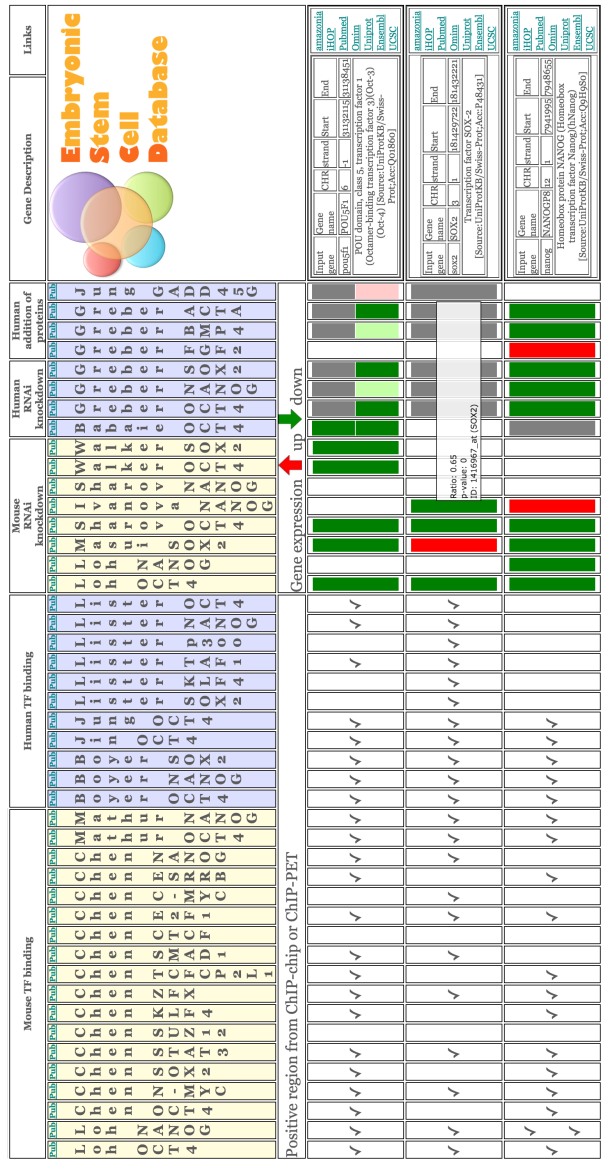


Figure 14: Screenshot of the Embryonic Stem Cell Database (March 2015). Datasets are described in columns, genes in rows. Datasets with yellow background are mouse specific, violet for human specific datasets. Each binding event from a ChIP experiment is marked with a ✓sign. Significant expression changes for perturbation experiments are expressed with colours (green – down-expression, red – up-expression).

We also included chromatin immunoprecipitation studies that were done using PET, array, or sequenced after pull-down. We did not re-analyse the data and used only the binding events defined by the original authors. For some of the experiments we only have binary information of the binding (for the majority of human ChIP experiments and (Loh *et al.*, 2006) for mouse), while for the others we also have a binding score provided in the data (e.g. (Chen *et al.*, 2008; Mathur *et al.*, 2008)). Binding events are illustrated in the web table with a ✓sign. In the datasets like (Loh *et al.*, 2006) where there can be more than one binding event per gene promoter, we mark the number of binding events with the corresponding number of ✓signs.

The database can be queried using a large variety of gene and protein identifiers thanks to the g:Convert tool that translates almost any gene identifier types to Ensembl gene IDs (Reimand *et al.*, 2007). We also support queries by Gene Ontology (Ashburner *et al.*, 2000) annotation identifiers. This allows the user to look for expression patterns and common regulators for genes that have been annotated into the same biological function, cellular component or are known to share molecular function.

We also provide an alternative view to the data through the *Gene filtering view* that enables to look for specific behaviour across datasets rather than query specific genes. It allows to set binding or expression direction filters to each individual dataset. For example, one can look for genes that are known to be down-regulated after OCT4 is knocked down in RNAi experiment but have not been detected in any of the OCT4 immunoprecipitation experiments. These hypothetically indirectly regulated genes are great candidates for finding intermediate transcription factors that mediate the signal from OCT4 to the downstream target genes. Such potential transcription factors can be picked from the other columns covering datasets that have no filters set.

Summary

In this work we analysed the OCT4 immunoprecipitation dataset in the context of previously published results. We identified transcription factor binding events using three distinct algorithms and treated them equally when defining the regions bound by OCT4. We further characterised the peak regions by matching OCT4 and SOX2 motifs. This led to the definition of six distinct modules of OCT4 target genes. For the module where neither OCT4 nor SOX2 motifs could be found, we proposed 4 alternative motifs and potential factors that could be regulating these genes. Finally, we gathered a compendium of publicly available data relevant for embryonic stem cell research. We created a freely available database that holds data from perturbation and immunoprecipitation experiments from mouse and human embryonic stem and carcinoma cell lines. The database can be queried in a gene oriented or a behaviour oriented manner, it promotes comparison of datasets and provides means for new hypotheses.

3.4 Identifying new key regulator supporting pluripotency in human embryonic stem cells using qualitative modelling (Ref. V)

The key regulators that keep embryonic stem cells in the pluripotent state are known (Boyer *et al.*, 2005). However, there still remain unanswered questions about the active mechanisms of gene regulation that maintain the pluripotent state. We also know little about the identity of factors capable of launching early differentiation.

Expanding the known core regulatory network

In order to study the human pluripotency regulatory network further, we first gathered public knowledge about the core regulatory network (Babaie *et al.*, 2007; Boyer *et al.*, 2005; Jung *et al.*, 2010; Matin *et al.*, 2004; Xu *et al.*, 2009). This led to a *literature network* consisting of 10 genes and 32 edges.

Next, we expanded the obtained network by incorporating 7 perturbation datasets (Babaie *et al.*, 2007; Greber, Lehrach, & Adjaye, 2007; Jung *et al.*, 2010) that we had collected earlier to the Embryonic Stem Cell Database (Jung *et al.*, 2010). The experiments we used were done using embryonic and embryonic carcinoma stem cell lines. The 7 perturbation experiments covered the three core regulators of hESCs – OCT4, SOX2 and NANOG; the critical growth factor added to the medium for keeping hESCs in the pluripotency state – FGF2; and factors responsible for early differentiation – BMP4, ACTA, GADD45G.

All the perturbation experiments were performed using microarrays. Fold change was calculated between the control and perturbed conditions. We included only genes that had at least 1.5 log2 fold change with p-value less than 1.0e-05 and detection p-value less than 1.0e-05. Each such gene was connected to its regulator with either a positive or a negative edge. This allowed us to expand the initial network of 10 nodes and 32 edges to a large dense hairball-like network consisting of 16395 edges and 7862 nodes.

Filtering the large regulatory network

The resulting network of such a large size complicates any kind of *in silico* modelling algorithms. Therefore, we had to apply different filtering approaches to reduce the network size. Our aim was to verify the genes showing the highest potential of being regulators of the pluripotent state. Thus, we concentrated on genes that would be the easiest to target experimentally. We chose to keep in the expanded network only genes that code proteins that are either located at the cell surface (receptors) or secreted into the media (ligands). We used Gene Ontology categories *cell surface receptor linked signaling pathway* (GO:0007166) and *receptor binding* (GO:0005102). After this filtering step we were left with 847 genes and 7862 edges in the network.

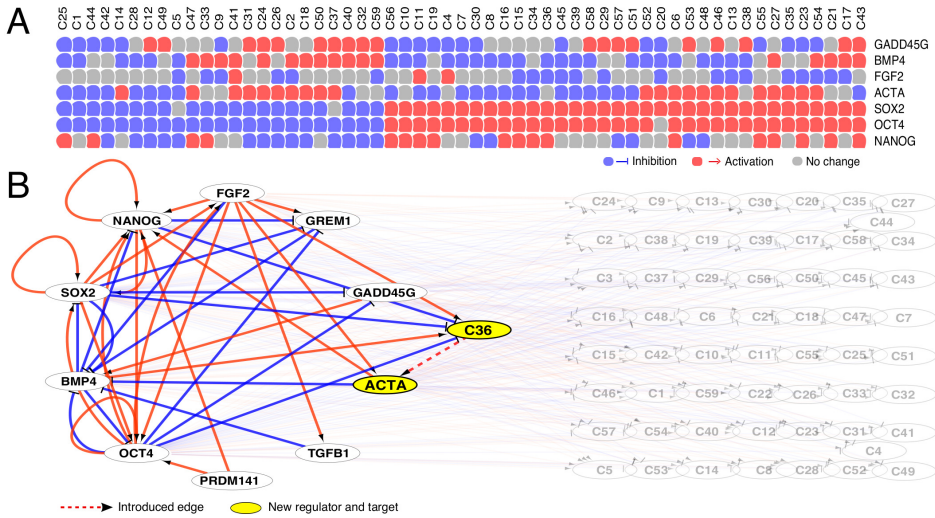


Figure 15: All genes that had an identical set of incoming edges were clustered together with all their edges described in the panel below (A). Finally, new regulatory edges were introduced creating feed-back loops from clusters to regulators (dashed red line edge from C36 to ACTA), thus producing individual regulatory models (B). Adapted from Figure 2 in Ref. V.

The added nodes derived from the perturbation experiments results can have 1 to 7 edges connecting them back to the core network. We reasoned that genes that are under more tight regulation, meaning they were perturbed in most of the experiments, could play a more important role in keeping the pluripotency state or initiating early differentiation. Thus, for further steps we kept only genes that were significantly up- or down-regulated in at least 5 out of the 7 perturbation experiments. This allowed us to remove 738 genes and 1471 edges from the network.

Using only a small number of datasets leads to a network where many of the individual genes in the network are regulated in an identical manner, and we cannot distinguish their individual effect in the given network. Therefore, to reduce the network size even more before starting the modelling, we clustered all the network nodes by their incoming edges, putting into a cluster genes that had identical regulatory edges (Figure 15). This allowed us to get the final network of 69 nodes and 347 edges. All the following *in silico* experiments were done using the network on (Figure 16).

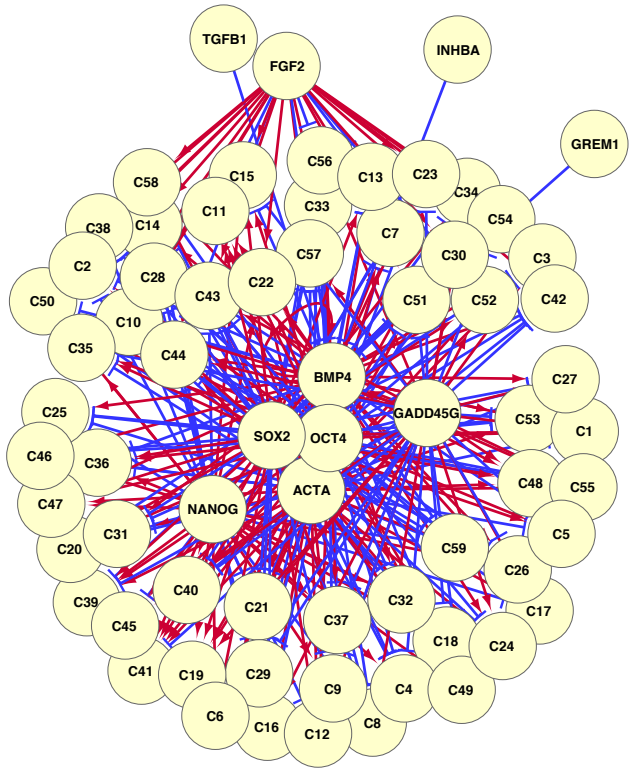


Figure 16: Hairball-like network that consists of only genes that are perturbed by at least 5 out of 7 experiments and are either receptors or ligands. Genes that had identical incoming edges are clustered together (e.g C36 is a cluster of two genes). Positive regulation is marked with red edges, negative with blue edges.

Modelling the network

The obtained network can already be used as an input for experimentalists to design validation experiments but that would lead to hundreds of experiments costing a lot of time and money. Therefore, our aim was to use *in silico* experiments to single out genes with the highest potential to support pluripotency in human embryonic stem cells.

In order to solve this problem we designed a computational approach that allows us to perform *in silico* perturbation experiments and measure the impact of the perturbed gene on the pluripotency state.

Edges in our network can represent either direct or indirect regulation as we are not able to distinguish the two types while using only the perturbation results. It was also out of the question to produce more experimental parameters that would be needed for quantitative modelling approaches such as differential equations. Thus, we selected qualitative (Boolean) modelling that deals with genes as binary switches, as best suited for our network type.

Biological networks are known to be rich in feed-back and feed-forward loops (Macarthur, Ma'ayan, & Lemischka, 2009; Milo *et al.*, 2002; Remy *et al.*, 2006) that provide the robustness to keep their biological state. Our combined network only had the initial clique of the three core regulators and few feed-back loops from the initial literature network. All the genes added from the perturbation experiments had only incoming regulatory edges meaning there was no feed-back from these genes to the core regulators.

To overcome this limitation, we artificially introduced such feed-back loops to the core regulators from one target node at a time. Each original network with an added edge forms a *model network*. We measure the impact of the introduced edge on the pluripotency state by calculating steady states of the model network.

We performed the *in silico* perturbation experiments using boolSim software (version 1.0) (Garg *et al.*, 2009). We ran boolSim in synchronous mode where all nodes change their activation states simultaneously in sequential time points. We let the program do the number of calculation steps needed to obtain the steady states. Every *in silico* perturbation experiment produced zero or more steady states.

We considered steady states to be both fixed and strongly connected steady states, where some of the network nodes were in cyclic mode. Each steady state indicates the potential activation and inactivation status of our *model network*. Each such state is described by a binary vector where 0 marks the passive and 1 the active mode of each node in the network.

Our original network had two steady states – one where the pluripotency markers (OCT4, SOX2, NANOG) where up-regulated and early differentiation markers (BMP4, GADD45G) were down-regulated at the same time; and the second state representing early differentiation where the activity levels were opposite to the previous steady state.

In total we had 1180 models as we introduced both a negative and a positive regulatory edge from each of the 59 terminal nodes to 10 core regulators. We did *in silico* knock-out and over-expression for each regulatory node included in the model (this covers 10 core regulators and a terminal node from which we created the feed-back loop) so each model network was perturbed 22 times in total. For further analysis we combined such binary vectors into one steady state matrix per model network. In the matrix each column represents a node in the network, and each row represents one steady state.

Evaluating the models

The steady state matrices were reduced by keeping only columns that corresponded to the pluripotency markers OCT4, NANOG and SOX2, and early differentiation markers GADD45G and BMP4. The resulting binary matrices were further dissected using principal component analysis (PCA) that allows to represent complex multidimensional data as two-dimensional plots.

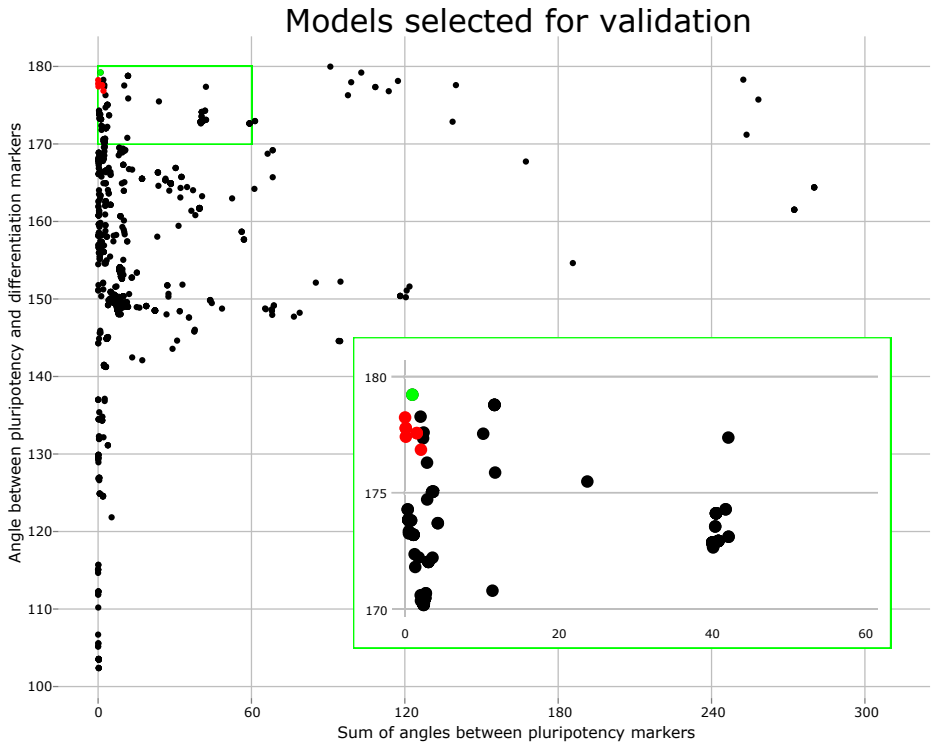


Figure 17: Model scoring. Each model was scored using the sum of angles between pluripotency markers (x-axis) and an angle between pluripotency and differentiation markers (y-axis). Models selected for validation (marked with red) had small angles between pluripotency markers and a large angle between pluripotency and differentiation markers. The validated model consisting of IL-11 is marked with green.

We evaluated each model using angles between eigenvectors defined by the first two principal components of the marker genes from the PCA analysis. Each eigenvector expresses the loadings in the first two principal components, and angles between such vectors represent the correlation between activity states of two genes. The final score for each model was based on an angle between the average eigenvector of pluripotency markers (OCT4, SOX2, NANOG) and early differentiation markers (BMP4, GADD45G). We looked for model networks that would produce a small angle between cellular state markers (either pluripotency or differentiation markers) and a large angle between different states. For experimental testing we chose markers that had a minimal angle between pluripotency markers and a maximal angle between differentiation markers and pluripotency markers (Figure 17).

Validation of candidate genes

We picked 15 candidate genes from the 8 top scoring models for experimental validation. For only 6 genes out of the 15 recombinant proteins or known inhibitors were available. So perturbations experiments were performed with CXCL12, EPHB2, EPHB3, IL-11, JUN and WNT3A at different molecular concentrations. While most of the candidate genes did not show a significant effect on the pluripotency state, the Interleukin-11 (IL-11) provided promising results.

The potential of IL-11 to keep hESCs in pluripotency state was validated by making cells grow without the standard medium containing bFGF but, instead, with supplemented IL-11. The cells were followed for 9 passages and immunohistochemistry results confirmed the expression of core hESCs markers OCT4, SOX2, NANOG and TRA1-60 (Figures 4 and 5 in Paper V).

Additional microarray experiments were carried out to compare the effect of bFGF addition and removal with the addition of IL-11. The correlation results of IL-11 treatment expression patterns place it between the deprived bFGF and bFGF treatment suggesting that FGF2 and IL-11 might control the pluripotency state through different mechanisms and with unequal strength.

IL-11 is a cytokine from the same IL-6 protein family as LIF which is necessary for maintaining pluripotency in mouse ESCs. Both cytokines operate through the same gp130 receptor (Nandurkar *et al.*, 1996; Timmermann *et al.*, 2000) interacting with Janus kinases. LIF is known to be a key factor keeping the mouse ESCs, that are known to be in more naïve state compared to the hESCs, in the pluripotent state. Based on our results we can hypothesise that IL-11 might play a role in converting the primed cells to a more naïve state, however it might need additional factors to achieve this.

Summary

In this work we showed how *in silico* experiments executed in a systematic manner can lead to identification of novel key regulators in a well studied biological system. We combined literature data with high-throughput perturbation datasets to study potential regulators of pluripotency and early differentiation in human embryonic stem cells. With different filtering approaches, we reduced the initial hairball-like network to a moderate size that we could model using qualitative modelling methods. Introducing feed-back loops between target genes and regulators *in silico*, we were able to measure the impact of such loops to the network status. All models were evaluated using eigenvectors representing cellular state markers. Based on the score, we chose 8 modules consisting of 15 genes. Out of the 6 actually tested genes the most promising effect was shown by Interleukin-11 that is capable of replacing bFGF known to be critical factor in the medium for keeping hESCs in pluripotency state.

CONCLUSIONS

The results of this thesis can be summarised as as follows.

I. Options for using only gene expression data for finding components of known pathways are limited.
We detected limitations of using only gene expression data to describe pathway membership in different biological pathway types. Based on our analysis, it makes more sense to use only a subset of strongly correlated pathway members for such tasks. Even when using only co-expression data, the novel candidate genes could be linked to the pathways they were predicted from.

II. Combining hierarchical clustering with automatic functional enrichment analysis helps to find co-expressed and co-functional gene sets.
Highlighting gene clusters that are both co-expressed and share functional annotations helps to get an initial overview of affected processes. Complicated clustering parameters are removed from the user and an easily interpretable overview is provided.

III. Combination of DNA motifs with DNA conservation data or chromatin immunoprecipitation identifies new regulatory information for stem cells.
Studying adipocyte differentiation using gene expression data and motif matching approaches we proposed that three transcription factors (MEIS1, CART1, PRRX2) could be new potential regulators of early adipocyte differentiation in mouse.
We defined six distinct modules of OCT4 target genes based on data from chromatin immunoprecipitation and motif matching experiments. For genes indirectly regulated by OCT4, we proposed 4 alternative motifs and potential factors that could be regulating these genes.

IV. Embryonic Stem Cells Database enables comparison of gene behaviour across large data compendia.
We gathered publicly available data relevant for the embryonic stem cell research into the Embryonic Stem Cell Database. The freely available database holds data from perturbation and immunoprecipitation experiments from mouse and human embryonic stem and carcinoma cell lines. The

database allows easy comparison of datasets and provides means for creating new hypotheses.

V. Integration of high-throughput data followed by *in silico* modelling allows to identify candidate genes for pluripotency regulation in human embryonic stem cells.
We combined literature data with high-throughput perturbation datasets and followed up with qualitative modeling. Using this approach, we identified IL-11 – a novel key regulator for pluripotency in human embryonic stem cells.

The main outcome of the thesis:

Integration of different high-throughput datasets enables establishing gene-gene relationships that would not be possible when looking at a single data type in isolation.

Bibliography

Adler, P.; Peterson, H.; Agius, P.; Reimand, J.; and Vilo, J. 2009. Ranking genes by their co-expression to subsets of pathway members. *Annals of the New York Academy of Sciences* 1158(1):1–13.

Allison, D. B.; Cui, X.; Page, G. P.; and Sabripour, M. 2006. Microarray data analysis: from disarray to consolidation and consensus. *Nature Reviews Genetics* 7(1):55–65.

Alon, U. 2007. Network motifs: theory and experimental approaches. *Nature Reviews Genetics* 8(6):450–461.

Andrews, P.; Matin, M.; Bahrami, A.; Damjanov, I.; Gokhale, P.; and Draper, J. 2005. Embryonic stem (ES) cells and embryonal carcinoma (EC) cells: opposite sides of the same coin. *Biochemical Society Transactions* 33(6):1526–1530.

Ashburner, M.; Ball, C. A.; Blake, J. A.; Botstein, D.; Butler, H.; Cherry, J. M.; Davis, A. P.; Dolinski, K.; Dwight, S. S.; Eppig, J. T.; Harris, M. A.; Hill, D. P.; Issel-Tarver, L.; Kasarskis, A.; Lewis, S.; Matese, J. C.; Richardson, J. E.; Ringwald, M.; Rubin, G. M.; and Sherlock, G. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics* 25(1):25–29.

Babaie, Y.; Herwig, R.; Greber, B.; Brink, T.; Wruck, W.; Groth, D.; Lehrach, H.; Burdon, T.; and Adjaye, J. 2007. Analysis of Oct4-dependent transcriptional networks regulating self-renewal and pluripotency in human embryonic stem cells. *Stem Cells* 25:500–510.

Bailey, T. L., and Elkan, C. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings of the International Conference on Intelligent Systems for Molecular Biology* 2:28–36.

Barberi, T.; Willis, L. M.; Socci, N. D.; and Studer, L. 2005. Derivation of multipotent mesenchymal precursors from human embryonic stem cells. *PLOS Medicine* 2(6):e161.

Barrett, T., and Edgar, R. 2006. Gene Expression Omnibus: Microarray data storage, submission, retrieval, and analysis. *Methods in Enzymology* 411:352–369.

Barrett, C. L.; Cho, B.-K.; and Palsson, B. O. 2011. Sensitive and accurate identification of protein–DNA binding events in ChIP-chip assays using higher order derivative analysis. *Nucleic Acids Research* 39(5):1656–1665.

Bauer, S.; Gagneur, J.; and Robinson, P. N. 2010. GOing bayesian: model-based

gene set analysis of genome-scale data. *Nucleic Acids Research* 38(11):3523–3532.

Benjamini, Y., and Hochberg, Y. 2000. On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics* 25(1):60–83.

Benjamini, Y., and Yekutieli, D. 2001. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics* 1165–1188.

Berriz, G. F., and Roth, F. P. 2008. The synergizer service for translating gene, protein and other biological identifiers. *Bioinformatics* 24(19):2272–2273.

Bieda, M.; Xu, X.; Singer, M.; Green, R.; and Farnham, P. 2006. Unbiased location analysis of E2F1-binding sites suggests a widespread role for E2F1 in the human genome. *Genome Research* 16:595–605.

Billon, N.; Iannarelli, P.; Monteiro, M. C.; Glavieux-Pardanaud, C.; Richardson, W. D.; Kessaris, N.; Dani, C.; and Dupin, E. 2007. The generation of adipocytes by the neural crest. *Development* 134(12):2283–2292.

Bolstad, B. M.; Irizarry, R. A.; Åstrand, M.; and Speed, T. P. 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 19(2):185–193.

Botquin, V.; Hess, H.; Fuhrmann, G.; Anastassiadis, C.; Gross, M. K.; Vriend, G.; and Schöler, H. R. 1998. New POU dimer configuration mediates antagonistic control of an osteopontin preimplantation enhancer by Oct-4 and Sox-2. *Genes & Development* 12(13):2073–2090.

Boyer, L.; Lee, T.; Cole, M.; Johnstone, S.; Levine, S.; Zucker, J.; Guenther, M.; Kumar, R.; Murray, H.; Jenner, R.; Gifford, D.; Melton, D.; Jaenisch, R.; and Young, R. 2005. Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* 122:947–956.

Breitling, R.; Armengaud, P.; Amtmann, A.; and Herzyk, P. 2004. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Letters* 573(1):83–92.

Bustin, S. 2002. Quantification of mRNA using real-time reverse transcription PCR (RT-PCR): trends and problems. *Journal of Molecular Endocrinology* 29(1):23–39.

Chambers, I., and Smith, A. 2004. Self-renewal of teratocarcinoma and embryonic stem cells. *Oncogene* 23(43):7150–7160.

Chavez, L.; Bais, A. S.; Vingron, M.; Lehrach, H.; Adjaye, J.; and Herwig, R. 2009. In silico identification of a core regulatory network of OCT4 in human embryonic stem cells using an integrated approach. *BMC Genomics* 10:314.

Chen, Z.; Torrens, J. I.; Anand, A.; Spiegelman, B. M.; and Friedman, J. M. 2005. Krox20 stimulates adipogenesis via C/EBPbeta-dependent and -independent mechanisms. *Cell Metabolism* 1(2):93–106.

Chen, X.; Xu, H.; Yuan, P.; Fang, F.; Huss, M.; Vega, V.; Wong, E.; Orlov, Y.; Zhang, W.; Jiang, J.; Loh, Y.; Yeo, H.; Yeo, Z.; Narang, V.; Govindarajan, K.;

Leong, B.; Shahab, A.; Ruan, Y.; Bourque, G.; Sung, W.; Clarke, N.; Wei, C.; and Ng, H. 2008. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* 133:1106–1117.

Chen, C.; Grennan, K.; Badner, J.; Zhang, D.; Gershon, E.; Jin, L.; and Liu, C. 2011. Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. *PLOS One* 6(2):e17238.

Chew, J. L.; Loh, Y. H.; Zhang, W.; Chen, X.; Tam, W. L.; Yeap, L. S.; Li, P.; Ang, Y. S.; Lim, B.; Robson, P.; and Ng, H. H. 2005. Reciprocal transcriptional regulation of Pou5f1 and Sox2 via the Oct4/Sox2 complex in embryonic stem cells. *Molecular and Cellular Biology* 25(14):6031–6046.

Chuaqui, R. F.; Bonner, R. F.; Best, C. J.; Gillespie, J. W.; Flaig, M. J.; Hewitt, S. M.; Phillips, J. L.; Krizman, D. B.; Tangrea, M. A.; Ahram, M.; Linehan, W. M.; Knezevic, V.; and Emmert-Buck, M. R. 2002. Post-analysis follow-up and validation of microarray experiments. *Nature Genetics* 32 Suppl:509–514.

Clark, A. T. 2007. The stem cell identity of testicular cancer. *Stem Cell Reviews and Reports* 3(1):49–59.

Croft, D.; Mundo, A. F.; Haw, R.; Milacic, M.; Weiser, J.; Wu, G.; Caudy, M.; Garapati, P.; Gillespie, M.; Kamdar, M. R.; Jassal, B.; Jupe, S.; Matthews, L.; May, B.; Palatnik, S.; Rothfels, K.; Shamovsky, V.; Song, H.; Williams, M.; Birney, E.; Hermjakob, H.; Stein, L.; and D'Eustachio, P. 2014. The Reactome pathway knowledgebase. *Nucleic Acids Research* 42(Database issue):D472–477.

Cui, X.; Kerr, M. K.; and Churchill, G. A. 2003. Transformations for cDNA microarray data. *Statistical Applications in Genetics and Molecular Biology* 2(1).

Damjanov, I.; Horvat, B.; and Gibas, Z. 1993. Retinoic acid-induced differentiation of the developmentally pluripotent human germ cell tumor-derived cell line, NCCIT. *Laboratory investigation; a journal of technical methods and pathology* 68(2):220–232.

Dani, C.; Smith, A. G.; Dessolin, S.; Leroy, P.; Staccini, L.; Villageois, P.; Darimont, C.; and Ailhaud, G. 1997. Differentiation of embryonic stem cells into adipocytes in vitro. *Journal of Cell Science* 110 (Pt 11):1279–1285.

Dixon, J. E.; Dick, E.; Rajamohan, D.; Shakesheff, K. M.; and Denning, C. 2011. Directed differentiation of human embryonic stem cells to interrogate the cardiac gene regulatory network. *Molecular Therapy* 19(9):1695–1703.

Draghici, S.; Khatri, P.; Eklund, A. C.; and Szallasi, Z. 2006. Reliability and reproducibility issues in DNA microarray measurements. *Trends in Genetics* 22(2):101–109.

Dudoit, S.; Shaffer, J. P.; and Boldrick, J. C. 2003. Multiple hypothesis testing in microarray experiments. *Statistical Science* 71–103.

Edgar, R.; Domrachev, M.; and Lash, A. E. 2002. Gene Expression Omnibus:

NCBI gene expression and hybridization array data repository. *Nucleic Acids Research* 30(1):207–210.

Eisen, M. B.; Spellman, P. T.; Brown, P. O.; and Botstein, D. 1998. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America* 95(25):14863–14868.

Fortini, M. E. 2009. Notch signaling: the core pathway and its posttranslational regulation. *Developmental Cell* 16(5):633–647.

Freude, K. K.; Penjwini, M.; Davis, J. L.; LaFerla, F. M.; and Blurton-Jones, M. 2011. Soluble amyloid precursor protein induces rapid neural differentiation of human embryonic stem cells. *Journal of Biological Chemistry* 286(27):24264–24274.

Garg, A.; Mohanram, K.; Di Cara, A.; De Micheli, G.; and Xenarios, I. 2009. Modeling stochasticity and robustness in gene regulatory networks. *Bioinformatics* 25:i101–109.

Gehlenborg, N.; O'Donoghue, S. I.; Baliga, N. S.; Goesmann, A.; Hibbs, M. A.; Kitano, H.; Kohlbacher, O.; Neuweger, H.; Schneider, R.; Tenenbaum, D.; and Gavin, A. C. 2010. Visualization of omics data for systems biology. *Nature Methods* 7(3 Suppl):56–68.

Gerin, I.; Bommer, G. T.; Lidell, M. E.; Cederberg, A.; Enerback, S.; and Macdougald, O. A. 2009. On the role of FOX transcription factors in adipocyte differentiation and insulin-stimulated glucose uptake. *The Journal of Biological Chemistry* 284(16):10755–10763.

Gibson, U.; Heid, C. A.; and Williams, P. M. 1996. A novel method for real time quantitative RT-PCR. *Genome Research* 6(10):995–1001.

Giralt, M., and Villarroya, F. 2013. White, brown, beige/brite: different adipose cells for different functions? *Endocrinology* 154(9):2992–3000.

Göke, J.; Chan, Y.-S.; Yan, J.; Vingron, M.; and Ng, H.-H. 2013. Genome-wide kinase-chromatin interactions reveal the regulatory network of ERK signaling in human embryonic stem cells. *Molecular Cell* 50(6):844–855.

Gordon, D. B.; Nekludova, L.; McCallum, S.; and Fraenkel, E. 2005. TAMO: a flexible, object-oriented framework for analyzing transcriptional regulation using DNA-sequence motifs. *Bioinformatics* 21(14):3164–3165.

Goto, T.; Adjaye, J.; Rodeck, C. H.; and Monk, M. 1999. Identification of genes expressed in human primordial germ cells at the time of entry of the female germ line into meiosis. *Molecular Human Reproduction* 5(9):851–860.

Gouzé, J.-L. 1998. Positive and negative circuits in dynamical systems. *Journal of Biological Systems* 6(01):11–15.

Greber, B.; Lehrach, H.; and Adjaye, J. 2007. Silencing of core transcription factors in human EC cells highlights the importance of autocrine FGF signaling for self-renewal. *BMC Developmental Biology* 7:46.

Greber, B.; Lehrach, H.; and Adjaye, J. 2008. Control of early fate decisions in human ES cells by distinct states of TGFbeta pathway activity. *Stem Cells and Development* 17(6):1065–1077.

Griffiths-Jones, S.; Grocock, R. J.; Van Dongen, S.; Bateman, A.; and Enright, A. J. 2006. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Research* 34(suppl 1):D140–D144.

Hanna, J.; Cheng, A. W.; Saha, K.; Kim, J.; Lengner, C. J.; Soldner, F.; Cassady, J. P.; Muffat, J.; Carey, B. W.; and Jaenisch, R. 2010. Human embryonic stem cells with biological and epigenetic characteristics similar to those of mouse ESCs. *Proceedings of the National Academy of Sciences of the United States of America* 107(20):9222–9227.

Hansen, K. D.; Irizarry, R. A.; and Zhijin, W. 2012. Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics* 13(2):204–216.

Harms, M., and Seale, P. 2013. Brown and beige fat: development, function and therapeutic potential. *Nature Medicine* 19(10):1252–1263.

Ho, J. W.; Bishop, E.; Karchenko, P. V.; Nègre, N.; White, K. P.; and Park, P. J. 2011. ChIP-chip versus ChIP-seq: lessons for experimental design and data analysis. *BMC genomics* 12(1):134.

Hochreiter, S.; Clevert, D.-A.; and Obermayer, K. 2006. A new summarization method for Affymetrix probe level data. *Bioinformatics* 22(8):943–949.

Huang, D. W.; Sherman, B. T.; Tan, Q.; Kir, J.; Liu, D.; Bryant, D.; Guo, Y.; Stephens, R.; Baseler, M. W.; Lane, H. C.; and Lempicki, R. A. 2007. DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic Acids Research* 35(Web Server issue):W169–175.

Hughes, J. D.; Estep, P. W.; Tavazoie, S.; and Church, G. M. 2000. Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *Journal of Molecular Biology* 296(5):1205–1214.

Ideker, T. E.; Thorsson, V.; and Karp, R. M. 2000. Discovery of regulatory interactions through perturbation: inference and experimental design. In *Pacific Symposium on Biocomputing*, volume 5, 302–313.

Inamura, M.; Kawabata, K.; Takayama, K.; Tashiro, K.; Sakurai, F.; Katayama, K.; Toyoda, M.; Akutsu, H.; Miyagawa, Y.; Okita, H.; Kiyokawa, N.; Umezawa, A.; Hayakawa, T.; Furue, M. K.; and Mizuguchi, H. 2011. Efficient generation of hepatoblasts from human ES cells and iPS cells by transient overexpression of homeobox gene HEX. *Molecular Therapy* 19(2):400–407.

Irizarry, R. A.; Hobbs, B.; Collin, F.; Beazer-Barclay, Y. D.; Antonellis, K. J.; Scherf, U.; and Speed, T. P. 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4(2):249–264.

Ivanova, N.; Dobrin, R.; Lu, R.; Kotenko, I.; Levorse, J.; DeCoste, C.; Schafer, X.; Lun, Y.; and Lemischka, I. R. 2006. Dissecting self-renewal in stem cells with RNA interference. *Nature* 442:533–538.

Jin, V.; O'Geen, H.; Iyengar, S.; Green, R.; and Farnham, P. 2007. Identification of an OCT4 and SRY regulatory module using integrated computational and experimental genomics approaches. *Genome Research* 17:807–817.

Johnson, J. M.; Edwards, S.; Shoemaker, D.; and Schadt, E. E. 2005. Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. *Trends in Genetics* 21(2):93–102.

Johnson, D. S.; Li, W.; Gordon, D. B.; Bhattacharjee, A.; Curry, B.; Ghosh, J.; Brizuela, L.; Carroll, J. S.; Brown, M.; Flicek, P.; Koch, C. M.; Dunham, I.; Bieda, M.; Xu, X.; Farnham, P. J.; Kapranov, P.; Nix, D. A.; Gingeras, T. R.; Zhang, X.; Holster, H.; Jiang, N.; Green, R. D.; Song, J. S.; McCuine, S. A.; Anton, E.; Nguyen, L.; Trinklein, N. D.; Ye, Z.; Ching, K.; Hawkins, D.; Ren, B.; Scacheri, P. C.; Rozowsky, J.; Karpikov, A.; Euskirchen, G.; Weissman, S.; Gerstein, M.; Snyder, M.; Yang, A.; Moqtaderi, Z.; Hirsch, H.; Shulha, H. P.; Fu, Y.; Weng, Z.; Struhl, K.; Myers, R. M.; Lieb, J. D.; and Liu, X. S. 2008. Systematic evaluation of variability in ChIP-chip experiments using predefined DNA targets. *Genome Research* 18(3):393–403.

Johnson, W. E.; Li, C.; and Rabinovic, A. 2007. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 8(1):118–127.

Josephson, R.; Ording, C. J.; Liu, Y.; Shin, S.; Lakshmipathy, U.; Toumadje, A.; Love, B.; Chesnut, J. D.; Andrews, P. W.; Rao, M. S.; and Auerbach, J. M. 2007. Qualification of embryonal carcinoma 2102Ep as a reference for human embryonic stem cell research. *Stem Cells* 25(2):437–446.

Jung, M.; Peterson, H.; Chavez, L.; Kahlem, P.; Lehrach, H.; Vilo, J.; and Adjaye, J. 2010. A data integration approach to mapping OCT4 gene regulatory networks operative in embryonic stem cells and embryonal carcinoma cells. *PLOS One* 5(5):e10709.

Kallas, A.; Pook, M.; Trei, A.; and Maimets, T. 2014. SOX2 is regulated differently from NANOG and OCT4 in human embryonic stem cells during early differentiation initiated with sodium butyrate. *Stem Cells International* 2014.

Kanehisa, M., and Goto, S. 2000. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* 28(1):27–30.

Kanehisa, M.; Goto, S.; Sato, Y.; Kawashima, M.; Furumichi, M.; and Tanabe, M. 2014. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Research* 42(Database issue):199–205.

Karolchik, D.; Kuhn, R. M.; Baertsch, R.; Barber, G. P.; Clawson, H.; Diekhans, M.; Giardine, B.; Harte, R. A.; Hinrichs, A. S.; Hsu, F.; Kober, K. M.; Miller, W.; Pedersen, J. S.; Pohl, A.; Raney, B. J.; Rhead, B.; Rosenbloom, K. R.; Smith, K. E.; Stanke, M.; Thakkapallayil, A.; Trumbower, H.; Wang, T.; Zweig,

A. S.; Haussler, D.; and Kent, W. J. 2008. The UCSC Genome Browser Database: 2008 update. *Nucleic Acids Research* 36(Database issue):D773–779.

Kaufmann, L. T., and Niehrs, C. 2011. Gadd45a and Gadd45g regulate neural development and exit from pluripotency in Xenopus. *Mechanisms of Development* 128(7-10):401–411.

Kaufmann, L. T.; Gierl, M. S.; and Niehrs, C. 2011. Gadd45a, Gadd45b and Gadd45g expression during mouse embryonic development. *Gene Expression Patterns* 11(8):465–470.

Kerrien, S.; Alam-Faruque, Y.; Aranda, B.; Bancarz, I.; Bridge, A.; Derow, C.; Dimmer, E.; Feuermann, M.; Friedrichsen, A.; Huntley, R.; Kohler, C.; Khadake, J.; Leroy, C.; Liban, A.; Lieftink, C.; Montecchi-Palazzi, L.; Orchard, S.; Risse, J.; Robbe, K.; Roechert, B.; Thorneycroft, D.; Zhang, Y.; Apweiler, R.; and Hermjakob, H. 2007. IntAct–open source resource for molecular interaction data. *Nucleic Acids Research* 35(Database issue):D561–565.

Kim, J.; Chu, J.; Shen, X.; Wang, J.; and Orkin, S. H. 2008. An extended transcriptional network for pluripotency of embryonic stem cells. *Cell* 132:1049–1061.

Kitano, H. 2004. Biological robustness. *Nature Reviews Genetics* 5(11):826–837.

Köhler, S.; Doelken, S. C.; Mungall, C. J.; Bauer, S.; Firth, H. V.; Bailleul-Forestier, I.; Black, G. C.; Brown, D. L.; Brudno, M.; Campbell, J.; FitzPatrick, D. R.; Eppig, J. T.; Jackson, A. P.; Freson, K.; Girdea, M.; Helbig, I.; Hurst, J. A.; Jahn, J.; Jackson, L. G.; Kelly, A. M.; Ledbetter, D. H.; Mansour, S.; Martin, C. L.; Moss, C.; Mumford, A.; Ouwehand, W. H.; Park, S. M.; Riggs, E. R.; Scott, R. H.; Sisodiya, S.; Van Vooren, S.; Wapner, R. J.; Wilkie, A. O.; Wright, C. F.; Vulto-van Silfhout, A. T.; de Leeuw, N.; de Vries, B. B.; Washingthon, N. L.; Smith, C. L.; Westerfield, M.; Schofield, P.; Ruef, B. J.; Gkoutos, G. V.; Haendel, M.; Smedley, D.; Lewis, S. E.; and Robinson, P. N. 2014. The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Research* 42(Database issue):D966–974.

Kubosaki, A.; Lindgren, G.; Tagami, M.; Simon, C.; Tomaru, Y.; Miura, H.; Suzuki, T.; Arner, E.; Forrest, A. R.; Irvine, K. M.; Schroder, K.; Hasegawa, Y.; Kanamori-Katayama, M.; Rehli, M.; Hume, D. A.; Kawai, J.; Suzuki, M.; Suzuki, H.; and Hayashizaki, Y. 2010. The combination of gene perturbation assay and ChIP-chip reveals functional direct target genes for IRF8 in THP-1 cells. *Molecular Immunology* 47(14):2295–2302.

Kull, M., and Vilo, J. 2008. Fast approximate hierarchical clustering using similarity heuristics. *BioData Mining* 1(1):9.

Łabaj, P. P.; Leparc, G. G.; Linggi, B. E.; Markillie, L. M.; Wiley, H. S.; and Kreil, D. P. 2011. Characterization and improvement of RNA-Seq precision in quantitative transcript expression profiling. *Bioinformatics* 27(13):i383–i391.

Landgrebe, J.; Wurst, W.; and Welzl, G. 2002. Permutation-validated principal components analysis of microarray data. *Genome Biology* 3(4):0019–0011.

Langfelder, P., and Horvath, S. 2012. Fast R functions for robust correlations and hierarchical clustering. *Journal of statistical software* 46(11).

Laudes, M.; Christodoulides, C.; Sewter, C.; Rochford, J. J.; Considine, R. V.; Sethi, J. K.; Vidal-Puig, A.; and O'Rahilly, S. 2004. Role of the POZ zinc finger transcription factor FBI-1 in human and murine adipogenesis. *The Journal of Biological Chemistry* 279(12):11711–11718.

Levsky, J. M., and Singer, R. H. 2003. Gene expression and the myth of the average cell. *Trends in Cell Biology* 13(1):4–6.

Lister, R.; Pelizzola, M.; Dowen, R. H.; Hawkins, R. D.; Hon, G.; Tonti-Filippini, J.; Nery, J. R.; Lee, L.; Ye, Z.; Ngo, Q. M.; Edsall, L.; Antosiewicz-Bourget, J.; Stewart, R.; Ruotti, V.; Millar, A. H.; Thomson, J. A.; Ren, B.; and Ecker, J. R. 2009. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* 462(7271):315–322.

Liu, X. S.; Brutlag, D. L.; and Liu, J. S. 2002. An algorithm for finding protein–DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nature Biotechnology* 20(8):835–839.

Lockhart, D. J.; Dong, H.; Byrne, M. C.; Follettie, M. T.; Gallo, M. V.; Chee, M. S.; Mittmann, M.; Wang, C.; Kobayashi, M.; Horton, H.; and Brown, E. L. 1996. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology* 14(13):1675–1680.

Loh, Y.; Wu, Q.; Chew, J.; Vega, V.; Zhang, W.; Chen, X.; Bourque, G.; George, J.; Leong, B.; Liu, J.; Wong, K.; Sung, K.; Lee, C.; Zhao, X.; Chiu, K.; Lipovich, L.; Kuznetsov, V.; Robson, P.; Stanton, L.; Wei, C.; Ruan, Y.; Lim, B.; and Ng, H. 2006. The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nature Genetics* 38:431–440.

Lukk, M.; Kapushesky, M.; Nikkila, J.; Parkinson, H.; Goncalves, A.; Huber, W.; Ukkonen, E.; and Brazma, A. 2010. A global map of human gene expression. *Nature Biotechnology* 28(4):322–324.

Macarthur, B. D.; Ma'ayan, A.; and Lemischka, I. R. 2009. Systems biology of stem cell fate and cellular reprogramming. *Nature Reviews Molecular Cell Biology* 10(10):672–681.

Mahony, S., and Benos, P. V. 2007. STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic Acids Research* 35(suppl 2):W253–W258.

Mangan, S., and Alon, U. 2003. Structure and function of the feed-forward loop network motif. *Proceedings of the National Academy of Sciences of United States of America* 100(21):11980–11985.

Mangan, S.; Zaslaver, A.; and Alon, U. 2003. The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks. *Journal of Molecular Biology* 334(2):197–204.

Marioni, J. C.; Mason, C. E.; Mane, S. M.; Stephens, M.; and Gilad, Y. 2008. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research* 18(9):1509–1517.

Masui, S.; Nakatake, Y.; Toyooka, Y.; Shimosato, D.; Yagi, R.; Takahashi, K.; Okochi, H.; Okuda, A.; Matoba, R.; Sharov, A. A.; Ko, M. S.; and Niwa, H. 2007. Pluripotency governed by Sox2 via regulation of Oct3/4 expression in mouse embryonic stem cells. *Nature Cell Biology.* 9(6):625–635.

Mathur, D.; Danford, T.; Boyer, L.; Young, R.; Gifford, D.; and Jaenisch, R. 2008. Analysis of the mouse embryonic stem cell regulatory networks obtained by ChIP-chip and ChIP-PET. *Genome Biology* 9:R126.

Matin, M. M.; Walsh, J. R.; Gokhale, P. J.; Draper, J. S.; Bahrami, A. R.; Morton, I.; Moore, H. D.; and Andrews, P. W. 2004. Specific knockdown of Oct4 and β2-microglobulin expression by RNA interference in human embryonic stem cells and embryonic carcinoma cells. *Stem Cells* 22(5):659–668.

Matys, V.; Kel-Margoulis, O. V.; Fricke, E.; Liebich, I.; Land, S.; Barre-Dirrie, A.; Reuter, I.; Chekmenev, D.; Krull, M.; Hornischer, K.; Voss, N.; Stegmaier, P.; Lewicki-Potapov, B.; Saxel, H.; Kel, A. E.; and Wingender, E. 2006. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Research* 34(Database issue):D108–110.

Metzker, M. L. 2010. Sequencing technologies—the next generation. *Nature Reviews Genetics* 11(1):31–46.

Milacic, M.; Haw, R.; Rothfels, K.; Wu, G.; Croft, D.; Hermjakob, H.; D'Eustachio, P.; and Stein, L. 2012. Annotating cancer variants and anti-cancer therapeutics in Reactome. *Cancers* 4(4):1180–1211.

Milo, R.; Shen-Orr, S.; Itzkovitz, S.; Kashtan, N.; Chklovskii, D.; and Alon, U. 2002. Network motifs: simple building blocks of complex networks. *Science* 298(5594):824–827.

Mitchell, P. J.; Timmons, P. M.; Hebert, J. M.; Rigby, P. W.; and Tjian, R. 1991. Transcription factor AP-2 is expressed in neural crest cell lineages during mouse embryogenesis. *Genes & Development* 5(1):105–119.

Molinelli, E. J.; Korkut, A.; Wang, W.; Miller, M. L.; Gauthier, N. P.; Jing, X.; Kaushik, P.; He, Q.; Mills, G.; Solit, D. B.; Pratilas, C. A.; Weigt, M.; Braunstein, A.; Pagnani, A.; Zecchina, R.; and Sander, C. 2013. Perturbation biology: inferring signaling networks in cellular systems. *PLOS Computational Biology* 9(12):e1003290.

Monteiro, M. C.; Sanyal, M.; Cleary, M. L.; Sengenes, C.; Bouloumie, A.; Bouloume, A.; Dani, C.; and Billon, N. 2011. PBX1: a novel stage-specific regulator of adipocyte development. *Stem Cells* 29(11):1837–1848.

Nandurkar, H. H.; Hilton, D. J.; Nathan, P.; Willson, T.; Nicola, N.; and Begley, C. G. 1996. The human IL-11 receptor requires gp130 for signalling: demonstration by molecular cloning of the receptor. *Oncogene* 12(3):585–593.

Nelson, J. D.; Denisenko, O.; and Bomsztyk, K. 2006. Protocol for the fast chromatin immunoprecipitation (ChIP) method. *Nature Protocols* 1(1):179–185.

Neph, S.; Stergachis, A. B.; Reynolds, A.; Sandstrom, R.; Borenstein, E.; and

Stamatoyannopoulos, J. A. 2012. Circuitry and dynamics of human transcription factor regulatory networks. *Cell* 150(6):1274–1286.

Nichols, J., and Smith, A. 2009. Naive and primed pluripotent states. *Cell Stem Cell* 4(6):487–492.

Parkinson, H.; Kapushesky, M.; Shojatalab, M.; Abeygunawardena, N.; Coulson, R.; Farne, A.; Holloway, E.; Kolesnykov, N.; Lilja, P.; Lukk, M.; Mani, R.; Rayner, T.; Sharma, A.; William, E.; Sarkans, U.; and Brazma, A. 2007. ArrayExpress–a public database of microarray experiments and gene expression profiles. *Nucleic Acids Research* 35(Database issue):D747–750.

Patterson, T. A.; Lobenhofer, E. K.; Fulmer-Smentek, S. B.; Collins, P. J.; Chu, T. M.; Bao, W.; Fang, H.; Kawasaki, E. S.; Hager, J.; Tikhonova, I. R.; Walker, S. J.; Zhang, L.; Hurban, P.; de Longueville, F.; Fuscoe, J. C.; Tong, W.; Shi, L.; and Wolfinger, R. D. 2006. Performance comparison of one-color and two-color platforms within the MicroArray Quality Control (MAQC) project. *Nature Biotechnology* 24(9):1140–1150.

Peri, S.; Navarro, J. D.; Amanchy, R.; Kristiansen, T. Z.; Jonnalagadda, C. K.; Surendranath, V.; Niranjan, V.; Muthusamy, B.; Gandhi, T. K.; Gronborg, M.; Ibarrola, N.; Deshpande, N.; Shanker, K.; Shivashankar, H. N.; Rashmi, B. P.; Ramya, M. A.; Zhao, Z.; Chandrika, K. N.; Padma, N.; Harsha, H. C.; Yatish, A. J.; Kavitha, M. P.; Menezes, M.; Choudhury, D. R.; Suresh, S.; Ghosh, N.; Saravana, R.; Chandran, S.; Krishna, S.; Joy, M.; Anand, S. K.; Madavan, V.; Joseph, A.; Wong, G. W.; Schiemann, W. P.; Constantinescu, S. N.; Huang, L.; Khosravi-Far, R.; Steen, H.; Tewari, M.; Ghaffari, S.; Blobe, G. C.; Dang, C. V.; Garcia, J. G.; Pevsner, J.; Jensen, O. N.; Roepstorff, P.; Deshpande, K. S.; Chinnaiyan, A. M.; Hamosh, A.; Chakravarti, A.; and Pandey, A. 2003. Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Research* 13(10):2363–2371.

Pesce, M., and Scholer, H. R. 2001. Oct-4: gatekeeper in the beginnings of mammalian development. *Stem Cells* 19(4):271–278.

Peterson, H.; Garg, A.; Wang, Y.; Vilo, J.; Xenarios, I.; and Adjaye, J. 2012. Qualitative modeling identifies IL-11 as a novel regulator in maintaining self-renewal in human pluripotent stem cells. *Frontiers in Physiology* 4:303–303.

Piasecka, B.; Robinson-Rechavi, M.; and Bergmann, S. 2012. Correcting for the bias due to expression specificity improves the estimation of constrained evolution of expression between mouse and human. *Bioinformatics* 28(14):1865–1872.

Pigolotti, S.; Krishna, S.; and Jensen, M. H. 2007. Oscillation patterns in negative feedback loops. *Proceedings of the National Academy of Sciences of United States of America* 104(16):6533–6537.

Ramani, A. K.; Bunescu, R. C.; Mooney, R. J.; and Marcotte, E. M. 2005. Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome. *Genome Biology* 6(5):R40.

Raouf, A.; Zhao, Y.; To, K.; Stingl, J.; Delaney, A.; Barbara, M.; Iscove, N.; Jones, S.; McKinney, S.; Emerman, J.; Aparicio, S.; Marra, M.; and Eaves, C. 2008. Transcriptome analysis of the normal human mammary cell commitment and differentiation process. *Cell Stem Cell* 3(1):109–118.

Reimand, J.; Arak, T.; and Vilo, J. 2011. g:Profiler–a web server for functional interpretation of gene lists (2011 update). *Nucleic Acids Research* 39(Web Server issue):W307–315.

Reimand, J.; Kull, M.; Peterson, H.; Hansen, J.; and Vilo, J. 2007. g:Profiler–a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Research* 35(Web Server issue):193–200.

Reimand, J.; Tooming, L.; Peterson, H.; Adler, P.; and Vilo, J. 2008. GraphWeb: mining heterogeneous biological networks for gene modules with functional significance. *Nucleic Acids Research* 36(Web Server issue):W452–459.

Reimers, M., and Weinstein, J. N. 2005. Quality assessment of microarrays: visualization of spatial artifacts and quantitation of regional biases. *BMC Bioinformatics* 6(1):166.

Remenyi, A.; Lins, K.; Nissen, L. J.; Reinbold, R.; Scholer, H. R.; and Wilmanns, M. 2003. Crystal structure of a POU/HMG/DNA ternary complex suggests differential assembly of Oct4 and Sox2 on two enhancers. *Genes & Development* 17(16):2048–2059.

Remy, E.; Ruet, P.; Mendoza, L.; Thieffry, D.; and Chaouiya, C. 2006. From logical regulatory graphs to standard petri nets: Dynamical roles and functionality of feedback citrcuits. In *LNCS*, 56–72. Springer-Verlag.

Ringnér, M. 2008. What is principal component analysis? *Nature Biotechnology* 26(3):303–304.

Rosen, E. D.; Sarraf, P.; Troy, A. E.; Bradwin, G.; Moore, K.; Milstone, D. S.; Spiegelman, B. M.; and Mortensen, R. M. 1999. PPARγ is required for the differentiation of adipose tissue in vivo and in vitro. *Molecular Cell* 4(4):611–617.

Rosen, E. D.; Walkey, C. J.; Puigserver, P.; and Spiegelman, B. M. 2000. Transcriptional regulation of adipogenesis. *Genes & Development* 14(11):1293–1307.

Ruepp, A.; Waegele, B.; Lechner, M.; Brauner, B.; Dunger-Kaltenbach, I.; Fobo, G.; Frishman, G.; Montrone, C.; and Mewes, H. W. 2010. CORUM: the comprehensive resource of mammalian protein complexes–2009. *Nucleic Acids Research* 38(Database issue):497–501.

Rustici, G.; Kolesnikov, N.; Brandizi, M.; Burdett, T.; Dylag, M.; Emam, I.; Farne, A.; Hastings, E.; Ison, J.; Keays, M.; Kurbatova, N.; Malone, J.; Mani, R.; Mupo, A.; Pedro Pereira, R.; Pilicheva, E.; Rung, J.; Sharma, A.; Tang, Y. A.; Ternent, T.; Tikhonov, A.; Welter, D.; Williams, E.; Brazma, A.; Parkinson, H.; and Sarkans, U. 2013. ArrayExpress update–trends in database

growth and links to data analysis tools. *Nucleic Acids Research* 41(Database issue):D987–990.

Sandelin, A.; Alkema, W.; Engstrom, P.; Wasserman, W. W.; and Lenhard, B. 2004. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Research* 32(Database issue):D91–94.

Schena, M.; Shalon, D.; Davis, R. W.; and Brown, P. O. 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270(5235):467–470.

Schlitt, T., and Brazma, A. 2007. Current approaches to gene regulatory network modelling. *BMC Bioinformatics* 8(Suppl 6):S9.

Schneider, T. D.; Stormo, G. D.; Gold, L.; and Ehrenfeucht, A. 1986. Information content of binding sites on nucleotide sequences. *Journal of Molecular Biology* 188(3):415–431.

Sharov, A. A.; Masui, S.; Sharova, L. V.; Piao, Y.; Aiba, K.; Matoba, R.; Xin, L.; Niwa, H.; and Ko, M. S. 2008. Identification of Pou5f1, Sox2, and Nanog downstream target genes with statistical confidence by applying a novel algorithm to time course microarray and genome-wide chromatin immunoprecipitation data. *BMC Genomics* 9:269.

Slonim, D. K. 2002. From patterns to pathways: gene expression data analysis comes of age. *Nature Genetics* 32:502–508.

Smith, A. D.; Sumazin, P.; Xuan, Z.; and Zhang, M. Q. 2006. DNA motifs in human and mouse proximal promoters predict tissue-specific expression. *Proceedings of the National Academy of Sciences of the United States of America* 103(16):6275–6280.

Smyth, G. K. 2004. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology* 3:Article3.

Snoussi, E. H. 1998. Necessary conditions for multistationarity and stable periodicity. *Journal of Biological Systems* 6(01):3–9.

Song, J.; Johnson, W.; Zhu, X.; Zhang, X.; Li, W.; Manrai, A.; Liu, J.; Chen, R.; and Liu, X. 2007. Model-based analysis of two-color arrays (MA2C). *Genome Biology* 8:R178.

Sopko, R.; Foos, M.; Vinayagam, A.; Zhai, B.; Binari, R.; Hu, Y.; Randklev, S.; Perkins, L. A.; Gygi, S. P.; and Perrimon, N. 2014. Combining genetic perturbations and proteomics to examine kinase-phosphatase networks in drosophila embryos. *Developmental Cell* 31(1):114–127.

Stormo, G. D. 2000. DNA binding sites: representation and discovery. *Bioinformatics* 16(1):16–23.

Tesar, P. J.; Chenoweth, J. G.; Brook, F. A.; Davies, T. J.; Evans, E. P.; Mack, D. L.; Gardner, R. L.; and McKay, R. D. 2007. New cell lines from mouse epiblast share defining features with human embryonic stem cells. *Nature* 448(7150):196–199.

Theunissen, T. W.; Powell, B. E.; Wang, H.; Mitalipova, M.; Faddah, D. A.; Reddy, J.; Fan, Z. P.; Maetzel, D.; Ganz, K.; Shi, L.; Lungjangwa, T.; Imsoonthornruksa, S.; Stelzer, Y.; Rangarajan, S.; D'Alessio, A.; Zhang, J.; Gao, Q.; Dawlaty, M. M.; Young, R. A.; Gray, N. S.; and Jaenisch, R. 2014. Systematic identification of culture conditions for induction and maintenance of naive human pluripotency. *Cell Stem Cell* 15(4):471–487.

Thomson, J. A.; Itskovitz-Eldor, J.; Shapiro, S. S.; Waknitz, M. A.; Swiergiel, J. J.; Marshall, V. S.; and Jones, J. M. 1998. Embryonic stem cell lines derived from human blastocysts. *Science* 282(5391):1145–1147.

Tibshirani, R.; Walther, G.; and Hastie, T. 2001. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63(2):411–423.

Timmermann, A.; Pflanz, S.; Grotzinger, J.; Kuster, A.; Kurth, I.; Pitard, V.; Heinrich, P. C.; and Muller-Newen, G. 2000. Different epitopes are required for gp130 activation by interleukin-6, oncostatin M and leukemia inhibitory factor. *FEBS Letters* 468(2-3):120–124.

Tomilin, A.; Reményi, A.; Lins, K.; Bak, H.; Leidel, S.; Vriend, G.; Wilmanns, M.; and Schöler, H. R. 2000. Synergism with the coactivator OBF-1 (OCA-B, BOB-1) is mediated by a specific POU dimer configuration. *Cell* 103(6):853–864.

Trapnell, C., and Salzberg, S. L. 2009. How to map billions of short reads onto genomes. *Nature Biotechnology* 27(5):455–457.

van Iersel, M. P.; Pico, A. R.; Kelder, T.; Gao, J.; Ho, I.; Hanspers, K.; Conklin, B. R.; and Evelo, C. T. 2010. The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services. *BMC Bioinformatics* 11(1):5.

Vandesompele, J.; De Preter, K.; Pattyn, F.; Poppe, B.; Van Roy, N.; De Paepe, A.; and Speleman, F. 2002. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biology* 3(7):research0034.1–0034.11.

Vaquerizas, J. M.; Kummerfeld, S. K.; Teichmann, S. A.; and Luscombe, N. M. 2009. A census of human transcription factors: function, expression and evolution. *Nature Reviews Genetics* 10(4):252–263.

Västrik, I.; D'Eustachio, P.; Schmidt, E.; Joshi-Tope, G.; Gopinath, G.; Croft, D.; de Bono, B.; Gillespie, M.; Jassal, B.; Lewis, S.; Matthews, L.; Wu, G.; Birney, E.; and Stein, L. 2007. Reactome: a knowledge base of biologic pathways and processes. *Genome Biology* 8(3):R39.

Wagner, A. 2002. Estimating coarse gene network structure from large-scale gene perturbation data. *Genome Research* 12(2):309–315.

Walker, E.; Ohishi, M.; Davey, R. E.; Zhang, W.; Cassar, P. A.; Tanaka, T. S.; Der, S. D.; Morris, Q.; Hughes, T. R.; Zandstra, P. W.; and Stanford, W. L. 2007.

Prediction and testing of novel transcriptional networks regulating embryonic stem cell self-renewal and commitment. *Cell Stem Cell* 1(1):71–86.

Wang, Y.; Barbacioru, C.; Hyland, F.; Xiao, W.; Hunkapiller, K. L.; Blake, J.; Chan, F.; Gonzalez, C.; Zhang, L.; and Samaha, R. R. 2006. Large scale real-time PCR validation on gene expression measurements from two commercial long-oligonucleotide microarrays. *BMC Genomics* 7(1):59.

Wang, Z.; Gerstein, M.; and Snyder, M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* 10(1):57–63.

Wang, J.; Levasseur, D.; and Orkin, S. 2008. Requirement of Nanog dimerization for stem cell self-renewal and pluripotency. *Proceedings of the National Academy of Sciences of the United States of America* 105:6326–6331.

Wasserman, W. W.; Palumbo, M.; Thompson, W.; Fickett, J. W.; and Lawrence, C. E. 2000. Human-mouse genome comparisons to locate regulatory sites. *Nature Genetics* 26(2):225–228.

Wilkes, T.; Laux, H.; and Foy, C. A. 2007. Microarray data quality-review of current developments. *OMICS A Journal of Integrative Biology* 11(1):1–13.

Williams, D. C.; Cai, M.; and Clore, G. M. 2004. Molecular basis for synergistic transcriptional activation by Oct1 and Sox2 revealed from the solution structure of the 42-kDa Oct1.Sox2.Hoxb1-DNA ternary transcription factor complex. *The Journal of Biological Chemistry* 279(2):1449–1457.

Wolfe, C. J.; Kohane, I. S.; and Butte, A. J. 2005. Systematic survey reveals general applicability of g̈uilt-by-associationẅithin gene coexpression networks. *BMC Bioinformatics* 6(1):227.

Wu, Z.; Rosen, E. D.; Brun, R.; Hauser, S.; Adelmant, G.; Troy, A. E.; McKeon, C.; Darlington, G. J.; and Spiegelman, B. M. 1999. Cross-regulation of C/EBPα and PPARγ controls the transcriptional pathway of adipogenesis and insulin sensitivity. *Molecular Cell* 3(2):151–158.

Wu, C.; Schwartz, J.-M.; and Nenadic, G. 2013. PathNER: a tool for systematic identification of biological pathway mentions in the literature. *BMC Systems Biology* 7(Suppl 3):S2.

Xu, Z.; Yu, S.; Hsu, C. H.; Eguchi, J.; and Rosen, E. D. 2008. The orphan nuclear receptor chicken ovalbumin upstream promoter-transcription factor II is a critical regulator of adipogenesis. *Proceedings of the National Academy of Sciences of the United States of America* 105(7):2421–2426.

Xu, N.; Papagiannakopoulos, T.; Pan, G.; Thomson, J. A.; and Kosik, K. S. 2009. MicroRNA-145 regulates OCT4, SOX2, and KLF4 and represses pluripotency in human embryonic stem cells. *Cell* 137:647–658.

Yeo, J.-C., and Ng, H.-H. 2013. The transcriptional regulation of pluripotency. *Cell research* 23(1):20–32.

Yuryev, A.; Mulyukov, Z.; Kotelnikova, E.; Maslov, S.; Egorov, S.; Nikitin, A.; Daraselia, N.; and Mazo, I. 2006. Automatic pathway building in biological association networks. *BMC Bioinformatics* 7(1):171.

Zhang, Y.; Szustakowski, J.; and Schinke, M. 2009. Bioinformatics analysis of microarray data. In *Cardiovascular Genomics*. Springer. 259–284.

SUMMARY IN ESTONIAN

Suuremahuliste andmete kasutamine geenidevaheliste seoste leidmiseks

Elusorganismi genoomis kodeeritud geenid määravad ära millistest RNA ja valgu molekulidest organism koosneb. Kuid geenide teadmine ei kirjelda veel ära, kuidas organism toimib, millal ja kuidas erinevad geenide produktid avalduvad ja mida need teevad. Geenide omavahelised seosed ilmnevad näiteks koos avaldumises, valkude vaheliste seostena, valk-DNA seostena ja erinevate bioloogiliste radade koostoimes. Kõigi nende seoste kirjeldamine ning võimaluse korral mudeldamine aitab meil sügavamalt mõista bioloogilise organismi olemust. Samuti on arvutuslike meetoditega võimalik katsetada, kas meie mudelid kirjeldavad bioloogilist uurimisobjekti piisavalt täpsed.

Seoses suure läbilaskevõimega tehnoloogiate odavnemisega on erinevate bioloogiliste protsesside mõõtmise kaudu saadavad andmemahud praegu pidevas kasvutrendis. Sellest tulenevalt on arvutusliku bioloogia üheks peamiseks eesmärgiks käesoleval ajal võimalikult automaatsete vahenditega leida keerukaid geenide vahelisi seoseid. Selleks saab ära kasutada laialdaselt levima hakanud suure andmemahuga eksperimentide tulemusi nagu geenikiipide abil mõõdetud geenisekspressiooni andmed, DNA ning RNA järjestamised ja valkude seondumised DNA-ga. Andmemahtude tohutu kasv nõuab uusi meetodeid toorandmete analüüsiks, omavaheliseks kombineerimiseks ning tulemuste visualiseerimiseks.

Käesoleva doktoritöö annab esmalt ülevaate geenide aktiivsuse mõõtmise eksperimentaalsetest ja arvutuslikest meetoditest ning geenihulkade funktsionaalsest kirjeldamisest. Töös käsitletakse lähemalt embrüonaalseid- ja rasvatüvirakke ning nende peamisi regulaatoreid. Samuti käsitletakse bioloogilisi võrgustikke, geenidevaheliste seoste leidmist ning enam levinud võrgu motiive. Lühidalt käsitletakse ka bioloogilisi radasid ning nende erinevust regulatoorsetest võrgustikest. Sellele järgneb võrgustike mudeldamise lühiülevaade. Kirjanduse ülevaate viimases osas käsitletakse suuremahuliste ning eriilmeliste andmete integreerimiseks vajalikke samme.

Eksperimentaalses osas on püstitatud järgmised eesmärgid:

1. analüüsida ja hinnata geenisekspressiooni andmestike kasutamise võimalusi bioloogiliste radade kirjeldamisel,

2. luua metoodika sarnase ekspressioonimustri ja funktsionaalsusega geenide automaatseks leidmiseks ning visualiseerimiseks suurtest geenidest ekspressiooni andmestikest
3. tuvastada embrüonaalsete tüvirakkude ja rasvarakkude regulatoorseid elemente kasutades erinevat tüüpi eksperimentaalsete andmete kombineerimist
4. prioritiseerida pluripotentsuse eest vastutavaid kandidaatgeene kasutades kvalitatiivset mudeldamist

Uurimustöö peamised tulemused on järgnevad:

1. *Bioloogiliste radade komponentide leidmiseks on piiratud võimalused, kui kasutakse ainult geenidest ekspressiooni andmeid.* Bioloogiliste radade taasloomisel, kasutades geenidest ekspressiooni andmestikke, on esmaseks piiranguks radade tüübid. Valgukomplekside ning metaboolsete radade komponentide ekspressiooni sarnasus on oluliselt suurem kui signaaliradade liikmete vahel. Oluline on kasutada mitte kõiki raja komponente vaid tasub keskenduda raja alamosadele, mis sagedamini on sarnaselt reguleeritud kui raja komponendid üldiselt. Eriti kehtib see suurte ning keeruliste signaaliradade puhul.

2. *Automaatne grupeerimis- ja visualiseerimismetoodika aitab leida funktsionaalselt sarnaseid geenide alamhulki suuremahulistest ekspressiooni andmestikest.* Suuremahuliste mikrokiibi andmete esmaseks automaatseks analüüsiks saab kasutada ligikaudse hierarhilise klasterdamise ja funktsionaalse rikastamise kombineerimist. Sel viisil on kerge välja tuua geenide alamhulki, mis on sarnase ekspressiooniga ning ühiste funktsionaalsete kirjeldustega. Seda tüüpi visuaalne ülevaade aitab märgata huvitavaid geenigruppe, mida teiste meetoditega edasi uurida.

3. *Kombineerides DNA järjestuste konserveerumisinfot arvutuslike motiiviotsimisalgoritmidega on võimalik esile tõsta rasvatüvirakkude kõige olulisemaid transkriptsiooni regulaatoreid.* Evolutsiooniliselt konserveerunud bioloogiliste funktsioonide uurimisel on kasulik arvesse võtta DNA konserveerumisinfot. Regulatoorsete motiivide kombineerimine konserveerumisinfoga võimaldas leida rasvatüvirakkude differentseerumise kulgu mõjutavaid regulaatoreid.

4. *Ühendades motiiviotsingud ja regulatoorsete valkude DNA-le seondumise info, on võimalik eristada otseseid ning kaudseid märklaudgeene.* Inimese embrüonaalsete tüvirakkude ühe keskse regulaatorvalgu OCT4 märklaudgeene saab jagada kuude peamisse gruppi. Enamasti seondub OCT4 otse sihtmärgile või seondub koos tuntud partneriga SOX2. Lisaks reguleerib OCT4 sihtmärke läbi vähelevinud homodimeersete komplekside. Osad sihtmärk geenid on reguleeritud tundmatu partneri või tundmatu OCT4 motiivi kaudu.

5. *Suuremahulistele andmetele on integreerimise ja omavahel võrreldavaks tegemisega võimalik anda lisaväärtust.* Koondades kokku ja tehes avalikkusele ligipääsetavaks erinevate publikatsioonide lisafailides avaldatud info, on võimalik leida geenide vahelisi seoseid, mida eraldiseisvaid andmeid analüüsides ei saa teostada.

6. *Andmete integreerimine ning arvutuslik mudeldamine võimaldab leida uusi olulisi geeniregulatsiooni võrgustiku regulaatoreid.* Kombineerides hästi annoteeritud geenidevaheliste seoste võrgustikku suuremahuliste häirituse katsetega, on võimalik koostada suuri geenivõrgustikke. Erinevate filtreerimis- ja grupeerimismeetoditega on võimalik saadud võrgustikku kärpida, keskendumaks hõlpsamini valideeritavatele geenidele. Mudeldamisega on võimalik lühikese aja jooksul viia läbi arvutuslikke rakkude mõjutamise eksperimente, mis laboratoorsete katsete puhul nõuaks palju ajalist ja rahalist resurssi.

   Käesolevas uurimistöös on näidatud erinevaid bioinformaatilisi viise kuidas suuremahuliste ning eritüübiliste eksperimentaalsete andmete kombineerimist saab rakendada geenidevaheliste seoste leidmiseks. Töö põhiline tulemus on, et erinevat tüüpi suuremahuliste andmestike integreerimine võimaldab leida selliseid geenidevahelisi seoseid, mida ei oleks võimalik leida kui analüüsiksime üht andmestikku korraga.

ACKNOWLEDGEMENTS

Bioinformatics is a fascinating world combining the great challenges of biology with capabilities of computer science. I am glad that I have had the chance to learn from the *Suur juht ja Õpetaja* Jaak how to survive in and drive the complicated world of research. I am also very grateful to professor Juhan Sedman who was willing to be my co-supervisor, so I could pursue my studies in bioinformatics at IMCB. I am thankful to professor Andres Metspalu for providing helpful suggestions at the start line of my road for becoming a scientist.

There would be no research without the adequate monetary support. The work in this thesis was supported by Estonian Science Foundation grant no. 7437 (MEM), Centre of Excellence in Computer Science (EXCS), the Estonian Ministry of Education and Research (Data Science Methods and Applications, IUT 34-4), target funding project SF0180008s12, Egide Parrot PHC Programme N^{o} 20679QJ, EU FP6 project ENFIN (LSHG-CT-2005-518254) and FP7 project AgedBrainSYS-BIO (FP7-HEALTH-2012-INNO-305299). I would like to also thank Marie Curie project STEMCAM (FP7-PEOPLE-2009-IAPP-251186) that supported my post-doctoral studies at the University of Geneva during 2011-2013.

I am especially thankful to European Commission for funding Experimental Network for Functional INtegration project called ENFIN in 2006. During that project I grew up as a scientist and learned the value of great collaborations and became friends with colleagues I enjoy working and communicating ever since.

I have spent endless days and nights around the globe with fellow grad students discussing the good and bad sides of science, grad school, bioinformatics and biology. It was a pleasure to academically grow up with you guys - Jüri , Raivo and Lemps! Since the beginning of 2014 I have worked in the best work environment ever and enjoyed lots of fun at work. Thanks for the friendly and supportive atmosphere, Ü2 team!

My dissertation is based on collaboration with Abhishek, Darja, Ioannis, James, Jüri, Kostja, Lemps, Marc, Meelis, Nathalie, Pascal, Phaedra and Raed from various labs around the world. Thanks for sharing ideas and knowledge, and making it happen.

I would like to sincerely thank my high school biology, chemistry and mathematics teachers Helina Metsalu, Ilme Pipar and Väino Soomets for encouraging my curiosity. These great teachers kept my restless mind busy with extra exercises, gave me the freedom of learning and heartened sharing my knowledge with peers. They started my love for seeking knowledge and set the cornerstones for me to become a scientist.

These thesis would not have been finished without Prof. Ioannis Xenarios' encouragement and everlasting support. I appreciate all the discussions, enthusiasm and motivation he has provided over the years. I am truly lucky to have such a Mentor. Merci beacoup, Ioannis!

I would like to thank Elena, Lemps, Sten and Katrin for pointing out the ambiguities and lack of detail in the draft version of these thesis. I am sincerely grateful to Liina. Her unhealthy interest towards commas and other nuances of grammar improved considerably readability of this thesis.

I am glad for my friends who have finished this endless misery before me and shown that graduation is indeed possible. Your examples fed the almost diminishing hope in me over the last years that seemed like a never-ending story. Thanks for showing the way Age, Anneleen, Dan, Hugo, Jüri, Katrin, Liina, Marcin, Meelis, Raivo, Sofie, Sten, Volli and Wijnand.

I am happy to have wonderful friends who were there to support me when an evil encephalitis carrying tick challenged my body and mind in spring 2009. The disease itself and recovery from it took awfully long time. Thanks to your support - Kata, Kairit & Volli, Lemps & Reet and Sten fighting with the disease was more joyful. Jaak, thanks for the confidence and giving me enough time to recover.

It is great to have friends outside of science who every now and then remind me that there is life beyond the endless grant applications, delivery documents, article writing and student supervision. All those waffle and board game evenings, geocaching trips and wandering in the mountains has helped me to keep the balance in life. Thanks for the companionship!

Last but not least, I owe a huge gratitude to my family who has supported me in all the crazy ideas, endless work weeks, countless trips (and provided the unofficial taxi rides from and to the Lennujaam) and accepted those too rare and too short visits.

My dearest ones, thanks a lot for always believing in me!

PUBLICATIONS

CURRICULUM VITAE

Name: Hedi Peterson
Date of birth: 12.04.1982
Citizenship: Estonian
Address: Institute of Computer Science
Liivi 2, Tartu 50409
E-mail: hedi.peterson@ut.ee

Education:

2006–….	University of Tartu, bioinformatics, PhD student
2005–2006	University of Uppsala, bioinformatics exchange student
2004–2006	University of Tartu, bioinformatics, M.Sc
2000–2004	University of Tartu, bioinformatics, B.Sc

Professional career:

2013–	University of Tartu, researcher and ELIXIR Estonia coordinator
2013–	Quretec, lead researcher
2011–2013	University of Geneva, Marie Curie Postdoctoral researcher
2008–2011	University of Tartu, extraordinary researcher
2007–2011	Quretec, researcher
2006–2007	EGeen, researcher
2006–2006	Estonian Biocentre, extraordinary researcher
2005–2005	Estonian Biocentre, programmer

Main Fields of Research:

Gene regulation mechanisms, network reconstruction and modeling methods. Gene regulation of embryonic stem cells and early differentiation. Toxicity testing using *in vitro* models. Bioinformatics tools and services for analysing high-throughput data.

Publications:

1. Érika Cosset, Yannick Martinez, Olivier Preynat-Seauve, Johannes-Alexander Lobrinus, Caroline Tapparel, Samuel Cordey, **Hedi Peterson**, Tom Petty, Marilena Colaianna, Vannary Tieng, Diderik Tirefort, Andras Dinnyes, Michel Dubois-Dauphin, Laurent Kaiser, Karl-Heinz Krause Human. Three-dimensional engineered neural tissue reveals cellular and molecular events following cytomegalovirus infection. (2015). *Biomaterials*, 53:296–308.
2. Zeynab Nayernia, Laurent Turchi, Erika Cosset, **Hedi Peterson**, Valérie Dutoit, Pierre-Yves Dietrich, Diderik Tirefort, Hervé Chneiweiss, Johannes-Alexander Lobrinus, Karl-Heinz Krause, Thierry Virolle, Olivier Preynat-Seauve. The relationship between brain tumor cell invasion of engineered

neural tissues and in vivo features of glioblastoma. (2013). *Biomaterials* 34.33:8279–8290.

3. **Hedi Peterson***, Raed Abu Dawud*, Abhishek Garg, Ying Wang, Jaak Vilo, Ioannis Xenarios, and James Adjaye. Qualitative modeling identifies IL-11 as a novel regulator in maintaining self-renewal in human pluripotent stem cells. (2013). *Frontiers in physiology* 4:303–303.
4. Anne K Krug*, Raivo Kolde*, John A Gaspar*, Eugen Rempel*, Nina V Balmer, Kesavan Meganathan, Kinga Vojnits, Mathurin Baquié, Tanja Waldmann, Roberto Ensenat-Waser, Smita Jagtap, Richard M Evans, Stephanie Julien, **Hedi Peterson**, Dimitra Zagoura, Suzanne Kadereit, Daniel Gerhard, Isaia Sotiriadou, Michael Heke, Karthick Natarajan, Margit Henry, Johannes Winkler, Rosemarie Marchan, Luc Stoppini, Sieto Bosgra, Joost Westerhout, Miriam Verwei, Jaak Vilo, Andreas Kortenkamp, Jürgen Hescheler, Ludwig Hothorn, Susanne Bremer, Christoph van Thriel, Karl-Heinz Krause, Jan G Hengstler, Jörg Rahnenführer, Marcel Leist, Agapios Sachinidis. Human embryonic stem cell-derived test systems for developmental neurotoxicity: a transcriptomics approach. (2013). *Archives of toxicology* 87.1:123–143.
5. Liina Tserel, Toomas Runnel, Kai Kisand, Maire Pihlap, Lairi Bakhoff, Raivo Kolde, **Hedi Peterson**, Jaak Vilo, Pärt Peterson, Ana Rebane. MicroRNA epxression profiles of human blood monocyte-derived dendritic cells and macrophages reveal miR-511 as putative positive regulator of Toll-like receptor 4. (2011*). The Journal of Biological Chemistry* 286(30): 26487–95
6. Liina Tserel, Raivo Kolde, Ana Rebane, Kai Kisand, Tõnis Org, **Hedi Peterson**, Jaak Vilo, Pärt Peterson. Gene Expression and Epigenetic Changes in Differentiation of Human Macrophages and Dendritic Cells. (2010). *Scandinavian Journal of Immunology* 71 (6), 526–526
7. Ana Rebane, Liina Tserel, Kai Kisand, Lairi Backhoff, Toomas Runnel, Raivo Kolde, **Hedi Peterson**, Jaak Vilo and Pärt Peterson. Balanced Expression of miRNA-s Secures Differentiation and Maturation of Human Blood Monocyte Derived Dendritic Cells. (2010). *Scandinavian Journal of Immunology*, 71(6), 512–513.
8. Nathalie Billon, Raivo Kolde*, Jüri Reimand*, Miguel C Monteiro, Meelis Kull, **Hedi Peterson**, Konstantin Tretyakov, Priit Adler, Brigitte Wdziekonski, Jaak Vilo and Christian Dani. Comprehensive transcriptome analysis of mouse embryonic stem cell adipogenesis unravels new processes of adipocyte development. (2010). *Genome Biology*, 11:R80
9. Liina Tserel, Raivo Kolde, Ana Rebane, Kai Kisand, Tõnis Org, **Hedi Peterson**, Jaak Vilo and Pärt Peterson. Genome-wide promoter analysis of histone modifications in human monocyte-derived antigen presenting cells. (2010). *BMC Genomics*, 11:642
10. Marc Jung*, **Hedi Peterson***, Lukas Chavez, Pascal Kahlem, Hans Lehrach, Jaak Vilo, James Adjaye. A data integration approach to mapping OCT4 gene regulatory networks operative in embryonic stem cells and

embryonal carcinoma cells. (2010). *PLOS ONE 5(5): e10709. doi:10.1371/ journal.pone.0010709*

11. Priit Adler*, Raivo Kolde*, Meelis Kull, Aleksander Tkachenko, **Hedi Peterson**, Jüri Reimand, Jaak Vilo. Mining for coexpression across hundreds of datasets using novel rank aggregation and visualization methods. (2009). *Genome Biology* 2009;10(12):R139.
12. Darja Krushevskaja, **Hedi Peterson**, Jüri Reimand, Meelis Kull, Jaak Vilo. Vishic – hierarchical functional enrichment analysis of microarray data. (2009). *Nucleic Acids Research,* Vol 37, W587–W592
13. Priit Adler*, **Hedi Peterson***, Phaedra Agius, Jüri Reimand, Jaak Vilo. Ranking genes by their co-expression to subsets of pathway members. (2009). *Annals of the New York Academy of Sciences* Mar;1158:1–13.
14. Jüri Reimand*, Laur Tooming*, **Hedi Peterson**, Priit Adler, Jaak Vilo: GraphWeb: mining heterogeneous biological networks for gene modules with functional significance. (2008). *Nucleic Acids Research.* Jul 1;36:W452–9.
15. Obi Griffith et al (The Open Regulatory Annotation Consortium, incl. **Hedi Peterson**, Priit Adler). OregAnno: an open-access community-driven resource for regulatory annotation. (2008). *Nucleic Acids Research*, Jan;36 D107–13
16. Priit Adler*, Jüri Reimand*, Jürgen Jänes, Raivo Kolde, **Hedi Peterson**, Jaak Vilo. KEGGanim: pathway animations for high-throughput data. (2008). *Bioinformatics*, Feb 15;24(4):588–90.
17. Jüri Reimand, Meelis Kull, **Hedi Peterson**, Jaanus Hansen, Jaak Vilo. g:Profiler – a web-based toolset for functional profiling of gene lists from large-scale experiments (2007) *Nucleic Acids Research*, Vol. 35, W193–W200.

* – Contributed equally

Scholarships:

1. EITSA mobility stipend for attending „Genome Informatics“ conference at CSHL 2008
2. Gertrud Thelin's scholarship for University of Tartu students for studing at University of Uppsala, spring semester 2006/2007
3. Kristjan-Jaak fellowship for studing at the University of Uppsala, 2005/06
4. Human Research Potential & the Socio-economic Knowledge Base: Access to Research Infrastructures (ARI), 09.03-10.04.2005 Uppsala
5. Estonian Olympic Commitee scholarship for Higher Education studies for National Level athletes, 2004-2006
6. Travel fellowship for attending ISMB conference at Glasgow, July 2004
7. Travel fellowship for attending workshop at Geneva, September 2003

Teaching:

1. ELIXIR g:Profiler hands-on course, Tartu, 05.12.2014
2. Lectures in bioinformatics course, Tartu, autumn 2010
3. BIIT software seminar City of Hope, USA, 23.11.2010
4. EMBO Practical Course on Analysis and Informatics of Microarray Data, Hinxton, UK 20.10.2010
5. EMBO Practical Course on Analysis and Informatics of Microarray Data, Hinxton, UK 15.10.2009
6. 9th BioSapiens European School of Bioinformatics, Brussels, Belgium 30.01.2009
7. EBI Roadshow 2009, Tartu 07.-10.09.2009 (main organizer)
8. ENFIN-EMBRACE Workshop, Helsingi, Soome 06.10.2009

Supervised thesis

1. Andreas Ellervee, bachelor thesis "Automatic detection of intermediate genes responsible for signal mediation from high-throughput data" 2015)
2. Henry Mägi, bachelor thesis "RNA-seq pipeline" (2014)
3. Kaur Alasoo, bachelor thesis "Combining Support Vector Machines to Predict Novel Angiogenesis Genes" (2010) (The thesis got the First prizes from the Estonian Ministry of Education and Science from the Estonian Academy of Sciences in 2010)
4. Sten Ilmjärv, master's thesis "Researching promoter sequences on the basis of gene expression dynamics in mouse BxD lines" (2009)
5. Priidik Vaikla, bachelor thesis "Project management system by the example of BIIT scientific group" (2008)
6. Sten Ilmjärv, bachelor thesis "Describing transcription start sites in Saccharomyces cerevisiae" (2007)
7. Marten Teino, bachelor thesis "Scientific literature database for a research group" (2007)

Membership of academic societies:

2014–...	International Society of Systems Biology (ISSB, founding committee member; member of Training and Student Council)
2008–...	Estonian Academy of Young Scientists (ENTA)
2005–...	Society for Bioinformatics in Northern Europe (SOCBIN)

Public and social activities:

2015–	Member of EDAM ontology Advisory Board
2014–	President of Mountaineering Club "Firn"
2014	Co-author of Concise Encyclopedia of Bioinformatics and Computational Biology (2nd Edition, Wiley, 2014)
2007–2009	Lectures at Kristiine Gymnasium, incl. "Back to School" project

ELULOOKIRJELDUS

Nimi: Hedi Peterson
Sünniaeg: 12.04.1982
Kodakondsus: eestlane
Aadress: Arvutiteaduse instituut
Liivi 2, Tartu 50409
E-mail: hedi.peterson@ut.ee

Haridus:

2006– ...	Tartu Ülikool, bioinformaatika, doktorantuur
2005–2006	Uppsala Ülikool, vahetustudeng bioinformaatikas
2004–2006	Tartu Ülikool, bioinformaatika, M.Sc
2000–2004	Tartu Ülikool, bioinformaatika, B.Sc

Teenistuskäik:

2013–	University of Tartu, teadur ja ELIXIR Eesti koordinaator
2013–	Quretec, juhtivteadur
2011–2013	University of Geneva, Marie Curie järeldoktorant
2008–2011	University of Tartu, erakorraline teadur
2007–2011	Quretec, teadur
2006–2007	EGeen, teadur
2006–2006	Estonian Biocentre, erakorraline teadur
2005–2005	Estonian Biocentre, programmeerija

Peamised uurimisvaldkonnad:
Geeniregulatsiooni mehhanismid, geenivõrgustike loomise- ja modelleerimismeetodid. Embrüonaalsete tüvirakkude pluripotentsuse ja varajase differentseerumise geeniregulatsioon. Toksikoloogiline testimine kasutades *in vitro* rakumudeleid. Bioinformaatiliste tööriistade ja veebirakenduste loomine suure tootlikusega bioloogiliste andmete analüüsiks.

Publikatsioonid:

1. Érika Cosset, Yannick Martinez, Olivier Preynat-Seauve, Johannes-Alexander Lobrinus, Caroline Tapparel, Samuel Cordey, **Hedi Peterson**, Tom Petty, Marilena Colaianna, Vannary Tieng, Diderik Tirefort, Andras Dinnyes, Michel Dubois-Dauphin, Laurent Kaiser, Karl-Heinz Krause Human. Three-dimensional engineered neural tissue reveals cellular and molecular events following cytomegalovirus infection. (2015). *Biomaterials*, 53:296–308.
2. Zeynab Nayernia, Laurent Turchi, Erika Cosset, **Hedi Peterson**, Valérie Dutoit, Pierre-Yves Dietrich, Diderik Tirefort, Hervé Chneiweiss, Johannes-Alexander Lobrinus, Karl-Heinz Krause, Thierry Virolle, Olivier Preynat-Seauve. The relationship between brain tumor cell invasion of

engineered neural tissues and in vivo features of glioblastoma. (2013). *Biomaterials* 34.33:8279–8290.

3. **Hedi Peterson***, Raed Abu Dawud*, Abhishek Garg, Ying Wang, Jaak Vilo, Ioannis Xenarios, and James Adjaye. Qualitative modeling identifies IL-11 as a novel regulator in maintaining self-renewal in human pluripotent stem cells. (2013). *Frontiers in physiology* 4:303–303.
4. Anne K Krug*, Raivo Kolde*, John A Gaspar*, Eugen Rempel*, Nina V Balmer, Kesavan Meganathan, Kinga Vojnits, Mathurin Baquié, Tanja Waldmann, Roberto Ensenat-Waser, Smita Jagtap, Richard M Evans, Stephanie Julien, **Hedi Peterson**, Dimitra Zagoura, Suzanne Kadereit, Daniel Gerhard, Isaia Sotiriadou, Michael Heke, Karthick Natarajan, Margit Henry, Johannes Winkler, Rosemarie Marchan, Luc Stoppini, Sieto Bosgra, Joost Westerhout, Miriam Verwei, Jaak Vilo, Andreas Kortenkamp, Jürgen Hescheler, Ludwig Hothorn, Susanne Bremer, Christoph van Thriel, Karl-Heinz Krause, Jan G Hengstler, Jörg Rahnenführer, Marcel Leist, Agapios Sachinidis. Human embryonic stem cell-derived test systems for developmental neurotoxicity: a transcriptomics approach. (2013) .*Archives of toxicology* 87.1:123–143.
5. Liina Tserel, Toomas Runnel, Kai Kisand, Maire Pihlap, Lairi Bakhoff, Raivo Kolde, **Hedi Peterson**, Jaak Vilo, Pärt Peterson, Ana Rebane. MicroRNA epxression profiles of human blood monocyte-derived dendritic cells and macrophages reveal miR-511 as putative positive regulator of Toll-like receptor 4. (2011). *The Journal of Biological Chemistry* 286(30): 26487–95
6. Liina Tserel, Raivo Kolde, Ana Rebane, Kai Kisand, Tõnis Org, **Hedi Peterson**, Jaak Vilo, Pärt Peterson. Gene Expression and Epigenetic Changes in Differentiation of Human Macrophages and Dendritic Cells. (2010). *Scandinavian Journal of Immunology* 71 (6), 526–526
7. Ana Rebane, Liina Tserel, Kai Kisand, Lairi Backhoff, Toomas Runnel, Raivo Kolde, **Hedi Peterson**, Jaak Vilo and Pärt Peterson. Balanced Expression of miRNA-s Secures Differentiation and Maturation of Human Blood Monocyte Derived Dendritic Cells. (2010). *Scandinavian Journal of Immunology*, 71(6), 512–513.
8. Nathalie Billon, Raivo Kolde*, Jüri Reimand*, Miguel C Monteiro, Meelis Kull, **Hedi Peterson**, Konstantin Tretyakov, Priit Adler, Brigitte Wdziekonski, Jaak Vilo and Christian Dani. Comprehensive transcriptome analysis of mouse embryonic stem cell adipogenesis unravels new processes of adipocyte development. (2010). *Genome Biology*, 11:R80
9. Liina Tserel, Raivo Kolde, Ana Rebane, Kai Kisand, Tõnis Org, **Hedi Peterson**, Jaak Vilo and Pärt Peterson. Genome-wide promoter analysis of histone modifications in human monocyte-derived antigen presenting cells. (2010). *BMC Genomics*, 11:642
10. Marc Jung*, **Hedi Peterson***, Lukas Chavez, Pascal Kahlem, Hans Lehrach, Jaak Vilo, James Adjaye. A data integration approach to mapping OCT4 gene regulatory networks operative in embryonic stem cells and

embryonal carcinoma cells. (2010). *PLOS ONE 5(5): e10709. doi:10.1371 /journal.pone.0010709*

11. Priit Adler*, Raivo Kolde*, Meelis Kull, Aleksander Tkachenko, **Hedi Peterson**, Jüri Reimand, Jaak Vilo. Mining for coexpression across hundreds of datasets using novel rank aggregation and visualization methods. (2009). *Genome Biology* 2009;10(12):R139.
12. Darja Krushevskaja, **Hedi Peterson**, Jüri Reimand, Meelis Kull, Jaak Vilo. Vishic – hierarchical functional enrichment analysis of microarray data. (2009). *Nucleic Acids Research*, Vol 37, W587–W592
13. Priit Adler*, **Hedi Peterson***, Phaedra Agius, Jüri Reimand, Jaak Vilo. Ranking genes by their co-expression to subsets of pathway members. (2009). *Annals of the New York Academy of Sciences* Mar; 1158:1–13.
14. Jüri Reimand*, Laur Tooming*, **Hedi Peterson**, Priit Adler, Jaak Vilo: GraphWeb: mining heterogeneous biological networks for gene modules with functional significance. (2008). *Nucleic Acids Research.* Jul 1;36:W452–9.
15. Obi Griffith et al (The Open Regulatory Annotation Consortium, incl. **Hedi Peterson**, Priit Adler). OregAnno: an open-access community-driven resource for regulatory annotation. (2008). *Nucleic Acids Research*, Jan;36 D107–13
16. Priit Adler*, Jüri Reimand*, Jürgen Jänes, Raivo Kolde, **Hedi Peterson**, Jaak Vilo. KEGGanim: pathway animations for high-throughput data. (2008). *Bioinformatics*, Feb 15;24(4):588–90.
17. Jüri Reimand, Meelis Kull, **Hedi Peterson**, Jaanus Hansen, Jaak Vilo. g:Profiler – a web-based toolset for functional profiling of gene lists from large-scale experiments (2007) *Nucleic Acids Research*, Vol. 35, W193-W200.

* – jagatud (esi)autorlus

Saadud stipendiumid:

1. EITSA mobiilsusstipendium osalemiseks “Genome Informatics” konverentsil CSHL-s 2008
2. Gertrud Thelin’s stipendium Tartu Ülikooli tudengitele õppimiseks Uppsala Ülikoolis, kevadsemester 2006/2007
3. Kristjan-Jaagu stipendium õppeaastaks 2005/2006 Uppsala Ülikoolis
4. Teadusuuringute inimpotentsiaali ja sotsiaal-majandusliku teabebaas: juurdepääs teadustöö infrastruktuuridele (ARI), 09.03–10.04.2005 Uppsala
5. Eesti olümpiakomitee kõrghariduse stipendium tippsportlastele, 2004–2006
6. Reisistipendium Glasgow ISMB konverentsil osalemiseks 2004. juulis
7. Reisistipendium Genfi töökojal osalemiseks 2003. septembris

Õppetöö:

1. ELIXIR g:Profiler kursus, Tartu, 05.12.2014
2. Loengud aines Bioinformaatika, Tartu, sügis 2010
3. BIIT tööriistade seminar City of Hope, USA, 23.11.2010
4. EMBO Praktiline kursus mikrokiibiandmete analüüsiks, Hinxton, Suurbrittannia 20.10.2010
5. EMBO Praktiline kursus mikrokiibiandmete analüüsiks, Hinxton, Suurbrittannia 15.10.2009
6. 9th BioSapiens Euroopa Bioinformaatika kool, Brüssel, Belgia 30.01. 2009
7. EBI Ringreis 2009, Tartu 07.–10.09.2009 (peaorganisaator)
8. ENFIN-EMBRACE töökoda, Helsingi, Soome 06.10.2009

Juhendatud väitekirjad

1. Andreas Ellervee, bakalaureusetöö "Geeniregulatsiooni signaali vahendavate geenide automaatne leidmine suuremahulistest andmetest" (2015)
2. Henry Mägi, bakalaureusetöö "RNA-seq andmete analüüsi töövoog" (2014)
3. Kaur Alasoo, bakalaureusetöö "Combining Support Vector Machines to Predict Novel Angiogenesis Genes" (2010) (The thesis got the First prizes from the Estonian Ministry of Education and Science from the Estonian Academy of Sciences in 2010)
4. Sten Ilmjärv, magistritöö "Promootorjärjestuste uurimine BxD hiireliini geenikspressiooni dünaamika alusel" (2009)
5. Priidik Vaikla, bakalaureusetöö "Projektijuhtimistarkvara loomine BIIT teadusrühma näitel" (2008)
6. Sten Ilmjärv, bakalaureusetöö "Transkriptsiooni algussaitide kirjeldamine pagaripärmil" (2007)
7. Marten Teino, bakalaureusetöö "Uurimisrühmasisene teaduskirjanduse andmebaas" (2007)

Kuulumine erialaliitudesse:

2014–...	Rahvusvaheline süsteemibioloogia ühing (ISSB, (taas)asutaja-komitee liige; Koolitus- ja üliõpilasesinduskomitee liige)
2008–...	Eesti Noorte Teadlaste Akadeemia liige (ENTA)
2005–...	Põhjamaade Bioinformaatika Ühingu liige (SOCBIN)

Ühiskondlik ja publitsistlik tegevus:

2015–	EDAM ontoloogia konsultatiivnõukogu liige
2014–	Alpiklubi "Firn" president
2014	Concise Encyclopedia of Bioinformatics and Computational Biology kaasautor (2nd Edition, Wiley, 2014)
2007–2009	Loengud Kristiine Gümnaasiumis, k.a. "Tagasi kooli" projekti raames

DISSERTATIONES BIOLOGICAE UNIVERSITATIS TARTUENSIS

1. **Toivo Maimets**. Studies of human oncoprotein p53. Tartu, 1991, 96 p.
2. **Enn K. Seppet**. Thyroid state control over energy metabolism, ion transport and contractile functions in rat heart. Tartu, 1991, 135 p.
3. **Kristjan Zobel**. Epifüütsete makrosamblike väärtus õhu saastuse indikaatoritena Hamar-Dobani boreaalsetes mägimetsades. Tartu, 1992, 131 lk.
4. **Andres Mäe**. Conjugal mobilization of catabolic plasmids by transposable elements in helper plasmids. Tartu, 1992, 91 p.
5. **Maia Kivisaar**. Studies on phenol degradation genes of *Pseudomonas* sp. strain EST 1001. Tartu, 1992, 61 p.
6. **Allan Nurk**. Nucleotide sequences of phenol degradative genes from *Pseudomonas sp.* strain EST 1001 and their transcriptional activation in *Pseudomonas putida.* Tartu, 1992, 72 p.
7. **Ülo Tamm**. The genus *Populus* L. in Estonia: variation of the species biology and introduction. Tartu, 1993, 91 p.
8. **Jaanus Remme**. Studies on the peptidyltransferase centre of the *E.coli* ribosome. Tartu, 1993, 68 p.
9. **Ülo Langel**. Galanin and galanin antagonists. Tartu, 1993, 97 p.
10. **Arvo Käärd**. The development of an automatic online dynamic fluorescense-based pH-dependent fiber optic penicillin flowthrought biosensor for the control of the benzylpenicillin hydrolysis. Tartu, 1993, 117 p.
11. **Lilian Järvekülg**. Antigenic analysis and development of sensitive immunoassay for potato viruses. Tartu, 1993, 147 p.
12. **Jaak Palumets**. Analysis of phytomass partition in Norway spruce. Tartu, 1993, 47 p.
13. **Arne Sellin**. Variation in hydraulic architecture of *Picea abies* (L.) Karst. trees grown under different enviromental conditions. Tartu, 1994, 119 p.
13. **Mati Reeben**. Regulation of light neurofilament gene expression. Tartu, 1994, 108 p.
14. **Urmas Tartes**. Respiration rhytms in insects. Tartu, 1995, 109 p.
15. **Ülo Puurand.** The complete nucleotide sequence and infections *in vitro* transcripts from cloned cDNA of a potato A potyvirus. Tartu, 1995, 96 p.
16. **Peeter Hõrak**. Pathways of selection in avian reproduction: a functional framework and its application in the population study of the great tit (*Parus major*). Tartu, 1995, 118 p.
17. **Erkki Truve**. Studies on specific and broad spectrum virus resistance in transgenic plants. Tartu, 1996, 158 p.
18. **Illar Pata**. Cloning and characterization of human and mouse ribosomal protein S6-encoding genes. Tartu, 1996, 60 p.
19. **Ülo Niinemets**. Importance of structural features of leaves and canopy in determining species shade-tolerance in temperature deciduous woody taxa. Tartu, 1996, 150 p.

20. **Ants Kurg**. Bovine leukemia virus: molecular studies on the packaging region and DNA diagnostics in cattle. Tartu, 1996, 104 p.
21. **Ene Ustav**. E2 as the modulator of the BPV1 DNA replication. Tartu, 1996, 100 p.
22. **Aksel Soosaar**. Role of helix-loop-helix and nuclear hormone receptor transcription factors in neurogenesis. Tartu, 1996, 109 p.
23. **Maido Remm**. Human papillomavirus type 18: replication, transformation and gene expression. Tartu, 1997, 117 p.
24. **Tiiu Kull**. Population dynamics in *Cypripedium calceolus* L. Tartu, 1997, 124 p.
25. **Kalle Olli**. Evolutionary life-strategies of autotrophic planktonic microorganisms in the Baltic Sea. Tartu, 1997, 180 p.
26. **Meelis Pärtel**. Species diversity and community dynamics in calcareous grassland communities in Western Estonia. Tartu, 1997, 124 p.
27. **Malle Leht**. The Genus *Potentilla* L. in Estonia, Latvia and Lithuania: distribution, morphology and taxonomy. Tartu, 1997, 186 p.
28. **Tanel Tenson**. Ribosomes, peptides and antibiotic resistance. Tartu, 1997, 80 p.
29. **Arvo Tuvikene**. Assessment of inland water pollution using biomarker responses in fish *in vivo* and *in vitro*. Tartu, 1997, 160 p.
30. **Urmas Saarma**. Tuning ribosomal elongation cycle by mutagenesis of 23S rRNA. Tartu, 1997, 134 p.
31. **Henn Ojaveer**. Composition and dynamics of fish stocks in the gulf of Riga ecosystem. Tartu, 1997, 138 p.
32. **Lembi Lõugas**. Post-glacial development of vertebrate fauna in Estonian water bodies. Tartu, 1997, 138 p.
33. **Margus Pooga**. Cell penetrating peptide, transportan, and its predecessors, galanin-based chimeric peptides. Tartu, 1998, 110 p.
34. **Andres Saag**. Evolutionary relationships in some cetrarioid genera (Lichenized Ascomycota). Tartu, 1998, 196 p.
35. **Aivar Liiv**. Ribosomal large subunit assembly *in vivo*. Tartu, 1998, 158 p.
36. **Tatjana Oja**. Isoenzyme diversity and phylogenetic affinities among the eurasian annual bromes (*Bromus* L., Poaceae). Tartu, 1998, 92 p.
37. **Mari Moora**. The influence of arbuscular mycorrhizal (AM) symbiosis on the competition and coexistence of calcareous grassland plant species. Tartu, 1998, 78 p.
38. **Olavi Kurina**. Fungus gnats in Estonia (*Diptera: Bolitophilidae, Keroplatidae, Macroceridae, Ditomyiidae, Diadocidiidae, Mycetophilidae*). Tartu, 1998, 200 p.
39. **Andrus Tasa**. Biological leaching of shales: black shale and oil shale. Tartu, 1998, 98 p.
40. **Arnold Kristjuhan.** Studies on transcriptional activator properties of tumor suppressor protein p53. Tartu, 1998, 86 p.

41. **Sulev Ingerpuu.** Characterization of some human myeloid cell surface and nuclear differentiation antigens. Tartu, 1998, 163 p.
42. **Veljo Kisand.** Responses of planktonic bacteria to the abiotic and biotic factors in the shallow lake Võrtsjärv. Tartu, 1998, 118 p.
43. **Kadri Põldmaa.** Studies in the systematics of hypomyces and allied genera (Hypocreales, Ascomycota). Tartu, 1998, 178 p.
44. **Markus Vetemaa.** Reproduction parameters of fish as indicators in environmental monitoring. Tartu, 1998, 117 p.
45. **Heli Talvik.** Prepatent periods and species composition of different *Oesophagostomum* spp. populations in Estonia and Denmark. Tartu, 1998, 104 p.
46. **Katrin Heinsoo.** Cuticular and stomatal antechamber conductance to water vapour diffusion in *Picea abies* (L.) karst. Tartu, 1999, 133 p.
47. **Tarmo Annilo.** Studies on mammalian ribosomal protein S7. Tartu, 1998, 77 p.
48. **Indrek Ots.** Health state indicies of reproducing great tits (*Parus major*): sources of variation and connections with life-history traits. Tartu, 1999, 117 p.
49. **Juan Jose Cantero.** Plant community diversity and habitat relationships in central Argentina grasslands. Tartu, 1999, 161 p.
50. **Rein Kalamees.** Seed bank, seed rain and community regeneration in Estonian calcareous grasslands. Tartu, 1999, 107 p.
51. **Sulev Kõks.** Cholecystokinin (CCK) — induced anxiety in rats: influence of environmental stimuli and involvement of endopioid mechanisms and serotonin. Tartu, 1999, 123 p.
52. **Ebe Sild.** Impact of increasing concentrations of O_3 and CO_2 on wheat, clover and pasture. Tartu, 1999, 123 p.
53. **Ljudmilla Timofejeva.** Electron microscopical analysis of the synaptonemal complex formation in cereals. Tartu, 1999, 99 p.
54. **Andres Valkna.** Interactions of galanin receptor with ligands and G-proteins: studies with synthetic peptides. Tartu, 1999, 103 p.
55. **Taavi Virro.** Life cycles of planktonic rotifers in lake Peipsi. Tartu, 1999, 101 p.
56. **Ana Rebane.** Mammalian ribosomal protein S3a genes and intron-encoded small nucleolar RNAs U73 and U82. Tartu, 1999, 85 p.
57. **Tiina Tamm.** Cocksfoot mottle virus: the genome organisation and translational strategies. Tartu, 2000, 101 p.
58. **Reet Kurg.** Structure-function relationship of the bovine papilloma virus E2 protein. Tartu, 2000, 89 p.
59. **Toomas Kivisild.** The origins of Southern and Western Eurasian populations: an mtDNA study. Tartu, 2000, 121 p.
60. **Niilo Kaldalu.** Studies of the TOL plasmid transcription factor XylS. Tartu 2000. 88 p.

61. **Dina Lepik.** Modulation of viral DNA replication by tumor suppressor protein p53. Tartu 2000. 106 p.
62. **Kai Vellak.** Influence of different factors on the diversity of the bryophyte vegetation in forest and wooded meadow communities. Tartu 2000. 122 p.
63. **Jonne Kotta.** Impact of eutrophication and biological invasionas on the structure and functions of benthic macrofauna. Tartu 2000. 160 p.
64. **Georg Martin.** Phytobenthic communities of the Gulf of Riga and the inner sea the West-Estonian archipelago. Tartu, 2000. 139 p.
65. **Silvia Sepp.** Morphological and genetical variation of *Alchemilla L.* in Estonia. Tartu, 2000. 124 p.
66. **Jaan Liira.** On the determinants of structure and diversity in herbaceous plant communities. Tartu, 2000. 96 p.
67. **Priit Zingel.** The role of planktonic ciliates in lake ecosystems. Tartu 2001. 111 p.
68. **Tiit Teder.** Direct and indirect effects in Host-parasitoid interactions: ecological and evolutionary consequences. Tartu 2001. 122 p.
69. **Hannes Kollist.** Leaf apoplastic ascorbate as ozone scavenger and its transport across the plasma membrane. Tartu 2001. 80 p.
70. **Reet Marits.** Role of two-component regulator system PehR-PehS and extracellular protease PrtW in virulence of *Erwinia Carotovora* subsp. *Carotovora*. Tartu 2001. 112 p.
71. **Vallo Tilgar.** Effect of calcium supplementation on reproductive performance of the pied flycatcher *Ficedula hypoleuca* and the great tit *Parus major*, breeding in Nothern temperate forests. Tartu, 2002. 126 p.
72. **Rita Hõrak.** Regulation of transposition of transposon Tn*4652* in *Pseudomonas putida*. Tartu, 2002. 108 p.
73. **Liina Eek-Piirsoo.** The effect of fertilization, mowing and additional illumination on the structure of a species-rich grassland community. Tartu, 2002. 74 p.
74. **Krõõt Aasamaa.** Shoot hydraulic conductance and stomatal conductance of six temperate deciduous tree species. Tartu, 2002. 110 p.
75. **Nele Ingerpuu.** Bryophyte diversity and vascular plants. Tartu, 2002. 112 p.
76. **Neeme Tõnisson.** Mutation detection by primer extension on oligonucleotide microarrays. Tartu, 2002. 124 p.
77. **Margus Pensa.** Variation in needle retention of Scots pine in relation to leaf morphology, nitrogen conservation and tree age. Tartu, 2003. 110 p.
78. **Asko Lõhmus.** Habitat preferences and quality for birds of prey: from principles to applications. Tartu, 2003. 168 p.
79. **Viljar Jaks.** p53 — a switch in cellular circuit. Tartu, 2003. 160 p.
80. **Jaana Männik.** Characterization and genetic studies of four ATP-binding cassette (ABC) transporters. Tartu, 2003. 140 p.
81. **Marek Sammul.** Competition and coexistence of clonal plants in relation to productivity. Tartu, 2003. 159 p

82. **Ivar Ilves.** Virus-cell interactions in the replication cycle of bovine papillomavirus type 1. Tartu, 2003. 89 p.
83. **Andres Männik.** Design and characterization of a novel vector system based on the stable replicator of bovine papillomavirus type 1. Tartu, 2003. 109 p.
84. **Ivika Ostonen.** Fine root structure, dynamics and proportion in net primary production of Norway spruce forest ecosystem in relation to site conditions. Tartu, 2003. 158 p.
85. **Gudrun Veldre.** Somatic status of 12–15-year-old Tartu schoolchildren. Tartu, 2003. 199 p.
86. **Ülo Väli.** The greater spotted eagle *Aquila clanga* and the lesser spotted eagle *A. pomarina*: taxonomy, phylogeography and ecology. Tartu, 2004. 159 p.
87. **Aare Abroi.** The determinants for the native activities of the bovine papillomavirus type 1 E2 protein are separable. Tartu, 2004. 135 p.
88. **Tiina Kahre.** Cystic fibrosis in Estonia. Tartu, 2004. 116 p.
89. **Helen Orav-Kotta.** Habitat choice and feeding activity of benthic suspension feeders and mesograzers in the northern Baltic Sea. Tartu, 2004. 117 p.
90. **Maarja Öpik.** Diversity of arbuscular mycorrhizal fungi in the roots of perennial plants and their effect on plant performance. Tartu, 2004. 175 p.
91. **Kadri Tali.** Species structure of *Neotinea ustulata*. Tartu, 2004. 109 p.
92. **Kristiina Tambets.** Towards the understanding of post-glacial spread of human mitochondrial DNA haplogroups in Europe and beyond: a phylogeographic approach. Tartu, 2004. 163 p.
93. **Arvi Jõers.** Regulation of p53-dependent transcription. Tartu, 2004. 103 p.
94. **Lilian Kadaja.** Studies on modulation of the activity of tumor suppressor protein p53. Tartu, 2004. 103 p.
95. **Jaak Truu.** Oil shale industry wastewater: impact on river microbial community and possibilities for bioremediation. Tartu, 2004. 128 p.
96. **Maire Peters.** Natural horizontal transfer of the *pheBA* operon. Tartu, 2004. 105 p.
97. **Ülo Maiväli.** Studies on the structure-function relationship of the bacterial ribosome. Tartu, 2004. 130 p.
98. **Merit Otsus.** Plant community regeneration and species diversity in dry calcareous grasslands. Tartu, 2004. 103 p.
99. **Mikk Heidemaa.** Systematic studies on sawflies of the genera *Dolerus*, *Empria*, and *Caliroa* (Hymenoptera: Tenthredinidae). Tartu, 2004. 167 p.
100. **Ilmar Tõnno.** The impact of nitrogen and phosphorus concentration and N/P ratio on cyanobacterial dominance and N_2 fixation in some Estonian lakes. Tartu, 2004. 111 p.
101. **Lauri Saks.** Immune function, parasites, and carotenoid-based ornaments in greenfinches. Tartu, 2004. 144 p.
102. **Siiri Rootsi.** Human Y-chromosomal variation in European populations. Tartu, 2004. 142 p.

103. **Eve Vedler.** Structure of the 2,4-dichloro-phenoxyacetic acid-degradative plasmid pEST4011. Tartu, 2005. 106 p.
104. **Andres Tover.** Regulation of transcription of the phenol degradation *pheBA* operon in *Pseudomonas putida*. Tartu, 2005. 126 p.
105. **Helen Udras.** Hexose kinases and glucose transport in the yeast *Hansenula polymorpha*. Tartu, 2005. 100 p.
106. **Ave Suija.** Lichens and lichenicolous fungi in Estonia: diversity, distribution patterns, taxonomy. Tartu, 2005. 162 p.
107. **Piret Lõhmus.** Forest lichens and their substrata in Estonia. Tartu, 2005. 162 p.
108. **Inga Lips.** Abiotic factors controlling the cyanobacterial bloom occurrence in the Gulf of Finland. Tartu, 2005. 156 p.
109. **Kaasik, Krista.** Circadian clock genes in mammalian clockwork, metabolism and behaviour. Tartu, 2005. 121 p.
110. **Juhan Javoiš.** The effects of experience on host acceptance in ovipositing moths. Tartu, 2005. 112 p.
111. **Tiina Sedman.** Characterization of the yeast *Saccharomyces cerevisiae* mitochondrial DNA helicase Hmi1. Tartu, 2005. 103 p.
112. **Ruth Aguraiuja.** Hawaiian endemic fern lineage *Diellia* (Aspleniaceae): distribution, population structure and ecology. Tartu, 2005. 112 p.
113. **Riho Teras.** Regulation of transcription from the fusion promoters generated by transposition of Tn*4652* into the upstream region of *pheBA* operon in *Pseudomonas putida*. Tartu, 2005. 106 p.
114. **Mait Metspalu.** Through the course of prehistory in india: tracing the mtDNA trail. Tartu, 2005. 138 p.
115. **Elin Lõhmussaar.** The comparative patterns of linkage disequilibrium in European populations and its implication for genetic association studies. Tartu, 2006. 124 p.
116. **Priit Kupper.** Hydraulic and environmental limitations to leaf water relations in trees with respect to canopy position. Tartu, 2006. 126 p.
117. **Heili Ilves.** Stress-induced transposition of Tn*4652* in *Pseudomonas Putida.* Tartu, 2006. 120 p.
118. **Silja Kuusk.** Biochemical properties of Hmi1p, a DNA helicase from *Saccharomyces cerevisiae* mitochondria. Tartu, 2006. 126 p.
119. **Kersti Püssa.** Forest edges on medium resolution landsat thematic mapper satellite images. Tartu, 2006. 90 p.
120. **Lea Tummeleht.** Physiological condition and immune function in great tits (*Parus major* l.): Sources of variation and trade-offs in relation to growth. Tartu, 2006. 94 p.
121. **Toomas Esperk.** Larval instar as a key element of insect growth schedules. Tartu, 2006. 186 p.
122. **Harri Valdmann.** Lynx (*Lynx lynx*) and wolf (*Canis lupus*) in the Baltic region: Diets, helminth parasites and genetic variation. Tartu, 2006. 102 p.

123. **Priit Jõers.** Studies of the mitochondrial helicase Hmi1p in *Candida albicans* and *Saccharomyces cerevisia*. Tartu, 2006. 113 p.
124. **Kersti Lilleväli.** Gata3 and Gata2 in inner ear development. Tartu, 2007. 123 p.
125. **Kai Rünk.** Comparative ecology of three fern species: *Dryopteris carthusiana* (Vill.) H.P. Fuchs, *D. expansa* (C. Presl) Fraser-Jenkins & Jermy and *D. dilatata* (Hoffm.) A. Gray (Dryopteridaceae). Tartu, 2007. 143 p.
126. **Aveliina Helm.** Formation and persistence of dry grassland diversity: role of human history and landscape structure. Tartu, 2007. 89 p.
127. **Leho Tedersoo.** Ectomycorrhizal fungi: diversity and community structure in Estonia, Seychelles and Australia. Tartu, 2007. 233 p.
128. **Marko Mägi.** The habitat-related variation of reproductive performance of great tits in a deciduous-coniferous forest mosaic: looking for causes and consequences. Tartu, 2007. 135 p.
129. **Valeria Lulla.** Replication strategies and applications of Semliki Forest virus. Tartu, 2007. 109 p.
130. **Ülle Reier**. Estonian threatened vascular plant species: causes of rarity and conservation. Tartu, 2007. 79 p.
131. **Inga Jüriado**. Diversity of lichen species in Estonia: influence of regional and local factors. Tartu, 2007. 171 p.
132. **Tatjana Krama.** Mobbing behaviour in birds: costs and reciprocity based cooperation. Tartu, 2007. 112 p.
133. **Signe Saumaa.** The role of DNA mismatch repair and oxidative DNA damage defense systems in avoidance of stationary phase mutations in *Pseudomonas putida*. Tartu, 2007. 172 p.
134. **Reedik Mägi**. The linkage disequilibrium and the selection of genetic markers for association studies in european populations. Tartu, 2007. 96 p.
135. **Priit Kilgas.** Blood parameters as indicators of physiological condition and skeletal development in great tits (*Parus major*): natural variation and application in the reproductive ecology of birds. Tartu, 2007. 129 p.
136. **Anu Albert**. The role of water salinity in structuring eastern Baltic coastal fish communities. Tartu, 2007. 95 p.
137. **Kärt Padari.** Protein transduction mechanisms of transportans. Tartu, 2008. 128 p.
138. **Siiri-Lii Sandre.** Selective forces on larval colouration in a moth. Tartu, 2008. 125 p.
139. **Ülle Jõgar.** Conservation and restoration of semi-natural floodplain meadows and their rare plant species. Tartu, 2008. 99 p.
140. **Lauri Laanisto.** Macroecological approach in vegetation science: generality of ecological relationships at the global scale. Tartu, 2008. 133 p.
141. **Reidar Andreson**. Methods and software for predicting PCR failure rate in large genomes. Tartu, 2008. 105 p.
142. **Birgot Paavel.** Bio-optical properties of turbid lakes. Tartu, 2008. 175 p.

143. **Kaire Torn.** Distribution and ecology of charophytes in the Baltic Sea. Tartu, 2008, 98 p.
144. **Vladimir Vimberg.** Peptide mediated macrolide resistance. Tartu, 2008, 190 p.
145. **Daima Örd.** Studies on the stress-inducible pseudokinase TRB3, a novel inhibitor of transcription factor ATF4. Tartu, 2008, 108 p.
146. **Lauri Saag.** Taxonomic and ecologic problems in the genus *Lepraria* (*Stereocaulaceae*, lichenised *Ascomycota*). Tartu, 2008, 175 p.
147. **Ulvi Karu.** Antioxidant protection, carotenoids and coccidians in greenfinches – assessment of the costs of immune activation and mechanisms of parasite resistance in a passerine with carotenoid-based ornaments. Tartu, 2008, 124 p.
148. **Jaanus Remm.** Tree-cavities in forests: density, characteristics and occupancy by animals. Tartu, 2008, 128 p.
149. **Epp Moks.** Tapeworm parasites *Echinococcus multilocularis* and *E. granulosus* in Estonia: phylogenetic relationships and occurrence in wild carnivores and ungulates. Tartu, 2008, 82 p.
150. **Eve Eensalu.** Acclimation of stomatal structure and function in tree canopy: effect of light and CO_2 concentration. Tartu, 2008, 108 p.
151. **Janne Pullat**. Design, functionlization and application of an *in situ* synthesized oligonucleotide microarray. Tartu, 2008, 108 p.
152. **Marta Putrinš.** Responses of *Pseudomonas putida* to phenol-induced metabolic and stress signals. Tartu, 2008, 142 p.
153. **Marina Semtšenko.** Plant root behaviour: responses to neighbours and physical obstructions. Tartu, 2008, 106 p.
154. **Marge Starast.** Influence of cultivation techniques on productivity and fruit quality of some *Vaccinium* and *Rubus* taxa. Tartu, 2008, 154 p.
155. **Age Tats.** Sequence motifs influencing the efficiency of translation. Tartu, 2009, 104 p.
156. **Radi Tegova.** The role of specialized DNA polymerases in mutagenesis in *Pseudomonas putida.* Tartu, 2009, 124 p.
157. **Tsipe Aavik.** Plant species richness, composition and functional trait pattern in agricultural landscapes – the role of land use intensity and landscape structure. Tartu, 2009, 112 p.
158. **Kaja Kiiver.** Semliki forest virus based vectors and cell lines for studying the replication and interactions of alphaviruses and hepaciviruses. Tartu, 2009, 104 p.
159. **Meelis Kadaja.** Papillomavirus Replication Machinery Induces Genomic Instability in its Host Cell. Tartu, 2009, 126 p.
160. **Pille Hallast.** Human and chimpanzee Luteinizing hormone/Chorionic Gonadotropin beta (*LHB/CGB*) gene clusters: diversity and divergence of young duplicated genes. Tartu, 2009, 168 p.
161. **Ain Vellak.** Spatial and temporal aspects of plant species conservation. Tartu, 2009, 86 p.

162. **Triinu Remmel.** Body size evolution in insects with different colouration strategies: the role of predation risk. Tartu, 2009, 168 p.
163. **Jaana Salujõe.** Zooplankton as the indicator of ecological quality and fish predation in lake ecosystems. Tartu, 2009, 129 p.
164. **Ele Vahtmäe.** Mapping benthic habitat with remote sensing in optically complex coastal environments. Tartu, 2009, 109 p.
165. **Liisa Metsamaa.** Model-based assessment to improve the use of remote sensing in recognition and quantitative mapping of cyanobacteria. Tartu, 2009, 114 p.
166. **Pille Säälik.** The role of endocytosis in the protein transduction by cell-penetrating peptides. Tartu, 2009, 155 p.
167. **Lauri Peil.** Ribosome assembly factors in *Escherichia coli.* Tartu, 2009, 147 p.
168. **Lea Hallik.** Generality and specificity in light harvesting, carbon gain capacity and shade tolerance among plant functional groups. Tartu, 2009, 99 p.
169. **Mariliis Tark.** Mutagenic potential of DNA damage repair and tolerance mechanisms under starvation stress. Tartu, 2009, 191 p.
170. **Riinu Rannap.** Impacts of habitat loss and restoration on amphibian populations. Tartu, 2009, 117 p.
171. **Maarja Adojaan.** Molecular variation of HIV-1 and the use of this knowledge in vaccine development. Tartu, 2009, 95 p.
172. **Signe Altmäe.** Genomics and transcriptomics of human induced ovarian folliculogenesis. Tartu, 2010, 179 p.
173. **Triin Suvi.** Mycorrhizal fungi of native and introduced trees in the Seychelles Islands. Tartu, 2010, 107 p.
174. **Velda Lauringson.** Role of suspension feeding in a brackish-water coastal sea. Tartu, 2010, 123 p.
175. **Eero Talts.** Photosynthetic cyclic electron transport – measurement and variably proton-coupled mechanism. Tartu, 2010, 121 p.
176. **Mari Nelis.** Genetic structure of the Estonian population and genetic distance from other populations of European descent. Tartu, 2010, 97 p.
177. **Kaarel Krjutškov.** Arrayed Primer Extension-2 as a multiplex PCR-based method for nucleic acid variation analysis: method and applications. Tartu, 2010, 129 p.
178. **Egle Köster.** Morphological and genetical variation within species complexes: *Anthyllis vulneraria* s. l. and *Alchemilla vulgaris* (coll.). Tartu, 2010, 101 p.
179. **Erki Õunap.** Systematic studies on the subfamily Sterrhinae (Lepidoptera: Geometridae). Tartu, 2010, 111 p.
180. **Merike Jõesaar.** Diversity of key catabolic genes at degradation of phenol and *p*-cresol in pseudomonads. Tartu, 2010, 125 p.
181. **Kristjan Herkül.** Effects of physical disturbance and habitat-modifying species on sediment properties and benthic communities in the northern Baltic Sea. Tartu, 2010, 123 p.

182. **Arto Pulk.** Studies on bacterial ribosomes by chemical modification approaches. Tartu, 2010, 161 p.
183. **Maria Põllupüü.** Ecological relations of cladocerans in a brackish-water ecosystem. Tartu, 2010, 126 p.
184. **Toomas Silla.** Study of the segregation mechanism of the Bovine Papillomavirus Type 1. Tartu, 2010, 188 p.
185. **Gyaneshwer Chaubey.** The demographic history of India: A perspective based on genetic evidence. Tartu, 2010, 184 p.
186. **Katrin Kepp.** Genes involved in cardiovascular traits: detection of genetic variation in Estonian and Czech populations. Tartu, 2010, 164 p.
187. **Virve Sõber.** The role of biotic interactions in plant reproductive performance. Tartu, 2010, 92 p.
188. **Kersti Kangro.** The response of phytoplankton community to the changes in nutrient loading. Tartu, 2010, 144 p.
189. **Joachim M. Gerhold.** Replication and Recombination of mitochondrial DNA in Yeast. Tartu, 2010, 120 p.
190. **Helen Tammert.** Ecological role of physiological and phylogenetic diversity in aquatic bacterial communities. Tartu, 2010, 140 p.
191. **Elle Rajandu.** Factors determining plant and lichen species diversity and composition in Estonian *Calamagrostis* and *Hepatica* site type forests. Tartu, 2010, 123 p.
192. **Paula Ann Kivistik.** ColR-ColS signalling system and transposition of Tn*4652* in the adaptation of *Pseudomonas putida.* Tartu, 2010, 118 p.
193. **Siim Sõber.** Blood pressure genetics: from candidate genes to genome-wide association studies. Tartu, 2011, 120 p.
194. **Kalle Kipper.** Studies on the role of helix 69 of 23S rRNA in the factor-dependent stages of translation initiation, elongation, and termination. Tartu, 2011, 178 p.
195. **Triinu Siibak.** Effect of antibiotics on ribosome assembly is indirect. Tartu, 2011, 134 p.
196. **Tambet Tõnissoo.** Identification and molecular analysis of the role of guanine nucleotide exchange factor RIC-8 in mouse development and neural function. Tartu, 2011, 110 p.
197. **Helin Räägel.** Multiple faces of cell-penetrating peptides – their intracellular trafficking, stability and endosomal escape during protein transduction. Tartu, 2011, 161 p.
198. **Andres Jaanus.** Phytoplankton in Estonian coastal waters – variability, trends and response to environmental pressures. Tartu, 2011, 157 p.
199. **Tiit Nikopensius.** Genetic predisposition to nonsyndromic orofacial clefts. Tartu, 2011, 152 p.
200. **Signe Värv.** Studies on the mechanisms of RNA polymerase II-dependent transcription elongation. Tartu, 2011, 108 p.
201. **Kristjan Välk.** Gene expression profiling and genome-wide association studies of non-small cell lung cancer. Tartu, 2011, 98 p.

202. **Arno Põllumäe.** Spatio-temporal patterns of native and invasive zooplankton species under changing climate and eutrophication conditions. Tartu, 2011, 153 p.
203. **Egle Tammeleht.** Brown bear (*Ursus arctos*) population structure, demographic processes and variations in diet in northern Eurasia. Tartu, 2011, 143 p.
205. **Teele Jairus.** Species composition and host preference among ectomycorrhizal fungi in Australian and African ecosystems. Tartu, 2011, 106 p.
206. **Kessy Abarenkov.** PlutoF – cloud database and computing services supporting biological research. Tartu, 2011, 125 p.
207. **Marina Grigorova.** Fine-scale genetic variation of follicle-stimulating hormone beta-subunit coding gene (*FSHB*) and its association with reproductive health. Tartu, 2011, 184 p.
208. **Anu Tiitsaar.** The effects of predation risk and habitat history on butterfly communities. Tartu, 2011, 97 p.
209. **Elin Sild.** Oxidative defences in immunoecological context: validation and application of assays for nitric oxide production and oxidative burst in a wild passerine. Tartu, 2011, 105 p.
210. **Irja Saar**. The taxonomy and phylogeny of the genera *Cystoderma* and *Cystodermella* (Agaricales, Fungi). Tartu, 2012, 167 p.
211. **Pauli Saag.** Natural variation in plumage bacterial assemblages in two wild breeding passerines. Tartu, 2012, 113 p.
212. **Aleksei Lulla.** Alphaviral nonstructural protease and its polyprotein substrate: arrangements for the perfect marriage. Tartu, 2012, 143 p.
213. **Mari Järve.** Different genetic perspectives on human history in Europe and the Caucasus: the stories told by uniparental and autosomal markers. Tartu, 2012, 119 p.
214. **Ott Scheler**. The application of tmRNA as a marker molecule in bacterial diagnostics using microarray and biosensor technology. Tartu, 2012, 93 p.
215. **Anna Balikova**. Studies on the functions of tumor-associated mucin-like leukosialin (CD43) in human cancer cells. Tartu, 2012, 129 p.
216. **Triinu Kõressaar.** Improvement of PCR primer design for detection of prokaryotic species. Tartu, 2012, 83 p.
217. **Tuul Sepp.** Hematological health state indices of greenfinches: sources of individual variation and responses to immune system manipulation. Tartu, 2012, 117 p.
218. **Rya Ero.** Modifier view of the bacterial ribosome. Tartu, 2012, 146 p.
219. **Mohammad Bahram.** Biogeography of ectomycorrhizal fungi across different spatial scales. Tartu, 2012, 165 p.
220. **Annely Lorents.** Overcoming the plasma membrane barrier: uptake of amphipathic cell-penetrating peptides induces influx of calcium ions and downstream responses. Tartu, 2012, 113 p.

221. **Katrin Männik.** Exploring the genomics of cognitive impairment: whole-genome SNP genotyping experience in Estonian patients and general population. Tartu, 2012, 171 p.
222. **Marko Prous.** Taxonomy and phylogeny of the sawfly genus *Empria* (Hymenoptera, Tenthredinidae). Tartu, 2012, 192 p.
223. **Triinu Visnapuu.** Levansucrases encoded in the genome of *Pseudomonas syringae* pv. tomato DC3000: heterologous expression, biochemical characterization, mutational analysis and spectrum of polymerization products. Tartu, 2012, 160 p.
224. **Nele Tamberg.** Studies on Semliki Forest virus replication and pathogenesis. Tartu, 2012, 109 p.
225. **Tõnu Esko.** Novel applications of SNP array data in the analysis of the genetic structure of Europeans and in genetic association studies. Tartu, 2012, 149 p.
226. **Timo Arula.** Ecology of early life-history stages of herring *Clupea harengus membras* in the northeastern Baltic Sea. Tartu, 2012, 143 p.
227. **Inga Hiiesalu.** Belowground plant diversity and coexistence patterns in grassland ecosystems. Tartu, 2012, 130 p.
228. **Kadri Koorem.** The influence of abiotic and biotic factors on small-scale plant community patterns and regeneration in boreonemoral forest. Tartu, 2012, 114 p.
229. **Liis Andresen.** Regulation of virulence in plant-pathogenic pectobacteria. Tartu, 2012, 122 p.
230. **Kaupo Kohv.** The direct and indirect effects of management on boreal forest structure and field layer vegetation. Tartu, 2012, 124 p.
231. **Mart Jüssi.** Living on an edge: landlocked seals in changing climate. Tartu, 2012, 114 p.
232. **Riina Klais.** Phytoplankton trends in the Baltic Sea. Tartu, 2012, 136 p.
233. **Rauno Veeroja.** Effects of winter weather, population density and timing of reproduction on life-history traits and population dynamics of moose (*Alces alces*) in Estonia. Tartu, 2012, 92 p.
234. **Marju Keis.** Brown bear (*Ursus arctos*) phylogeography in northern Eurasia. Tartu, 2013, 142 p.
235. **Sergei Põlme.** Biogeography and ecology of *alnus*- associated ectomycorrhizal fungi – from regional to global scale. Tartu, 2013, 90 p.
236. **Liis Uusküla.** Placental gene expression in normal and complicated pregnancy. Tartu, 2013, 173 p.
237. **Marko Lõoke.** Studies on DNA replication initiation in *Saccharomyces cerevisiae.* Tartu, 2013, 112 p.
238. **Anne Aan.** Light- and nitrogen-use and biomass allocation along productivity gradients in multilayer plant communities. Tartu, 2013, 127 p.
239. **Heidi Tamm.** Comprehending phylogenetic diversity – case studies in three groups of ascomycetes. Tartu, 2013, 136 p.

240. **Liina Kangur.** High-Pressure Spectroscopy Study of Chromophore-Binding Hydrogen Bonds in Light-Harvesting Complexes of Photosynthetic Bacteria. Tartu, 2013, 150 p.
241. **Margus Leppik.** Substrate specificity of the multisite specific pseudouridine synthase RluD. Tartu, 2013, 111 p.
242. **Lauris Kaplinski.** The application of oligonucleotide hybridization model for PCR and microarray optimization. Tartu, 2013, 103 p.
243. **Merli Pärnoja.** Patterns of macrophyte distribution and productivity in coastal ecosystems: effect of abiotic and biotic forcing. Tartu, 2013, 155 p.
244. **Tõnu Margus.** Distribution and phylogeny of the bacterial translational GTPases and the Mqsr/YgiT regulatory system. Tartu, 2013, 126 p.
245. **Pille Mänd**. Light use capacity and carbon and nitrogen budget of plants: remote assessment and physiological determinants. Tartu, 2013, 128 p.
246. **Mario Plaas.** Animal model of Wolfram Syndrome in mice: behavioural, biochemical and psychopharmacological characterization. Tartu, 2013, 144 p.
247. **Georgi Hudjašov.** Maps of mitochondrial DNA, Y-chromosome and tyrosinase variation in Eurasian and Oceanian populations. Tartu, 2013, 115 p.
248. **Mari Lepik**. Plasticity to light in herbaceous plants and its importance for community structure and diversity. Tartu, 2013, 102 p.
249. **Ede Leppik**. Diversity of lichens in semi-natural habitats of Estonia. Tartu, 2013, 151 p.
250. **Ülle Saks.** Arbuscular mycorrhizal fungal diversity patterns in boreonemoral forest ecosystems. Tartu, 2013, 151 p.
251. **Eneli Oitmaa**. Development of arrayed primer extension microarray assays for molecular diagnostic applications. Tartu, 2013, 147 p.
252. **Jekaterina Jutkina.** The horizontal gene pool for aromatics degradation: bacterial catabolic plasmids of the Baltic Sea aquatic system. Tartu, 2013, 121 p.
253. **Helen Vellau.** Reaction norms for size and age at maturity in insects: rules and exceptions. Tartu, 2014, 132 p.
254. **Randel Kreitsberg.** Using biomarkers in assessment of environmental contamination in fish – new perspectives. Tartu, 2014, 107 p.
255. **Krista Takkis.** Changes in plant species richness and population performance in response to habitat loss and fragmentation.Tartu, 2014, 141 p.
256. **Liina Nagirnaja.** Global and fine-scale genetic determinants of recurrent pregnancy loss. Tartu, 2014, 211 p.
257. **Triin Triisberg.** Factors influencing the re-vegetation of abandoned extracted peatlands in Estonia. Tartu, 2014, 133 p.
258. **Villu Soon.** A phylogenetic revision of the *Chrysis ignita* species group (Hymenoptera: Chrysididae) with emphasis on the northern European fauna. Tartu, 2014, 211 p.

259. **Andrei Nikonov.** RNA-Dependent RNA Polymerase Activity as a Basis for the Detection of Positive-Strand RNA Viruses by Vertebrate Host Cells. Tartu, 2014, 207 p.
260. **Eele Õunapuu-Pikas**. Spatio-temporal variability of leaf hydraulic conductance in woody plants: ecophysiological consequences. Tartu, 2014, 135 p.
261. **Marju Männiste**. Physiological ecology of greenfinches: information content of feathers in relation to immune function and behavior. Tartu, 2014, 121 p.
262. **Katre Kets.** Effects of elevated concentrations of CO_2 and O_3 on leaf photosynthetic parameters in *Populus tremuloides*: diurnal, seasonal and interannual patterns. Tartu, 2014, 115 p.
263. **Külli Lokko.** Seasonal and spatial variability of zoopsammon communities in relation to environmental parameters. Tartu, 2014, 129 p.
264. **Olga Žilina**. Chromosomal microarray analysis as diagnostic tool: Estonian experience. Tartu, 2014, 152 p.
265. **Kertu Lõhmus**. Colonisation ecology of forest-dwelling vascular plants and the conservation value of rural manor parks. Tartu, 2014, 111 p.
266. **Anu Aun.** Mitochondria as integral modulators of cellular signaling. Tartu, 2014, 167 p.
267. **Chandana Basu Mallick.** Genetics of adaptive traits and gender-specific demographic processes in South Asian populations. Tartu, 2014, 160 p.
268. **Riin Tamme.** The relationship between small-scale environmental heterogeneity and plant species diversity. Tartu, 2014, 130 p.
269. **Liina Remm.** Impacts of forest drainage on biodiversity and habitat quality: implications for sustainable management and conservation. Tartu, 2015, 126 p.
270. **Tiina Talve.** Genetic diversity and taxonomy within the genus *Rhinanthus*. Tartu, 2015, 106 p.
271. **Mehis Rohtla**. Otolith sclerochronological studies on migrations, spawning habitat preferences and age of freshwater fishes inhabiting the Baltic Sea. Tartu, 2015, 137 p.
272. **Alexey Reshchikov.** The world fauna of the genus *Lathrolestes* (Hymenoptera, Ichneumonidae). Tartu, 2015, 247 p.
273. **Martin Pook.** Studies on artificial and extracellular matrix protein-rich surfaces as regulators of cell growth and differentiation. Tartu, 2015, 142 p.
274. **Mai Kukumägi.** Factors affecting soil respiration and its components in silver birch and Norway spruce stands. Tartu, 2015, 155 p.
275. **Helen Karu.** Development of ecosystems under human activity in the North-East Estonian industrial region: forests on post-mining sites and bogs. Tartu, 2015, 152 p.