DISSERTATIONES
BIOLOGICAE
UNIVERSITATIS
TARTUENSIS
277

PRIIT ADLER

Analysis and visualisation of large scale microarray data

DISSERTATIONES BIOLOGICAE UNIVERSITATIS TARTUENSIS
277

DISSERTATIONES BIOLOGICAE UNIVERSITATIS TARTUENSIS
277

PRIIT ADLER

Analysis and visualisation of large scale microarray data

UNIVERSITY OF TARTU
PRESS
1632

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

Dissertation is accepted for the commencement of the degree of Doctor of Philosophy in bioinformatics at University of Tartu on 19th of June 2015 by the Council of the Institute of Molecular and Cellular Biology, University of Tartu.

Supervisors:

Prof. Jaak Vilo, PhD
Institute of Computer Science
University of Tartu
Tartu, Estonia

Prof. Juhan Sedman, PhD
Institute of Molecular and Cell Biology
University of Tartu
Tartu, Estonia

Opponent:

Gabriella Rustici, PhD
School of the Biological Sciences
University of Cambridge
Cambridge, United Kingdom

Commencement:

Room No 105, 23B Riia St, Tartu, on August 26th,
2015, at 10:15

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

ISSN 1024-6479
ISBN 978-9949-32-873-4 (print)
ISBN 978-9949-32-874-1 (pdf)

University of Tartu Press
www.tyk.ee

“An education was a bit like a communicable sexual disease. It made you unsuitable for a lot of jobs and then you had the urge to pass it on”

Terry Pratchett

TABLE OF CONTENTS

LIST OF ORIGINAL PUBLICATIONS

This thesis is based on the following original publications which will be referred to in the text by their Roman numerals:

I **Adler, P.**[*], Reimand, J.[*], Jänes, J., Kolde, R., Peterson, H., and Vilo, J. (2008). KEGGanim: pathway animations for high-throughput data. *Bioinformatics*, 24(4):588–90.

II **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*, 1158:1–13.

III **Adler, P.**[*], Kolde, R.[*], Kull, M., Tkachenko, A., Peterson, H., Reimand, J., and Vilo, J. (2009). Mining for coexpression across hundreds of datasets using novel rank aggregation and visualization methods. *Genome Biology*, 10(12):R139.

IV Kolde, R., Laur, S., **Adler, P.**, and Vilo, J. (2012). Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics*, 28(4):573–580.

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

My contribution to these articles:

Ref. I – Designed and implemented the visualisation framework for KEGG pathways and implemented web application. Prepared expression data used as examples. Participated in writing the manuscript.

Ref. II – Co-conducted the study, managed high-throughput expression data and performed cross-validation analysis on Reactome pathways and participated in interpreting the results. Participated in writing the manuscript.

Ref. III – Designed and implemented Multi Experiment Matrix (MEM) tool and its web interface. Downloaded and prepared high-throughput expression data used by the application. Participated in developing the rank aggregation algorithm. Performed one of the proof of principle analyses in the article. Participated in writing the manuscript.

Ref. IV – Performed one of the proof of principle analyses for the study.

[*] - Equal contribution.

LIST OF ABBREVIATIONS

CGI	Common Gateway Interface
DNA	Deoxyribonucleic Acid
ES cells	Embryonic Stem cells
FARMS	Factor Analysis for Robust Microarray Summarisation
FC	Fold Change
FDR	False Discovery Rate
FTP	File Transfer Protocol
GEO	Gene Expression Omnibus
GIF	Graphics Interchange Format
GO	Gene Ontology
HPO	Human Phenotype Ontology
ID	Identifier
KEGG	Kyoto Encyclopedia of Genes and Genomes
MAQC	MicroArray Quality Control
MAS5.0	Affymetrix Microarray Suite 5
MCM	Mini Chromosome Maintenance
MEM	Multi Experiment Matrix
NCDF	Network Common Data Format
PCA	Principal Component Analysis
RMA	Robust Multi-array Average
RNA	Ribonucleic Acid
RRA	Robust Rank Aggregation
RT-PCR	Reverse Transcription Polymerase Chain Reaction
SOFT	Simple Omnibus Format in Text
SWOG	Simple Web Object Graphics
TF	Transcription Factor
TIFF	Tagged Image File Format

INTRODUCTION

High-throughput gene expression data has been generated across the globe for almost two decades. A wealth of publicly available data has been gathered into large database such as ArrayExpress or GEO. Although once analysed, the data still contain answers to questions unexplored by others. As new methods of data analysis are developed and innovative visualisations become possible, a systematic approach to revisit and reanalyse existing data might reveal new knowledge.

In the first part of this thesis we have a short overview of high-throughput gene expression data, introduce common analysis and visualisation methods for single datasets and cover relevant meta-analysis pipelines. Beside public gene expression databases, we also provide overview of pathway databases KEGG and Reactome, which are extensively used within publications that are part of this thesis.

In the practical part of this thesis, we first demonstrate how it is possible to visualise and animate high-throughput expression data using KEGG pathways. Visualisation of expression data in the context of KEGG pathway and observing the expression dynamics across samples enables more detailed interpretation of experimental results. To make it accessible to wider audience we have implemented KEGGanim web tool.

KEGG, nor any other public, pathway database does not cover entire genome. Only roughly one third of all genes are annotated to biological pathways. We present a study where we measured the predictive power of high-throughput gene expression data to reconstruct Reactome pathways and to propose potential new candidates. A high-throughput public data-collection with more than 6000 samples was used to perform cross-validation on 35 Reactome pathways. We give overview of the results and discuss observed benefit of using only a subset of pathway genes in the analysis as they might be more tightly co-regulated than entire pathway.

Similarly can be argued about gene expression data, that only subset of expression data should be used to study condition-specific co-expression patterns of related genes. It is proposed that only approximately one fifth of all genes are at once expressed in any biological condition. We describe a framework where co-expression queries can be performed across hundreds of publicly available high-throughput gene expression datasets. Relevant datasets are first selected based on standard deviation of the query gene. In each dataset co-expression values are calculated and all genes are ranked based on found distances. Finally, novel statistical rank aggregation approach is used to create a unified prioritised list of globally co-expressed genes. Method has been implemented in Multi Experiment Matrix (MEM) web tool.

Described rank aggregation method is suitable to solve problems also outside MEM framework and has been published as R package. We provide an overview of some of the other experimental settings with real and simulated data to highlight the features of the presented robust rank aggregation method.

I. REVIEW OF LITERATURE

In eukaryotic cells the hereditary information is stored as long sequences of deoxyribonucleic acid (DNA) molecules. The long DNA molecules are also referred as polynucleotides as they contain single nucleotides in repetition. The order of nucleotide molecules in these long chains defines the information they contain. Regions within DNA, that are used to encode other types of functional molecular polymers are referred as genes. Hence, the overall sum of DNA molecules is also called genome. In human genome the total length of DNA molecules is approximately three billion bases. It is organised into individual molecules, 22 autosomal chromosomes, which are represented by 2 copies – one copy from mother and the other from father, and two sex chromosomes. All together there are 46 DNA molecules per cell.

There are approximately 22000 genes defined in human genome. Genetic information is read from the DNA through process called transcription. The transcription process yields messenger ribonucleic acid (messenger RNA or mRNA) which is another type of polynucleotide. It is similar to DNA, but instead of deoxyribose it has ribose and instead of thymine is has uracil. Messenger RNA is used to transport genetic code out of the nucleus. In cell cytoplasm there are molecular machineries called ribosomes that process mRNA to produce proteins through translation. Proteins are the main building blocks of the cells. They participate in reactions as enzymes and signalling agents and also take part in transcriptional regulation of genes. Each protein can have very specific task or several depending on its configuration and post-processing. Compared to 21855 protein coding genes there are 86434 proteins defined for human in Ensembl database version 80 (Cunningham et al., 2015). For each gene there is a number of options how the mRNA can be alternatively spliced (Modrek and Lee, 2002).

Although only 1.5% of the entire genome is covered by protein-coding genes, a recent study states that more than 75% of the genome is covered by other transcriptional activity (Kellis et al., 2014), most of it is very rare. This percentage might still be an underestimate as only a selection of cell types was covered.

In human body there are hundreds of different types of tissues and cells. Although each cell contains the same DNA, the way how information is read and processed will lead to different cell types and different stages in cell lifecycle. Malfunctions in DNA reading or gene regulation can lead to various diseases including cancer. Gene regulation is a complicated process and consists of many steps. One of the more straightforward steps is the regulation through transcription. The existence and quantity of mRNA molecules are the first prerequisites for protein production. There are no cost effective high-throughput methods to quantify protein levels in cells, but there are high-throughput methods to quantify mRNA levels.

In this thesis we focus on characterisation of gene expression on transcriptional level as this can be performed in high-throughput manner and has been done so for the last two decades (DeRisi et al., 1997; Lashkari et al., 1997).

1.1. High-throughput expression data

The advances in biotechnology have given rise to microarrays. Microarrays are glass slides, or other hard surface slides, that are covered by small oligonucleotide molecules (probes). The oligonucleotide molecules are attached to the microarray surface by one end. Their sequence is complementary to a sequence of a specific gene (Lockhart et al., 1996). Microarrays allow to quantify mRNA levels for many thousands of genes simultaneously from a biological sample. First mRNA is extracted from the biological sample and converted into complementary DNA (cDNA) by reverse transcriptase. Probes catch cDNA molecules from the sample solution in sequence specific manner. Each microarray can contain hundreds of thousands different probes corresponding to different genes, covering vast majority of genes for an organism. This kind of technology allows to take transcriptional still images of cellular activity. More images lead to better understanding of underlying processes and help us to decipher cellular functions.

The microarrays discussed within this thesis are *gene expression* microarrays. There are also other types of microarrays, for example, genotyping or next generation sequencing that are also performed in a microarray format, but these are not the focus of the current thesis.

1.1.1. Normalisation

Generating the data is only the first step in the whole experiment. Methods to process, normalise and analyse are essential to interpret the gene expression microarray data. Raw microarray data is considered to be noisy (Bolstad et al., 2003). There are two principal sources of noise: biological and technical. Both type of noise can be controlled or tested by generating more biological and technical replicate samples (Klebanov and Yakovlev, 2007). Still, in the raw format the data is rarely suitable for interpretation. Statistical methods are used to transform the data so, that it would meet the requirements of the analysis methods, while still retaining its biological signal. This process is called normalisation. The objective of normalisation is to make separate samples comparable to each other within the experiment. Many normalisation applications also transform the data so, that signal value distribution would look normally distributed.

Different microarray technologies have different standards. Several normalisation methods have been developed to meet the design of Affymetrix GeneChips, for example. Robust multi-array average (RMA) (Irizarry et al., 2003), MAS5.0 (Hubbell et al., 2002), FARMS (Hochreiter et al., 2006) to name a few.

On Affymetrix GeneChip platforms a probe set is small collection of probes that represent the same transcript. There can be more than one probe set representing a single gene. The result of preprocessing of gene expression microarray data is a numeric matrix – expression matrix, where columns represent different samples and each row represents expression values summarised on a probe set or a gene level. A row in this matrix and a column are referred as probe set and sample expression profile, respectively.

Most widely used normalisation method, to date, is RMA. It uses log transformation and quantile normalisation between samples. Distribution quantile values are made equal between signals across all samples and signals from individual samples. That ensures that all individual samples follow the same signal value distribution and therefore are more comparable to each other.

For strong biological signal the choice of normalisation method does not matter, but for weaker signals different approaches may give slightly different results as shown by Millenaar et al. (2006). They compared different normalisation pipelines on differential expression outcome. While all methods found similar number of differentially expressed genes in given experimental setup, roughly only one third of the genes was common for all methods compared. From the methods used, RMA was highlighted as the one showing the highest correlation coefficient with Real Time RT-PCR results, which were used as "gold standard" (Millenaar et al., 2006).

Probe sets of the most popular Affymetrix GeneChip platforms were developed before human genome was fully sequenced. Due to that, some of the probe sets are not useful and many of them have been reannotated. Some research groups have gone further and disregarded the initial Affymetrix defined probe sets all together. Instead, probes are individually remapped to human transcriptome and new custom probe sets on various specificity level are designed. BrainArray customCDFs[1] provide annotations on gene or transcript level, which in many cases allow to overcome the problem of ambiguous probe set annotations (Dai et al., 2005). It has also shown that the downstream analysis benefits from improved probe set definitions (Sandberg and Larsson, 2007).

1.1.2. Quality assessment

Assessment of data quality is essential for microarray data analysis (Kauffmann et al., 2009). Analysis methods make an assumption that samples are comparable. Outliers or failed samples in the data can lead to severe bias in analysis results.

Microarray experiments have many aspects and comprehensive single value quality assessment would not be feasible. Defects, such as saturation or lack of signal, spacial defects on microarrays (such as fingerprints, scraping marks), RNA degradation rate are quantified and evaluated separately. Spacial defects are best

[1] CDF - *.CEL* definition file, *.CEL* is Affymetrix specific data format

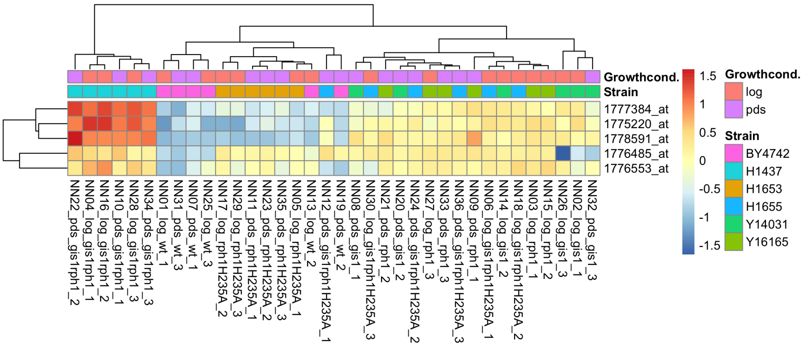


Figure 1. R and *pheatmap* package applied to visualise five yeast probe sets across several yeast strains. Rows and columns of the heat map are hierarchically clustered based on gene and sample expression profile similarity. Image has only illustrative purpose.

interpreted visually. There are several R (R Core Team, 2014) packages that provide summarised report across number of quality criteria measured (Kauffmann et al., 2009; Parman et al., 2005). However batch effects may be more difficult to discover. One option is to use principal component analysis (PCA), as done by *arrayQualityMetrics* (Kauffmann et al., 2009), to summarise the variation of high dimensional data into fewer dimensions. PCA components are linear combinations of original data. They describe directions in which the data shows highest variance. Each following component is fitted on the data from which previous components have been subtracted. First few principal components can then be plotted on two-dimensional scatterplots. Batch effect may arise from different time of day when the microarrays were processed or slight alterations in sample preparation kits. Once detected they can be removed using dedicated software as for example *ComBat* (Johnson et al., 2007).

1.1.3. Visualisation

Human vision is not adapted to read large numeric matrices. Therefore, for visualisation purposes, it is more reasonable to display expression matrices as either heat maps (Figure 1) or line graphs (Figure 2). Heat map conveys numerical information via colours. Human eyes can separate coloured spots only few pixels wide, so expression data can be presented in much more compact manner. Usage of hierarchical clustering and functional cluster detection algorithms can increase the compactness effect even more (Krushevskaya et al., 2009). In expression heat maps different colour codes are used to indicate down and up regulation, negative and positive expression respectively (Figure 1).

Line graphs are suitable to visualise vectors. Line graph is a two dimensional plot where one axis indicates individual samples (usually x axis), the other axis indicates expression signal values (y axis). Individual lines depict expression dynamics for one probe set across samples (Figure 2).

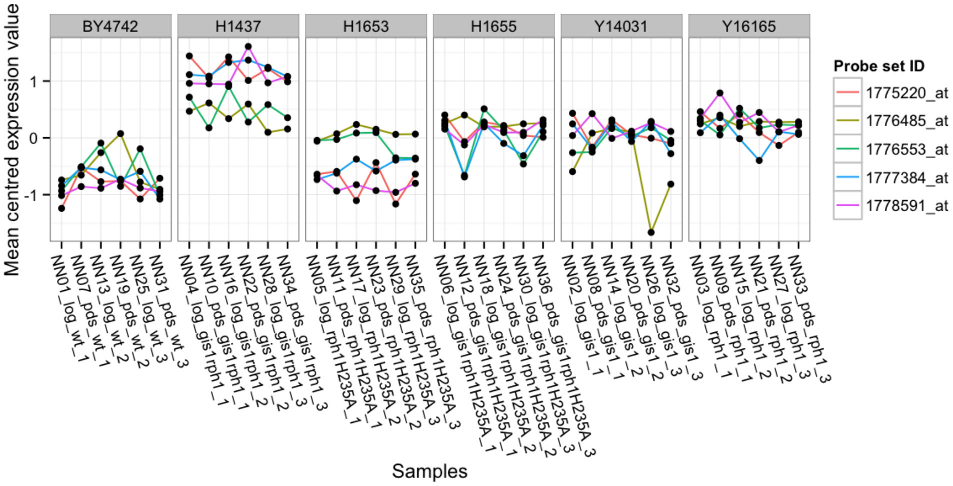


Figure 2. R and *ggplot2* package applied to visualise five yeast probe sets across several yeast strains. Here, individual probe sets are indicated with different colours. Individual yeast strains are separated into panels. Lines are coupled with points to highlight individual sample values. Image has only illustrative purpose.

Alternatively, scatterplots can be used to compare two probe sets or two samples against each other. In that case both axes would indicate expression signal values and points on the plot depict individual samples or probe sets respectively (see Figure 3).

1.1.4. Public data

Tens of thousands of microarray experiments have been conducted across the globe. The implications and value of the data generated mostly exceeds the experiment in question. After researchers have received the answers they were seeking for from the data and published their article(s), they also release the data into public domain for two principal reasons. First, that others could validate their results. And secondly, so that others could reuse their data to answer new questions. Over the years there have formed several specific and general databases to store high-throughput gene expression data. Such databases are for example GEO (Edgar et al., 2002) and ArrayExpress (Brazma et al., 2003) for general purposes or more specifically themed databases such as “The Cancer Genome Atlas” (Collins and Barker, 2007), “Connectivity map” (Lamb et al., 2006), “Cancer Cell Line Encyclopedia” (Barretina et al., 2012).

To ensure data reusability after initial publication the data needs to be annotated with at least minimal information about experimental setup and sample descriptions (Brazma et al., 2001). Minimum Information About a Microarray Experiment (MIAME) framework does provide such standards and is enforced while publishing data to any public database. The better the experiments and samples are described, the easier it is to be reused by an external researcher. However, demanding too rigorous data upload standards diminishes the overall data submission rates (Rung and Brazma, 2013).

ArrayExpress and GEO are the two oldest and most widely used general purpose high-throughput data collecting databases. Besides array based data they also accept sequencing data. During the time period from 2010 to 2012 ArrayExpress data volume grew three times, from 370000 assays and 13000 experiments to over million assays and 30000 experiments (Rustici et al., 2013). In recent years the proportion of high-throughput sequencing data has grown. Although the methods published within this thesis are built mostly for the microarray technology, they are not limited to it. Expression matrix can be extracted from sequencing data with similar properties as for microarrays (Guo et al., 2013). RNA sequencing allows to measure and detect gene expression in greater detail. It has been shown to correlate with existing Affymetrix GeneChip measurements very well (Marioni et al., 2008).

1.2. Data analysis

Broadly we can divide high-throughput data analysis into two categories. Firstly, differential expression analysis where two or more groups of samples are compared. Mean expression value of a gene is estimated based on the samples in each group and then compared. Genes that expression levels are statistically different are usually sought. The other type of analysis is a co-expression analysis where genes or samples are grouped based on their expression profile similarity. Expression profile is a vector of numbers that represent expression values across all samples for a gene or across all genes for a sample.

1.2.1. Differential gene expression analysis

Microarrays provide invaluable resource to study gene expression in different biological conditions. Response to drug treatment, healthy *vs.* disease, comparison of tissues or different time points – these are only small number of examples that high-throughput expression analysis helps us to analyse. Differential expression analysis determines which genes have altered expression level between sample groups. In case of, for example, cancer we look for genes that are up-regulated or down-regulated to study about the potential cause or mechanism of the condition and to identify potential drug targets. To understand the effects of the drug we want to identify the genes that have been affected by the treatment.

Mean expression value for a gene is calculated within both groups and the difference is often presented as a fold change (FC). Fold change “2” means that the measured mean expression value in target sample group is twice as high compared to the reference sample group in original signal scale. Alternatively logFC can be used to express log transformed expression difference in linear scale for both positive and negative direction.

Comparing the mean expression level alone does not suffice to conclude that a gene was differentially expressed. The assumption is that gene expression levels or values across all samples are normally distributed or rather sampled from normal distribution. Normal distribution can be described by its mean value and standard deviation. For each sample group we try to estimate the mean value of underlying distribution. How confidently we can estimate the mean value depends on the variance of the data and number of samples within the group. The fewer samples and higher standard deviation, the less confidently we can estimate the mean value. Slightly different mean values for groups do not mean that they can not be sampled from the same normal distribution. The statistical significance of the difference is typically determined with the *t-test* or *moderated t-test* (Smyth, 2004).

High-throughput gene expression experiments contain thousands of individual genes. The differential expression statistic is calculated for each gene. As *t-test* assumes normal distribution and treats observed data as sampled from larger underlying population, the intrinsic variation of this kind of sampling should be considered. When performing thousand *t-tests* on two vectors sampled from the same underlying normal distribution, by average 50 individual *t-test* p-values would still be smaller than standard 0.05 significance threshold. The p-value is a measure of probability that null hypothesis is true, in this case that all values in these two vectors come from the same underlying distribution. By using significance threshold 0.05 we accept that there is 5% probability that the observed results are false positives, as they are in this example.

In statistics multiple testing correction is used to alleviate the discovery of false positive results. Popular methods for multiple testing correction are Bonferroni correction and false discovery rate (FDR) (Benjamini and Hochberg, 1995). Out of the two Bonferroni is more conservative. Each individual p-value need to satisfy following criteria: $p \leq \alpha/n$, where α is significance threshold (standard 0.05) and n is the number of related tests performed. The FDR correction is less conservative. In increasingly ordered p-values vector, individual values need to satisfy $p_k \leq \frac{k}{n}\alpha$, where k is the index of p-value being corrected, n and α are the same as in case of Bonferroni correction.

Very popular statistical computing platform R has many packages to work with high-throughput microarray data, one of the more popular ones is *limma* that uses linear models and empirical Bayes (referred as the *moderated t-test*) to assess differential expression (Smyth, 2005). The *limma* packages is available though *Bioconductor* initiative (Gentleman et al., 2004).

A classical differential expression experiment setup is perturbation experiment that allows to infer direct and indirect targets of a transcription factor. In perturbation experiments the expression of a transcription factor is altered in controlled conditions. For example a transcription factor or gene of interest (or its exon) is deleted from the genome (knockout) (Hu et al., 2007) or over-expressed (Butcher et al., 2006). Other genes that are affected by the absence or over-expression of the transcription factor are identified by comparing samples where it is not perturbed.

The result of differential expression analysis is a list of genes that are differentially expressed. They can be ordered based on statistical significance or fold change. First inquiry would be to learn any prior knowledge about the genes in the list. What are their known functions and whether they can, and if yes, then how, explain the results of the current experiment. It is reasonable to look for up-regulated and down-regulated genes separately.

Over the years knowledge learned about the genes have been gathered into database called Gene Ontology (Ashburner et al., 2000). Gene Ontology (GO) has become invaluable tool to analyse and interpret gene set enrichment analysis. Gene Ontology has a tree like structure – a directed acyclic graph. Terms, that group together functionally related genes, start out to describe the generic functionality and with depth grow to more specific. All genes related to a term are also part of the parent term. A gene can belong to many terms. The Gene Ontology is divided into three trees: molecular function, cellular component and biological process.

There are two principal ways to conduct such functional enrichment analysis. First is by the use of hypergeometric overlap statistic as done in g:Profiler toolset (Reimand et al., 2011). The other option is to use the weighted Kolmogorov-Smirnov-like statistic as showed by Subramanian et al. (2005). Either case multiple testing correction should be applied to reduce the discovery of spurious terms. Gene set ontologies in g:Profiler include GO, biological pathways (KEGG (Kanehisa and Goto, 2000), Reactome (Joshi-Tope et al., 2005)), regulatory motif annotations (TRANSFAC (Matys et al., 2006), miRBase (Griffiths-Jones et al., 2006)), CORUM protein complexes (Ruepp et al., 2008), Human phenotype ontology (HPO) (Kohler et al., 2014) and BioGRID protein-protein interactions (Stark et al., 2006).

g:Profiler toolset includes also g:Convert tool, that enables seamless conversion from any popular gene namespace to specified target namespace (Reimand et al., 2011). That is essential to incorporate resources that originally use different namespaces for gene identification. In g:Profiler the centric namespace is Ensembl gene ID. For namespace conversions g:Convert uses Ensembl BioMart (Cunningham et al., 2015; Kinsella et al., 2011) database and cross-references therein. As g:Profiler has been developed within our work group, we have extensively used it in our research and it is also incorporated into tools presented in this thesis.

1.2.2. Gene co-expression analysis

Co-expression measures similarity between genes. Genes are compared to each other by their expression profiles across samples. Genes that share similar profiles are indicated to be co-expressed. This has been exploited to infer function to a novel gene by its expressional similarity to a known or a group of known genes.

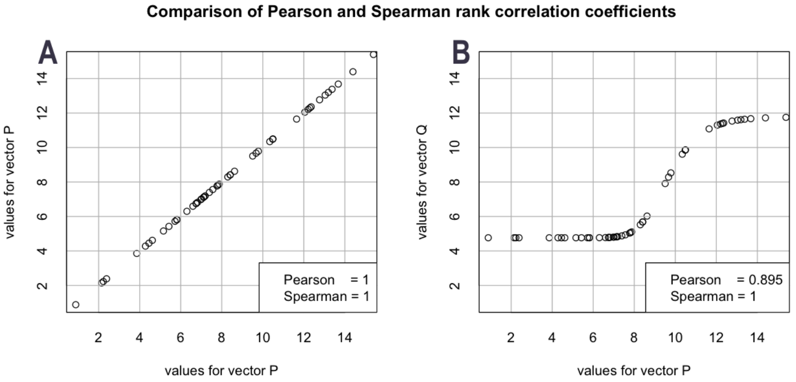


Figure 3. Example of different correlation interpretation between Pearson correlation and Spearman rank correlation coefficients. A) Scatterplot of vector *P*, that is sampled from normal distribution, against itself. Both correlation coefficients recognise it as perfect correlation. B) Scatterplot of vector *P* against vector *Q*, here Spearman rank correlation coefficient shows perfect correlation, while Pearson correlation coefficient does not.

One of the first experiments that used high-throughput expression data and co-expression analysis was published by Spellman et al. (1998). They used Fourier transformation and Pearson correlation coefficient values between candidates and known cell cycle genes to identify periodically regulated genes during cell cycle.

Wolfe et al. (2005) used five co-expression networks and the Gene Ontology to validate the general principle of “Guilt by association”. They found strong tendency for transcriptional co-expression in well over 900 Gene Ontology terms.

Clustering is often used to group together genes that behave similarly. Most common clustering methods are hierarchical clustering and K-means clustering.

Hierarchical clustering creates a cluster for each object – either gene or sample. Correlations or distances are calculated for all against all clusters. Two clusters that are the closest or most correlated are merged. These steps are iterated until only one cluster remains. This not only assigns all genes (or samples) to a cluster, but shows hierarchical relations between all clusters. It is often presented with a dendrogram as shown in, for example, VisHiC tool (Krushevskaya et al., 2009).

While merging clusters in hierarchical clustering there are three principal options: complete, average and single linkage. They translate to comparing the most distant members in both clusters, comparing average of all against all distances between clusters and comparing closest members between two clusters respectively.

K-means clustering requires *a priori* number of clusters it tries to fit on the data. Given the number of clusters, K-means randomly assigns objects as centroids (initial cluster centres) and then all the rest of the objects are assigned to the closest cluster. Then new cluster centres are calculated as the mean across all the objects in the cluster. These steps are iterated either number of user specified times or until the clustering converges.

One alternative to standard K-means clustering is fuzzy K-means clustering, where objects are not assigned strictly to one cluster, but rather given a score of their probability to belong to any given cluster (Gasch and Eisen, 2002).

Hierarchical and K-means clustering can also be used in combination to alleviate the shortcomings of both methods, as shown by Chen et al. (2005) for example.

1.2.3. Distance metrics

When comparing genes or samples to each other it is very important to understand what is expected out of the comparison. The resulting clusters and interpretation of the results depends heavily on distance metrics used to compare expression profiles. There are number of metrics to calculate either distance or correlation coefficient between numeric vectors. Distance is non-negative real number, the value 0 indicates identical vectors. Correlation values range between -1 and 1, where 1 depicts perfect correlation, -1 perfect anti-correlation and 0 shows no correlation between the two vectors. Correlation can always be transformed into distance measure by subtracting it from 1 (*i.e.* $d = 1 - r$, where d is distance, and r is correlation coefficient).

The simplest distance between two vectors with equal length is Euclidean distance. For vectors $P = (p_1, p_2, ..., p_n)$ and $Q = (q_1, q_2, ..., q_n)$ Euclidean is calculated by following formula:

$$d_{P,Q} = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}. \tag{1.1}$$

For microarray data where expression levels of individual genes depend on many technical and normalisation artefacts, Euclidean distance may produce large distances between genes that are still similar in terms of correlation. For correlation two most popular metrics are Pearson correlation and Spearman rank correlation coefficients. Pearson correlation coefficient measures correlation or dynamics of two vectors – do they go up and down synchronously. Pearson correlation coefficient measures linear dependence between two variables. Spearman rank correlation coefficient measures monotonic dependence between two variables which does not have to be linear, see Figure 3. In other words, Pearson correlation coefficient assumes normal distribution, while Spearman does not. Both metrics use the same core formula to calculate correlation:

$$r_{P,Q} = \frac{\sum_{i=1}^{n} (p_i - \bar{p})(q_i - \bar{q})}{\sqrt{\sum_{i=1}^{n} (p_i - \bar{p})^2 \sum_{i=1}^{n} (q_i - \bar{q})^2}}. \tag{1.2}$$

The difference is that while Pearson correlation coefficient is calculated on original values, Spearman rank correlation coefficient first assigns ranks to the values and then the formula is applied on rank values. See Figure 3 for different interpretation between Pearson correlation and Spearman rank correlation coefficients.

A special case of correlation is absolute correlation, where only the strength of linear (or monotonic) dependence is measured, regardless its direction – absolute value of the correlation coefficient.

Yona et al. (2006) compared most popular distance measure for microarray data. They showed that Euclidean distance is indeed consistently outperformed by correlation based measures. Spearman rank correlation coefficient was highlighted as consistently performing better than other classical metrics. However, in one out of four datasets tested, Spearman was outperformed by all others and Euclidean showed very good results. It shows that the choice of distance metrics should be highly data oriented – how data is generated, normalised and scaled, whether it is mean or median centred, *etc* – all these factors should be considered.

There is also a number of modern distance and correlation measures, for example Mutual information distance (Priness et al., 2007) and Maximal information coefficient (Reshef et al., 2011). In addition to linear dependence or simple distance between vectors, these measures are capable finding other types of dependencies or patterns in the data. However, they are computationally more expensive to calculate and have yet to gain their popularity within bioinformatic community.

1.3. Data analysis across datasets

There has been several studies that explore how much and to which extent individual gene expression datasets and different platforms can be combined. Here we highlight only few.

Irizarry et al. (2005) have in addition to platform design taken into account the laboratory effect. They demonstrate using experimental data from 10 different laboratories and 3 different platform technologies that while the analysis results generally agree, there are still relatively large differences in laboratory to laboratory correlations. In this study, the results from the best laboratory agree better across different platforms that to other laboratories. This highlights the importance of laboratory installation, equipment quality and technician experience and skills.

The general concurrence of results between datasets has been also shown on data provided by MicroArray Quality Control (MAQC) consortium (Shi et al., 2006). They note that reproducibility of differentially expressed gene lists across datasets can be improved when ranked fold change with non-stringent p-value cutoff is used instead of ranked *t-test* p-value (Shi et al., 2006).

1.3.1. Meta-analysis

Analyses performed across multiple independent datasets with integrated results are referred as meta-analysis. In the examples here we review only meta-analyses performed on high-throughput gene expression datasets.

Chen et al. (2013) showed that differential expression analysis across many related datasets can also be performed using one unifying Bayesian integrated modelling. The constraint here, and in any method that analyses differential expression across datasets, is that sample class labels need to be provided and only studies comparing relatively similar biological conditions can be integrated in such manner (Chen et al., 2013).

Most genes are active only in very specific stage of organismal development, cell cycle phase or environmental condition. In case of microarrays we can always observe expression profile of any gene in all datasets, regardless whether it is truly expressed or not. We can use this gene profile to find other genes that behave similarly, but this similarity might not be very consistent from dataset to dataset. The observed similarity in single or few datasets can occur by random chance or be caused by very specific experimental condition. While observing gene to gene similarity across larger collection of datasets we can be increasingly more confident of found similarity.

Lukk et al. (2010) combined data from 206 public microarray gene expression datasets into single large scale expression matrix. All datasets were from Affymetrix GeneChip Human Genome HG-U133A platform that was the most popular gene expression platform at the time. All data was quality controlled and normalised using RMA normalisation method. Almost half of the samples were excluded after quality control and duplicate sample removal. Still, it is one of the largest gene expression datasets to date. The samples in final dataset were manually curated and annotated to specified meta-groups with various degree of detail. They used mainly principal component analysis to show how different biological and experimental conditions are related to each other. In the article they show how the first two principal components can be viewed as hematopoietic and malignancy components respectively and how different cell line expression profiles are from normal and diseased samples expression profiles (Lukk et al., 2010).

Raw data normalisation across many experiments allows better sample to sample comparison and augmentation of weak signals by combining class labels from different labs. Yet, while doing so, one needs to keep in mind that such raw data normalisation is unable to remove undesired features from the signal such as laboratory or batch effects (Rung and Brazma, 2013). Depending on analysis it might be safer to leave independent experiments intact to have better control of between-laboratory heterogeneity and use summary-level meta-analysis instead (Rung and Brazma, 2013).

Another issue is data sample annotations. Poor sample characterisation within public data makes large scale differential expression meta-analysis infeasible. In such cases co-expression analysis between genes can be used instead. As it does not set any demands on sample labels it makes incorporating large collection of data into single analysis easier. There are three principal methods to apply co-expression in meta-analysis across datasets.

First option is to calculate all against all distances and use statistical heuristics to identify significant correlations within the data. Data randomisation or simulation analysis can be used to derive such heuristics. This option was first introduced by Lee et al. (2004) and later applied in web-based tool GEMMA published by Zoubarev et al. (2012). The meta-analysis part here is to count in how many datasets the observed connection between gene pair is deemed significant. It was also concluded that a connection can be considered trustworthy if it appears at least in three independent datasets (Lee et al., 2004).

Second option is to normalise within dataset distances so, that they would become comparable across datasets. Huttenhower et al. (2006) calculated all against all distances within each dataset and used Fisher's Z-transformation. After subtracting the mean and dividing with dataset standard deviation the resulted Z-scores are normally distributed and suitable for across dataset analysis. Hibbs et al. (2007) applied this methodology on yeast expression datasets and developed web tool SPELL. SPELL allows multiple genes in the query and assigns weights to individual datasets based on average co-expression values between individual query genes. The final order of similar genes is decided by their weighted average correlation across all datasets for all query genes. This has been further developed and applied to human data by Zhu et al. (2015). In addition to dataset weighting as used in SPELL they introduce hubbiness correction. From each query to target gene co-expression Z-score they subtract average co-expression Z-score between target gene and the rest of the genome. This then shows relative similarity between the query gene and the target gene (Zhu et al., 2015).

Third option is to calculate co-expression values between query and all the other genes, order them based on the observed co-expression value, assign ranks and use a rank aggregation approach to obtain a global similarity metrics. We demonstrate this approach in Sections 3.3 and 3.4. Rank aggregation avoids dealing with individualities of the datasets by dealing only with the final product – ordered gene lists (Pihur et al., 2009).

1.4. Pathway databases

Over the years many protein-protein, gene-gene or metabolite-enzyme relations have been characterised. This gathered knowledge is systemised in pathway databases. Pathway databases provide excellent framework to interpret and propose hypothesis for new biological high-throughput experiments. Pathways consist of a group of related genes and chemical compounds (metabolites for example), based on common task or process. Typical layout is a graph, where pathway members are shown as nodes and specific interactions are depicted by edges between corresponding nodes.

1.4.1. KEGG

Kyoto Encyclopedia of Genes and Genomes (KEGG) contains mainly text mining based interactions built into pathways (Kanehisa and Goto, 2000). Pathway entity is global across all organisms. Referring to evolutionary aspect – centric function has evolved in some point of time and for most organisms the backbone will remain the same across evolution with slight variation in participants. Path-

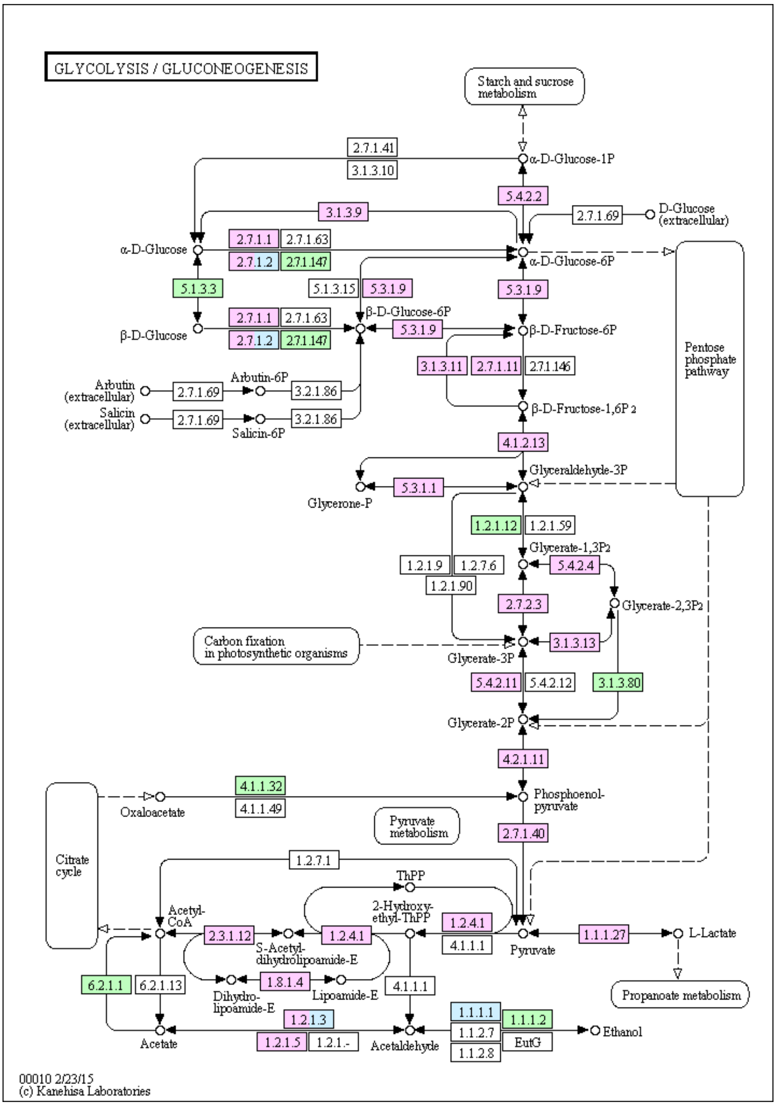


Figure 4. Example of KEGG *Glycolysis / Gluconeogenesis* pathway map. Metabolites are indicated with circles, enzymes with rectangles, arrows and lines between pathway members indicate reactions. Enzymes are identified by their nomenclature name (NC-IUBMB, for example EC 2.4.2 – Pentosyltransferases). KEGG pathway maps cover many organisms, in this example human specific enzymes are highlighted using colour codes: light-green – present in human, pink – present in human and associated with a disease and light blue – present in human and target of a drug. Hyperlinks to related pathways or additional information about pathway entities are provided when viewed online on KEGG website[2].

[2] http://www.genome.jp/kegg-bin/show_pathway?org_name=hsadd&mapno=00010

way images are hand curated graphical pictures. Organism specific pathways are highlighted using organism specific image annotation files and colour codes. For a long time KEGG has been the largest publicly available pathway database. They provide good coverage of core metabolism and signalling pathways. In addition, they have a variety of disease and drug related pathways. They have well documented technical information regarding the pathways which makes them computationally easy to use. KEGG is also known for its compound database, which are also integrated into pathways. An example of KEGG *Glycolysis / Gluconeogenesis* pathway is shown on Figure 4. In our tool KEGGanim, we provide a possibility to visualise the quantity of compounds along side with expression levels of genes or proteins, see Section 3.1.

1.4.2. Reactome

Reactome is a database where experimentally validated interactions are built into hierarchical pathways (Joshi-Tope et al., 2005). All interactions in Reactome pathway database are reviewed by experts in the field. In addition to literature evidence, reactions are also linked to experimental data. Interactions in the pathways are viewed as chemical reactions. Updates in the pathways are done in close collaboration between experts in biology and Reactome curators (Milacic et al., 2012). The hierarchical structure allows to have a general overview of biochemical process as well have more detailed view of individual sub processes. Crosslinks are provided to other bioinformatic databases that provide additional annotations and information about reaction components.

Reactome has very good high-quality coverage of the core metabolism and signalling pathways. Recently they have put more effort to characterise disease and cancer related signalling pathways. The emphasis is to show how disease state and wild type state diverge (Milacic et al., 2012).

All pathways can also be viewed as ontology terms and have been used as such in gene set enrichment analysis methods and tools, for example in g:Pofiler (Reimand et al., 2011).

II. AIMS OF THE PRESENT STUDY

The aim of current thesis is to develop methods for public high-throughput data analysis and visualisation. The specific aims of this thesis are the following:

Ref. I To develop high-throughput expression data visualisation framework for KEGG pathways and package it as a web tool for public usage.

Ref. II To study the extent that public high-throughput gene expression data can be applied to biological pathway reconstruction and augmentation.

Ref. III To develop statistical method for data driven query based co-expression analysis across hundreds of public gene expression datasets and provide methods of visualisation.

Ref. IV To study further properties of the proposed rank aggregation method, highlight its features and its applicability to various biological experimental setups.

III. RESULTS AND DISCUSSION

3.1. Pathway animations of high-throughput data (*Ref I*)

We have developed KEGGanim (`biit.cs.ut.ee/kegganim`) web tool to combine previous knowledge on cellular mechanisms and high-throughput experimental data. KEGGanim makes possible to visually observe gene and protein activity within KEGG pathways (Kanehisa et al., 2014). High-throughput gene expression experiments allow to take snapshots of living organisms transcriptional state, a cell profile. Pathways represent our best understanding of genetic and metabolic interactions and enable us to have a focused view to a specific group of interactions within larger data.

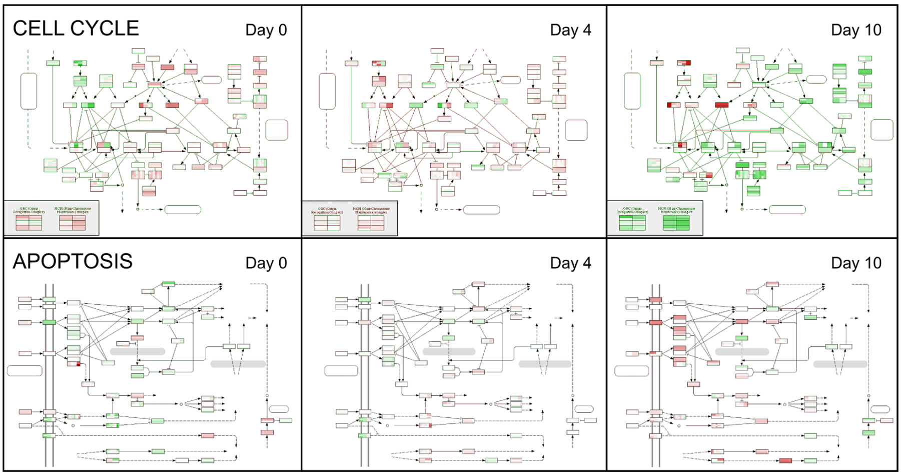
At the time of the initial development of KEGGanim web tool, KEGG was the largest database with most genes and organisms annotated to its pathways. KEGG pathway database has intuitive pathway images and corresponding configuration files. Textual description of pathway image layout – the position of proteins, interactions and links are provided in tab separated text file format. This is considerable asset as most pathway databases do not provide portable annotated graphical images of pathways. Presently KEGG database has approximately seven thousand genes annotated to pathways. That is still less than one third of all genes annotated to human genome so far. By combining high-throughput expression data and KEGG pathways we can understand better depicted cellular processes and also spot gaps in our knowledge about genetic networks.

Our examples are concentrated on high-throughput gene expression platforms (microarrays) as these are most widely used to study gene expression and provide information simultaneously for many genes (up to entire genome).

Implementation

KEGG provides its source data via File Transfer Protocol (FTP). Database is updated on daily basis, as new knowledge comes available the relevant pathways are updated. For each pathway in each supported organism, there are two files: an image file with graphical layout and annotation file binding gene (rectangle) and metabolite (circle) annotations with respective shape coordinates on the image. We use *Perl* script to parse the annotation files and retrieve relevant information.

Perl has also good web application support via Common Gateway Interface (*CGI*) module. For graphics we use Simple Web Object Graphics (*SWOG*) language and its *Perl* module that has been developed by Jaanus Hansen (Hansen, 2005). Its seamless compatibility with *Perl* allows it to be integrated into compu-



	Cell Cycle	Apoptosis
Day 0	Bub1b, Bub2, Cadd45a, Ccna1, Ccnb2, Ccne1, Cdc20, Cdc25a, Cdc25b, Cdc2a, Cdc45l, Cdc6, Cdc7, Cdk2, Cdk7, Chek1, Chek2, Crebbp, Dbf4, E2F1, E2F4, Espl1, Fzr1, Hdac1, Mad1l1, Mad2l1, Mad2l2, Mcm2, Mcm3, Mcm4, Mcm5, Mcm6, Mcm7, Mdm2, Orc1l, Orc2l, Orc5l, Orc6l, Pcna, Pkmyt1, Plk1, PPtg1, Prkdc, Sfn, Smc1a, Smc1b, Wee1	Atm, Bax, Bid, Chuk, Dffa, Dffb, Irak3, Trp53
Day 10	Cadd45b, Cadd45g, Ccnd2, Ccnd3, Cdc16, Cdkn1a, Cdkn1e, Cdkn1b, Cdkn2b, Cdkn2d, Rb1, Rbl1	Apaf1, Bad, Birc3, Capn2, Casp6, Casp7, Casp8, Casp12, Cf1ar, Csf2rb1, Csf2rb2, Endog, Fas, Il1b, Il1r1, Il3ra, Irak2, Map3k14, Myd88, Nfkbia, Nkfb1, Nkfb2, Rela, Ripk1, Tnfrsf1a, Tnfrsf10b, Tradd

Figure 5. Image depicts *Cell Cycle* and *Apoptosis* KEGG pathway dynamics during mouse embryonic development, adapted from Schulz et al. (2009). Gene lists in the bottom of the image represent pathway members that are either up regulated (orange) or down regulated (blue). On the pathway image, down and up regulated genes are indicated with green to red colour scale, respectively. KEGGanim "Cinefilm" tool was used for composing this image. The image was compiled by Raivo Kolde.

tational analysis and visualisation pipelines (Adler et al., 2009a; Reimand et al., 2011). *SWOG* allows to read in an image and to modify and add layers to existing and new images. We apply *SWOG* to combine KEGG pathway images and high-throughput expression data. Expression values are converted into colour codes and plotted over relevant pathway components on KEGG pathway image.

KEGGanim output is an animated Graphics Interchange Format (*.GIF*) image. Animation is achieved by looping images depicting individual samples. Each image becomes a frame in the animation. User can download the animation and display it inside slide presentations or in a web report. As the animated *.GIF* image would not be a suitable illustration for publication, we also provide an alternative static output method called "Cinefilm". This creates a single image where user selected frames are stitched together side by side. Annotated and graphically appealing visual aids can help to make intricate biological network relations and concepts easier to follow and to understand.

Users can upload their own data. The interpretation of the numeric values depends on the intent of the user. For example, it might be preferable to display relative fold change for a genes across sample groups instead of individual expression levels.

User uploaded private data is stored in password protected folders. File and user management is organised by g:Pedam function. g:Pedam is compatible with native GEO Simple Omnibus Format in Text (SOFT) to encourage public data re-usage. g:Pedam was implemented by Jüri Reimand.

The data upload to KEGGanim itself is not limited to any specific data type. The format of the data is a straightforward numeric matrix, where columns correspond to samples and rows to genomic features. For example tab separated file (*.TSV*) with a header row to describe sample annotations and first column containing common gene or protein names is a frequently used format. Any new dataset can be easily converted to such a format and visualised in KEGGanim framework.

In addition to user own uploaded data we provide 36 example public datasets to be visualised with KEGGanim tool. This serves as a showcase for new users.

Over the years there have been developed many different namespaces for representing gene IDs. KEGG database mostly uses Entrez Gene IDs, however there are exceptions. To make the general mapping as robust as possible we use Ensembl namespace as the common ground. Genomic feature IDs from data and from KEGG database are mapped to Ensembl gene IDs using g:Convert framework (Reimand et al., 2011), many to many mappings are allowed and visualised accordingly. Genes sharing the same enzyme annotation divide corresponding rectangle horizontally. In case the same gene is represented by multiple features, for example microarray probe sets, the rectangle is divided vertically. Any combination of above is allowed, each sub-rectangle displays then the expression level of a single feature from the data.

Unfortunately for metabolites there are no unambiguous way to map IDs from different sources, thus we leave it up to the user to map their metabolite IDs to

KEGG namespace. Metabolites also have their own, per metabolite, colour scale and is comparable only to itself between samples.

Use cases

One of the examples of KEGGanim usage is motivated by the question of what biological processes are affected in a specific study. Usually first part of an analysis is gene expression comparison measured in two conditions. Differential expression analysis enables to identify more interestingly behaving genes within the experiment. Enrichment analysis can be used to identify biological processes that are most perturbed in the experiment. Although enrichment analysis can identify relevant KEGG pathways that are more interesting, it does not provide insight into how specified enriched genes are related to each other within the pathway. Proper visualisation helps to understand pathway layout and positional effect of the enriched genes. KEGGanim enables such visualisation to illustrate expression dynamics within pathways between the various conditions.

KEGGanim is suitable to visualise time series experiments. For example, FunGenES project expression data analysis of mouse embryonic development revealed symmetric expression changes in cell cycle and apoptosis pathways (Schulz et al., 2009, Figure 5). While in early development cell cycle is active, it will be down regulated during later phases of embryonic development as the organism gets closer to birth. On the opposite side apoptosis will be up regulated during later phases of embryonic development as tissues mature and organism moves toward balance in cell recycling.

Mashanov et al. (2014) used KEGGanim tool to show transcriptional changes during organ regeneration in sea cucumber. The visualisations of *Pathways in cancer* and *Focal Adhesion* pathways in three time points (days 2, 12 and 20 post-injury) are provided as Tagged Image File Format *.TIFF* images in publication supplementary material.

Another example is by Altmäe et al. (2012), where they have uploaded differential expression analysis results into KEGGanim tool and share the folder name and password with the readers. They compared day 3 *vs.* day 5 embryos and proliferative *vs.* midsecretory endometrial tissues to identify transcriptional changes that occur during embryo implantation.

Summary

Here, we demonstrated KEGGanim tool, that combines high-throughput expression data and KEGG pathway images. KEGGanim generates interactive animations across individual samples of the high-throughput data. Animations are suitable to be used in slide presentations or on the web. Individual frames of the animation can be stitched together using “Cinefilm” function. Such still images are suitable for publications.

3.2. Ranking genes by their co-expression to subset of pathway members (*Ref II*)

Pathway databases contain collected knowledge about genetic and metabolic pathways. Although vast, this knowledge is still far from complete. Out of approximately 22000 coding genes in human genome only one third is described in either KEGG (Kanehisa et al., 2014) or Reactome (Croft et al., 2014) databases. At the time of the publication Reactome database described the connections for 1804 protein coding genes and KEGG database for 4220 genes.

High-throughput experiments provide a lot of potential to identify interactions between genes. The principle of “Guilt by association” (Wolfe et al., 2005) has been repeatedly applied to infer functional annotations to poorly annotated genes as well to validate existing interactions between genes or proteins.

In our analysis we look whether and to what extent the principle is applicable in Reactome pathways and how using subset of pathway genes (sub-pathway) might improve the results. Reactome was chosen as a benchmark database as its networks were thoroughly validated.

Experimental setup

Many thousands of high-throughput gene expression experiments have been performed worldwide. Each such experiment contains information of many genes (up to entire annotated genome) across variation of biological conditions. In this study we used compilation of public gene expression data (Lukk et al., 2010). This is a vast collection of biological samples from Affymetrix GeneChip Human Genome HG-U133A platform. All data is gathered as raw *.CEL* files, which are quality checked and then normalised using RMA algorithm. The dataset is a collection of 206 studies generated in 163 separate laboratories. Out of initial 9004 samples 5372 remained after quality control and sample duplication removal. At the time of our analysis the dataset was still work in progress and also included 736 duplicate samples.

Affymetrix GeneChip Human Genome HG-U133A platform consists of 22283 unique probe sets which at the time of the study represented 12580 Ensembl genes. Some of the genes may have multiple probe sets mapped to its coordinates on the genome. However, these may or may not represent the same or overlapping transcript(s). In our study we used the most favourable probe set for such cases. For a gene pair we chose the probe sets that showed highest correlation. Ambiguous probe sets that mapped to multiple Ensemble gene IDs were omitted from the study.

The aim of our analysis was to study the predictive power of gene co-expression to infer connections between genes and biological pathways. For 35 selected Reactome pathways we performed exhaustive *leave-one-out* analysis. Iteratively leaving one gene out of the pathway we used remaining pathway members to predict the association between the rest of the genes on the platform and the pathway. For that we calculated average correlation coefficient between the pathway genes and each gene not in the pathway, including the left-out gene. Genes not in the pathway were then ordered based on observed average correlation coefficient and the rank of the left-out gene was determined. Rank 1 would indicate that the left-out gene is the closest gene to the rest of the pathway.

We were also interested whether there exists a sub-pathway that is more related to the left-out gene than entire pathway. Sub-pathway is here purely co-expression based measurement without taking into account interactions defined in Reactome database. We used correlation based threshold to define the active sub-pathway. Only those pathway members which correlation coefficient was higher than threshold t were used to calculate average correlation coefficient score between a gene and the pathway. As it was exploratory study, we tested different thresholds. For each *leave-one-out* iteration we optimised the threshold t so that the left out gene would get the highest rank possible.

In this study we used Pearson correlation coefficient, which is the most frequently used correlation metrics in biological studies. In biological data Pearson correlation coefficient measures expression dynamics similarity between gene expression profiles as discussed in Section 1.2.3. This would allow to identify gene groups and connections where the regulation of several genes are guided by the same factors. As an alternative we also tested absolute Pearson correlation coefficient.

Results

We evaluated the results for a pathway based on number of left-out genes that were retrieved within top n genes across individual iterations. We tested four n values: 1, 10, 100 and 1000, denoted by T1, T10, T100 and T1000. While using entire pathway to calculate average correlation, no pathway had more than 8% in T1 group. However while using sub-pathway we observed 3 pathways with more than 20% of left-out genes in T1 group: *Translation*, *Pyruvate metabolism and TCA cycle* and *Metabolism of noncoding RNA*.

For T100 group *Translation*, *Pyruvate metabolism and TCA cycle* and *Electron transport chain* pathways retrieved the left-out member close to 70% of the cases.

Interestingly we found that for some of the pathways absolute Pearson correlation coefficient performed comparatively better. For example the overall highest performance was shown in *Translation* pathway with 31% and 58% of pathway members seen in T1 and T10 groups while using absolute Pearson correlation coefficient. Although in general the difference between Pearson correlation coefficient and absolute Pearson correlation coefficient was not consistent enough to be significant.

The least performing pathways were all Reactome signalling pathways. We conclude that the regulation of signalling pathways is fairly independent of the transcriptional regulation. There may be pathway components where transcriptional regulation is involved, but more widely the signalling is conducted by protein modifications and enzymatic activities. Molecular protein complexes on the other hand have more straightforward transcriptional regulation. Mostly the parts required to build a protein complex are needed simultaneously. Nevertheless, our understanding of transcriptional co-expression landscape is still vague. Almost two thirds of all known genes are not placed into any pathway. This leaves ample room to study further the genes that are not annotated to any pathway, but expression wise are consistently more similar to the pathway than its members.

Our analysis revealed several candidate genes that were consistently very similar to *Translation* pathway genes. Upon closer look we could identify literature evidence between several candidates and pathway genes.

Summary

Biological pathways are an attempt to map cellular interactions into intelligible system. As new knowledge is gathered about the underlying reactions, the topology of pathways may change and evolve. One contributor to this change will also be high-throughput transcription analysis, that helps to characterise exiting interactions inside pathways and propose new ones.

In this study we did not consider pathways as undivided entities, but as largest allowed gene set. While performing exhaustive *leave-one-out* cross-validation we proved that using more tightly co-regulated sub-pathways shows more promise, than using the entire pathway. Also we used single gene expression matrix, although a large one with 6108 biological samples. Not all genes in all conditions are similarly regulated. It is possible, that two genes in one condition are co-regulated, in another are anti-regulated (in opposite directions) and in most conditions are not regulated at all. To describe and understand this selective co-regulation is the objective of the next publication in this thesis.

3.3. Mining co-expression across many experiments (*Ref III*)

The microarray gene expression data in public domain grows constantly. In this section we introduce Multi Experiment Matrix (MEM, `biit.cs.ut.ee/mem`) tool and show how to utilise this increasing resource to perform gene co-expression queries across thousands of datasets simultaneously.

There are different ways to combine data from different sources and it depends on the questions asked and the technical possibilities. Data from the same or very similar experimental platforms can be viewed together after proper normalisation (Lukk et al., 2010; Rung and Brazma, 2013). Laboratory specific batch effect should be considered as discussed in Section 1.1.2, but otherwise concatenated data can be viewed as single dataset and analysed as such. Wider spectrum of repeated experimental conditions can produce more stable results.

If data concatenation is not possible or not desired then normalisation and data analysis is performed independently in each data source and the results are combined (Hibbs et al., 2007; Huttenhower et al., 2006; Lee et al., 2004; Zhu et al., 2015; Zoubarev et al., 2012). These examples are more explained in Section 1.3.1.

To have a global overview of gene co-expression is a rather difficult task. Each individual dataset addresses defined experimental setup and is designed to measure genes' response to given stimuli or profile their expression across specified conditions. Using more datasets together would allow to study wider spectrum of conditions and gain more statistical power for the analysis. In previous publication (Adler et al., 2009b, *Ref II*) we used compilation of public data sets, a collection of 206 independent studies (Lukk et al., 2010).

In this section we explore an alternative option where co-expression is evaluated in many individual datasets and then combined into a single result by rank aggregation method. There are several advantages of rank aggregation over the concatenation approach. Direct comparison of individual datasets allows for better evaluation of co-expression stability across datasets. Most genes are not always expressed, there are some that are more widely expressed (often referred as housekeeping genes) (Thellin et al., 1999) and there are genes that are expressed rarely or in fewer conditions (Wang et al., 2009). When comparing the co-expression results from individual datasets we can observe in how many and in which datasets the genes are co-expressed. In addition combining more data together will make the whole analysis more robust against the potential noise in individual datasets.

Implementation

Lets denote the query gene g^* and the rest of the genes in the datasets as g_i, where i is their respective index. In each individual dataset MEM algorithm calculates pair-wise co-expression values between the g^* and all g_i. The final co-expression score between the g^* and g_i is achieved using rank aggregation method.

In every dataset included into the analysis all genes are sorted and ranked based on their co-expression value to the g^*. The most similar gene to the g^* will be assigned rank 2, second most similar rank 3 and so on for all genes in the dataset.

For each gene we get a rank vector containing a rank value from each corresponding dataset, denoted by $r(g^*, g_i) = [r^i_1, ..., r^i_m]$, where m is total number of datasets. We can state a null hypothesis that assumes no connection between the the g^* and a g_i. In such case the ranks in the rank vector observed would appear uniformly distributed. Alternatively, when the ranks appear not to follow uniform distribution, we have to reject the null hypothesis and can state with observed probably that there is a connection between the g^* and a g_i.

After sorting the normalised ranks in $r(g^*, g_i) \in (0,1)$ to get order statistic we can model the probability of observing a given rank value in any position in the vector using beta distribution. The beta distribution is described by two shape parameters α and β. Here α denotes the modelled position in sorted rank vector and $\alpha + \beta - 1$ is the length of the vector. We use beta one sided test to inquire how probable it is that observed rank value belongs to uniform distribution. Values that are smaller than significance threshold are of interest as they indicate that given g^* and g_i are consistently more similar in a number of datasets, denoted by the value of α. The lowest p-value in observed rank vector is treated as MEM similarity score.

The above can be summarised using $p_k = pBeta(r^i_k, k, (m - k + 1))$, where $k \in [1, m]$. The final score between g^* and g_i is then $sc(g^*, g_i) = \min\{p_1, ..., p_m\}$.

In the end we get a list of genes ordered by MEM similarity scores. These scores can be, after proper multiple testing correction, used also as *p-values*. Smaller score indicates more datasets where the g^* and g_i are consistently more similar to each other than would be expected by the null hypothesis. MEM workflow also records the ranks for g_i in each dataset. More often than not, after the appropriate clustering of datasets a pattern may emerge that highlights a group of genes that are similarly ranked to the the g^* in related datasets.

Not all the datasets may be relevant for given query gene g^*. In some of the datasets g^* might not be expressed at all or the expression of gene g^* is not affected by the experimental conditions studied. In case of gene expression microarrays this is typically reflected in lack of variance of expression values. We therefore have implemented predefined filter to exclude datasets from the analysis where the g^* does not display enough variance. Empirical testing showed that for our largest collection at the time of the study this threshold was 0.29. All dataset where the standard deviation value for gene g^* is lower than set threshold are omitted from the analysis. The threshold seems to apply well on all log transformed expression data, regardless of the collection size or platform (data not shown). This is reasonable as it excludes datasets with less meaningful correlation values. Other genes, that are not expressed or not affected by experimental setup, would in this case score high correlation values, adding potential noise to overall analysis.

As the data collections have grown in size, often a quick first look would be more engaging for new users. In the initial query we display the results using only up to 100 datasets where the query gene g^* shows highest standard deviation values. Unless the default standard deviation filter 0.29 is not reached first.

The validity of such initial restriction is justified by the mini chromosome maintenance (MCM) subunit correlation analysis. In the experiment we measured median rank distance between MCM subunits. We demonstrated that relative stability in median rank distances was already achieved with 75 datasets in the analysis. Of course, more datasets in the analysis will increase the performance and the threshold of 100 is foremost considered as a measure for providing quick glance on initial results.

Gene expression data

MEM co-expression queries are data driven. This means that the quality and quantity of the data are important for the analysis. One of the key issues here is the initial choice of data to be included into single analysis. All analyses in MEM tool are currently performed across single Affymetrix platform. This ensures that all datasets used in a single analysis measure transcriptomic entities in consistent manner. Some Affymetrix platforms are sufficiently popular and provide thousands of publicly available and variable datasets.

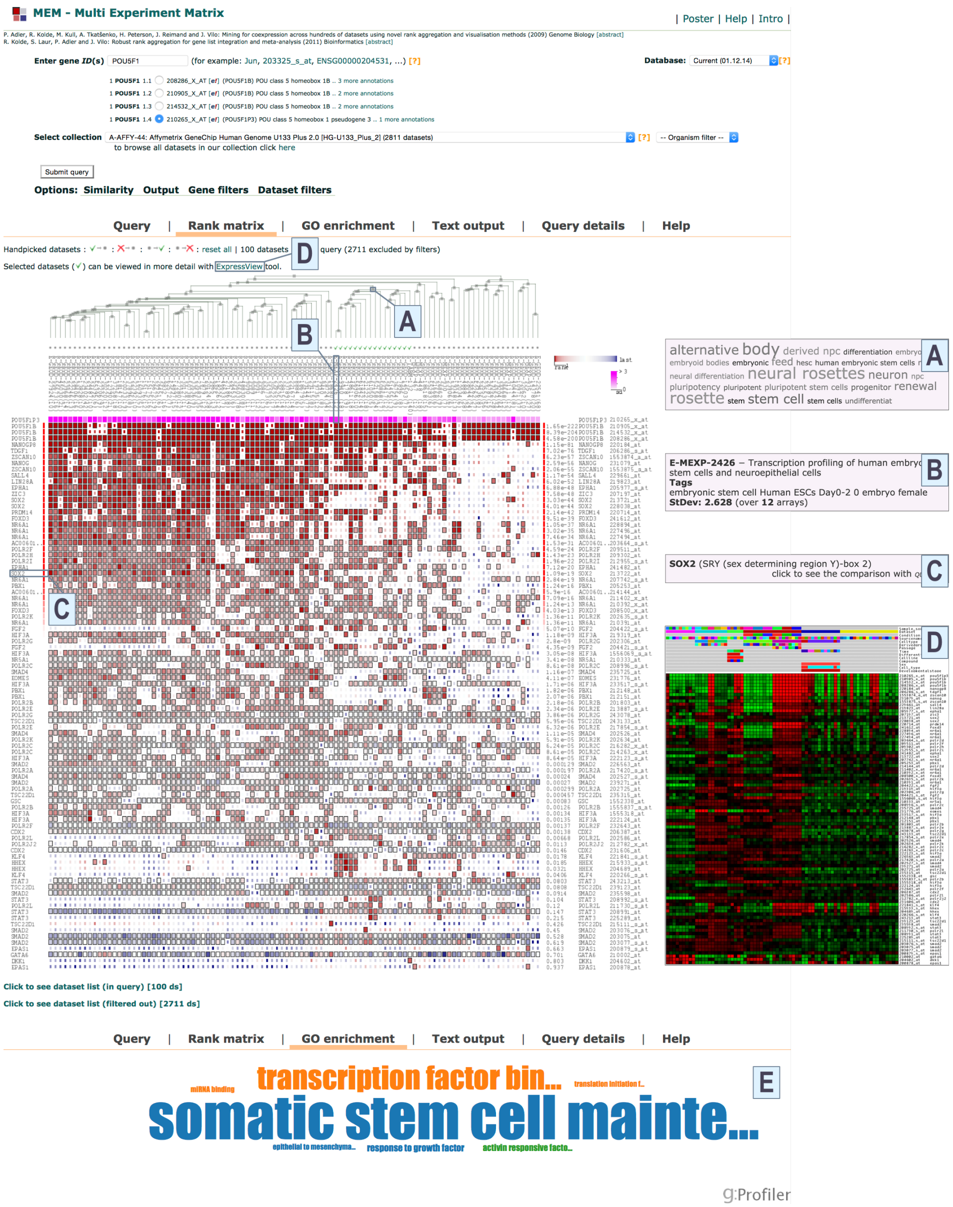
For our data collections we have downloaded all datasets from ArrayExpress database (Brazma et al., 2003), that met our requirements. We required the data to be from Affymetrix expression platforms and have raw data available. In the original collection we had data across 10 organisms, altogether 2467 datasets and more than 60000 samples. The latest data collection available at the time of writing this thesis has 30 organisms, 13252 datasets and more than 335000 samples (01.12.14).

We downloaded and normalised raw data on our own. Doing the preprocessing steps ourselves ensures more consistent quality and compatibility of the data. Publicly available processed datasets are often derived using independent normalisation and pre-processing pipelines, that are not easily compared (Rung and Brazma, 2013). We use the same steps and normalisation parameters across processed datasets. Normalised expression data and sample meta-data is preserved in Network Common Data Format (NCDF) files. NCDF is flexible binary data type, that allows rapid access to the data and thus enables online calculations. Meta-data can conveniently be stored in the same format and in the same file as expression data.

We aim to periodically update our collections as new expression data becomes available, covering more experimental conditions. Lately there has been significant shift toward sequencing data. That too, will be incorporated into future MEM data collections.

User interface

A web-based tool can more easily be used by wider audience. The large database behind the scenes (latest collection 102 GB, all collections 263 GB, 01.12.14) makes packaging not reasonable for the entire tool. The front-end web page is implemented as a series of *Perl CGI* scripts. Computationally heavy co-expression



MEM - Multi Experiment Matrix

| Poster | Help | Intro |

P. Adler, R. Kolde, M. Kull, A. Tkatšenko, H. Peterson, J. Reimand and J. Vilo: Mining for coexpression across hundreds of datasets using novel rank aggregation and visualisation methods (2009) Genome Biology [abstract]
R. Kolde, S. Laur, P. Adler and J. Vilo: Robust rank aggregation for gene list integration and meta-analysis (2011) Bioinformatics [abstract]

Enter gene *ID*(s) POU5F1 (for example: Jun, 203325_s_at, ENSG00000204531, ...) [?]

Database: Current (01.12.14) [?]

1 **POU5F1** 1.1 ○ 208286_X_AT [e!] (POU5F1B) POU class 5 homeobox 1B .. 3 more annotations
1 **POU5F1** 1.2 ○ 210905_X_AT [e!] (POU5F1B) POU class 5 homeobox 1B .. 2 more annotations
1 **POU5F1** 1.3 ○ 214532_X_AT [e!] (POU5F1B) POU class 5 homeobox 1B .. 2 more annotations
1 **POU5F1** 1.4 ● 210265_X_AT [e!] (POU5F1P3) POU class 5 homeobox 1 pseudogene 3 .. 1 more annotations

Select collection A-AFFY-44: Affymetrix GeneChip Human Genome U133 Plus 2.0 [HG-U133_Plus_2] (2811 datasets) [?] -- Organism filter --
to browse all datasets in our collection click here

Submit query

Options: Similarity Output Gene filters Dataset filters

Query | Rank matrix | GO enrichment | Text output | Query details | Help

Handpicked datasets : ✓→● : ✗→● : ●→✓ : ●→✗ : reset all | 100 datasets in query (2711 excluded by filters)

Selected datasets (✓) can be viewed in more detail with ExpressView tool.

alternative body derived npc differentiation embryo embryoid bodies embryonic feed hesc human embryonic stem cells neural differentiation neural rosettes neuron npc pluripotency pluripotent pluripotent stem cells progenitor renewal rosette stem stem cell stem cells undifferentiat

E-MEXP-2426 – Transcription profiling of human embryo stem cells and neuroepithelial cells
Tags
embryonic stem cell Human ESCs Day0-2 0 embryo female
StDev: 2.628 (over **12** arrays)

SOX2 (SRY (sex determining region Y)-box 2)
click to see the comparison with q

Click to see dataset list (in query) [100 ds]

Click to see dataset list (filtered out) [2711 ds]

Query | Rank matrix | GO enrichment | Text output | Query details | Help
E
transcription factor bin...
miRNA binding
translation initiation f...
somatic stem cell mainte...
epithelial to mesenchyma...
response to growth factor
activin responsive facto...
g:Profiler

Figure 6. Screenshot of MEM web interface and analysis output for embryonic stem cell regulator POU5F1(OCT4) gene. The top part of the image shows user interface for query specification. Displayed heat map highlights co-expression rank values between POU5F1 gene and Reactome *Transcriptional regulation of pluripotent stem cells* pathway. Analysis output was filtered using *Gene filters* option. Textual annotations on the image (A-D) highlight interactive features of MEM that allow to get additional information about the genes and experiments. (E) In the bottom of image g:Profiler (Reimand et al., 2011) enrichment analysis is visualised for outputted genes.

calculation and scoring is implemented in *C++* for speed. Since publication we have optimised, as well parallelised the *C++* backend, to cope with the growing data size.

Web interface allows, besides hosting the calculation and centralised data collection, interactive visualisation and cross references to related tools and resources. The emphasis of the MEM tool is to visualise the observed ranks for found globally similar genes relative to the query gene g^* as a heat map. The heat map uses red-white-blue scale to show which genes in given dataset were similar (red) or distant (blue). MEM rank aggregation method identifies in which datasets g^* and g_i were more similar than expected by chance. For each gene g_i in the results it can be different set of datasets. We have highlighted relevant datasets with a black frame on the heat map image for respective gene (Figure 6).

Datasets are clustered using the rank matrix and Happieclust (Kull and Vilo, 2008) hierarchical clustering algorithm. It is possible to display enriched keywords as a tagcloud for a cluster of datasets by hovering on the nodes in the clustering dendrogram. It provides additional insight into functional characterisation of experimental conditions studied in selected datasets. For example a group of genes is consistently co-expressed in datasets describing embryonic stem cells. Figure 6 provides illustration of MEM web interface and expected results.

We use extensively g:Convert tool for flexible gene ID identification (Reimand et al., 2011). User queries are converted from any known gene or protein ID into selected platform ID. In the output we perform the inverse. Often cryptic platform IDs are converted to common gene name and also gene description is retrieved.

The principal output of the tool, the list of significantly similar genes to the query gene g^*, can be further characterised using g:GOSt tool (Reimand et al., 2011).

Use cases

MEM tool enables to compare genes co-expression across datasets. Observe whether genes that are tightly co-expressed in certain experimental condition are so also in other tissue, clinical diagnose or cell line experiments. Some genes display co-expression across wide variety of public datasets, others in only few, biologically related, datasets.

MEM tool has found usage by other researchers to solve their independent problems or use as comparison while developing similar resource.

For example Altmäe et al. (2012) used MEM global co-expression analysis to evaluate the co-expression significance of physically interacting proteins.

Sircoulomb et al. (2011) used MEM to find co-expressed genes to a gene of their interest (ZNF703) on popular Affymetrix GeneChip Human Genome U133 Plus 2.0 microarray platform.

Ivanov et al. (2013) used MEM to characterise the co-expression partners for SOX10 gene in different cancer types.

Zhu et al. (2015) have developed SEEK tool that similarly to MEM performs co-expression analyses over large collection of human high-throughput gene expression datasets and was discussed in section 1.3.1. Noteworthy here is that they used MEM methodology as a benchmark to evaluate the performance of their own method. This is a recognition as well as challenge at the same time.

Summary

Here, we demonstrated MEM tool to perform global co-expression queries across hundreds of public datasets in single analysis. MEM allows to perform convenient and interactive gene co-expression queries. Co-expression results from individual datasets are combined into single global prioritised gene list by novel rank aggregation algorithm. We have published the rank aggregation method separately as R package (Kolde et al., 2012) and discuss if further in the next section of this thesis.

3.4. Robust rank aggregation for gene list integration and meta-analysis (*Ref IV*)

In previous publication (Adler et al., 2009a, *Ref III*) we demonstrated web tool which principal contribution was to make possible gene co-expression queries across many hundreds of public gene expression datasets. Important component of this tool was the rank aggregation algorithm used to combine results from individual datasets. In this paper we explain in greater detail how rank aggregation works and how it can improve meta-analysis across multiple data sources. Robust Rank Aggregation (RRA) method is packaged into R library, and thus making it available and usable in other analysis pipelines beside MEM.

Common conception is that converting similarity or distance measures into ranks causes loss of information. Indeed, measure density and distribution parameters are lost when converting values into ranks. However, comparing and aggregating values with different distribution background may yield even more problematic results (Obayashi and Kinoshita, 2009). We show here that the loss of information by using rank aggregation methods is insignificant compared to the added value one would get by merging information from multiple sources.

In the publication (Kolde et al., 2012, *Ref IV*) we set out to demonstrate three principal features of RRA method: it is robust to noise, can handle incomplete rankings and assigns a significance score to each element in the resulting ranking. In addition, it is efficient to compute, which is not a trivial issue while performing large scale analysis.

Study with simulated data

We performed several experiments to stress these features. Simulated data allowed to test different scenarios under controlled conditions. Robustness to noise, incomplete rankings, as well the significance score were tested comprehensively this way. We performed comparisons against two other methods. The first method calculates average rank value across all rankings for each entity. The other method we called Stuart by the lead author when it was first published by Stuart et al. (2003). In the analysis we used optimised version of the algorithm published by Aerts et al. (2006).

Simulated data consisted of ten lists with implanted positive elements. For noise tolerance analysis different number of randomly shuffled lists were added. All lists consisted of 1000 elements.

Both order statistics based methods (RRA and Stuart) separated the true positive and the noise better than average based scoring. Both methods also produce score that can be interpreted as significance p-value for elements. We used FDR to correct for multiple testing. Compared to RRA, that did not retrieve any false positives, Stuart method is statistically less stringent and deemed also large number of true negatives significant.

In our opinion we consider the robustness to noise one of the key features. In real life experiments we usually do not know the portion of the noise, nor can control that. With 25% of the input lists randomly shuffled, RRA still recovered more than 50% of the true positives. RRA consistently outperformed average method in this case, Stuart method could not be compared as it is unable to reliably separate implanted elements from the noise.

In case of incomplete data, RRA started to show diminishing results only after 95% of low ranking genes in each input were removed. In real life data, it should never get that extreme, but this feature is welcome while aggregating results from different biological platforms.

Study with Biological data

Experiments with biological data allowed to study how RRA performs in real life conditions. We performed two experiments with real life high-throughput data.

Hu et al. (2007) have done extensive yeast transcription factor (TF) knockout experiment. It was reanalysed by Reimand et al. (2010) to identify target genes most affected by the knockout of a TF. This kind of experimental setup sheds light to underlying biological networks allowing to identify putative targets for a TF. We performed proof of principle experiment where we tried to retrieve Gene Ontology term members using individual ranked gene lists associated with TFs belonging to the term. For comparison we also calculated Fisher exact p-value between individual TF list and GO term. In most cases aggregated results outperform individual lists. In case of *response to chemical stimulus* only 6 individual TF based gene lists out of 39 showed significant association with the term. Aggregated results however overcame the noise and showed yet stronger enrichment to the term than any of the individual lists.

In the second experiment we used high-throughput gene expression data available via MEM tool (Adler et al., 2009a, *ref III*) and ChIP-seq study performed by Chen et al. (2008). The study covered 15 embryonic stem (ES) cell related TFs. In this experiment they identified genes that were regulated by these TFs using ChIP-seq promoter analysis. We were interested whether we can use gene co-expression analysis on high-throughput gene expression data to infer the same connections between regulators and their targeted genes. As the original analysis was performed on mouse ES cells, we used mouse Affymetrix GeneChip Mouse Genome 430 2.0 gene expression platform. From the available data we selected only datasets that mentioned expression profiling from ES cell types. There were 12 such datasets available at the time. For each of the TFs we calculated area under the curve for predicting gene associations with the regulator in each dataset separately. We used RRA to predict gene associations with the TF across all 12 datasets and compared to the results from individual datasets.

The results reveal, that while in some datasets few of the queries on individual datasets outperform aggregated results, the aggregated results show consistently good approximation to the best performing datasets. Mostly it is difficult to guess *a priory* which of the available dataset would be relevant. Questionable might be the quality of the data or its connection to the biological question at hand.

Summary

Here we show the benefit of a rank aggregation method that can take many ordered inputs lists and highlight only the most relevant bit of information contained in the combined data. The studies with biological data presented here also show how useful public gene expression data re-usage can be to predict connections or interactions between genes. We packaged the method as *RobustRankAggreg* R library to make possible its integration into custom analysis pipelines by external researchers.

CONCLUSIONS

In this thesis I have introduced methods to analyse and visualise high-throughput datasets as well methods of meta-analysis. The results of the initial analysis are better interpreted if viewed in the context of prior biological knowledge. Biological pathway databases and Gene Ontology are typically used to provide such context. Here, we have developed KEGGanim web tool that allows to visualise and animate high-throughput expression data on top of KEGG pathway images. The tool is publicly accessible to anyone who would like to use it. User own expression data can be uploaded and visualised. Animations can be used in slide presentations as illustrations or still images can be generated for publications, using "Cinefilm" function.

There is a wealth of publicly available gene expression data. We used a collection of human gene expression data across 6108 samples to measure the predictive power of co-expression analysis to reconstruct existing Reactome pathways. This study had two principal conclusions. First, we observed that co-expression analysis does not yield good results on pathways that do not need its members to be present in consistent manner. Such are, for example, signalling pathways, where the signal is transmitted by protein modifications and not via transcriptional control. On the other hand, pathways that require its components to be present in close proximity and employ transcriptional regulation to achieve that, showed more favourable results. Secondly, we observed that using only subset of genes from the pathway will improve the prediction of closely related pathway components. Different pathway components may be active in different cellular states and therefore not related in terms of co-expression.

The most important outcome of this thesis is Multi Experiment Matrix (MEM) tool. We have collected and processed large collection of publicly available high-throughput gene expression datasets. MEM allow to perform query based co-expression analyses across hundreds of datasets together. This makes it possible to re-use already existing expression data and allows to discover signals that would otherwise be difficult to find from a single dataset. The main output of the tool is heat map type rank matrix, from where it is possible to observe the emerged gene co-expression patterns between genes and datasets. MEM has interactive web interface to define queries, visualise results and provide crosslinks to further characterise found gene lists, learn more about individual genes and datasets. We have developed a novel rank aggregation method to compile the final prioritised gene list.

Finally, to better demonstrate the proposed rank aggregation method we performed several experiments with simulated data and compared our method against two other rank aggregation methods. We showed that our proposed method is robust to noise, can handle missing data and separates true signal from background noise better than the compared alternatives. We have packaged the proposed method as *RobustRankAggreg* (RRA) R library. This enables to incorporate RRA method into custom meta-analysis pipelines by external researchers.

BIBLIOGRAPHY

Adler, P., Kolde, R., Kull, M., Tkachenko, A., Peterson, H., Reimand, J., and Vilo, J. (2009a). Mining for coexpression across hundreds of datasets using novel rank aggregation and visualization methods. *Genome Biol.*, 10(12):R139.

Adler, P., Peterson, H., Agius, P., Reimand, J., and Vilo, J. (2009b). Ranking genes by their co-expression to subsets of pathway members. *Ann. N. Y. Acad. Sci.*, 1158:1–13.

Aerts, S., Lambrechts, D., Maity, S., Van Loo, P., Coessens, B., De Smet, F., Tranchevent, L. C., De Moor, B., Marynen, P., Hassan, B., Carmeliet, P., and Moreau, Y. (2006). Gene prioritization through genomic data fusion. *Nat. Biotechnol.*, 24(5):537–544.

Altmäe, S., Reimand, J., Hovatta, O., Zhang, P., Kere, J., Laisk, T., Saare, M., Peters, M., Vilo, J., Stavreus-Evers, A., and Salumets, A. (2012). Research resource: interactome of human embryo implantation: identification of gene expression pathways, regulation, and integrated regulatory networks. *Mol. Endocrinol.*, 26(1):203–217.

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. *Nat. Genet.*, 25(1):25–29.

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., Wilson, C. J., Lehar, J., Kryukov, G. V., Sonkin, D., Reddy, A., Liu, M., Murray, L., Berger, M. F., Monahan, J. E., Morais, P., Meltzer, J., Korejwa, A., Jane-Valbuena, J., Mapa, F. A., Thibault, J., Bric-Furlong, E., Raman, P., Shipway, A., Engels, I. H., Cheng, J., Yu, G. K., Yu, J., Aspesi, P., de Silva, M., Jagtap, K., Jones, M. D., Wang, L., Hatton, C., Palescandolo, E., Gupta, S., Mahan, S., Sougnez, C., Onofrio, R. C., Liefeld, T., MacConaill, L., Winckler, W., Reich, M., Li, N., Mesirov, J. P., Gabriel, S. B., Getz, G., Ardlie, K., Chan, V., Myer, V. E., Weber, B. L., Porter, J., Warmuth, M., Finan, P., Harris, J. L., Meyerson, M., Golub, T. R., Morrissey, M. P., Sellers, W. R., Schlegel, R., and Garraway, L. A. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603–607.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300.

Bolstad, B. M., Irizarry, R. A., Astrand, 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.

Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C. A., Causton, H. C., Gaasterland, T., Glenisson, P., Holstege, F. C., Kim, I. F., Markowitz, V., Matese, J. C., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J., and Vingron, M. (2001). Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet.*, 29(4):365–371.

Brazma, A., Parkinson, H., Sarkans, U., Shojatalab, M., Vilo, J., Abeygunawardena, N., Holloway, E., Kapushesky, M., Kemmeren, P., Lara, G. G., Oezcimen, A., Rocca-Serra, P., and Sansone, S. A. (2003). ArrayExpress–a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.*, 31(1):68–71.

Butcher, R. A., Bhullar, B. S., Perlstein, E. O., Marsischky, G., LaBaer, J., and Schreiber, S. L. (2006). Microarray-based method for monitoring yeast overexpression strains reveals small-molecule targets in TOR pathway. *Nat. Chem. Biol.*, 2(2):103–109.

Chen, M., Zang, M., Wang, X., and Xiao, G. (2013). A powerful Bayesian meta-analysis method to integrate multiple gene set enrichment studies. *Bioinformatics*, 29(7):862–869.

Chen, T.-S., Tsai, T.-H., Chen, Y.-T., Lin, C.-C., Chen, R.-C., Li, S.-Y., and Chen, H.-Y. (2005). A combined k-means and hierarchical clustering method for improving the clustering efficiency of microarray. In *Intelligent Signal Processing and Communication Systems, 2005. ISPACS 2005. Proceedings of 2005 International Symposium on*, pages 405–408. IEEE.

Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V. B., Wong, E., Orlov, Y. L., Zhang, W., Jiang, J., Loh, Y. H., Yeo, H. C., Yeo, Z. X., Narang, V., Govindarajan, K. R., Leong, B., Shahab, A., Ruan, Y., Bourque, G., Sung, W. K., Clarke, N. D., Wei, C. L., and Ng, H. H. (2008). Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, 133(6):1106–1117.

Collins, F. S. and Barker, A. D. (2007). Mapping the cancer genome. Pinpointing the genes involved in cancer will help chart a new course across the complex landscape of human malignancies. *Sci. Am.*, 296(3):50–57.

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 Res.*, 42(Database issue):D472–477.

Cunningham, F., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., Gil, L., Giron,

C. G., Gordon, L., Hourlier, T., Hunt, S. E., Janacek, S. H., Johnson, N., Juettemann, T., Kahari, A. K., Keenan, S., Martin, F. J., Maurel, T., McLaren, W., Murphy, D. N., Nag, R., Overduin, B., Parker, A., Patricio, M., Perry, E., Pignatelli, M., Riat, H. S., Sheppard, D., Taylor, K., Thormann, A., Vullo, A., Wilder, S. P., Zadissa, A., Aken, B. L., Birney, E., Harrow, J., Kinsella, R., Muffato, M., Ruffier, M., Searle, S. M., Spudich, G., Trevanion, S. J., Yates, A., Zerbino, D. R., and Flicek, P. (2015). Ensembl 2015. *Nucleic Acids Res.*, 43(Database issue):D662–669.

Dai, M., Wang, P., Boyd, A. D., Kostov, G., Athey, B., Jones, E. G., Bunney, W. E., Myers, R. M., Speed, T. P., Akil, H., Watson, S. J., and Meng, F. (2005). Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res.*, 33(20):e175.

DeRisi, J. L., Iyer, V. R., and Brown, P. O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 278(5338):680–686.

Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, 30(1):207–210.

Gasch, A. P. and Eisen, M. B. (2002). Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. *Genome Biol.*, 3(11):research0059.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y., and Zhang, J. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, 5(10):R80.

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 Res.*, 34(Database issue):D140–144.

Guo, Y., Sheng, Q., Li, J., Ye, F., Samuels, D. C., and Shyr, Y. (2013). Large scale comparison of gene expression levels by microarrays and RNAseq using TCGA data. *PLoS ONE*, 8(8):e71462.

Hansen, J. (2005). Graphics language swog. *Bachelor thesis, University of Tartu.*

Hibbs, M. A., Hess, D. C., Myers, C. L., Huttenhower, C., Li, K., and Troyanskaya, O. G. (2007). Exploring the functional landscape of gene expression: directed search of large microarray compendia. *Bioinformatics*, 23(20):2692–2699.

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

Hu, Z., Killion, P. J., and Iyer, V. R. (2007). Genetic reconstruction of a functional transcriptional regulatory network. *Nat. Genet.*, 39(5):683–687.

Hubbell, E., Liu, W. M., and Mei, R. (2002). Robust estimators for expression analysis. *Bioinformatics*, 18(12):1585–1592.

Huttenhower, C., Hibbs, M., Myers, C., and Troyanskaya, O. G. (2006). A scalable method for integration and functional analysis of multiple microarray datasets. *Bioinformatics*, 22(23):2890–2897.

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.

Irizarry, R. A., Warren, D., Spencer, F., Kim, I. F., Biswal, S., Frank, B. C., Gabrielson, E., Garcia, J. G., Geoghegan, J., Germino, G., Griffin, C., Hilmer, S. C., Hoffman, E., Jedlicka, A. E., Kawasaki, E., Martinez-Murillo, F., Morsberger, L., Lee, H., Petersen, D., Quackenbush, J., Scott, A., Wilson, M., Yang, Y., Ye, S. Q., and Yu, W. (2005). Multiple-laboratory comparison of microarray platforms. *Nat. Methods*, 2(5):345–350.

Ivanov, S. V., Panaccione, A., Nonaka, D., Prasad, M. L., Boyd, K. L., Brown, B., Guo, Y., Sewell, A., and Yarbrough, W. G. (2013). Diagnostic SOX10 gene signatures in salivary adenoid cystic and breast basal-like carcinomas. *Br. J. Cancer*, 109(2):444–451.

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.

Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., de Bono, B., Jassal, B., Gopinath, G. R., Wu, G. R., Matthews, L., Lewis, S., Birney, E., and Stein, L. (2005). Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res.*, 33(Database issue):D428–432.

Kanehisa, M. and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, 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 Res.*, 42(Database issue):199–205.

Kauffmann, A., Gentleman, R., and Huber, W. (2009). arrayQualityMetrics–a bioconductor package for quality assessment of microarray data. *Bioinformatics*, 25(3):415–416.

Kellis, M., Wold, B., Snyder, M. P., Bernstein, B. E., Kundaje, A., Marinov, G. K., Ward, L. D., Birney, E., Crawford, G. E., Dekker, J., Dunham, I., Elnitski, L. L., Farnham, P. J., Feingold, E. A., Gerstein, M., Giddings, M. C., Gilbert, D. M., Gingeras, T. R., Green, E. D., Guigo, R., Hubbard, T., Kent, J., Lieb, J. D., Myers, R. M., Pazin, M. J., Ren, B., Stamatoyannopoulos, J. A., Weng, Z., White, K. P., and Hardison, R. C. (2014). Defining functional DNA elements in the human genome. *Proc. Natl. Acad. Sci. U.S.A.*, 111(17):6131–6138.

Kinsella, R. J., Kahari, A., Haider, S., Zamora, J., Proctor, G., Spudich, G., Almeida-King, J., Staines, D., Derwent, P., Kerhornou, A., Kersey, P., and Flicek, P. (2011). Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database (Oxford)*, 2011:bar030.

Klebanov, L. and Yakovlev, A. (2007). How high is the level of technical noise in microarray data? *Biol. Direct*, 2:9.

Kohler, 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 Res.*, 42(Database issue):D966–974.

Kolde, R., Laur, S., Adler, P., and Vilo, J. (2012). Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics*, 28(4):573–580.

Krushevskaya, D., Peterson, H., Reimand, J., Kull, M., and Vilo, J. (2009). VisHiC–hierarchical functional enrichment analysis of microarray data. *Nucleic Acids Res.*, 37(Web Server issue):W587–592.

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

Lamb, J., Crawford, E. D., Peck, D., Modell, J. W., Blat, I. C., Wrobel, M. J., Lerner, J., Brunet, J. P., Subramanian, A., Ross, K. N., Reich, M., Hieronymus, H., Wei, G., Armstrong, S. A., Haggarty, S. J., Clemons, P. A., Wei, R., Carr, S. A., Lander, E. S., and Golub, T. R. (2006). The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, 313(5795):1929–1935.

Lashkari, D. A., DeRisi, J. L., McCusker, J. H., Namath, A. F., Gentile, C., Hwang, S. Y., Brown, P. O., and Davis, R. W. (1997). Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proc. Natl. Acad. Sci. U.S.A.*, 94(24):13057–13062.

Lee, H. K., Hsu, A. K., Sajdak, J., Qin, J., and Pavlidis, P. (2004). Coexpression analysis of human genes across many microarray data sets. *Genome Res.*, 14(6):1085–1094.

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. *Nat. Biotechnol.*, 14(13):1675–1680.

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

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 Res.*, 18(9):1509–1517.

Mashanov, V. S., Zueva, O. R., and Garcia-Arraras, J. E. (2014). Transcriptomic changes during regeneration of the central nervous system in an echinoderm. *BMC Genomics*, 15:357.

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 Res.*, 34(Database issue):D108–110.

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 (Basel)*, 4(4):1180–1211.

Millenaar, F. F., Okyere, J., May, S. T., van Zanten, M., Voesenek, L. A., and Peeters, A. J. (2006). How to decide? Different methods of calculating gene expression from short oligonucleotide array data will give different results. *BMC Bioinformatics*, 7:137.

Modrek, B. and Lee, C. (2002). A genomic view of alternative splicing. *Nat. Genet.*, 30(1):13–19.

Obayashi, T. and Kinoshita, K. (2009). Rank of correlation coefficient as a comparable measure for biological significance of gene coexpression. *DNA Res.*, 16(5):249–260.

Parman, C., Halling, C., and Gentleman, R. (2005). affyQCReport: QC report generation for affyBatch objects. *R package version*, 1(0):1.

Pihur, V., Datta, S., and Datta, S. (2009). RankAggreg, an R package for weighted rank aggregation. *BMC Bioinformatics*, 10:62.

Priness, I., Maimon, O., and Ben-Gal, I. (2007). Evaluation of gene-expression clustering via mutual information distance measure. *BMC Bioinformatics*, 8:111.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

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

Reimand, J., Vaquerizas, J. M., Todd, A. E., Vilo, J., and Luscombe, N. M. (2010). Comprehensive reanalysis of transcription factor knockout expression data in Saccharomyces cerevisiae reveals many new targets. *Nucleic Acids Res.*, 38(14):4768–4777.

Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., Lander, E. S., Mitzenmacher, M., and Sabeti, P. C. (2011). Detecting novel associations in large data sets. *Science*, 334(6062):1518–1524.

Ruepp, A., Brauner, B., Dunger-Kaltenbach, I., Frishman, G., Montrone, C., Stransky, M., Waegele, B., Schmidt, T., Doudieu, O. N., Stumpflen, V., and Mewes, H. W. (2008). CORUM: the comprehensive resource of mammalian protein complexes. *Nucleic Acids Res.*, 36(Database issue):D646–650.

Rung, J. and Brazma, A. (2013). Reuse of public genome-wide gene expression data. *Nat. Rev. Genet.*, 14(2):89–99.

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 Res.*, 41(Database issue):D987–990.

Sandberg, R. and Larsson, O. (2007). Improved precision and accuracy for microarrays using updated probe set definitions. *BMC Bioinformatics*, 8:48.

Schulz, H., Kolde, R., Adler, P., Aksoy, I., Anastassiadis, K., Bader, M., Billon, N., Boeuf, H., Bourillot, P. Y., Buchholz, F., Dani, C., Doss, M. X., Forrester, L., Gitton, M., Henrique, D., Hescheler, J., Himmelbauer, H., Hubner, N., Karantzali, E., Kretsovali, A., Lubitz, S., Pradier, L., Rai, M., Reimand, J., Rolletschek, A., Sachinidis, A., Savatier, P., Stewart, F., Storm, M. P., Trouillas, M., Vilo, J., Welham, M. J., Winkler, J., Wobus, A. M., and Hatzopoulos, A. K. (2009). The FunGenES database: a genomics resource for mouse embryonic stem cell differentiation. *PLoS ONE*, 4(9):e6804.

Shi, L., Reid, L. H., Jones, W. D., Shippy, R., Warrington, J. A., Baker, S. C., Collins, P. J., de Longueville, F., Kawasaki, E. S., Lee, K. Y., Luo, Y., Sun, Y. A., Willey, J. C., Setterquist, R. A., Fischer, G. M., Tong, W., Dragan, Y. P., Dix, D. J., Frueh, F. W., Goodsaid, F. M., Herman, D., Jensen, R. V., Johnson, C. D., Lobenhofer, E. K., Puri, R. K., Schrf, U., Thierry-Mieg, J., Wang, C., Wilson, M., Wolber, P. K., Zhang, L., Amur, S., Bao, W., Barbacioru, C. C., Lucas, A. B., Bertholet, V., Boysen, C., Bromley, B., Brown, D., Brunner, A., Canales, R., Cao, X. M., Cebula, T. A., Chen, J. J., Cheng, J., Chu, T. M., Chudin, E., Corson, J., Corton, J. C., Croner, L. J., Davies, C., Davison, T. S., Delenstarr, G., Deng, X., Dorris, D., Eklund, A. C., Fan, X. H., Fang, H., Fulmer-Smentek, S., Fuscoe, J. C., Gallagher, K., Ge, W., Guo, L., Guo, X., Hager, J., Haje, P. K., Han, J., Han, T., Harbottle, H. C., Harris, S. C., Hatchwell, E., Hauser, C. A., Hester, S., Hong, H., Hurban, P., Jackson, S. A., Ji, H., Knight, C. R., Kuo, W. P., LeClerc, J. E., Levy, S., Li, Q. Z., Liu, C., Liu, Y., Lombardi, M. J., Ma, Y., Magnuson, S. R., Maqsodi, B., McDaniel, T., Mei, N., Myklebost, O., Ning, B., Novoradovskaya, N., Orr, M. S., Osborn, T. W., Papallo, A., Patterson, T. A., Perkins, R. G., Peters, E. H., Peterson, R., Philips, K. L., Pine, P. S., Pusztai, L., Qian, F., Ren, H., Rosen, M., Rosenzweig,

B. A., Samaha, R. R., Schena, M., Schroth, G. P., Shchegrova, S., Smith, D. D., Staedtler, F., Su, Z., Sun, H., Szallasi, Z., Tezak, Z., Thierry-Mieg, D., Thompson, K. L., Tikhonova, I., Turpaz, Y., Vallanat, B., Van, C., Walker, S. J., Wang, S. J., Wang, Y., Wolfinger, R., Wong, A., Wu, J., Xiao, C., Xie, Q., Xu, J., Yang, W., Zhang, L., Zhong, S., Zong, Y., and Slikker, W. (2006). The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat. Biotechnol.*, 24(9):1151–1161.

Sircoulomb, F., Nicolas, N., Ferrari, A., Finetti, P., Bekhouche, I., Rousselet, E., Lonigro, A., Adelaide, J., Baudelet, E., Esteyries, S., Wicinski, J., Audebert, S., Charafe-Jauffret, E., Jacquemier, J., Lopez, M., Borg, J. P., Sotiriou, C., Popovici, C., Bertucci, F., Birnbaum, D., Chaffanet, M., and Ginestier, C. (2011). ZNF703 gene amplification at 8p12 specifies luminal B breast cancer. *EMBO Mol Med*, 3(3):153–166.

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*, 3:Article3.

Smyth, G. K. (2005). Limma: linear models for microarray data. In *Bioinformatics and computational biology solutions using R and Bioconductor*, pages 397–420. Springer.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., and Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell*, 9(12):3273–3297.

Stark, C., Breitkreutz, B. J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. (2006). BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.*, 34(Database issue):D535–539.

Stuart, J. M., Segal, E., Koller, D., and Kim, S. K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 302(5643):249–255.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.*, 102(43):15545–15550.

Thellin, O., Zorzi, W., Lakaye, B., De Borman, B., Coumans, B., Hennen, G., Grisar, T., Igout, A., and Heinen, E. (1999). Housekeeping genes as internal standards: use and limits. *J. Biotechnol.*, 75(2-3):291–295.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, 10(1):57–63.

Wolfe, C. J., Kohane, I. S., and Butte, A. J. (2005). Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks. *BMC Bioinformatics*, 6:227.

Yona, G., Dirks, W., Rahman, S., and Lin, D. M. (2006). Effective similarity measures for expression profiles. *Bioinformatics*, 22(13):1616–1622.

Zhu, Q., Wong, A. K., Krishnan, A., Aure, M. R., Tadych, A., Zhang, R., Corney, D. C., Greene, C. S., Bongo, L. A., Kristensen, V. N., Charikar, M., Li, K., and Troyanskaya, O. G. (2015). Targeted exploration and analysis of large cross-platform human transcriptomic compendia. *Nat. Methods*, 12(3):211–214.

Zoubarev, A., Hamer, K. M., Keshav, K. D., McCarthy, E. L., Santos, J. R., Van Rossum, T., McDonald, C., Hall, A., Wan, X., Lim, R., Gillis, J., and Pavlidis, P. (2012). Gemma: a resource for the reuse, sharing and meta-analysis of expression profiling data. *Bioinformatics*, 28(17):2272–2273.

SUMMARY IN ESTONIAN

Paljude mikrokiibi andmestike suuremahuline analüüsimine ja visualiseerimine

Suuremahulised geeniekspressiooni mikrokiibid on olnud viimased kaks aastakümmet peamiseks vahendiks, et uurida ja analüüsida erinevaid bioloogilisi tingimusi. Geeniekspressiooni mikrokiibid võimaldavad korraga kvantifitseerida väga suure hulga geenide transkriptsioonilist aktiivsust. Aastate jooksul tehtud eksperimentide andmed on kogutud suurtesse andmebaasidesse nagu näiteks GEO või ArrayExpress. Sealt on võimalik andmeid kätte saada ja taaskasutada sarnaste bioloogiliste tingimuste uurimisel või suuremates meta-analüüsides üle paljude andmestike. Siin esitatud töös rõhutame just andmete taaskasutamist ja meta-analüüsi olulisust tundlikumate ja täpsemate tulemuse saamiseks. Tulemuste tõlgendamisel on tihti oluline roll ka asjakohase visualiseerimise meetodi valikul.

Esitleme KEGGanim tööriista, mis võimaldab siduda suuremahulisi ekspressiooniandmeid ja KEGG bioloogiliste radade andmebaasi. KEGGanim loob ekspressiooniandmetest rajapõhise animatsiooni, kus on võimalik jälgida raja komponentide ekpressiooni dünaamikat üle erinevate katsete. Selline konkreetse bioloogilise raja konteksti fokusseeritud vaade võimaldab paremini eksperimendi tulemusi interpreteerida ja visualiseerida. KEGGanimi animatsioonid sobivad kasutamiseks ettekannete slaididel. “Cinefilm” funktsioon võimaldab genereerida pilte üle mitme animatsiooni kaadri. Sellised pildid sobivad kasutamiseks publikatsioonides.

Radade andmebaasid ei kirjelda kõiki geenide vahelisi seoseid. Vaid üks kolmandik kõikidest geenidest on annoteeritud KEGGi või Reactome’i radade andmebaasidesse. Meid huvitas, kui edukalt on võimalik ära kasutada suuremahulisi geeniekspressiooni andmeid olemasolevate bioloogiliste radade ja nende liikmete omavahelise seose kirjeldamisel. Selleks viisime läbi geenide koos-ekspressiooni analüüsi Reactome’i andmebaasi radadel kasutades ristvalideerimise meetodit. Tehtud tööl oli kaks peamist järeldust. Esiteks ei sobi geenide koos-ekspressiooni analüüs signaaliradade kirjeldamiseks. Signaaliradade komponendid enamasti ei sõltu koordineeritud transkriptsiooni regulatsioonist. Rajad, mille komponendid on vajalikud näiteks kindlal rakutsükli etapil või moodustavad suuremaid valgukogumikke, on palju paremini kirjeldatavad kasutades geeniekspressiooni sarnasuse analüüsi. Teiseks, leidsime, et kasutades vaid alamosa raja geenidest, on võimalik oluliselt parandada analüüsi tulemusi. Erinevad raja osad võivad olla funktsionaalselt küllalti erinevad, mistõttu peaks neid ka analüüsima eraldi.

Esitleme Multi Experiment Matrix (MEM) tööriista, mis võimaldab teha geenide koos-ekpressiooni päringuid üle paljude andmestike korraga. Paljud geenid on ekspresseerunud vaid kindlates koetüüpides või kindlatel tingimustel. Ka koos-ekspressioon võib ilmneda erinevate geenide vahel erinevates tingimustes. MEM-tööriist võimaldab hõlpsalt selliseid seoseid tuvastada ja interaktiivne veebilahendus pakub võimalusi leidude täpsemaks edasiseks analüüsiks viidates teistele, seotud tööriistadele.

Me oleme laadinud ArrayExpressi andmebaasist alla kõik avalikud Affymetrix platvormil olevad geeniekpressiooni andmestike toorandmed. Selliseid andmestikke on (seisuga 01.12.14) rohkem kui 13000, sisaldades enam kui 330000 individuaalset mikrokiibi katset. Uute andmestike arv andmebaasis kasvab pidevalt ja uued andmestikud lisatakse perioodiliselt ka MEM-tööriista.

Igas etteantud andmestikus järjestatakse geenid vastavalt nende sarnasusele päringugeeniga. Kõikidele geenidele omistatakse neile vastav astak-väärtus. Individuaalsetest andmestikest pärit astak-väärtused seotakse ühiseks globaalseks prioritiseeritud geenide järjekorraks kasutades statistilist astakute agregeerimise meetodit.

Viimaks tutvustame arendatud astakute agregeerimese meetodit – Robust Rank Aggregation (RRA). Me võrdlesime RRA meetodit teiste astakute agregreerimise meetoditega, kasutades selleks simuleeritud andmeid. Me näitasime, et RRA on võimeline eraldama sisendandmetest olulise informatsiooni isegi kõrge müra fooni või osaliselt puuduvate andmete puhul. Samuti demonstreerime, kuidas on RRA meetodit võimalik kasutada bioloogilistes meta-analüüsi eksperimentides väljaspool MEM paradigmat. RRA on pakendatud R paketina (*RobustRankAggreg*), et seda oleks lihtsam integreerida teistesse analüüsi töövoogudesse.

ACKNOWLEDGEMENTS

I'm grateful to my supervisor Prof. Jaak Vilo, who has given me opportunity to study and work in such a wonderful environment. I'm proud to belong to BIIT group, it will always be my *home*. I also give my thanks in everybody who is and has been in the group over the years.

I also thank Prof. Juhan Sedman for being my co-supervisor.

In IMCB, in room 304, I grew up (or maybe not just yet). I had the best of friends along to the ride. Thank you Hedi, Priit, Jüri, Sten, Triinu and Eero.

The last year and a half I have enjoyed the best work environment possible. The people here make it all the better. Thank you Hedi, Elena, Dima, Mari-Liis, Anti, Liis, Tõnis, Gea, Martin and all the people in STACC who are stuck behind the glass door with post-its.

There is a shortlist of people who have contributed to the publications within this thesis. Thank you Hedi, Raivo, Jüri, Meelis, Phaedra, Sven, Jürgen, Aleksandr and of course Jaak.

Connected to this list I also thank Tambet, who is wizard if it comes to programming.

A special thanks goes to those who pushed me to write this terribly difficult thing called thesis and helped along the way. Thank you Hedi and Reet. There is also a number of people who read it and provided much appreciated feedback and guidance. Thank you Hedi, Reet, Jaak, Liis, Elena, Dima and Heino.

I thank my parents, by grandparents and my brothers for being there for me, for raising me or growing with me. I also thank my extended family who have provided a lot of support and also been there for me and our children.

Most of all I thank my own family. My very special Reet, Iona and Stig. You are the centre of my life!

PUBLICATIONS

CURRICULUM VITAE

Name:	Priit Adler
Date of birth:	06.09.1982
Citizenship:	Estonian
Address:	Institute of Computer Science Liivi 2, Tartu 50409
E-mail:	adler@ut.ee

Education:

2007–	University of Tartu, bioinformatics, PhD student
2005–2007	University of Tartu, Gene technology/bioinformatics, M.Sc
2001–2005	University of Tartu, Gene technology/bioinformatics, B.Sc

Professional career:

2014–	Quretec, programmer
2010–	University of Tartu, programmer
2006–2007	AS FibroTx, programmer/bioinformatician
2004–2005	Estonian Biocentre, programmer

Main Fields of Research:

Large-scale gene expression analysis – generic gene expression description between tissues and conditions. Co- and differential expression analysis.
Gene expression in biological pathway context – gene behaviour in biological pathways, dynamics between genes and conditions. Augmentation of biological pathways by finding new putative candidate genes.

Publications:

1. Örd, T., Örd, D., **Adler, P.**, Vilo, J., Örd, T. (2015) TRIB3 enhances cell viability during glucose deprivation in HEK293-derived cells by upregulating IGFBP2, a novel nutrient deficiency survival factor. *BBA - Molecular Cell Research*, Accepted, in press
2. Lees, J. G.*, Heriche, J. K.*, Morilla, I., Fernandez, J. M., **Adler, P.**, Krallinger, M., Vilo, J., Valencia, A., Ellenberg, J., Ranea, J. A., and Orengo, C. (2015). Fun-L: Gene prioritization for RNAi screens. *Bioinformatics*, 31(12):2052–2053.

3. Heriche, J. K., Lees, J. G., Morilla, I., Walter, T., Petrova, B., Roberti, M. J., Hossain, M. J., **Adler, P.**, Fernandez, J. M., Krallinger, M., Haering, C. H., Vilo, J., Valencia, A., Ranea, J. A., Orengo, C., and Ellenberg, J. (2014). Integration of biological data by kernels on graph nodes allows prediction of new genes involved in mitotic chromosome condensation. *Molecular Biology of the Cell*, 25(16):2522–2536.
4. Kolde, R., Laur, S., **Adler, P.**, and Vilo, J. (2012). Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics*, 28(4):573–580.
5. Billon, N., Kolde, R., Reimand, J., Monteiro, M. C., Kull, M., Peterson, H., Tretyakov, K., **Adler, P.**, Wdziekonski, B., Vilo, J., and Dani, C. (2010). Comprehensive transcriptome analysis of mouse embryonic stem cell adipogenesis unravels new processes of adipocyte development. *Genome Biology*, 11(8):R80.
6. **Adler, P.***, Kolde, R.*, Kull, M., Tkachenko, A., Peterson, H., Reimand, J., and Vilo, J. (2009). Mining for coexpression across hundreds of datasets using novel rank aggregation and visualization methods. *Genome Biology*, 10(12):R139.
7. Schulz, H.*, Kolde, R.*, **Adler, P.**, Aksoy, I., Anastassiadis, K., Bader, M., Billon, N., Boeuf, H., Bourillot, P.-Y., Buchholz, F., Dani, C., Doss, M. X., Forrester, L., Gitton, M., Henrique, D., Hescheler, J., Himmelbauer, H., Hübner, N., Karantzali, E., Kretsovali, A., Lubitz, S., Pradier, L., Rai, M., Reimand, J., Rolletschek, A., Sachinidis, A., Savatier, P., Stewart, F., Storm, M. P., Trouillas, M., Vilo, J., Welham, M. J., Winkler, J., Wobus, A. M., and and, A. K. H. (2009). The FunGenES database: a genomics resource for mouse embryonic stem cell differentiation. *PLoS ONE*, 4(9):e6804.
8. **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*, 1158:1–13.
9. **Adler, P.***, Reimand, J.*, Jänes, J., Kolde, R., Peterson, H., and Vilo, J. (2008). KEGGanim: pathway animations for high-throughput data. *Bioinformatics*, 24(4):588–90.
10. 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(suppl 2):W452–W459.

Scholarships:

1. ISMB/ECCB Travel Fellowship, Multi Experiment Matrix – web tool for mining co-expressed genes over hundreds of datasets, 2009
2. Artur Lind fellowship, 2008

Teaching:

1. ELIXIR RNAseq analysis course, Tartu, 12.06.2015
2. ELIXIR Linux command line course, Tartu, 15.05.2015
3. ELIXIR MEM hands-on course, Tartu, 09.12.2014
4. Lectures in Bioinformatics course, Tartu, autumn 2010
5. EBI roadshow 2009, Tartu 07.-10.09.2009

Supervised theses:

1. Tõnis Tasa, master's thesis "Re-using public RNA-Seq data" (2015)
2. Aleksei Panarin, bachelor thesis "Pathway-specific Gene Expression Data Analysis" (2009)

Participated courses and workshops:

1. "e-Infrastructure for Massively Parallel Sequencing" UPPNEX Workshop 2015 (Uppsala)
2. "Best practices for training in Next Generation Sequencing (NGS) Analysis" TGAC/Elixir workshop 2015 (Cambridge)
3. "Bioinformatics for Systems and Synthetic Biology" Bologna Winter School 2007
4. "Advanced Analysis and Informatics of Microarray Data" EMBO Practical Course 2007 (Hinxton)

Membership to academic societies:

1. International Society for Computational Biology (ISCB), since 2007
2. ISCB Student Council, since 2007

ELULOOKIRJELDUS

Nimi:	Priit Adler
Sünniaeg:	06.09.1982
Kodakondsus:	Eesti
Aadress:	Arvutiteaduse instituut Liivi 2, Tartu 50409
E-mail:	adler@ut.ee

Haridus:

2007–	Tartu Ülikool, bioinformaatika, doktorantuur
2005–2007	Tartu Ülikool, Geeni tehnoloogia/bioinformaatika, M.Sc
2001–2005	Tartu Ülikool, Geeni tehnoloogia/bioinformaatika, B.Sc

Teenistuskäik:

2014–	Quretec, programmeerija
2010–	Tartu Ülikool, programmeerija
2006–2007	AS FibroTx, programmeerija/bioinformaatik
2004–2005	Eesti Biokeskus, programmeerija

Peamised uurimisvaldkonnad:

Suuremahuliste geeniekspressiooni andmete analüüs – geeniekspressiooni mustrite uurimine erinevate tingimuste ja kudede vahel. Geenide iseloomustamine nende koos-ekspressiooni alusel.

Geeni ekspressiooni uurimine bioloogiliste radade kontekstis – geenide ekspressiooni dünaamika uurimine. Ekspressiooni dünaamika raja sees ja erinevate tingimuste vahel. Uute raja kandidaatide leidmine kasutades ekspressiooni sarnasust.

Publikatsioonid:

1. Örd, T., Örd, D., **Adler, P.**, Vilo, J., Örd, T. (2015) TRIB3 enhances cell viability during glucose deprivation in HEK293-derived cells by upregulating IGFBP2, a novel nutrient deficiency survival factor. *BBA - Molecular Cell Research*, Accepted, in press
2. Lees, J. G.*, Heriche, J. K.*, Morilla, I., Fernandez, J. M., **Adler, P.**, Krallinger, M., Vilo, J., Valencia, A., Ellenberg, J., Ranea, J. A., and Orengo, C. (2015).

Fun-L: Gene prioritization for RNAi screens. *Bioinformatics*, 31(12):2052–2053.

3. Heriche, J. K., Lees, J. G., Morilla, I., Walter, T., Petrova, B., Roberti, M. J., Hossain, M. J., **Adler, P.**, Fernandez, J. M., Krallinger, M., Haering, C. H., Vilo, J., Valencia, A., Ranea, J. A., Orengo, C., and Ellenberg, J. (2014). Integration of biological data by kernels on graph nodes allows prediction of new genes involved in mitotic chromosome condensation. *Molecular Biology of the Cell*, 25(16):2522–2536.
4. Kolde, R., Laur, S., **Adler, P.**, and Vilo, J. (2012). Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics*, 28(4):573–580.
5. Billon, N., Kolde, R., Reimand, J., Monteiro, M. C., Kull, M., Peterson, H., Tretyakov, K., **Adler, P.**, Wdziekonski, B., Vilo, J., and Dani, C. (2010). Comprehensive transcriptome analysis of mouse embryonic stem cell adipogenesis unravels new processes of adipocyte development. *Genome Biology*, 11(8):R80.
6. **Adler, P.***, Kolde, R.*, Kull, M., Tkachenko, A., Peterson, H., Reimand, J., and Vilo, J. (2009). Mining for coexpression across hundreds of datasets using novel rank aggregation and visualization methods. *Genome Biology*, 10(12):R139.
7. Schulz, H.*, Kolde, R.*, **Adler, P.**, Aksoy, I., Anastassiadis, K., Bader, M., Billon, N., Boeuf, H., Bourillot, P.-Y., Buchholz, F., Dani, C., Doss, M. X., Forrester, L., Gitton, M., Henrique, D., Hescheler, J., Himmelbauer, H., Hübner, N., Karantzali, E., Kretsovali, A., Lubitz, S., Pradier, L., Rai, M., Reimand, J., Rolletschek, A., Sachinidis, A., Savatier, P., Stewart, F., Storm, M. P., Trouillas, M., Vilo, J., Welham, M. J., Winkler, J., Wobus, A. M., and and, A. K. H. (2009). The FunGenES database: a genomics resource for mouse embryonic stem cell differentiation. *PLoS ONE*, 4(9):e6804.
8. **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*, 1158:1–13.
9. **Adler, P.***, Reimand, J.*, Jänes, J., Kolde, R., Peterson, H., and Vilo, J. (2008). KEGGanim: pathway animations for high-throughput data. *Bioinformatics*, 24(4):588–90.
10. 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(suppl 2):W452–W459.

Saadud stipendiumid:

1. Reisistipendium ISMB/ECCB konverentsil osalemiseks, 2009
2. Artur Linnu stipendium, 2008

Õppetöö:

1. ELIXIR RNAseq analüüs kursus, Tartu, 12.06.2015
2. ELIXIR Linux käsurida kursus, Tartu, 15.05.2015
3. ELIXIR MEM kursus, Tartu, 09.12.2014
4. Loengud aines Bioinformaatika, Tartu, sügis 2010
5. EBI ringreis 2009, Tartu 07.-10.09.2009

Juhendatud väitekirjad:

1. Tõnis Tasa, magistriöö "Avalike RNA-Seq andmete taaskasutamine" (2015)
2. Aleksei Panarin, bakalaureusetöö "Bioloogiliste radade spetsiifiline geeniekspressiooni andmete analüüs" (2009)

Osaletud kursused ja töötoad:

1. "e-Infrastructure for Massively Parallel Sequencing" UPPNEX Töötuba 2015 (Uppsala)
2. "Best practices for training in Next Generation Sequencing (NGS) Analysis" TGAC/Elixir Töötuba 2015 (Cambridge)
3. "Bioinformatics for Systems and Synthetic Biology" Bologna Talvekool 2007
4. "Advanced Analysis and Informatics of Microarray Data" EMBO Praktiline kursus 2007 (Hinxton)

Kuulumine erialaliitudesse:

1. Rahvusvaheline arvutusliku bioloogia ühing (ISCB), since 2007
2. ISCB tundengite ühing, since 2007

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.
276. **Hedi Peterson.** Exploiting high-throughput data for establishing relationships between genes. Tartu, 2015, 186 p.