Realia et naturalia

DISSERTATIONES
MATHEMATICAE
UNIVERSITATIS
TARTUENSIS
90

RAIVO KOLDE

Methods for re-using public gene expression data

TARTU ÜLIKOOL
1632
UNIVERSITAS TARTUENSIS

DISSERTATIONES MATHEMATICAE UNIVERSITATIS TARTUENSIS
90

DISSERTATIONES MATHEMATICAE UNIVERSITATIS TARTUENSIS
90

RAIVO KOLDE

Methods for re-using public gene expression data

UNIVERSITY OF TARTU
PRESS
1632

Institute of Computer Science, Faculty of Mathematics and Computer Science, University of Tartu, Estonia

Dissertation is accepted for the commencement of the degree of Doctor of Philosophy (PhD) on April 29th, 2014 by the Council of the Institute of Computer Science, University of Tartu.

Supervisor:

Prof. PhD.	Jaak Vilo University of Tartu Tartu, Estonia

Opponents:

Prof. PhD.	Lars Juhl Jensen University of Copenhagen Copenhagen, Denmark
PhD.	Kimmo Palin University of Helsinki Helsinki, Finland

The public defense will take place on June 16th, 2014 at 14:00 in Liivi 2-404.

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

ISSN 1024-4212
ISBN 978-9949-32-550-4 (print)
ISBN 978-9949-32-551-1 (pdf)

University of Tartu
www.tyk.ee

To my children: Artur and Katariina

Contents

PUBLICATIONS INCLUDED IN THIS THESIS

PUBLICATIONS INCLUDED IN HERE

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

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

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

4. R. Kolde and J. Vilo. GOsummaries: an R Package for Visual Functional Annotation of Experimental Data. in preparation.

PUBLICATIONS NOT INCLUDED IN THIS THESIS

1. P. Adler, J. Reimand, J. Jänes, R. Kolde, H. Peterson, and J. Vilo. KEGGanim: pathway animations for high-throughput data. *Bioinformatics*, 24 (4):588–590, Feb. 2008.

2. M. P. Storm, B. Kumpfmueller, B. Thompson, R. Kolde, J. Vilo, O. Hummel, H. Schulz, and M. J. Welham. Characterization of the phosphoinositide 3-kinase-dependent transcriptome in murine embryonic stem cells: identification of novel regulators of pluripotency. *Stem Cells*, 27(4):764–775, Apr. 2009.

3. M. Trouillas, C. Saucourt, B. Guillotin, X. Gauthereau, L. Ding, F. Buchholz, M. X. Doss, A. Sachinidis, J. Hescheler, O. Hummel, N. Huebner, R. Kolde, J. Vilo, H. Schulz, and H. Boeuf. Three LIF-dependent signatures

and gene clusters with atypical expression profiles, identified by transcriptome studies in mouse ES cells and early derivatives. *BMC Genomics*, 10 (1):73, 2009.

4. E. Maron, K. Kallassalu, A. Tammiste, R. Kolde, J. Vilo, I. Tõru, V. Vasar, J. Shlik, and A. Metspalu. Peripheral gene expression profiling of CCK-4-induced panic in healthy subjects. *American Journal of Medical Genetics. Part B*, 153B(1):269–274, Jan. 2010.

5. T. Vooder, K. Välk, R. Kolde, R. Roosipuu, J. Vilo, and A. Metspalu. Gene Expression-Based Approaches in Differentiation of Metastases and Second Primary Tumour. *Case Reports in Oncology*, 3(2):255–261, 2010.

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

7. M. X. Doss, V. Wagh, H. Schulz, M. Kull, R. Kolde, K. Pfannkuche, T. Nolden, H. Himmelbauer, J. Vilo, J. Hescheler, and A. Sachinidis. Global transcriptomic analysis of murine embryonic stem cell-derived brachyury (T) cells. *Genes to Cells*, 15(3):209–228, Feb. 2010.

8. K. Välk, T. Vooder, R. Kolde, M.-A. Reintam, C. Petzold, J. Vilo, and A. Metspalu. Gene expression profiles of non-small cell lung cancer: survival prediction and new biomarkers. *Oncology*, 79(3-4):283–292, 2010.

9. L. Tserel, R. Kolde, A. Rebane, K. Kisand, T. Org, H. Peterson, J. Vilo, and P. Peterson. Genome-wide promoter analysis of histone modifications in human monocyte-derived antigen presenting cells. *BMC Genomics*, 11(1): 642, 2010.

10. U. Võsa, T. Vooder, R. Kolde, K. Fischer, K. Välk, N. Tõnisson, R. Roosipuu, J. Vilo, A. Metspalu, and T. Annilo. Identification of miR-374a as a prognostic marker for survival in patients with early-stage nonsmall cell lung cancer. *Genes, Chromosomes & Cancer*, 50(10):812–822, Oct. 2011.

11. Y. Chen, M. Jørgensen, R. Kolde, X. Zhao, B. Parker, E. Valen, J. Wen, and A. Sandelin. Prediction of RNA Polymerase II recruitment, elongation and stalling from histone modification data. *BMC Genomics*, 12(1):544, 2011.

12. L. Tserel, T. Runnel, K. Kisand, M. Pihlap, L. Bakhoff, R. Kolde, H. Peterson, J. Vilo, P. Peterson, and A. Rebane. MicroRNA expression profiles

of human blood monocyte-derived dendritic cells and macrophages reveal miR-511 as putative positive regulator of Toll-like receptor 4. *The Journal of Biological Chemistry*, 286(30):26487–26495, July 2011.

13. K. Lokk, T. Vooder, R. Kolde, K. Välk, U. Võsa, R. Roosipuu, L. Milani, K. Fischer, M. Koltsina, E. Urgard, T. Annilo, A. Metspalu, and N. Tõnisson. Methylation markers of early-stage non-small cell lung cancer. *PLoS ONE*, 7(6):e39813, 2012.

14. J. A. Gaspar, M. X. Doss, J. Winkler, V. Wagh, J. Hescheler, R. Kolde, J. Vilo, H. Schulz, and A. Sachinidis. Gene expression signatures defining fundamental biological processes in pluripotent, early, and late differentiated embryonic stem cells. *Stem Cells and Development*, 21(13):2471–2484, Sept. 2012.

15. A. K. Krug, R. Kolde, J. A. Gaspar, E. Rempel, N. V. Balmer, K. Meganathan, K. Vojnits, M. Baquié, T. Waldmann, R. Ensenat-Waser, S. Jagtap, R. M. Evans, S. Julien, H. Peterson, D. Zagoura, S. Kadereit, D. Gerhard, I. Sotiriadou, M. Heke, K. Natarajan, M. Henry, J. Winkler, R. Marchan, L. Stoppini, S. Bosgra, J. Westerhout, M. Verwei, J. Vilo, A. Kortenkamp, J. Hescheler, L. Hothorn, S. Bremer, C. van Thriel, K.-H. Krause, J. G. Hengstler, J. Rahnenführer, M. Leist, and A. Sachinidis. Human embryonic stem cell-derived test systems for developmental neurotoxicity: a transcriptomics approach. *Archives of Toxicology*, 87(1):123–143, Jan. 2013.

16. U. Võsa, T. Vooder, R. Kolde, J. Vilo, A. Metspalu, and T. Annilo. Meta-analysis of microRNA expression in lung cancer. *International Journal of Cancer*, 132(12):2884–2893, June 2013.

17. T. Waldmann, E. Rempel, N. V. Balmer, A. König, R. Kolde, J. A. Gaspar, M. Henry, J. Hescheler, A. Sachinidis, J. Rahnenführer, J. G. Hengstler, and M. Leist. Design principles of concentration-dependent transcriptome deviations in drug-exposed differentiating stem cells. *Chemical Research in Toxicology*, 27(3):408–420, Mar. 2014.

18. S. Ilmjärv, C. A. Hundahl, R. Reimets, M. Niitsoo, R. Kolde, J. Vilo, E. Vasar, and H. Luuk. Estimating differential expression from multiple indicators. *Nucleic Acids Research*, page gku158, Feb. 2014

ABSTRACT

Public gene expression databases contain data about more than million biological samples, from hundreds of tissues and diseases. In principle, we know the expression patterns for all genes in these samples. By re-using and integrating these datasets it is possible to tackle novel biological problems with only computational means. To take maximal advantage of public gene expression data, there is a need for appropriate statistical methodology, user friendly software for exploring the data, novel ways of visualisation, etc. This thesis presents several articles that address these problems.

In many cases even making the data more easily accessible can be useful. As a first project, we built a web based data atlas for a large consortium studying embryonic stem cells. The atlas consisted of series of interactive visualisations that presented the data from various angles. The goal of this work was to facilitate collaboration between the partners and to give the public an easy way to access the data.

Often it is possible to increase the power of an analysis by performing it in many datasets simultaneously and then integrating the results. Our web server MEM takes advantage of this idea, by allowing to search for genes with similar expression pattern over hundreds of datasets. By using many experiments we increase the reliability of results.

Using many datasets creates a need to integrate the results. For MEM we created novel method RRA for integrating ranked gene lists. It works well on noisy input data, since it is robust and evaluates also the significance of the results. These features, however, make the method useful also in many other bioinformatic applications, where data has to be integrated from multiple experiments.

Visual inspection of results is critical in every step of the microarray analysis, since the data is complex and can contain unexpected patterns. This thesis presents a method for visualising the results of functional analysis of gene lists as word clouds. It allows to combine the annotations of multiple gene lists and present them together with experimental data. This way it is possible to create concise and effective summaries for common analyses on gene expression data.

CHAPTER 1

INTRODUCTION

The invention of gene expression microarrays has given molecular biologists the ability to measure expression of thousands of genes in parallel with a relatively low effort. The results from such experiments carry a wealth of information, but the sheer size of the data has created numerous computational and statistical challenges. This has attracted many computer scientists and statisticians to work on this topic. They study how to effectively store and handle the data, how to normalise and preprocess it and, finally, how to interpret it statistically. By the midpoint of previous decade, main components of the analysis pipeline for a single experiment had emerged. The software and infrastructure has evolved significantly from then, but methods for basic analyses such as normalisation, clustering and statistical testing have not seen major improvements.

At the same time new bioinformatic challenges emerged from terabytes of data that had been gathered into public gene expression databases. The interesting aspect about the gene expression data is that the measurements are relatively comprehensive. They yield information also about the genes that are not in the immediate interest of the researchers who conducted the experiment. This information, however, can be relevant in some other context. Therefore, the existing genomic data can be re-used to find answers to new biological questions. There have been numerous studies since then employing this approach, for example, to predict gene function or to identify disease specific genes.

Still, there is big potential in re-use of the existing data. Databases contain tens of thousands of experiments describing hundreds of tissues and diseases. When presented appropriately these datasets can assist in day-to-day lab work: in designing experiments, in prioritising candidate genes for experimental analysis or in viewing the obtained results in a wider context. Although the data can be downloaded freely from public databases, it is not easy to extract the relevant information from there. Useful bits of information are hidden in huge amounts of data and what can be called "useful" depends heavily on the question at hand. Thus, each

biological question requires customised approach and there is no single solution that would take care of all the problems related to re-use of gene expression data. Rather the solution is to develop many approaches that can reveal different aspects of the data.

The main goal for this work is to create methods and tools that help to take advantage of public gene expression data. This thesis covers four articles that tackle this problem from different angles.

One of the main obstacles for re-using gene expression data is its accessibility. Even though there are huge collections of data freely available in dedicated databases, it still takes considerable effort to download the data, perform proper pre-processing and actual analysis. Moreover, people who would need the results the most, lack the specific statistical and computational skills that the analysis requires. One way to improve the accessibility of the data is to create web based environments for that make certain types of analysis on public data more user friendly. This thesis covers two articles that take this approach.

- **Article I** - "*The FunGenES database: a genomics resource for mouse embryonic stem cell differentiation.*" - describes a database for Functional Genomics in Embryonic Stem Cells (FunGenES) consortium data, where we provided tools to interactively mine the gene expression data generated by the consortium.

- **Article II** - "*Mining for coexpression across hundreds of datasets using novel rank aggregation and visualization methods.*" - describes a web based tool called Multi Experiment Matrix (MEM) that allows to search genes with correlated expression patterns over large collections of public gene expression data.

Often when re-using gene expression data there is a need to integrate results from multiple sources. For example, there can be multiple similar studies or several experimental approaches to study the same question. There are many methods for data integration, but often the analysis reduces to comparison of gene lists. While this task is common, there are not many techniques that are designed specifically with lists of genes in mind. This thesis presents a rank aggregation method that can handle many of the practical problems that integration of gene lists presents.

- **Article III** - "*Robust Rank Aggregation for gene list integration and meta-analysis.*" - describes the Robust Rank Aggregation algorithm in detail.

While re-using expression data and working with multiple datasets it is important to obtain good overview about each dataset. Knowing features like data

quality or most important experimental factors could have substantial impact for the downstream analysis. Various visualisation methods in combination with data mining approaches are the key to summarise the large datasets in compact format. Last article in this thesis presents a novel visualisation method that can quickly create high level functional overviews of complex datasets and, thus, improve the quality of subsequent analysis.

- **Article IV** - "*GOsummaries: an R package for visual functional annotation of experiemtnal data*" - describes an R package GOsummaries package that allows to summarise the results of common analysis methods in compact manner, by representing functional annotations of gene lists as word clouds.

Thesis itself is organised as follows. First, the Preliminaries chapter gives a short background of the methods in molecular biology and bioinformatics that are required to understand the rest of the thesis. Next chapters give more specific context and summarise shortly the articles I-IV. As several of those are written in collaboration with many co-authors, then the summaries are concentrating more on my contribution to the studies. Copies of the articles I-IV are included at the end of this dissertation.

CHAPTER 2

PRELIMINARIES

2.1 Gene expression and its regulation

The heritable information in most living organisms is stored in long molecules of deoxyribonucleic acid (DNA). These molecules are chains of four nucleic acids: guanine, adenine, thymine, and cytosine (G, A, T and C). The order of the nucleic acids encodes instructions on how a cell is built and how does it work. Most importantly, regions in the genome, called genes, encode the sequence of amino acids of all the proteins that can be found in the cell. Proteins are vital parts of a cell that participate in almost every process taking place in there. Humans have roughly 20000 protein-coding genes (Consortium et al., 2012).

The information in DNA is turned into functional proteins in a two step process. First, the sequence of the DNA is transcribed into a complementary sequence of ribonucleic acid (RNA), called messenger RNA (mRNA). Then these molecules are transported to ribosomes, where the RNA sequence is translated into amino acid sequence that forms a protein. This process of converting information in a gene into a protein is called gene expression

In principle, the genomic sequence is the same in every cell of the organism. The diversity in the build-up and function of the cells only becomes possible through the differential regulation of gene expression. There is a number of mechanisms that control the gene expression.

An important class of proteins that has the ability to bind to DNA in a sequence-specific manner is called transcription factors. When bound, they can attract RNA polymerase protein that performs the transcription, or conversely block the transcription. All of the transcribed RNA is not used to make proteins. Some RNA itself is used for regulating gene expression. For example, small microRNA (miRNA) molecules bind to complementary mRNA molecules and suppress their translation into proteins.

The gene expression and the described regulatory mechanisms represent only

a subset of all the events happening in a cell. However, these processes are the most relevant in the context of this work, since they can be characterised in a high-throughput manner.

2.2 Measuring gene expression

Measuring the expression of genes and the activity of the regulatory mechanisms often boils down to quantifying the abundance of different RNA/DNA molecules. The amount of a specific protein, can be approximated by the number of the corresponding mRNA molecules in the cell.

The ideal way to quantify RNA/DNA molecules would be sequencing them individually and counting afterwards. Until recently such an approach was prohibitively slow and expensive. As a substitute high-throughput technique, microarrays were developed. Microarrays are experimental devices that contain single stranded DNA molecules with predefined sequence to capture RNA molecules with complementary sequence from the solution of interest. The molecules in the sample are labeled with a fluorescent dye, so the abundance of molecules that are bound to a specific set of probes on the array can be estimated by the intensity of light they emit. Typical microarray can fit in the order of 10000 - 1000000 different probes. Therefore, it is possible to fit a probe or even multiple probes for every gene to a microarray. This makes microarrays convenient for measuring gene expression.

Much of the work in this thesis is devoted to the analysis of gene expression microarray data. Thanks to all the new developments in sequencing, imaging and other high-throughput methods, the importance of gene expression microarrays in molecular biology is declining. However, most of the methodology that was developed for analysing such data will remain relevant as it can be easily adapted to newer technologies.

2.3 Gene expression data and common analysis approaches

At first glance a dataset with information about all genes offers a large number of possibilities for analysis, but in most cases the analysis follows the steps shown in the Figure 2.1.

First, the measurements are obtained from the machine in a raw format. The format and content of these files depends a lot of the technology used. The first goal of the analysis is to convert the data into a matrix format where rows correspond to genes, columns to samples and the values in the matrix show the expres-

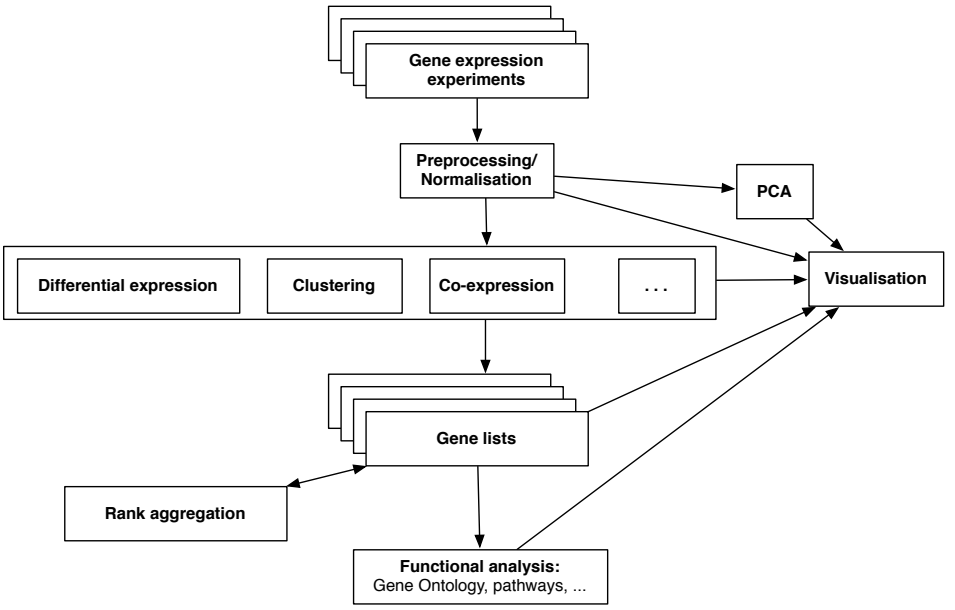

Figure 2.1: Overview of the typical analysis pipeline of gene expression microarrays.

sion levels. This step again depends on the technology, but usually involves image processing, background correction, removal of technical biases, etc. A great deal of research has been conducted on pre-processing and normalisation of the raw data (Amaratunga and Cabrera, 2001; Bolstad et al., 2003; Huber et al., 2002). However, in this thesis we focus on analysis that comes after we have obtained the gene expression matrix.

To get a first glance of the data, people often use Principal Component Analysis (PCA). PCA reduces dimensionality of the data while preserving relationships between the observations or samples. In case of expression data, it allows to shrink thousands of gene expression values into only a few principal components that can be visualised as a scatterplot. The distances between the samples in this plot approximate the actual distances in the high dimensional space, giving a good overview of their relationships. This type of visualisation is popular, since it summarises the data concisely in one plot, does not require any additional data or pose any constraints to the experimental setup. A more thorough overview about PCA is given in section 6.2.

The most important approach in genome-wide gene expression analysis is finding differentially expressed genes between experimental groups. In the simplest case of two groups of samples, healthy and diseased for example, we can apply a t-test across every row of the expression matrix. As a result we obtain a list of genes where the difference in average expression between the experimental

groups is statistically significant. Depending on the experimental setup one can apply also more complex statistical algorithms than t-test, such as more sophisticated linear models (Smyth, 2004), non-parametric tests (Breitling et al., 2004) or Bayesian approaches (Baldi and Long, 2001). Differential expression analysis allows to discover genes that are related to various diseases, specific to certain tissues, etc.

Another central concept in gene expression analysis is the co-expression between the genes. Most fundamentally, if two genes have a similar expression profile through several biological conditions, they can be expected to be similar also at functional level. For example, they can be regulated by the same factors or participate in similar processes. There are various ways how the similarity information is used in the bioinformatic analyses. For example, the similarly expressed genes have been used to assign functions to the unannotated genes (Huttenhower et al., 2009), to prioritise the disease related genes (Aerts et al., 2006), to verify predicted protein-protein interactions (Kemmeren et al., 2002), etc.

The most popular approach using similarity of expression profiles is clustering. Since the number of distinct patterns in the data is usually much lower than the number of genes then it makes sense to group together genes with similar profiles. There are two main types of clustering. Hierarchical clustering creates a tree using the similarity scores. K-means and other similar algorithms divide the data into predefined number of groups. Clustering of genes is mostly used for visualisation purposes, to identify major trends in the dataset with more than two or three experimental groups. The columns of the expression matrix can be clustered as well, this reveals the similarity structure between the biological samples. This has been used, for example, to define subtypes of cancer that have similar expression profiles (Sørlie et al., 2001).

Clustering, co-expression analysis, differential expression analysis and many other methods yield (ranked) lists of genes as a result. A gene list in turn is a natural entry format for many downstream analyses that try to integrate the findings with existing information. A common approach is to search biological processes and pathways that are enriched in a list of genes. This approach is called Gene Ontology enrichment analysis and is described in detail in section 6.1. It allows to convert the gene names into more general and understandable functional terms. But lists of genes can also be viewed in the context of existing gene networks to see if they are functionally related or can have protein level interactions (Szklarczyk et al., 2011).

An important part of any analysis is data visualisation. The gene expression data contains information about tens of thousands of genes in many samples. Thus, it is not possible to comprehend the information in the data by just looking at the numbers. Visualisation usually accompanies every part of the analysis

described in Figure 2.1, aiding the understanding of the results, providing sanity checks for the methods and helping to identify outliers and dominant patterns in the data. The most common way to display the gene expression data is a clustered heatmap (Eisen et al., 1998). It is a great fit, since it is designed to display larger numeric matrices, just like gene expression matrix.

When the results of a gene expression experiment are published, the raw data is usually uploaded to some public database like Gene Expression Omnibus (GEO) (Barrett et al., 2009) or ArrayExpress (Parkinson et al., 2009). This allows to independently validate the analytic procedures and the claims of the authors.

2.4 Re-analysis of collections of gene expression data

Storing data in public gene expression databases has also created the opportunity to re-use the datasets in other contexts. In the beginning of 2013 there was data from over 30 000 experiments covering over million samples in ArrayExpress database (Rustici et al., 2013). It means that wide spectrum of tissues, diseases and treatments are covered already with existing microarray data. Therefore, it is a great opportunity for the bioinformatics community to re-analyse or create tools that allow re-analysing the growing amount of public experimental data. A good overview about the usage of public data can be found in (Rung and Brazma, 2013). The goal of this thesis was to create tools and methods that would facilitate re-use of public data.

One aspect that complicates the re-use of expression data is the poor availability of annotations describing the original experiment in appropriate detail. Even though there are rigorous standards for annotating microarray experiments (Brazma et al., 2001), in most cases it is still hard to understand the experimental setup and the origin of samples without reading the original article. Performing an analysis on larger set of data requires a lot of re-curation to unify all annotations. It has been done on several occasions (Lukk et al., 2010; Rhodes et al., 2004b), but it is often not practical.

There are many things that can be done with collections of gene expression data with adequate annotations. One common approach has been to gather datasets from specific domains of molecular biology, such as stem cells or cancer and to present them in the web in a more accessible format. Examples include Stemformatics (Wells et al., 2013), Stembase (Sandie et al., 2009), ESCDb (Jung et al., 2010), etc. These databases are aimed to be used by biologists who need to compare their own results against existing data and who want to learn more about the behaviour of specific genes in a larger set of relevant conditions. Even without complicated analyses, just by making the data more easily accessible, it is possible to create useful resources for the experimental community.

A common approach is to scale up single dataset analysis methodologies. For example, the co-expression analysis can be used in multiple dataset approach. We can be more confident that co-expression indicates functional relationship between specific genes, if it persists across many conditions. Therefore, including multiple datasets in the analysis should considerably improve the biological interpretation of the results. In co-expression studies we also do not have to worry too much about re-curation of the datasets, since correlations between the expression values can be calculated without knowing anything about the samples. There are many examples that use co-expression information in this way. Most importantly it has been used for predicting gene function (Huttenhower et al., 2009; Hibbs et al., 2007; Wolfe et al., 2005; Sardiello et al., 2009).

Other popular approach is to collect experiments with similar goals, perform differential expression analysis and try to aggregate the results in a meta-analysis (Ramasamy et al., 2008). For example, there are many studies identifying lung cancer specific genes in different cohorts, therefore, it is reasonable to base the conclusions on larger number of samples. This has been done, for example, for thyroid (Griffith et al., 2006), breast (Wirapati et al., 2008) and colorectal cancers (Chan et al., 2008). A web based resource Oncomine (Rhodes et al., 2007) has been created that collects differentially expressed gene lists from public cancer data and allows researchers easily select and combine results of the experiments.

In both differential expression meta-analysis and co-expression studies methodological questions arise on how to integrate information from many sources. There are many options and the choice depends heavily on the nature and quality of the data. An overview about the integration methods for differential expression meta-analysis can be found in section 5.1.

CHAPTER 3

FUNGENES DATABASE

The FunGenES consortium (2004-2007) consisted of 15 research groups and was founded to study basic biological properties of embryonic stem cells and their differentiation. Among other experiments, consortium partners created many gene expression datasets describing various lineages of cell differentiation. Our role as the bioinformatics partner in this collaboration was mainly to provide bioinformatic support in analysing and interpreting the data. As the experiments explored closely related topics and were often complementary, it was also important to facilitate sharing of the results and data between research groups. Most importantly, there was little embryonic stem cell data in public domain at that time, therefore, we also had to create a publicly available resource that allowed convenient access to the FunGenES experiments.

As a result, we contributed to several individual studies by analyzing the data (Billon et al., 2010; Trouillas et al., 2009; Storm et al., 2009; Gaspar et al., 2012; Doss et al., 2010). However, to facilitate collaboration within the project and share the data with community we ended up building a set of simple web based tools and interfaces available at http://biit.cs.ut.ee/fungenes/. These are introduced in more detail below.

3.1 FunGenES database - article I

A common scenario within the consortium was that one group identified a set of genes with interesting expression pattern in one experiment and then wanted to check how these genes are behaving in related experiments. Based on raw data, such analysis would take more time and effort than the expected result is worth.

Expressview: web based clustering and visualisation tool

To make this analysis easier we created a simple web service. It included all the FunGenES data and people could enter genes of interest and see their expression as a clustered heatmap on selected dataset. Despite being a simple tool, it gave the biologists an access to data generated within the consortium and a way to put their own results into a wider context.

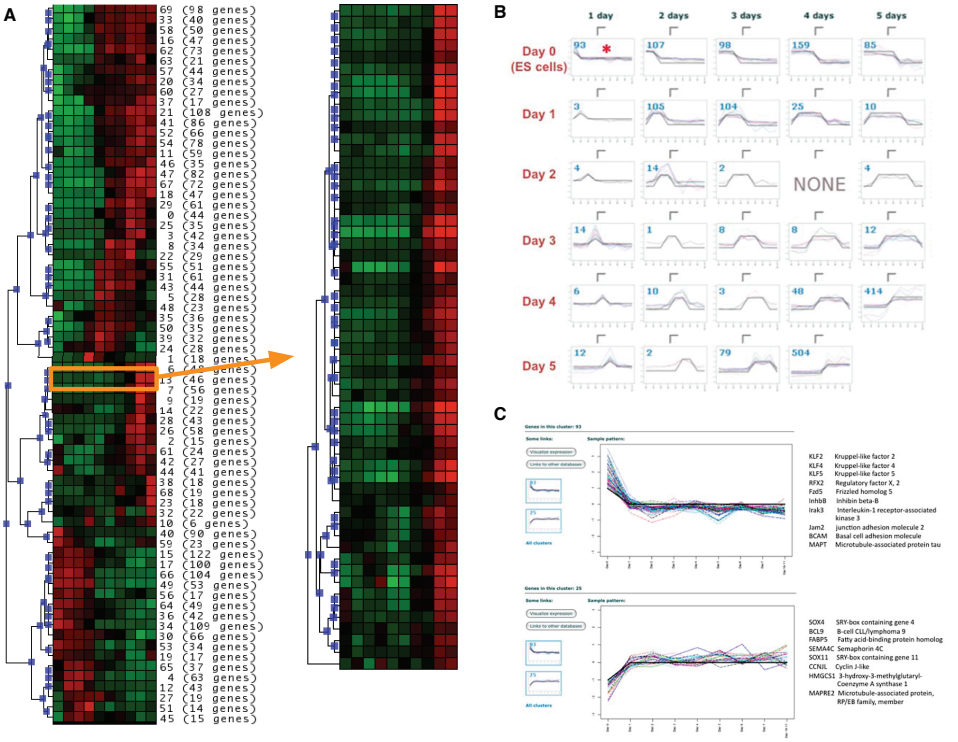

Figure 3.1: Panel A shows the result of two level clustering on the time course data. Each row in the left heatmap corresponds to a cluster of genes. The clusters centers are in turn clustered hierarchically. In the web interface one can zoom into clusters or combine clusters together by clicking either on the heatmap rows or the nodes of the tree. The right heatmap displays the expression of gene from cluster 13 that is marked with the orange square. Panel B shows the overview of the gene expression "wave" analysis, where we associated genes with prespecified array of patterns. Panel C shows an overview of on specific wave (marked with red star in B).

This tool gave an overview of only small part of the dataset. Some of the datasets were rather complex, covering several time points and treatments. To generate sensible hypotheses it was important to understand the general structure of the dataset in terms of prevalent trends and patterns. This type of analysis is

usually done by clustering, however, the ordinary methods have several shortcomings. The k-means type of methods divide the genes into predefined number of groups. These methods are fast and can handle all the genes easily, but it is hard to optimise the number of clusters. The distribution of genes in gene expression space is rather continuous and their division into non-overlapping groups is rather arbitrary. Ideally, we would also like to take into account the functional similarity of the genes, vary the level of granularity of the clusters based on the interestingness of the patterns, etc. Therefore, the results generated using an automatic methods often just do not "feel right". Hierarchical clustering is much better in that sense. Based on the tree it creates, it is easy to select the sub-trees that are interesting. However, hierarchical clustering has both computational and perceptional problems with larger datasets. Its complexity grows quadratically as the number of genes increases. It is possible to speed up the computations to be able to cluster large datasets (Kull and Vilo, 2008), but even then it is hard to visually inspect a clustering tree that covers thousands of genes.

To build on the strengths of both approaches, we decided to merge them and provide a web interface where one could interactively browse and re-organise the clusters. At first we used the k-means clustering of the data, using an arbitrary number of clusters, for example 50. Then we performed hierarchical clustering of the centres of these pre-formed clusters and displayed the result on a heatmap. In the web interface it is possible to merge the clusters that are similar to each other according to the hierarchical clustering tree. Also, it is possible to zoom into the clusters to see individual genes. An example of this visualisation can be seen on Figure 3.1A.

The original k-means clusters give an unbiased general overview of all the patterns present in the data. However, owing to the subsequent hierarchical clustering we excluded the problem of the number of clusters being too big, since redundant clusters could always be merged afterwards.

This type of clustering turned out to be useful for displaying and analysing the stem cell differentiation time series included in the project. This data contained many temporal patterns related to the specific stages of differentiation. Statistical tests fitting regressions or group-wise comparisons would quantify the presence of predefined patterns. However, the number of possible patterns in such setting is too high to be tested individually. Ordinary k-means clustering would identify the most common patterns, but there would be questions about the correct number of clusters and other parameters that influence the result. The hierarchical display of the k-means results and interactive features of our web interface allow to zoom into clusters and combine them to create an optimal grouping of the genes.

Time series clustering with templates

Using the clustering approach, we found out that the dominant patterns in the data were "waves" of genes switched on and off at different points of differentiation. To quantify these patterns more rigorously, we created a set of wave like template patterns and searched for most correlated genes with every template. We created an interactive web interface to display these gene groups. It can be seen in Figure 3.1B and 3.1C. In this page, it is possible to see the prevalence of various patterns and study the relevant genes together with their functional annotations.

The FunGenES project also spurred development of some standalone tools in our group, such as g:Profiler (Reimand et al., 2011) for GO enrichment analysis and KEGGAnim (Adler et al., 2008) for overlaying expression data to the KEGG pathway maps. Although the use of each individual tool is fairly limited, when combined together, they can give a rather comprehensive overview of the available data and provide an easy way for access. We created a separate webpage that brought all the analyses together and added also quick links to external databases. For example, with only few clicks, it is possible to start from a group of genes with interesting wave pattern, inspect how these genes behave in other FunGenES datasets, annotate them functionally with g:Profiler or visualise their relations in gene networks using external tool like STRING (Szklarczyk et al., 2011).

Summary and impact

By creating FunGenES database, we have generated a resource that adds value to the body of data that arose from the consortium. Several studies (Trouillas et al., 2009; Billon et al., 2010) from the FunGenES consortium and elsewhere (Singh et al., 2012; Davis and Summers, 2012; Tanwar et al., 2014) have used these database features extensively, for viewing their results in the context of FunGenES data. The tools have been used by our group in several subsequent collaboration projects and are still being developed further.

Contribution

In this project, I developed the heatmap-based visualisation and clustering tools, the time series clustering and put together the final website. Additionally, I took part of writing the article and provided several case studies for it.

CHAPTER 4

CO-EXPRESSION QUERIES ON LARGE COLLECTIONS OF DATA

Within FunGenES consortium, we created a web-based resource that presents extensive microarray data in a more accessible format. The users can now browse and ask simple questions without the need to deal with the technical analysis details of microarray data. Such approach can be developed further, if we include more data from pubic domain and implement more demanding data analysis pipelines.

Co-expression analysis is one of the analytical techniques that benefits greatly from the usage of multiple experiments (see also Chapter 2). Since co-expression can reveal functional relations between the genes, it is relevant in many applications. Even though in co-expression analysis one does not have to worry about the data annotations too much, it is still a lot of work. One has to download all the datasets, perform proper normalisation, decide on the aggregation strategy, perform the actual analysis, etc.

To make co-expression analysis on public microarray data easier, we have created a web based resource named Multi-Experiment Matrix (MEM) that is located at http://biit.cs.ut.ee/mem/. We have downloaded and pre-processed large amounts of microarray data and built a query interface that can identify co-expressed genes in a given set of experiments.

4.1 MEM - article II

The goal of MEM is to find genes that are co-expressed with a given gene across many gene expression experiments. The co-expression search is performed in several steps:

- in every dataset, search for genes with profiles similar to a gene of interest;

- aggregate the individual gene lists;
- visualise the results in an interactive manner.

The result will be a ranked list of genes. In the next sections each of those steps is discussed in more detail

Similarity search

First, MEM performs the similarity search in every dataset. In principle, any distance measure can be used in this step to measure the co-expression. Several of those are also implemented in MEM. However, the most natural metric for co-expression is the correlation distance, since we are interested in gene pairs with similar expression dynamic. Some biological questions also require other type of distances. For example, if the query gene is a suppressor of transcription, then it would be interesting to search for genes that are anti-correlated with it. Some regulators can both induce and repress gene expression, thus, both correlated and anti-correlated genes are interesting. For these reasons, we have also the anti-correlation and absolute correlation implemented in MEM

Gene list aggregation

In the next step of the analysis, the information from all the individual similarity searches is aggregated. We considered using the actual similarity scores, like correlation values, for aggregation, but this demonstrated to have several drawbacks. First, we would have had to create a custom aggregation scheme for every metric, since their characteristics are quite different. More importantly, we discovered that the distribution of the similarity scores depended heavily on the structure of the dataset, the experimental design, number of samples, etc. Therefore, the similarity scores from different datasets were not directly comparable and we decided to aggregate the data on the ranked gene list level.

An obvious solution would have been to re-order the genes by their average rank or use some rank aggregation method. However, all the existing methods exhibit the same problem, they take into account all the information in the input rankings and, thus, are sensitive to noisy inputs. There were not many gene pairs that were co-expressed universally in every dataset. Even with gene pairs with strong co-expression there was a large proportion of datasets where the co-expression was non-existent. Therefore, a good rank aggregation algorithm for our setting had to be sensitive in a situation where many of the inputs could be considered as noise.

To find a more sensitive method, we introduced a new probabilistic approach for aggregating ranks. Instead of trying to find genes that are consistently similar

to the query gene in all the datasets, we tried to find the ones that are similar in unexpectedly many datasets. Since we published this method in a separate article it is described in more detail in the next chapter.

Figure 4.1: Example of MEM user interface and output for embryonic stem cell regulator Nanog. The top part shows the user interface for specifying the query and bottom part displays the results. The most co-expressed genes are displayed in the rows and the matrix shows the correlation ranks in individual datasets. The visual cues A-D, highlight the interactive features of MEM that allow to get more information about the genes and experiments.

User interface

The MEM analysis is performed in a web based user interface. User has to enter the name of a gene of interest and optionally select the datasets to run the analysis. The resulting gene list and additional information is shown as a heatmap type plot. Figure 4.1 shows an example of MEM output for well known embryonic stem cell regulator Nanog. The heatmap shows the top genes from the query

and their positions in every dataset as a matrix. For example, in Figure 4.1 we can see that most of the top genes are co-expressed with Nanog only in a small number of datasets. Several interactive features allow more in-depth study of these results. For example, one can get more detailed information about the experiments and genes (Figure 4.1B and 4.1C). The datasets with similar results can also be characterised using text mining approaches. The resulting word cloud can be seen in Figure 4.1A. We have also integrated the heatmap drawing web interface that was developed for FunGenES to MEM and this can be used to visualise the expression of the resulting genes in the underlying datasets (Figure 4.1D).

Dataset selection

The core algorithm of MEM works as described above. However, we discovered that we can obtain more meaningful results by performing dataset selection, prior to performing any analysis. In theory, adding datasets should increase the breadth of the search and, thus, improve the results of co-expression analysis. In practice, however, a gene might not always be expressed or its expression is the same across samples and its co-expression with other genes is random. Also the presence of various regulatory mechanisms means that the co-expression patterns between tissues can be different. For solving these problems we have implemented two dataset selection mechanisms. To concentrate on a specific tissue, disease or cell type, one can manually select the datasets using the dataset annotations as a guide. To remove noisy datasets from the search, MEM will only use datasets where the standard deviation of the query gene is above some threshold. We argue that in case of small variation the changes in expression correspond to random fluctuations, but larger variation can indicate some biologically meaningful signal.

To find a suitable default threshold, we performed an experiment. We ran MEM queries on 2000 genes. In each query we recorded the number of genes above the same significance limit. Then we checked if the number of results in a MEM query is correlated with the number of datasets where standard deviation of the query gene was over some threshold. Indeed, we found correlation. Therefore, we get more results from a MEM query if the variation of a query gene is over some threshold in more datasets. To find the best standard deviation threshold, we found this correlation with many potential thresholds between 0 and 1 and found that values around 0.3 gave the best correlation and, thus, serves as the best threshold. This number is valid for Affymetrix data that is normalised with RMA method (Irizarry et al., 2003), since this is the main type of data that is used by MEM.

Empirical observations confirmed the utility of selecting the datasets for a query gene in such a way. For example, in Figure 4.1 the genes that were identified as co-expressed with Nanog, an embryonic stem cell regulator, were also related

to embryonic stem cells. However, if we performed this query on all datasets, then the results were rather generic and did not display strong enrichment of any specific functional category.

Summary and impact

With MEM we created a unique tool that enables analyses what in most cases would be practically infeasible. The amount of raw data processed for the original publication was large, around half a terabyte. At the time of publishing, at the end of 2009, it was one of the largest gene expression data collections that can be interactively queried and mined. MEM has already been used in several studies. For example, to add evidence that two genes are functionally related (Chen et al., 2013; Tabach et al., 2013; Sircoulomb et al., 2011; Schraenen et al., 2010). In some studies the gene lists that were identified by MEM were used to infer functional context of certain genes or gene groups (Lacunza et al., 2013; Ivanov et al., 2013).

Contribution

The large amount of data that was included into MEM created various challenges. We had to develop suitable methodology and implement it in an efficient manner to make the queries possible. Also, we had to create a user interface that would allow to customise queries and display the results in a compact but informative manner. Tackling such challenges takes a group effort. I was mainly responsible for the methodological part, but also performed several case studies and drafted the article.

CHAPTER 5

ROBUST RANK AGGREGATION

In MEM, the co-expression query was performed in individual datasets and the ranked gene lists from these queries were aggregated into one ranking afterwards. Since the inputs were noisy, the traditional rank aggregation methods were not the best choice. Thus, we came up with a statistical rank aggregation method - Robust Rank Aggregation (RRA) - that is more tolerant to noise and also adds statistical confidence to the re-ranked elements.

Rank aggregation, in principle, could have many applications in bioinformatics, especially when re-using public expression data. Ranked lists of genes are a common output type for many bioinformatic analysis pipelines (see Figure 2.1 from Chapter 2) and rank aggregation methods are a natural fit for integrating data from multiple sources. However, the same problems that we had in MEM, are present in other bioinformatic applications. The gene lists tend to be noisy and in many cases there might be nothing relevant in the output at all. Also there can be problems with missing information. We recognised that RRA with some modifications has several features that make it a great fit for this setting and, thus, can be practical in many other situations beyond MEM.

The most obvious application of this method is differential expression meta-analysis. Next section gives a brief background on the problems and methods related to such analysis. Although, it is only one possible scenario for using RRA, much of this applies to other use-cases as well.

5.1 Meta-analysis of gene expression data

Differential expression meta-analysis integrates differentially expressed gene lists from multiple experiments with similar goals, for example, mis-regulated genes in different cancer cohorts.

Ramasamy et al. (2008) list the best practices to perform such study. It would be ideal to download raw data for all the experiments, normalise it, put it all together in one table and perform differential expression analysis using a suitable model. However, in practice the experiments are done using different platforms, covering differing sets of genes and using incompatible technologies. More importantly, many of the published studies do not include raw data and the published list of genes is the only available result. Therefore, first, the results have to be acquired from the individual studies and then aggregated somehow.

For aggregation there are several options. Simplest is maybe vote counting (Rhodes et al., 2004a; Griffith et al., 2006), where genes are ranked by the number of input lists where they are present. The statistical significance can be associated with the results using permutations. Rank aggregation approaches take more information into account, as they consider also the ordering of the significant results. Several studies apply the classical rank aggregation methods (DeConde et al., 2006; Pihur et al., 2008), but some have devised novel strategies that take the biological setting more into account (Zintzaras and Ioannidis, 2008; Hong and Breitling, 2008). Finally, one can also include the information about the effect sizes and p-values and use a technique like Fisher sum of logs (Rhodes et al., 2002) for combining p-values or inverse variance technique to combine effect sizes (Choi et al., 2003).

In theory, the latter methods are preferred, but in practice we might not be able to apply them because there are several common problems with the data. First, the effect size and p-value information might be incomparable between datasets, since initially different statistical tests were used. Second, several methods require full gene lists to be able to work properly, but if the raw data is not available, we can only rely on the lists of significant genes. Third, some genes are missing from some lists, since they were not measured by a certain platform. It is common that newer arrays can cover several times as many genes as the older ones. This type of structurally missing data should be handled properly, otherwise the results are biased towards genes that are represented in more lists. Finally, it is entirely plausible that even the top results of the aggregation are not relevant at all, if the input lists correspond to different biological questions, the underlying cohorts are incompatible or have just poor quality. Therefore, it is critical that the aggregation method would assess the significance of the results.

If the critical statistical information, such as effect sizes and p-values for all genes, is missing or incompatible, then it is not possible to use the methods like Fisher sum of logs. However, even if they are available, simpler rank aggregation methods can provide a better fit. The p-values and effect sizes are sensitive to study design and structure of the cohorts (Hong and Breitling, 2008). The ranking of the genes tends to be more stable between comparisons than the p-values.

On the other hand, vote counting methods are robust and can almost always be applied, since they need only gene lists that even do not have to be ranked. But these methods have several other problems. First, the results will be granular, especially, when the number of lists to be aggregated is small. Second, the results depend on rather arbitrary significance thresholds and studies with larger number of significant genes may dominate overall results. Most importantly, it is hard to assess significance of the results. In principle, it can be done using permutations. However, it gets complex if the number of the significant genes differs between the lists and there is structurally missing information.

Rank aggregation methods have the potential to be a good compromise between the two options. They take into account the ordering information, but are resistant to the noise in the actual p-values. Available methods, however, have several problems. The classical rank aggregation methods (DeConde et al., 2006; Pihur et al., 2008), do not assign significance to the results, do not take the structurally missing information into account properly and are not robust enough, to be really practical. Other rank aggregation methods (Zintzaras and Ioannidis, 2008; Hong and Breitling, 2008) require full rankings that are often unavailable. The Robust Rank Aggregation (RRA) method that we developed for MEM, however, fits this setting well. On one hand it is robust and can measure significance of the final results, and on the other hand it copes with most of the practical problems.

5.2 Robust Rank Aggregation method - article III

Unlike the classical rank aggregation methods, RRA is based on statistical model. It means that we have described an uninteresting scenario, a null model, and try to re-rank genes based on how much do they deviate from this. In the null model, all the input lists would be random permutations of the same set of genes. That means, if we take one gene and extract its positions from all the input lists, then the distribution of them should be uniform. However, we are interested in genes that preferentially are ranked at the top of the list. Therefore, we have to look for the genes that have more of small ranks than would be expected by the uniform distribution. The difference from uniform distribution can be tested statistically and the test scores can be used for re-ranking the genes.

Algorithm

First, we normalise the ranks in each input list by dividing them with the total number of genes. This converts them to the range from 0 to 1. Now we have to compare for each gene the actual normalised rank distribution with standard uniform distribution. A simple option would be to set an arbitrary threshold, for

example 10%, and use the number of times a gene appeared at the top 10% as a test statistic. Under null hypothesis of uniform distribution this number has a binomial distribution. Knowing this, it is possible to calculate a p-value for every gene and re-rank them according to the p-values.

Of course the 10% cutoff would be rather arbitrary and is not always optimal. For example, in some cases the ranks are most enriched at top 5% and in other at top-20%. Therefore, we can try several cutoffs and select the one with the best p-value. In RRA we use each individual rank in the rank vector as a cutoff, calculate a p-value and report the smallest of them. After using Bonferroni correction we can use this score as a p-value.

An illustration of this algorithm is shown in Figure 5.1. There are examples with two genes, one with many ranks at the top and other with more uniform distribution. The panel B shows how the p-values change with different cutoffs and where they reach the minimum. In case of the first gene, we can see that the p-value drops rapidly, since there are many ranks close to 0. For the second gene none of the p-values get too small since the ranks are distributed more or less according to the null (uniform) distribution.

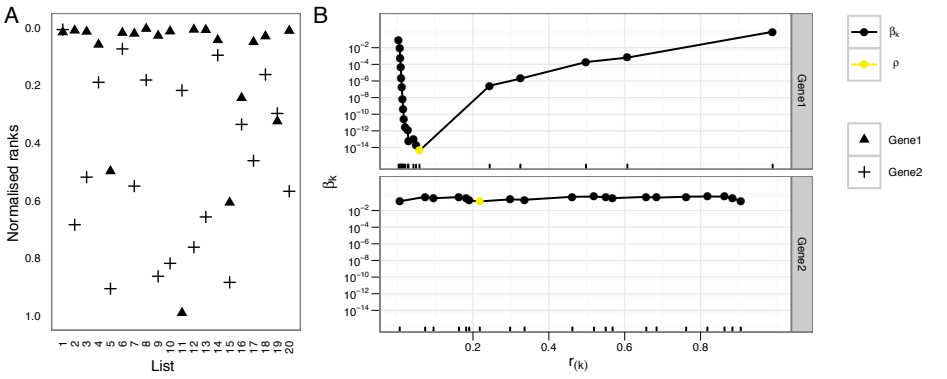

Figure 5.1: Illustraton of RRA principle on two genes with different rank distribution. Panel A shows the distribution of the ranks for two genes in 20 input rankings. Panel B, shows the p-values for each of the ranks. Yellow dot marks the minimal value that is used as the final score for the gene.

The RRA method satisfies most of the requirements for gene expression meta-analysis that were listed above. First it is robust by design. The scores depend only on the best ranks and do not take into account the worse ones. Thus, adding a totally random list to the inputs would not change the results much. The robustness was confirmed by the simulations and case studies in the article.

Second, RRA assigns a p-value to every gene. This p-value shows if a gene is ranked more often at the top than expected by the null model. Having such scores

is important in practice, since it gives an indication of how many genes at the top of the aggregated list we can trust.

Missing data

Often we have only the top rankings or only the significant results available. For RRA this is not a big problem, since it takes into account only the best ranks anyway. Therefore, we can replace the unknown ranks with the worst possible rank. Simulations show that this schema is efficient. There is no big difference in the number of significant genes if we have either full rankings available or only the top lists.

The RRA framework can also handle properly the structurally missing information, *i.e.* situations where some genes are measured in only a subset of input lists. This requires small adjustments to the algorithm. When the number of genes measured by different platforms has large differences, the comparison of raw ranks is not entirely fair. However, RRA uses normalised ranks and taking into account the total number of genes measured by a certain platform. When calculating the p-values it is possible to leave out all the lists where a gene was not measured. As a result we get a valid p-value that measures the statistical evidence of a gene being ranked more preferentially at the top. As such, it is correct to compare these p-values, even if they are based on wildly differing numbers of rankings.

Example: meta-analysis of miRNAs in cancer

Simulations and case studies showed that in terms of pure rank aggregation performance it is comparable or slightly better than alternative methods. However, its real value lies in its versatility and ability to handle the most common practical problems. Here we take a look at a meta-analysis study that employed the RRA algorithm and where many of its features proved to be critical.

MicroRNAs (miRNAs) are interesting class of small RNA molecules that are not translated into proteins, but are involved in the regulation of the messenger RNA translation. The role of miRNAs in cancer is reported in many articles, concentrating mainly on the miRNA expression differences between cancerous and normal tissue. In every study the results are somewhat different, due to differences in patient cohorts, sample collection and analysis methodologies. Urmo Võsa and Dr Tarmo Annilo from the Institute of Molecular and Cell Biology at University of Tartu decided to summarise the already published results in a meta-analysis (Võsa et al., 2013).

More specifically, they identified 20 articles where miRNA expression in non-small lung cancer was compared to normal samples and tried to create a meta

signature of up- and down-regulated miRNAs. Since in many cases the raw data was not available, they had to resort to aggregating published gene lists. When analysing these gene lists, they faced many of the problems that were described above.

For example, the number of samples in these studies varied wildly. Seven studies relied on less than ten samples, whereas two studies had more than hundred samples. It shows that some of the inputs might have been of much poorer quality than the others. Therefore, the robustness of the approach was important. In many cases, the raw data and full rankings were not available, so one had to use the top lists of miRNAs published in the original articles. Different number of miRNAs measured by different platforms imposed a bigger problem. Earlier arrays measured only as little as 228 different miRNA while later ones measured almost four times as many. Therefore, for some miRNAs much more evidence was available than for the others and the ability to handle such situation properly was critical.

Using the RRA method it was possible to identify a meta-signature of 15 significant miRNAs: 7 up- and 8 down-regulated. Downstream analysis showed that the targets of the identified miRNAs were enriched for signalling and cancer related genes. It was important that the RRA assigns significance to every miRNA, since otherwise it would have been hard to know how many of them were actually relevant.

Summary and impact

In RRA we have created a novel rank aggregation methodology that is particularly well suited for genomic setting due to its robustness and versatility. It has already been used in several studies showing the wide array of potential use-cases. As described above, it has been employed in differential expression meta-analysis projects (Võsa et al., 2013; Ma et al., 2013; Frampton et al., 2014), but also on results from siRNA screens (Widau et al., 2014) and significantly enriched pathways (Sun et al., 2013).

Contribution

In large part the algorithm was designed by me. Co-authors contributed with proper mathematical treatment of missing data and also helped with case-studies for the article.

CHAPTER 6

GOSUMMARIES PACKAGE FOR VISUALISING GENOMIC ANALYSIS RESULTS

While working with multiple public datasets it is helpful to know as much as possible about each dataset. Being aware of potential problems, influential factors and dominant expression patterns helps to formulate more appropriate research questions and select optimal analysis strategy. Thorough analysis of each dataset can be a time consuming undertaking. Running a typical analysis pipeline as depicted in Figure 2.1 in Chapter 2 can be relatively fast. However, it is not easy to visualise the results in concise manner and, thus, the study and comparison of datasets can take a lot of time.

From all the steps in the analysis pipeline, visualising the results of functional analysis is the most complicated. The long tables of functional category names do not lend themselves to compact visual representations. At the same time, functional analysis is a critical part of any gene expression analysis pipeline, as most of the analyses converge on this step (see Figure 2.1). Thus, improvements for visualising functional enrichment analysis results can improve the interpretability of of results from several different analysis methods.

This chapter presents a visualisation method GOsummaries that allows to create easily readable visual summaries of functional enrichment analysis results. But before introducing the method itself, next section gives an overview about the functional enrichment analysis and existing visualisation methods.

6.1 Gene Ontology enrichment analysis and visualisation

Functional enrichment analysis of gene lists is usually performed using data from Gene Ontology (GO) project (Gene Ontology Consortium, 2001). This is a curated resource, where genes are associated with functional terms. Gene Ontology vocabulary of functional terms is defined in a tree-like structure . The general terms, like "Growth", "Signalling" or "Metabolic process", are in the top and specific, like "heart capillary growth" or "neuronal signal transduction" in the bottom of the concept tree (acyclic directed graph). Its main difference from a tree is that every term can have multiple parents. Genes are associated with the functional terms by special teams of curators who extract up-to-date information from publications and computer algorithms that extrapolate existing associations. The tree-like structure is leveraged, by propagating associations from children to parent terms.

Typical use for GO is searching for enriched GO terms in a list of genes. First, the number of overlapping genes is found between the gene lists of interest and every GO term. The significance of the overlaps is usually assessed using Fisher exact test. The result will be a list of significantly enriched GO terms with the corresponding p-values. Such analysis can be performed using web-based tools, like g:Profiler (Reimand et al., 2011) or DAVID (Dennis et al., 2003). If the gene lists of interest is ranked, for example a list of differentially expressed genes, one can use also the Gene Set Analysis (GSA) analysis algorithms, like GSEA (Subramanian et al., 2005) or GAGE (Luo et al., 2009). These use rank based statistical tests to find GO terms with genes preferentially in the top of the ranked list of interest. However, the result looks the same: a subset of significant GO terms with p-values.

As the GO vocabulary contains many related terms that share the large proportion of the genes, the number of results is usually large, but rather redundant. In case of one or two gene lists the interpretation of enrichment results is rather straightforward. However, typical gene expression analysis workflow (Figure 2.1) usually yields more than one or two lists. For example, in differential expression the up- and down-regulated genes in every comparison are viewed separately and there are usually more than one comparison done. Additionally, the results of GO analysis cannot be interpreted without considering them in a wider context of experimental lists p and biological background of the gene lists. Thus, good visualisation methods are needed, to summarise GO enrichment analysis results effectively, to allow comparison between the lists and to put the results back into their biological context.

Inherently, the results, a collection of terms with p-values, are not easy to vi-

sualise. Still, there are some methods available. Several tools use the hierarchical structure of GO to visualise the results: g:Profiler (Reimand et al., 2011) uses this to group the significant results, GOrilla (Eden et al., 2009) overlays the GO graph with enrichment scores, Enrichment Map (Merico et al., 2010) visualises the results as a network and REVIGO (Supek et al., 2011) displays significant categories as a treemap.

All these tools concentrate on visualising GO enrichment results describing one gene list. They become less effective if the number of gene lists grows. One promising approach is representing the data as word clouds. Using word cloud it is possible to create a compact view of textual expressions, by highlighting the most important ones in size and colour. The word clouds can be easily arranged together in one figure to give overview of GO annotations of multiple gene lists and it is possible to also add the graphs of experimental data. Several tools offer the word cloud visualisation feature such as GeneCodis3 (Tabas-Madrid et al., 2012), REVIGO (Supek et al., 2011) and Cytoscape word cloud plugin (Oesper et al., 2011). But in all the cases the visualisation concentrates on one gene list and does not fully exploit the option to combine several word clouds together.

For that reason, we created an R package GOsummaries that can create word clouds from GO enrichment analysis results and arrange them together to summarise results across several gene lists. To ease the interpretation in the biological context we can also include the figures describing the experimental data behind the gene lists. As such, GOsummaries can be used to visualise the results of several gene analysis methods, like k-means clustering or Principal Component Analysis. The latter is a popular approach to visualise microarray data and we provide more details about how it works and how its results can be interpreted in the following section.

6.2 Principal Component Analysis

Let us have a $n \times m$ data matrix X. Principal Component Analysis (PCA) is an orthogonal linear transformation of the data into a new coordinate system. The first coordinate, or principal component as it is called, is chosen to maximise the variance of the projection of the data. Mathematically, the projection is done by multiplying X with an m element unit vector. Therefore, the first component is defined by a vector

$$\mathbf{w}_{(1)} = \arg \max_{||w||=1} Var(Xw).$$

The next components are defined in a similar manner, but with additional constraint that the new weight vector was orthogonal to the previous ones. Vector $\mathbf{w}_{(k)}$ is called a weight or loadings vector, as each of its elements shows how

much a feature in the original data contributes to the principal component. The number of principal components is equal to $\min(m, n)$.

The euclidean distance between two data points that are projected to the principal components will be the same as in the original dataset. Usually, in real data first few components describe most of the variation, thus, the distance between data points in the space of first two or three principal components can be used as an approximation of their actual distance. Therefore, it is possible to use PCA to reduce the dimensionality of the data without losing too much of the original information.

This property makes PCA extremely useful in many situations. In gene expression analysis it is common to use PCA to visualise the relationships between the samples. The PCA is applied to transposed gene expression matrix, where rows represent biological samples and columns represent genes. If we drew the scatterplot of the samples on the first two principal components, then we could immediately see how the samples cluster, or if there are some outliers, etc. If these principal components explain large enough proportion of variation then the structure seen on the plot would be rather close to the actual clustering pattern of the original data.

Another question is how to interpret the meaning of the components. Standard approach is to study the loadings vector $\mathbf{w}_{(k)}$ to see what are the features that have the largest positive and negative loadings, as these have the most profound impact on the observation distribution on the principal component. This approach has been used, for example, to interpret and name the "Big 5" of independent personality traits in psychology (Goldberg, 1990). In the gene expression applications, however, it is not so common to interpret the loadings of a principal component. One reason might be that the loading vectors might be just too long for reasonable interpretation. However, in this case it is possible to perform GO enrichment analysis on the genes that have the loadings with largest impact. This is exactly what GOsummaries does.

6.3 GOsummaries package - article IV

Construction of word clouds

The most important input for GOsummaries is a group of gene lists. These are annotated functionally using g:Profiler web tool (Reimand et al., 2011). To filter the results GOsummaries uses several additional parameters. It removes both too generic and too specific terms from the results, by constraining the size of GO terms. To remove redundant terms GOsummaries uses the hierarchical filtering option by g:Profiler. It overlays the enriched terms on GO graph and from each connected component selects the category with strongest enrichment.

The remaining GO terms with p-values are visualised as word clouds. The size of the words is proportional to the $-log_{10}$ of p-values. The sizes are not comparable between word clouds of the same plot, since GOsummaries tries to use the available space as efficiently as possible. However, the absolute scale of p-values is expressed as the colour of the words.

Layout of a GOsummaries plot

For displaying the word clouds GOsummaries has defined a special structure of a plot. GOsummaries shows the word clouds together with information describing the underlying gene list, like the name and size of it. Most importantly it is also possible to add plots describing the gene list. For example, GOsummaries can display the expression values of the underlying genes next to the word clouds. By displaying the expression patterns together with functional annotations we can create concise summaries of the common analysis pipelines. In the package we have defined GOsummaries plots for three different analysis methods: differential expression, clustering and PCA.

In general, the figure consists of blocks that represent either one or two closely related gene lists, such as a cluster or up- and down-regulated genes from a differential expression analysis. Each block contains the word cloud(s), name, size and plot with background information about the list. On Figure 6.1 one can see three blocks, one for every analysis type. In principle, the graph slot can contain any type of plot and users can define their own. However, currently GOsummaries implements few options that depend on the underlying data. If there is no additional data besides the gene list, it shows the number of genes in the list as a bars. If there is also expression data available, it displays the expression of the genes as boxplot (see Figures 6.1A and 6.1B). In these figures each box represents one sample and they show the distribution of expression values on y-axis. By adding the expression values it is possible to immediately relate the GO annotations to the actual gene expression patterns, which is important, for example, in case of clustering.

Principal Component Analysis

For clustering and differential expression, the GOsummaries provides just a concise representation of the results. It makes the comparison of annotations easier, but does not provide qualitatively novel insights. With PCA our approach is a bit more unconventional. Usually, PCA of gene expression data is visualised as a 2D scatterplot of samples projected onto the first 2 principal components. This visualisation allows to study the global similarity and dissimilarity of the samples. It is possible to see if there are any outliers or whether the samples with similar

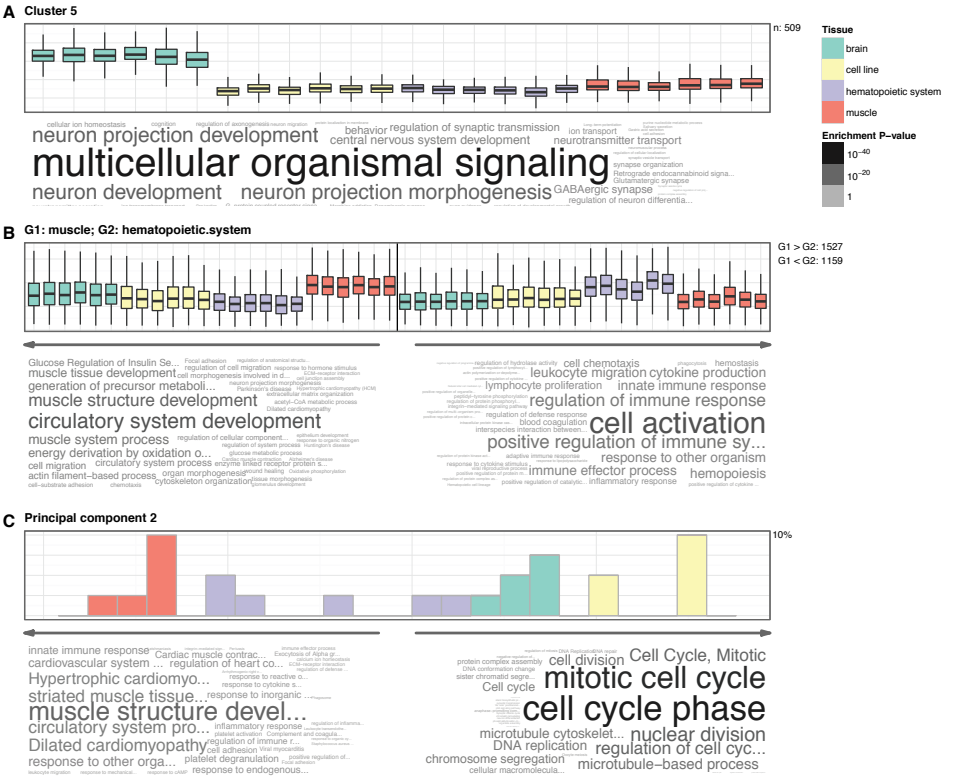

Figure 6.1: An example of GOsummaries output using expression data from 4 tissues. The panels correspond to different input types: k-means clustering (A), differential expression (B) and PCA (C). The plots above the word clouds display the underlying experimental data. For clustering (A) and differential expression (B), we show the distribution of expression values as boxplots, where each box corresponds to one sample. In case of PCA (C) the histogram displays the projection of samples to a principal component.

annotation group together. But this is basically all that we can conclude from such plots.

The GOsummaries visualisation of PCA is different in several aspects and can be considered complementary to the usual 2D display. Each block shows one component. The panel displays the distribution of samples on each principal component separately as a histogram, that is coloured based on sample annotations. The GO annotations are based on 500 genes with largest positive and negative loadings. Example of this visualisation can be seen in Figure 6.1C. Figure 6.2 shows how the GOsummaries representation of PCA is related to the ordinary depiction of PCA results.

When 2D scatterplot tries to show the clustering structure of the samples, then

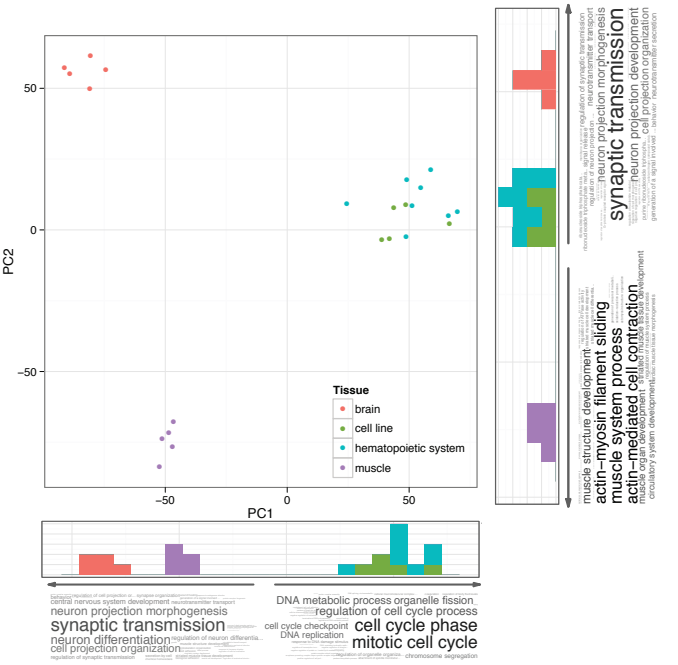

Figure 6.2: Ordinary 2D representation of PCA and corresponding components as displayed by GOsummaries.

GOsummaries visualisation helps to dive into and interpret the meaning of every single component. The histogram display shows the biological groups that are distinguished by the component. The word cloud identifies the functional categories of genes that either drive the samples right or left on the principal axis. This allows us to say, for example based on Figure 6.1C, that cell lines and muscle cells differ mostly by higher expression of cell cycle genes in cell lines and higher expression of muscle structure development related genes in muscle cells. This type of annotation of the principal components allows to reveal the underlying biological processes responsible for the structure of the clusters, that would remain unnoticed in a typical PCA plot analysis.

Another advantage of GOsummaries display is that it makes easy to study more than two or three principal components. In more complex datasets, components beyond the first few can reveal interesting factors influencing the dataset. With GOsummaries it is possible to draw a plot with even 10 or 20 components and to check quickly if there is anything relevant.

Summary

The GOsummaries package improves the presentation of gene expression analysis results. It combines the patterns that were found from data with their functional annotation into easy-to-read figures. This approach can considerably speed up the process of interpretation of the results, generation of new hypotheses and identification of the inconsistencies in the data. Although the package is best suited for gene expression data analysis, the methods apply nicely to many other scenarios where the output of the analysis can be represented as a list of genes.

Contribution

The method was designed and implemented by me.

CONCLUSION

The gene expression microarrays allow creating comprehensive snapshots of cellular states, by measuring the expression of all genes in parallel. As these datasets have accumulated in databases, there is possibility to re-use this information to test novel hypotheses. Taking advantage of the public data is mostly a bioinformatic challenge. It involves developing more appropriate statistical methods, insightful ways to view the data and user friendly software that makes the data and the analysis methods more accessible. The work presented in this thesis covers all these aspects, with a general aim to make re-use of gene expression data more accessible, effective and insightful.

The most significant methodological contribution here is the Robust Rank Aggregation (RRA) algorithm for integrating ranked lists of genes. It is common to have multiple experiments or data sources that describe the same phenomena and there is a need to aggregate the results. The integration is often done with *ad hoc* methods. RRA is an algorithm for the integration task that has a solid statistical background and takes into account common problems with gene lists. The method was designed for MEM web server, but has been subsequently used in several other studies, most notably, for differential expression meta-analysis.

Other projects have also methodological components, but they are more closely related to visualisation. In FunGenES database we combined two common clustering techniques: k-means and hierarchical clustering into one approach that allows to create high level overviews of large datasets. With GOsummaries package it is possible to summarise complex analysis pipelines in concise and easily readable figures. This allows to rapidly study and compare the biological content of multiple datasets. Most important innovation in GOsummaries is the visualisation of Principal Component Analysis (PCA) results. PCA is used ubiquitously for analysing all types of high-throughput data, but information the typical results convey is relatively limited. By associating functional annotations to the principal components, GOsummaries can make the PCA analysis results much more revealing and insightful.

Finally, much of the focus in this work has been on creating user friendly tools that would make analysis methods and the public data more accessible. Both RRA

and GOsummaries are implemented as add-on packages to statistical software R, which is the most popular platform for gene expression analysis. In FunGenES database and MEM we did not just implement an algorithm, but built full web-based user interfaces for accessing and analysing the data. The tools developed for FunGenES formed a seed for a suite of web-based data analysis tools that is still in use and being developed further in our working group. In MEM we created a unique resource that allows to search information interactively from thousands of datasets and perform an analysis that would be in most cases practically infeasible.

Bibliography

P. Adler, J. Reimand, J. Jänes, R. Kolde, H. Peterson, and J. Vilo. KEGGanim: pathway animations for high-throughput data. *Bioinformatics*, 24(4):588–590, Feb. 2008.

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

S. Aerts, D. Lambrechts, S. Maity, and P. Van Loo. Gene prioritization through genomic data fusion. *Nature Biotechnology*, 2006.

D. Amaratunga and J. Cabrera. Analysis of data from viral DNA microchips. *Journal of the American Statistical Association*, 96(456):1161–1170, 2001.

P. Baldi and A. D. Long. A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. *Bioinformatics*, 17(6):509–519, May 2001.

T. Barrett, D. B. Troup, S. E. Wilhite, P. Ledoux, D. Rudnev, C. Evangelista, I. F. Kim, A. Soboleva, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, R. N. Muertter, and R. Edgar. NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Research*, 37(Database issue):D885–90, 2009.

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

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

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

R. Breitling, P. Armengaud, A. Amtmann, and P. Herzyk. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Letters*, 573(1-3):83–92, Aug. 2004.

S. K. Chan, O. L. Griffith, I. T. Tai, and S. J. M. Jones. Meta-analysis of colorectal cancer gene expression profiling studies identifies consistently reported candidate biomarkers. *Cancer Epidemiology, Biomarkers & Prevention*, 17(3): 543–552, Mar. 2008.

D. Chen, P. Bhat-Nakshatri, C. Goswami, S. Badve, and H. Nakshatri. ANTXR1, a stem cell-enriched functional biomarker, connects collagen signaling to cancer stem-like cells and metastasis in breast cancer. *Cancer Research*, 73(18): 5821–5833, Sept. 2013.

Y. Chen, M. Jørgensen, R. Kolde, X. Zhao, B. Parker, E. Valen, J. Wen, and A. Sandelin. Prediction of RNA Polymerase II recruitment, elongation and stalling from histone modification data. *BMC Genomics*, 12(1):544, 2011.

J. K. Choi, U. Yu, S. Kim, and O. J. Yoo. Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics*, 19 Suppl 1:i84–90, 2003.

E. P. Consortium, B. E. Bernstein, E. Birney, I. Dunham, E. D. Green, C. Gunter, and M. Snyder. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, Sept. 2012.

M. R. Davis and K. M. Summers. Molecular Genetics and Metabolism. *Molecular Genetics and Metabolism*, 107(4):635–647, Dec. 2012.

R. P. DeConde, S. Hawley, S. Falcon, N. Clegg, B. Knudsen, and R. Etzioni. Combining results of microarray experiments: a rank aggregation approach. *Statistical Applications in Genetics and Molecular Biology*, 5:Article15, 2006.

G. Dennis, B. T. Sherman, D. A. Hosack, J. Yang, W. Gao, H. C. Lane, and R. A. Lempicki. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biology*, 4(5):P3, 2003.

M. X. Doss, V. Wagh, H. Schulz, M. Kull, R. Kolde, K. Pfannkuche, T. Nolden, H. Himmelbauer, J. Vilo, J. Hescheler, and A. Sachinidis. Global transcriptomic analysis of murine embryonic stem cell-derived brachyury (T) cells. *Genes to Cells*, 15(3):209–228, Feb. 2010.

E. Eden, R. Navon, I. Steinfeld, D. Lipson, and Z. Yakhini. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics*, 10:48, 2009.

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

A. E. Frampton, E. Giovannetti, N. B. Jamieson, J. Krell, T. M. Gall, J. Stebbing, L. R. Jiao, and L. Castellano. A microRNA meta-signature for pancreatic ductal adenocarcinoma. *Expert Review of Molecular Diagnostics*, pages 1–5, Feb. 2014.

J. A. Gaspar, M. X. Doss, J. Winkler, V. Wagh, J. Hescheler, R. Kolde, J. Vilo, H. Schulz, and A. Sachinidis. Gene expression signatures defining fundamental biological processes in pluripotent, early, and late differentiated embryonic stem cells. *Stem Cells and Development*, 21(13):2471–2484, Sept. 2012.

Gene Ontology Consortium. Creating the gene ontology resource: design and implementation. *Genome Research*, 11(8):1425–1433, Aug. 2001.

L. R. Goldberg. An alternative "description of personality": the big-five factor structure. *Journal of Personality and Social Psychology*, 59(6):1216–1229, Dec. 1990.

O. L. Griffith, A. Melck, S. J. M. Jones, and S. M. Wiseman. Meta-analysis and meta-review of thyroid cancer gene expression profiling studies identifies important diagnostic biomarkers. *Journal of Clinical Oncology*, 24(31):5043–5051, Nov. 2006.

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

F. Hong and R. Breitling. A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments. *Bioinformatics*, 24 (3):374–382, Jan. 2008.

W. Huber, A. von Heydebreck, H. Sültmann, A. Poustka, and M. Vingron. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18 Suppl 1:S96–104, 2002.

C. Huttenhower, E. M. Haley, M. A. Hibbs, V. Dumeaux, D. R. Barrett, H. A. Coller, and O. G. Troyanskaya. Exploring the human genome with functional maps. *Genome Research*, 19(6):1093–1106, June 2009.

S. Ilmjärv, C. A. Hundahl, R. Reimets, M. Niitsoo, R. Kolde, J. Vilo, E. Vasar, and H. Luuk. Estimating differential expression from multiple indicators. *Nucleic Acids Research*, page gku158, Feb. 2014.

R. A. Irizarry, B. M. Bolstad, F. Collin, L. M. Cope, B. Hobbs, and T. P. Speed. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Research*, 31(4):e15, Feb. 2003.

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

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

P. Kemmeren, N. L. van Berkum, J. Vilo, T. Bijma, R. Donders, A. Brazma, and F. C. P. Holstege. Protein interaction verification and functional annotation by integrated analysis of genome-scale data. *Molecular Cell*, 9(5):1133–1143, June 2002.

R. Kolde and J. Vilo. GOsummaries: an R Package for Visual Functional Annotation of Experimental Data. in preparation.

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

A. K. Krug, R. Kolde, J. A. Gaspar, E. Rempel, N. V. Balmer, K. Meganathan, K. Vojnits, M. Baquié, T. Waldmann, R. Ensenat-Waser, S. Jagtap, R. M. Evans, S. Julien, H. Peterson, D. Zagoura, S. Kadereit, D. Gerhard, I. Sotiriadou, M. Heke, K. Natarajan, M. Henry, J. Winkler, R. Marchan, L. Stoppini, S. Bosgra, J. Westerhout, M. Verwei, J. Vilo, A. Kortenkamp, J. Hescheler, L. Hothorn, S. Bremer, C. van Thriel, K.-H. Krause, J. G. Hengstler, J. Rahnenführer, M. Leist, and A. Sachinidis. Human embryonic stem cell-derived test

systems for developmental neurotoxicity: a transcriptomics approach. *Archives of Toxicology*, 87(1):123–143, Jan. 2013.

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

E. Lacunza, M. E. Rabassa, R. Canzoneri, M. Pellon-Maison, M. V. Croce, C. M. Aldaz, and M. C. Abba. Identification of signaling pathways modulated by RHBDD2 in breast cancer cells: a link to the unfolded protein response. *Cell stress & chaperones*, pages 1–10, Sept. 2013.

K. Lokk, T. Vooder, R. Kolde, K. Välk, U. Võsa, R. Roosipuu, L. Milani, K. Fischer, M. Koltsina, E. Urgard, T. Annilo, A. Metspalu, and N. Tõnisson. Methylation markers of early-stage non-small cell lung cancer. *PLoS ONE*, 7(6): e39813, 2012.

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

W. Luo, M. S. Friedman, K. Shedden, K. D. Hankenson, and P. J. Woolf. GAGE: generally applicable gene set enrichment for pathway analysis. *BMC Bioinformatics*, 10(1):161, 2009.

M.-Z. Ma, X. Kong, M.-Z. Weng, K. Cheng, W. Gong, Z.-W. Quan, and C.-H. Peng. Candidate microRNA biomarkers of pancreatic ductal adenocarcinoma: meta-analysis, experimental validation and clinical significance. *Journal of Experimental & Clinical Cancer Research*, 32(1):1–1, Sept. 2013.

E. Maron, K. Kallassalu, A. Tammiste, R. Kolde, J. Vilo, I. Tõru, V. Vasar, J. Shlik, and A. Metspalu. Peripheral gene expression profiling of CCK-4-induced panic in healthy subjects. *American Journal of Medical Genetics. Part B*, 153B(1): 269–274, Jan. 2010.

D. Merico, R. Isserlin, O. Stueker, A. Emili, and G. D. Bader. Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. *PLoS ONE*, 5(11):e13984, 2010.

L. Oesper, D. Merico, R. Isserlin, and G. D. Bader. WordCloud: a Cytoscape plugin to create a visual semantic summary of networks. *Source Code for Biology and Medicine*, 6:7, 2011.

H. Parkinson, M. Kapushesky, N. Kolesnikov, G. Rustici, M. Shojatalab, N. Abeygunawardena, H. Berube, M. Dylag, I. Emam, A. Farne, E. Holloway, M. Lukk,

J. Malone, R. Mani, E. Pilicheva, T. F. Rayner, F. Rezwan, A. Sharma, E. Williams, X. Z. Bradley, T. Adamusiak, M. Brandizi, T. Burdett, R. Coulson, M. Krestyaninova, P. Kurnosov, E. Maguire, S. G. Neogi, P. Rocca-Serra, S.-A. Sansone, N. Sklyar, M. Zhao, U. Sarkans, and A. Brazma. ArrayExpress update–from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Research*, 37(Database issue):D868–72, Jan. 2009.

V. Pihur, S. Datta, and S. Datta. Finding common genes in multiple cancer types through meta-analysis of microarray experiments: a rank aggregation approach. *Genomics*, 92(6):400–403, Dec. 2008.

A. Ramasamy, A. Mondry, C. Holmes, and D. Altman. Key Issues in Conducting a Meta-Analysis of Gene Expression Microarray Datasets. *PLoS Medicine*, 5 (9):e184, Sept. 2008.

J. Reimand, T. Arak, and J. Vilo. g:Profiler–a web server for functional interpretation of gene lists (2011 update). *Nucleic Acids Research*, 39(Web Server issue): W307–15, July 2011.

D. R. Rhodes, T. R. Barrette, M. A. Rubin, D. Ghosh, and A. M. Chinnaiyan. Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. *Cancer Research*, 62(15): 4427–4433, Aug. 2002.

D. R. Rhodes, J. Yu, K. Shanker, N. Deshpande, R. Varambally, D. Ghosh, T. Barrette, A. Pandey, and A. M. Chinnaiyan. Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proceedings of the National Academy of Sciences of the United States of America*, 101(25):9309–9314, June 2004a.

D. R. Rhodes, J. Yu, K. Shanker, N. Deshpande, R. Varambally, D. Ghosh, T. Barrette, A. Pandey, and A. M. Chinnaiyan. ONCOMINE: a cancer microarray database and integrated data-mining platform. *Neoplasia*, 6(1):1–6, 2004b.

D. R. Rhodes, S. Kalyana-Sundaram, V. Mahavisno, R. Varambally, J. Yu, B. B. Briggs, T. R. Barrette, M. J. Anstet, C. Kincead-Beal, P. Kulkarni, S. Varambally, D. Ghosh, and A. M. Chinnaiyan. Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. *Neoplasia*, 9(2):166–180, Feb. 2007.

J. Rung and A. Brazma. Reuse of public genome-wide gene expression data. *Nature Reviews Genetics*, 14(2):89–99, Feb. 2013.

G. Rustici, N. Kolesnikov, M. Brandizi, T. Burdett, M. Dylag, I. Emam, A. Farne, E. Hastings, J. Ison, M. Keays, N. Kurbatova, J. Malone, R. Mani, A. Mupo, R. Pedro Pereira, E. Pilicheva, J. Rung, A. Sharma, Y. A. Tang, T. Ternent, A. Tikhonov, D. Welter, E. Williams, A. Brazma, H. Parkinson, and U. Sarkans. ArrayExpress update–trends in database growth and links to data analysis tools. *Nucleic Acids Research*, 41(Database issue):D987–90, Jan. 2013.

R. Sandie, G. A. Palidwor, M. R. Huska, C. J. Porter, P. M. Krzyzanowski, E. M. Muro, C. Perez-Iratxeta, and M. A. Andrade-Navarro. Recent developments in StemBase: a tool to study gene expression in human and murine stem cells. *BMC Research Notes*, 2(1):39, 2009.

M. Sardiello, M. Palmieri, A. di Ronza, D. L. Medina, M. Valenza, V. A. Gennarino, C. Di Malta, F. Donaudy, V. Embrione, R. S. Polishchuk, S. Banfi, G. Parenti, E. Cattaneo, and A. Ballabio. A gene network regulating lysosomal biogenesis and function. *Science*, 325(5939):473–477, July 2009.

A. Schraenen, G. de Faudeur, L. Thorrez, K. Lemaire, G. Van Wichelen, M. Granvik, L. Van Lommel, P. in't Veld, and F. Schuit. mRNA expression analysis of cell cycle genes in islets of pregnant mice. *Diabetologia*, 53(12): 2579–2588, Dec. 2010.

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

S. K. Singh, B. L. Veo, M. N. Kagalwala, W. Shi, S. Liang, and S. Majumder. Dynamic Status of REST in the Mouse ESC Pluripotency Network. *PLoS ONE*, 7(8):e43659, Aug. 2012.

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

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

T. Sørlie, C. M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, T. Thorsen, H. Quist, J. C. Matese, P. O. Brown, D. Botstein, P. E. Lønning, and A. L. Børresen-Dale. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences of the United States of America*, 98(19):10869–10874, Sept. 2001.

M. P. Storm, B. Kumpfmueller, B. Thompson, R. Kolde, J. Vilo, O. Hummel, H. Schulz, and M. J. Welham. Characterization of the phosphoinositide 3-kinase-dependent transcriptome in murine embryonic stem cells: identification of novel regulators of pluripotency. *Stem Cells*, 27(4):764–775, Apr. 2009.

A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550, Oct. 2005.

H. Sun, H. Wang, R. Zhu, K. Tang, Q. Gong, J. Cui, Z. Cao, and Q. Liu. iPEAP: integrating multiple omics and genetic data for pathway enrichment analysis. *Bioinformatics*, 30(5):737–739, Oct. 2013.

F. Supek, M. Bošnjak, N. Škunca, and T. Šmuc. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS ONE*, 6(7):e21800, 2011.

D. Szklarczyk, A. Franceschini, M. Kuhn, M. Simonovic, A. Roth, P. Minguez, T. Doerks, M. Stark, J. Muller, P. Bork, L. J. Jensen, and C. von Mering. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Research*, 39(Database issue):D561–8, Jan. 2011.

Y. Tabach, T. Golan, A. Hernández-Hernández, A. R. Messer, T. Fukuda, A. Kouznetsova, J.-G. Liu, I. Lilienthal, C. Levy, and G. Ruvkun. Human disease locus discovery and mapping to molecular pathways through phylogenetic profiling. *Molecular Systems Biology*, 9:692, 2013.

D. Tabas-Madrid, R. Nogales-Cadenas, and A. Pascual-Montano. GeneCodis3: a non-redundant and modular enrichment analysis tool for functional genomics. *Nucleic Acids Research*, 40(Web Server issue):W478–83, July 2012.

V. Tanwar, J. B. Bylund, J. Hu, J. Yan, J. M. Walthall, A. Mukherjee, W. H. Heaton, W.-D. Wang, F. Potet, M. Rai, S. Kupershmidt, E. W. Knapik, and A. K. Hatzopoulos. Gremlin 2 promotes differentiation of embryonic stem cells to atrial fate by activation of the JNK signaling pathway. *Stem Cells*, pages n/a–n/a, Mar. 2014.

M. Trouillas, C. Saucourt, B. Guillotin, X. Gauthereau, L. Ding, F. Buchholz, M. X. Doss, A. Sachinidis, J. Hescheler, O. Hummel, N. Huebner, R. Kolde, J. Vilo, H. Schulz, and H. Boeuf. Three LIF-dependent signatures and gene clusters with atypical expression profiles, identified by transcriptome studies in mouse ES cells and early derivatives. *BMC Genomics*, 10(1):73, 2009.

L. Tserel, R. Kolde, A. Rebane, K. Kisand, T. Org, H. Peterson, J. Vilo, and P. Peterson. Genome-wide promoter analysis of histone modifications in human monocyte-derived antigen presenting cells. *BMC Genomics*, 11(1):642, 2010.

L. Tserel, T. Runnel, K. Kisand, M. Pihlap, L. Bakhoff, R. Kolde, H. Peterson, J. Vilo, P. Peterson, and A. Rebane. MicroRNA expression profiles of human blood monocyte-derived dendritic cells and macrophages reveal miR-511 as putative positive regulator of Toll-like receptor 4. *The Journal of Biological Chemistry*, 286(30):26487–26495, July 2011.

K. Välk, T. Vooder, R. Kolde, M.-A. Reintam, C. Petzold, J. Vilo, and A. Metspalu. Gene expression profiles of non-small cell lung cancer: survival prediction and new biomarkers. *Oncology*, 79(3-4):283–292, 2010.

T. Vooder, K. Välk, R. Kolde, R. Roosipuu, J. Vilo, and A. Metspalu. Gene Expression-Based Approaches in Differentiation of Metastases and Second Primary Tumour. *Case Reports in Oncology*, 3(2):255–261, 2010.

U. Võsa, T. Vooder, R. Kolde, K. Fischer, K. Välk, N. Tõnisson, R. Roosipuu, J. Vilo, A. Metspalu, and T. Annilo. Identification of miR-374a as a prognostic marker for survival in patients with early-stage nonsmall cell lung cancer. *Genes, Chromosomes & Cancer*, 50(10):812–822, Oct. 2011.

U. Võsa, T. Vooder, R. Kolde, J. Vilo, A. Metspalu, and T. Annilo. Meta-analysis of microRNA expression in lung cancer. *International Journal of Cancer*, 132 (12):2884–2893, June 2013.

T. Waldmann, E. Rempel, N. V. Balmer, A. König, R. Kolde, J. A. Gaspar, M. Henry, J. Hescheler, A. Sachinidis, J. Rahnenführer, J. G. Hengstler, and M. Leist. Design principles of concentration-dependent transcriptome deviations in drug-exposed differentiating stem cells. *Chemical Research in Toxicology*, 27(3):408–420, Mar. 2014.

C. A. Wells, R. Mosbergen, O. Korn, J. Choi, N. Seidenman, N. A. Matigian, A. M. Vitale, and J. Shepherd. Stemformatics: visualisation and sharing of stem cell gene expression. *Stem Cell Research*, 10(3):387–395, May 2013.

R. C. Widau, A. D. Parekh, M. C. Ranck, D. W. Golden, K. A. Kumar, R. F. Sood, S. P. Pitroda, Z. Liao, X. Huang, T. E. Darga, D. Xu, L. Huang, J. Andrade, B. Roizman, R. R. Weichselbaum, and N. N. Khodarev. RIG-I-like receptor LGP2 protects tumor cells from ionizing radiation. *Proceedings of the National Academy of Sciences of the United States of America*, 111(4):E484–91, Jan. 2014.

P. Wirapati, C. Sotiriou, S. Kunkel, P. Farmer, S. Pradervand, B. Haibe-Kains, C. Desmedt, M. Ignatiadis, T. Sengstag, F. Schütz, D. R. Goldstein, M. Piccart, and M. Delorenzi. Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures. *Breast Cancer Research*, 10(4):R65, 2008.

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

E. Zintzaras and J. P. A. Ioannidis. Meta-analysis for ranked discovery datasets: theoretical framework and empirical demonstration for microarrays. *Computational Biology and Chemistry*, 32(1):38–46, Feb. 2008.

ACKNOWLEDGEMENTS

First and foremost I would like to thank my wife Anastassia who has always supported me in all my endeavours. This work would not be possible without my supervisor Jaak Vilo who has given me the opportunity to work on interesting problems, trusted me and created a wonderful environment for conducting research. It is hard to do relevant bioinformatic research without good collaborators, therefore, I'm grateful to all my co-authors for involving me in their projects and their patience while working with me. A special thanks goes to Sven Laur, Hedi Peterson, Konstantin Tretyakov, Jüri Reimand, Kersti Jääger and Rein Prank who took their time to read the thesis and gave me valuable feedback. Finally, I would like to thank my co-workers and friends in BIIT working group who have made the journey enjoyable.

The work in this thesis was supported in parts by European Social Fund's Doctoral Studies and Internationalisation Programme DoRa, which is carried out by Foundation Archimedes, Tiger University Program of the Estonian Information Technology Foundation, Estonian Doctoral School of Information and Communication Technology (IKTDK), Centre of Excellence in Computer Science (EXCS), Estonian Science Foundation grant no. 7437 (MEM), target funding project SF0180008s12 and EU Framework 7 projects ESNATS (HEALTH-F5-2008-201619) and DETECTIVE (HEALTH-F5-2010-266838).

KOKKUVÕTE
(SUMMARY IN ESTONIAN)

MEETODID AVALIKE GEENIEKSPRESSIOONI ANDMETE TAASKASUTAMISEKS

Geeniekspressiooni mikrokiipide leiutamine on andnud bioloogidele võimaluse mõõta tuhandete geenide avaldumist paralleelselt. Taolistest eksperimentidest saadud tulemused sisaldavad rohkelt huvitavat informatsiooni, kuid andmete rohkus omakorda esitab mitmeid arvutuslikke ja statistilisi väljakutseid. Seetõttu on paljud statistikud ja arvutiteadlased hakanud nende probleemidega tegelema. Uuritakse, kuidas säilitada ja käidelda andmeid efektiivselt, kuidas töödelda ja normaliseerida ning lõpuks kuidas oleks kõige õigem neid statistiliselt tõlgendada.

Eelmise kümnendi keskpaigaks olid geeniekspressiooni andmete analüüsi peamised sammud paika pandud. Analüüsi tarkvara ja taristu on sellest ajast oluliselt arenenud, kuid peamised meetodid ühe andmestiku analüüsiks on samad. Küll aga on uued arvutuslikud väljakutsed esile kerkinud seoses andmete kogunemisega avalikesse andmebaasidesse. Kuna geeniekspression mõõtmised hõlmavad kogu genoomi, siis mõõdetakse igas eksperimendis ka nende geenide avaldumist, mis otseselt antud kontekstis huvi ei paku. Need väärtused võivad aga huvitavad olla mõnes teises olukorras, kus nende põhjal võib leida vastuseid uutele bioloogilistele küsimustele. Taolist lähenemist on kasutatud paljudes uuringutes, näiteks ennustamaks geenide funktsioone või leidmaks uusi haigustega seotud geene.

Sellegipoolest on neis andmetes palju kasutamata potentsiaali. Andmebaasides on kümneid tuhandeid eksperimente, mis kirjeldavad sadu erinevaid koetüüpe ja haiguseid. Sobivalt esitatuna võivad need andmed aidata katsete planeerimisel, kandidaatgeenide prioritiseerimisel, tulemuste laiemasse konteksti panekul jne. Olgugi, et geeniekspressiooni andmed on vabalt alla laetavad, pole neist vajaliku

informatsiooni eraldamine sugugi lihtne. Iga huvitava informatsiooni killu kohta on kordades rohkem ebahuvitavat ning see, mis informatsioon on parajasti huvitav, sõltub küsimusest mida uuritakse. Seetõttu ei ole olemas ühte meetodit või tööriista, mis lahendaks kõik andmete taaskasutamisega seotud probleemid. Pigem on lahendus töötada välja palju erinevaid lähenemisi, mis näitavad andmeid erinevatest vaatenurkadest.

Antud töö peamiseks eesmärgiks ongi muuta bioloogilisi uuringuid efektiivsemaks läbi olemasolevate andmete taaskasutamise. Väitekiri koosneb viiest peatükist, millest esimene kirjeldab bioloogilist tausta kui ka olulisemaid bioinformaatilisi meetodeid geeniekspressiooni andmete analüüsiks. Järgmised neli aga annavad igaüks ülevaate ühest konkreetsest artiklist.

Üks peamisi takistusi andmete uuestikasutamisel on nende ligipääsetavus. Andmete alla laadimine, eeltöötlus ja analüüs on suhteliselt ajamahukas ettevõtmine, mis nõuab ka oskust kasutada suhteliselt spetsiifilisi statistilisi ja arvutuslikke meetodeid. Paljudel teadlastel, kellel oleks taolise analüüsi tulemustest kasu, sellised oskused puuduvad. Üks viis andmete ligipääsetavust parandada on luua veebikeskkondi, mis võimaldavad lihtsa vaevaga jooksutada avalikel andmetel konkreetseid analüüse. Antud töös on kaks artiklit pühendatud sellele teemale.

Artikkel I kirjeldab veebipõhist analüüsi keskkonda mis võimaldab uurida ühe üleeuroopalise konsortsiumi tekitatud ekspresiooni andmid. See koosnes mitmest tööriistast, mis ühest küljest andsid üldise pildi andmetes toimuvast, kuid samas võimaldasid andmetest väljatulevaid konkreetseid mustreid detailselt kirjeldada.

Artikkel II kirjeldab veebipõhist tööriista MEM, mis keskendub sarnase ekspressiooni mustriga geenide otsimisele üle sadade või isegi tuhandete avalike andmestike. Selleks laadisime me alla ligi pool terabaiti andmeid ja viisime need analüüsiks sobivale kujule. MEM ise on veebiserver, mis võimaldab sellel andmestikul teha interaktiivseid päringuid. Avaldamise ajal oli MEM üks suuremaid, kui mitte suurim, geeni ekspressiooni andmekogu, mida oli võimalik veebis interaktiivselt kaevandada. Kokkuvõttes on MEM unikaalne tööriist, mis võimaldab teha päringuid, mis oleks muude vahenditega liiga töömahukad olemaks praktilised.

Tihti on geeniekspressiooni andmete taaskasutamisel eesmärgiks integreerida olemasolevad andmed mitmetest allikatest, näiteks tulemused mitmetest sarnastest uuringutest või erinevat tüüpi mõõtmised sama bioloogilise fenomeni kohta. Meetodeid andmete integreerimiseks on palju, kuid tihtipeale taandub see analüüs geeni nimekirjade võrdlemise peale. Olgugi, et see ülesanne on väga tavaline, siis olemasolevad metoodikad ei võta hästi arvesse geeni nimekirjade eripärasid. Antud töö tutvustab ühte lähenemist geenide nimekirjade agregeerimiseks.

Artikkel III kirjeldab üldist astakute agregeerimise algoritmi RRA, mis on eriti sobiv just geenide nimekirjade jaoks. RRA kõige olulisem omadus on tema mü-

rakindlus, mis on oluline, sest ülegenoomseid mõõtmisi võivad mõjutada paljud tehnilised faktorid, mis konkreetsete geenide puhul muudavad andmed ebausaldusväärseks. Seetõttu on tähtis, et RRA mõõdab ka tulemuste statistilist olulisust, mis võimaldab aru saada, kas agregeeritavates nimekirjades üldse oli midagi ühist. Lisaks suudab RRA hästi toime tulla ka olukordades, kus geenide nimekiri katab ainult kõige olulisemad tulemused või kus paljude geenide puhul pole isegi mõõtmisi teostatud. Praktikas on mõlemad olukorrad väga tavalised, kuid enamus alternatiivseid meetodeid ei võta neid oma arvutustes arvesse.

Andmete taaskasutamisel tuleb enda küsimusele vastuse leidmiseks läbi vaadata mitmeid andmestikke. Seetõttu on oluline, et neist oleks võimalik kiirelt ülevaade saada, nägemaks kas andmed on kvaliteetsed, mis tüüpi mustreid nad sisaldavad, jne. Siin tulevad appi erinevad andmete visualiseerimise võtted, mis kombineerituna andmekaeve meetoditega võimaldavad esitada suurte andmestike kohta kompaktseid graafilisi ülevaateid. Viimane artikkel antud töös käsitleb just genoomsete andmete visualiseerimise teemat.

Artikkel IV tutvustab andmete visualiseerimise meetodit ja R paketti nimega GOsummaries, mis annab võimaluse näidata koos geenide funktsionaalseid annotatsioone ja ekspressiooni tasemeid. Peamine idee on esitada geenide nimekirja iseloomustavaid bioloogiliste protsesside nimesid kompaktse sõnapilvena. Tavaliselt esitatakse neid tulemusi pikkade tabelitena, mida on väga raske omavahel võrrelda. Sõnapilvi aga on lihtne kokku panna ning saab ka lisada graafikuid, mis kirjeldavad antud geenide käitumist ekspressiooni andmetes. Nii on võimalik kiiresti võrrelda funktsionaalseid annotatsioone erinevate nimekirjade vahel ning seostada neid vastavate bioloogiliste mustritega. Kasutades GOsummaries paketti saab tekitada kokkuvõtlikke graafikuid erinevate geenide ekspressiooni analüüsi meetodite kohta, sest enamusel neist on tulemuseks just nimekirjad geenidest. Taoliseid joonised on võimalik tekitada vaid mõne rea koodiga. Seega annab GOsummaries võimaluse kiiresti, kuid samas sisukalt, uurida andmestikke, mida plaanitakse oma töös kasutada.

PUBLICATIONS

CURRICULUM VITAE

Personal data

Name	Raivo Kolde
Birth	April 21st, 1983, Tallinn, Estonia
Citizenship	Estonian
Marital Status	Married
Languages	Estonian, English
Address	Savi 6-12, 50405 Tartu, Estonia
Contact	raivo.kolde@eesti.ee

Education

2009–	University of Tartu, Ph.D. candidate in Computer Science
2005–2008	University of Tartu, M.Sc. in Mathematical Statistics
2001–2005	University of Tartu, B.Sc. in Mathematical Statistics
1998–2001	Tallinn Secondary Science School, secondary education
1989–1998	Tallinn Õismäe School of Humanities, primary education

Employment

2013–	University of Tartu, Institute of Computer Science, researcher
2012–2013	University of Tartu, Institute of Computer Science, programmer
2008–2009	University of Tartu, Institute of Computer Science, extraordinary researcher
2007–	OÜ Quretec, researcher
2006–2007	AS EGeen, bioinformatician

ELULOOKIRJELDUS

Isikuandmed

Nimi	Raivo Kolde
Sünniaeg ja -koht	21. Aprill 1983 Tallinn, Eesti
Kodakondsus	eestlane
Perekonnaseis	abielus
Keelteoskus	eesti, inglise
Aadress	Savi 6-12, 50405 Tartu, Eesti
Kontaktandmed	+372 50 67 961 raivo.kolde@eesti.ee

Haridustee

2009–	Tartu Ülikool, informaatika doktorant
2005–2008	Tartu Ülikool, MSc matemaatilises statistikas
2001–2005	Tartu Ülikool, BSc matemaatilises statistika
1998–2001	Tallinna Reaalkool, keskharidus
1989–1998	Tallinna Õismäe Humanitaar Gümnaasium, põhiharidus

Teenistuskäik

2013–	Tartu Ülikool, Arvutiteaduse instituut, teadur
2012–2013	Tartu Ülikool, Arvutiteaduse instituut, programmeerija
2008–2009	Tartu Ülikool, Arvutiteaduse instituut, erakorraline teadur
2007–	OÜ Quretec, teadur
2006–2007	AS EGeen, bioinformaatik

DISSERTATIONES MATHEMATICAE UNIVERSITATIS TARTUENSIS

1. **Mati Heinloo.** The design of nonhomogeneous spherical vessels, cylindrical tubes and circular discs. Tartu, 1991, 23 p.
2. **Boris Komrakov.** Primitive actions and the Sophus Lie problem. Tartu, 1991, 14 p.
3. **Jaak Heinloo.** Phenomenological (continuum) theory of turbulence. Tartu, 1992, 47 p.
4. **Ants Tauts.** Infinite formulae in intuitionistic logic of higher order. Tartu, 1992, 15 p.
5. **Tarmo Soomere.** Kinetic theory of Rossby waves. Tartu, 1992, 32 p.
6. **Jüri Majak.** Optimization of plastic axisymmetric plates and shells in the case of Von Mises yield condition. Tartu, 1992, 32 p.
7. **Ants Aasma.** Matrix transformations of summability and absolute summability fields of matrix methods. Tartu, 1993, 32 p.
8. **Helle Hein.** Optimization of plastic axisymmetric plates and shells with piece-wise constant thickness. Tartu, 1993, 28 p.
9. **Toomas Kiho.** Study of optimality of iterated Lavrentiev method and its generalizations. Tartu, 1994, 23 p.
10. **Arne Kokk.** Joint spectral theory and extension of non-trivial multiplicative linear functionals. Tartu, 1995, 165 p.
11. **Toomas Lepikult.** Automated calculation of dynamically loaded rigid-plastic structures. Tartu, 1995, 93 p, (in Russian).
12. **Sander Hannus.** Parametrical optimization of the plastic cylindrical shells by taking into account geometrical and physical nonlinearities. Tartu, 1995, 74 p, (in Russian).
13. **Sergei Tupailo.** Hilbert's epsilon-symbol in predicative subsystems of analysis. Tartu, 1996, 134 p.
14. **Enno Saks.** Analysis and optimization of elastic-plastic shafts in torsion. Tartu, 1996, 96 p.
15. **Valdis Laan.** Pullbacks and flatness properties of acts. Tartu, 1999, 90 p.
16. **Märt Põldvere.** Subspaces of Banach spaces having Phelps' uniqueness property. Tartu, 1999, 74 p.
17. **Jelena Ausekle.** Compactness of operators in Lorentz and Orlicz sequence spaces. Tartu, 1999, 72 p.
18. **Krista Fischer.** Structural mean models for analyzing the effect of compliance in clinical trials. Tartu, 1999, 124 p.

19. **Helger Lipmaa.** Secure and efficient time-stamping systems. Tartu, 1999, 56 p.
20. **Jüri Lember.** Consistency of empirical k-centres. Tartu, 1999, 148 p.
21. **Ella Puman.** Optimization of plastic conical shells. Tartu, 2000, 102 p.
22. **Kaili Müürisep.** Eesti keele arvutigrammatika: süntaks. Tartu, 2000, 107 lk.
23. **Varmo Vene.** Categorical programming with inductive and coinductive types. Tartu, 2000, 116 p.
24. **Olga Sokratova.** Ω-rings, their flat and projective acts with some applications. Tartu, 2000, 120 p.
25. **Maria Zeltser.** Investigation of double sequence spaces by soft and hard analitical methods. Tartu, 2001, 154 p.
26. **Ernst Tungel.** Optimization of plastic spherical shells. Tartu, 2001, 90 p.
27. **Tiina Puolakainen.** Eesti keele arvutigrammatika: morfoloogiline ühestamine. Tartu, 2001, 138 p.
28. **Rainis Haller.** *M(r,s)*-inequalities. Tartu, 2002, 78 p.
29. **Jan Villemson.** Size-efficient interval time stamps. Tartu, 2002, 82 p.
30. **Eno Tõnisson.** Solving of expession manipulation exercises in computer algebra systems. Tartu, 2002, 92 p.
31. **Mart Abel.** Structure of Gelfand-Mazur algebras. Tartu, 2003. 94 p.
32. **Vladimir Kuchmei.** Affine completeness of some ockham algebras. Tartu, 2003. 100 p.
33. **Olga Dunajeva.** Asymptotic matrix methods in statistical inference problems. Tartu 2003. 78 p.
34. **Mare Tarang.** Stability of the spline collocation method for volterra integro-differential equations. Tartu 2004. 90 p.
35. **Tatjana Nahtman.** Permutation invariance and reparameterizations in linear models. Tartu 2004. 91 p.
36. **Märt Möls.** Linear mixed models with equivalent predictors. Tartu 2004. 70 p.
37. **Kristiina Hakk.** Approximation methods for weakly singular integral equations with discontinuous coefficients. Tartu 2004, 137 p.
38. **Meelis Käärik.** Fitting sets to probability distributions. Tartu 2005, 90 p.
39. **Inga Parts.** Piecewise polynomial collocation methods for solving weakly singular integro-differential equations. Tartu 2005, 140 p.
40. **Natalia Saealle.** Convergence and summability with speed of functional series. Tartu 2005, 91 p.
41. **Tanel Kaart.** The reliability of linear mixed models in genetic studies. Tartu 2006, 124 p.
42. **Kadre Torn.** Shear and bending response of inelastic structures to dynamic load. Tartu 2006, 142 p.

43. **Kristel Mikkor.** Uniform factorisation for compact subsets of Banach spaces of operators. Tartu 2006, 72 p.
44. **Darja Saveljeva.** Quadratic and cubic spline collocation for Volterra integral equations. Tartu 2006, 117 p.
45. **Kristo Heero.** Path planning and learning strategies for mobile robots in dynamic partially unknown environments. Tartu 2006, 123 p.
46. **Annely Mürk.** Optimization of inelastic plates with cracks. Tartu 2006. 137 p.
47. **Annemai Raidjõe.** Sequence spaces defined by modulus functions and superposition operators. Tartu 2006, 97 p.
48. **Olga Panova.** Real Gelfand-Mazur algebras. Tartu 2006, 82 p.
49. **Härmel Nestra.** Iteratively defined transfinite trace semantics and program slicing with respect to them. Tartu 2006, 116 p.
50. **Margus Pihlak.** Approximation of multivariate distribution functions. Tartu 2007, 82 p.
51. **Ene Käärik.** Handling dropouts in repeated measurements using copulas. Tartu 2007, 99 p.
52. **Artur Sepp.** Affine models in mathematical finance: an analytical approach. Tartu 2007, 147 p.
53. **Marina Issakova.** Solving of linear equations, linear inequalities and systems of linear equations in interactive learning environment. Tartu 2007, 170 p.
54. **Kaja Sõstra.** Restriction estimator for domains. Tartu 2007, 104 p.
55. **Kaarel Kaljurand.** Attempto controlled English as a Semantic Web language. Tartu 2007, 162 p.
56. **Mart Anton.** Mechanical modeling of IPMC actuators at large deformations. Tartu 2008, 123 p.
57. **Evely Leetma.** Solution of smoothing problems with obstacles. Tartu 2009, 81 p.
58. **Ants Kaasik.** Estimating ruin probabilities in the Cramér-Lundberg model with heavy-tailed claims. Tartu 2009, 139 p.
59. **Reimo Palm.** Numerical Comparison of Regularization Algorithms for Solving Ill-Posed Problems. Tartu 2010, 105 p.
60. **Indrek Zolk.** The commuting bounded approximation property of Banach spaces. Tartu 2010, 107 p.
61. **Jüri Reimand.** Functional analysis of gene lists, networks and regulatory systems. Tartu 2010, 153 p.
62. **Ahti Peder.** Superpositional Graphs and Finding the Description of Structure by Counting Method. Tartu 2010, 87 p.
63. **Marek Kolk.** Piecewise Polynomial Collocation for Volterra Integral Equations with Singularities. Tartu 2010, 134 p.

64. **Vesal Vojdani.** Static Data Race Analysis of Heap-Manipulating C Programs. Tartu 2010, 137 p.
65. **Larissa Roots.** Free vibrations of stepped cylindrical shells containing cracks. Tartu 2010, 94 p.
66. **Mark Fišel.** Optimizing Statistical Machine Translation via Input Modification. Tartu 2011, 104 p.
67. **Margus Niitsoo**. Black-box Oracle Separation Techniques with Applications in Time-stamping. Tartu 2011, 174 p.
68. **Olga Liivapuu.** Graded q-differential algebras and algebraic models in noncommutative geometry. Tartu 2011, 112 p.
69. **Aleksei Lissitsin.** Convex approximation properties of Banach spaces. Tartu 2011, 107 p.
70. **Lauri Tart.** Morita equivalence of partially ordered semigroups. Tartu 2011, 101 p.
71. **Siim Karus.** Maintainability of XML Transformations. Tartu 2011, 142 p.
72. **Margus Treumuth.** A Framework for Asynchronous Dialogue Systems: Concepts, Issues and Design Aspects. Tartu 2011, 95 p.
73. **Dmitri Lepp.** Solving simplification problems in the domain of exponents, monomials and polynomials in interactive learning environment T-algebra. Tartu 2011, 202 p.
74. **Meelis Kull.** Statistical enrichment analysis in algorithms for studying gene regulation. Tartu 2011, 151 p.
75. **Nadežda Bazunova.** Differential calculus $d^3 = 0$ on binary and ternary associative algebras. Tartu 2011, 99 p.
76. **Natalja Lepik.** Estimation of domains under restrictions built upon generalized regression and synthetic estimators. Tartu 2011, 133 p.
77. **Bingsheng Zhang.** Efficient cryptographic protocols for secure and private remote databases. Tartu 2011, 206 p.
78. **Reina Uba.** Merging business process models. Tartu 2011, 166 p.
79. **Uuno Puus.** Structural performance as a success factor in software development projects – Estonian experience. Tartu 2012, 106 p.
80. **Marje Johanson.** $M(r, s)$-ideals of compact operators. Tartu 2012, 103 p.
81. **Georg Singer.** Web search engines and complex information needs. Tartu 2012, 218 p.
82. **Vitali Retšnoi.** Vector fields and Lie group representations. Tartu 2012, 108 p.
83. **Dan Bogdanov.** Sharemind: programmable secure computations with practical applications. Tartu 2013, 191 p.
84. **Jevgeni Kabanov.** Towards a more productive Java EE ecosystem. Tartu 2013, 151 p.
85. **Erge Ideon**. Rational spline collocation for boundary value problems. Tartu, 2013, 111 p.
86. **Esta Kägo**. Natural vibrations of elastic stepped plates with cracks. Tartu, 2013, 114 p.

87. **Margus Freudenthal.** Simpl: A toolkit for Domain-Specific Language development in enterprise information systems. Tartu, 2013, 151 p.
88. **Boriss Vlassov.** Optimization of stepped plates in the case of smooth yield surfaces. Tartu, 2013, 104 p.
89. **Elina Safiulina.** Parallel and semiparallel space-like submanifolds of low dimension in pseudo-Euclidean space. Tartu, 2013, 85 p.