Department of Computer Science
Series of Publications A
Report A-2010-1

Construction of a global map of human gene expression: the process, tools and analysis

Margus Lukk

To be presented, with the permission of the Faculty of Science of the University of Helsinki, for public criticism in Auditorium XV, University Main Building, on August 19th 2010, at 12 o'clock noon.

University of Helsinki
Finland

Contact information

Postal address:
Department of Computer Science
P.O. Box 68 (Gustaf Hällströmin katu 2b)
FI-00014 University of Helsinki
Finland

Email address: postmaster@cs.helsinki.fi (Internet)

URL: http://www.cs.Helsinki.FI/

Telephone: +358 9 1911

Telefax: +358 9 191 51120

ISSN 1238-8645
ISBN 978-952-10-6370-1 (paperback)
ISBN 978-952-10-6371-8 (PDF)
Computing Reviews (1998) Classification: A.0, C.0.0
Helsinki 2010
Helsinki University Print

Construction of a global map of human gene expression: the process, tools and analysis

Margus Lukk

Department of Computer Science
P.O. Box 68, FI-00014 University of Helsinki, Finland
lukk@cs.helsinki.fi
http://www.cs.helsinki.fi/u/lukk/

PhD Thesis, Series of Publications A, Report A-2010-1
Helsinki, August 2010, 120 pages
ISSN 1238-8645
ISBN 978-952-10-6370-1 (paperback)
ISBN 978-952-10-6371-8 (PDF)

Abstract

This thesis studies human gene expression space using high throughput gene expression data from DNA microarrays. In molecular biology, high throughput techniques allow numerical measurements of expression of tens of thousands of genes simultaneously. In a single study, this data is traditionally obtained from a limited number of sample types with a small number of replicates. For organism-wide analysis, this data has been largely unavailable and the global structure of human transcriptome has remained unknown.

This thesis introduces a human transcriptome map of different biological entities and analysis of its general structure. The map is constructed from gene expression data from the two largest public microarray data repositories, GEO and ArrayExpress. The creation of this map contributed to the development of ArrayExpress by identifying and retrofitting the previously unusable and missing data and by improving the access to its data. It also contributed to creation of several new tools for microarray data manipulation and establishment of data exchange between GEO and ArrayExpress.

The data integration for the global map required creation of a new large ontology of human cell types, disease states, organism parts and cell lines. The ontology was used in a new text mining and decision tree based method for automatic conversion of human readable free text microarray data annotations into categorised format. The data comparability and minimisation of the systematic measurement errors that are characteristic to each laboratory in this large cross-laboratories integrated dataset, was ensured by computation of a range of microarray data quality metrics and exclusion of incomparable data. The structure of a global map of human gene expression was then explored by principal component analysis and hierarchical clustering using heuristics and help from another purpose built sample ontology.

A preface and motivation to the construction and analysis of a global map of human gene expression is given by analysis of two microarray datasets of human malignant melanoma. The analysis of these sets incorporate indirect comparison of statistical methods for finding differentially expressed genes and point to the need to study gene expression on a global level.

Computing Reviews (1998) Categories and Subject Descriptors:

H.2.8	Database Management: Database Applications - Data mining
I.1.5.3	Pattern recognition: Custering - Algorithms, Similarity measures
J.3	Life and medical science

General Terms:
bioinformatics, DNA microarrays, data analysis, systems biology

Additional Key Words and Phrases:
gene expression, data integration, exploratory data analysis, clustering, principal component analysis, distance metrics, malignant melanoma, human

Acknowledgements

My first and foremost gratitude goes to my advisor Professor Esko Ukkonen for accepting me as a student in ComBi graduate school and supporting me through out the years of my development as bioinformatician. I am equally grateful to Dr. Alvis Brazma for having me as a permanent visitor in his group at European Bioinformatics Institute (EBI) and for becoming my second unofficial adviser. My gratitude also goes to Dr. Erkki Hölttä for his guidance and interesting discussions about melanoma biology.

I would like to thank the staff of the department of Computer Science for the excellent education they provide. A warm thank you to my office mate Dr. Kimmo Palin who served as a faithful companion and an ultimate information source of computer science, statistics, mathematics and general information of the department and life in Helsinki.

At EBI, I would like to thank Dr. Helen Parkinson and her curation team teaching me about ArrayExpress database and for creating the best imaginable working environment. I would also like to thank Dr. Parkinson for reading and commenting on the manuscript of this thesis. My thanks also goes to all members and visitors of the ArrayExpress team over the years for their influence to the work in this thesis. I am thankful to Dr. Samuel Kaski for his encouragement in thinking towards future scientific issues.

I want thank my collaborators in Finland, UK and Germany throughout the years. I am especially grateful for Dr. Janne Nikkilä who introduced me to R/Bioconductor and to Dr. Tim Rayner for his influence to my Perl skills. I would like to thank all of the funding bodies on both sides of English channel that supported the work.

My gratitude goes to ComBi coordinator Heikki Lokki for his help in administrative matters over email while I was living in UK. I would like to thank Dr. Katja Kivinen for helping me with the language for the Finnish abstract of this thesis. Similarly, I am grateful to the BREW workshops for providing invaluable experience in scientific writing and presentation skills in the earlier years of my studies.

My gratitude goes to my parents Anne and Heino, and my brothers Toomas and Tiit for their support and encouragements throughout this thesis. Finally, I want to thank my dear wife Dawn-Louise and our two sons Daniel and Andreas for their constant and never ending love.

Cambridge, July 15, 2010
Margus Lukk

This thesis is based on the following papers, which are referred in the text by their Roman numbers:

I E Kääriäinen, P Nummela, J Soikkeli, M Yin, M Lukk, T Jahkola, S Virolainen, A Ora, E Ukkonen, O Saksela and E Hölttä. Switch to an invasive growth phase in melanoma is associated with tenascin-C, fibronectin, and procollagen-I forming specific channel structures for invastion. *The Journal of Pathology* 2006 Oct;210(2):181-91.

II J Soikkeli[1], M Lukk[1], P Nummela, S Virolainen, T Jahkola, R Katainen, L Harju, E Ukkonen, O Saksela and E Hölttä. Systematic search for the best gene expression markers for melanoma micrometastasis detection. *The Journal of Pathology* 2007 Oct;213(2):180-9.

III H. Parkinson, M. Kapushesky, M. Shojatalab, N. Abeygunawardena, R. Coulson, A. Farne, E. Holloway, N. Kolesnykov, P. Lilja, M. Lukk, R, Mani, T. Rayner, A. Sharma, E. Williams, U. Sarkans and A. Brazma. ArrayExpress–a public database of microarray experiments and gene expression profiles. *Nucleic Acids Research* 2007 Jan;35 (Database issue):D747-50.

IV Tim F. Rayner, Faisal Ibne Rezwan, Margus Lukk, Xiangqun Zheng Bradley, Anna Farne, Ele Holloway, James Malone, Eleanor Williams and Helen Parkinson. MAGETabulator, a suite of tools to support the microarray data format MAGE-TAB. *Bioinformatics* 2009 Jan 15;25(2):279-80.

V Audrey Kauffmann, Tim F. Rayner, Helen Parkinson, Misha Kapushesky, Margus Lukk, Alvis Brazma and Wolfgang Huber. Importing ArrayExpress datasets into R/Bioconductor. *Bioinformatics* 2009 Aug 15;25(16):2092-4.

VI Margus Lukk, Misha Kapushesky, Janne Nikkilä, Helen Parkinson, Angela Goncalves, Wolfgang Huber, Esko Ukkonen and Alvis Brazma. A global map of human gene expression. *Nature Biotechnology* 2010 Apr;28(4):322-324

[1] These authors contributed equally to this work.

Contents

Abbreviations

ANOVA - Analysis of variance
cDNA - Complementary deoxyribonucleic acid
CIBEX - Center for information biology gene expression database
cRNA - Complementary ribonucleic acid
Cy3 - Cyanine 3
Cy5 - Cyanine 5
DDBJ - DNA data bank of Japan
DNA - Deoxyribonucleic acid
EBI - European bioinformatics institute
EFO - Experimental factor ontology
GEO - Gene expression omnibus
MAGE - Microarray and gene expression
MAGE-ML - Microarray gene expression - markup language
MAGE-OM - Microarray gene expression - object model
MAGE-TAB - Microarray gene expression tabular
MAGEstk - Microarray gene expression software toolkit
MAQC - Microarray quality control
MAS5 - Microarray suite version 5
MGED - Microarray gene expression data society
MIAME - Minimum information about a microarray experiment
miRNA - Micro ribonucleic acid
MM - Mismatch
mRNA - Messenger ribonucleic acid
NCBI - National centre for biotechnology information
NUSE - Normalised unscaled standard error
PCA - Principal component analysis
PCR - Polymerase chain reaction
PLM - Probe level model
PM - Perfect match
RLE - Relative log expression
RMA - Robust multi-array average
RNA - Ribonucleic acid
RNA-seq - Ribonucleic acid sequencing
rRNA - Ribosomal ribonucleic acid
RT-PCR - Reverse transcription-polymerase chain reaction
S2N - Signal to noise
SAM - Significance analysis of microarrays
SOFT - Simple omnibus format in text
SOM - Self-organizing map

tRNA - Transfer ribonucleic acid
UHTS - Ultra high throughput sequencing
UML - Unified modelling language
UMLS - Unified medical language system
WTSS - Whole transcriptome shotgun sequencing
XML - Extensible markup language

Chapter 1

Introduction

> ”In theory there is no difference between theory and practice.
> In practice there is.”
>
> *Yogi Berra*

1.1 Motivation

This thesis is about processing and analysis of microarray data. Microarray data analysis is a branch of bioinformatics which is an application of information technology and computer science to the field of molecular biology. The biological questions that motivate this thesis include: (1) which of the genes are involved in the escape of melanoma cells from the primary tumour during metastasation?; (2) which of the genes could be used to detect the progression of melanoma cells to the sentinel lymph nodes?; and ultimately, (3) how are different cell types organised in the human gene expression space?

These biological questions have scientific value for the following reasons: Firstly, the formation of metastasis in malignant melanoma is not well understood. It is known that the cells of a primary tumour undergo a transition which allow them to leave via lymphatic system. The mechanism and genes it involves have remained unknown. Secondly, the early detection of invasive melanoma cells in sentinel lymph nodes has an important impact on survival. Many genes with melanoma specific expression have been proposed for detection of the presence of melanoma cells in the sentinel lymph nodes; but no consensus of which genes should be used in clinical tests has been established. Most importantly, no transcriptome wide search for such genes has been carried out. Finally, similarly to the first two questions, transcriptional states of human cells are largely studied only in the context of a small number of other, and often related, cell and tissue types e.g. disease vs. normal. A global view of which genes are expressed where and how the expression differs between different cell states in the context of a whole organism is non-existent. The data for answering such question has not been available to date. This thesis addresses this need.

The first two motivating questions relate to melanoma biology while the last one is more general. The link between these questions is the microarray data and the use of similar analysis methods. However, it was the research undertaken with the first two questions that raised the issues of the third. Addressing complete gene expression between different cell types is important but more so is to compare a cell type against all other cell and tissue types in the organism. It is this question of the global structure of human gene expression which is the most complex and underpins this thesis.

Primarily, this thesis aims to address these biological questions through the use of computational analysis of microarray data. How can these answers be found? Which data processing and analysis methods should be used? What alterations to the methods are required to apply these to quantities of data on which they have never been applied before? Where can we obtain more microarray data, how can it be integrated for co-analysis, and how to assure its comparability? How can we visualise the results? These are just some of the problems for which this introduction tries to provide answers.

1.2 Summaries of original publications

Short summaries of original publications included in this thesis are given below.

Paper I

Paper I is the study of molecular mechanisms behind the escape of melanoma cells from the site of primary tumour. We start by measuring gene expression in samples of benign nevi, primary melanomas, and melanomas with known metastasis using different microarray platforms. The pre-processing of the numerical data and subsequent search for genes which are highly up-regulated in the invasive phase of the cancer is carried out for each microarray platform separately. The three extracellular matrix proteins encoding genes tenascin-C (TN-C), fibronectin (FN) and procollagen-I (PCOL-I) identified during the analysis are then used for immunohistochemical staining to localise their expression in the tumour. We show that all three proteins co-localise in the tumour and form tubular meshworks and channels ensheathing the melanoma cells.

Paper II

The focus of Paper II is the search and selection of genes that are specifically expressed in metastatic melanoma for construction of RT-PCR based clinical assay that could be used for detection of invaded melanoma cells in sentinel lymph nodes either in parallel with, or as replacement of, the immunohistochemical examination.

We measure the gene expression of biological replicates of groups of benign nevi as well as normal, micro metastatic and macro metastatic lymph nodes using two different microarray platforms. The expression values from both platforms are quantified and merged into one virtual microarray using a small set of control transcripts shared between both of the arrays. The data is processed further by a version of quantile normalisation specially implemented for this paper. Finally the data is filtered to include genes with sufficient group wise separation.

The robustness of the set of differentially expressed candidate genes suitable for the clinical assay is ensured by the parallel use of multiple statistical tests and methods suggested for these types of analysis in the literature. A tool for computing signal to noise (S2N) values and associated p-values using permutation test in this paper was implemented according to [GST$^+$99].

We then test the set of candidate genes for the assay parallel to immunohistochemical analysis on a body of sentinel lymph nodes. We conduct a statistical analysis of the results and show that two of the genes tested in the reverse transcription-polymerase chain reaction (RT-PCR) assay perform better, as assessed by disease recurrence, than histological and immunohistochemical examination of the lymph nodes. The statistical analysis also identifies two further melanoma specific marker genes which are capable for differentiating between melanoma cells and cells of benign nevi in the sentinel lymph nodes.

Paper III

In Paper III we present an update on ArrayExpress which is one of the leading gene expression data repositories in the world [ItHV+08]. The update introduces a new query interface for simpler data search and retrieval both in ArrayExpress repository and in data warehouse; and new tools developed for easier data submission. The update includes also a retrofit of majority of Affymetrix native format data files which had been corrupted while being loaded to ArrayExpress.

Paper IV

Paper IV is a report of a set of software tools to support new microarray gene expression tabular (MAGE-TAB) data format. The software suite includes tools for preparation, syntactic and semantic validation of the data presented in MAGE-TAB format, visualisation of investigation designs coded in MAGE-TAB, conversion of MAGE-TAB documents to older microarray gene expression markup language (MAGE-ML) format, conversion of Gene Expression Omnibus (GEO) Simple Omnibus Format in Text (SOFT) format data files to MAGE-TAB and post hoc addition of ontology terms to existing MAGE-TAB documents.

The significance of the tools presented in this paper lies in the fact that MAGE-TAB format has been proposed by the microarray community to replace the more complex MAGE-ML data format. The tools in this software suite have also made it possible to start public microarray data exchange between GEO and ArrayExpress. MAGE-TAB has already been established as new official data export format in ArrayExpress and will be the main microarray data import/export format for the next generation ArrayExpress database.

Paper V

In Paper V we present a new software package for direct access of ArrayExpress data from R statistical language environment. The package allows users to browse and work with ArrayExpress data in R environment without external downloading and complex data formatting. The package uses MAGE-TAB formatted data from Array-

Express ftp site and converts it to R data objects; a convenient input to many microarray data analysis tools available in R/Bioconductor project.

Paper VI

In Paper VI we present a global map of human gene expression. We study the difference and distance of the gene expression between a large number of cell states, types and lines. We visualise this information by the first principal components of the data and call the identified global level results a global map of human gene expression. We identify six major 'continents' on this map and characterise their content and relations.

The data for the map was collected from public microarray data in ArrayExpress and GEO. Strict data quality control was performed and in the case of GEO, the data was subjected to a semi-automatic text mining based meta data re-annotation. To present and group the data in a more meaningful way, we organised the sample annotations into a custom ontology and used this in data visualisation by hierarchical clustering and principal component analysis (PCA). As some of the tools and methods that we used were created for datasets several magnitudes smaller than in this paper, the computation and visualisation was possible only with the help of various heuristics. We also analysed several subgroups of the data both in sample and in gene dimension. We computed gene ontology enrichments for selected gene groups and identified differentially expressed genes between groups of samples. Finally, we created a web interface with a set of query options for more specific exploration of the data.

1.3 Contributions of the author

Main contributions of this thesis in publication order:

I The paper identifies three extracellular matrix protein encoding genes which play a role in the escape of the melanoma cells from the primary tumour. The paper also proposes a new concept of structured tumour cell spreading in melanoma via formation of special channel-like structures observed by usage of immunohistochemistry. The author contributed to the paper via processing and analysis of different types of the melanoma gene expression data which lead to identification of these three genes.

II The paper reports the first genome-wide search for the best gene expression markers to detect melanoma micrometastasis. The paper confirms the good performance of genes TYR and MLANA in clinical RT-PCR based metastatic melanoma cell detection assay and identifies two other genes PRAME and SPP1 as novel malignancy indicators and ideal therapy targets in metastatic melanoma. The author performed the microarray data processing and analysis to identify good candidate genes for the assay; and the statistical analysis of the RT-PCR data. Together with Dr. Erkki Hölttä, the author interpreted the biological data and selected the candidate genes for the RT-PCR assay.

III The paper presents the update of the latest improvements in ArrayExpress which include introduction of the new user friendly database browse and query interface, a report of starting to serve the data in the database ftp site for bulk downloads and introduction of new data submission tools. The main contribution of the author was taking a lead in the re-design of the ArrayExpress web interface for improved data access and display. The author also contributed through analysis of ArrayExpress user behaviour using database web server log files, through retrofit of majority of the Affymetrix CEL files in the database from various sources as these had been corrupted while being loaded to the database; and through testing of the new interfaces and data submission tools.

IV The paper introduces a series of MAGE-TAB data format related software tools which are critical for the every day operation of ArrayExpress. The main contribution of the author was the design and implementation of the custom ontology based text mining application to discover microarray gene expression data society ontology (MGED ontology) based category values in GEO free text annotations. The author also provided input to the development of GEO to ArrayExpress data importer which is continuously used for mirroring GEO data in ArrayExpress.

V The paper introduces a software package for data retrieval from ArrayExpress to R statistical language environment. The author contributed to the testing of the software and adjusted the software for the internal use of ArrayExpress.

VI The paper introduces and characterises a first global map of human gene expression and provides a web interface for detailed exploration of the underlying data. The presented work is also an example of successful large scale microarray data integration. The contributions of the author include choosing the topic, creating all necessary tools for the data integration, integrating the data, performing the data quality check and filtering of the data based on the quality and performing majority of the presented data analysis and visualisation.

* * *

The rest of the thesis is organised as follows. Introduction to gene expression and gene expression data is given in Chapter 2. Microarray data integration with its challenges in data selection, integration and quality assessment for data comparability are discussed in Chapter 3. Microarray data analysis methods used in different parts of this thesis are outlined in Chapter 4. The initial work with melanoma data which motivated the larger part of the included work and became itself secondary is introduced in Chapter 5. Finally, the introductory part of the thesis is concluded by Chapter 6 which summarises the main results of this work. The original publications I-VI are reprinted at the end of the thesis.

Chapter 2

Gene expression profiling

This chapter is an introduction to genes, gene expression and gene expression data. Section 2.1 provides a short introduction to genes and gene expression. Section 2.2 describes two different types of microarrays used in Papers I, II and VI. Definitions and an overview of microarray data standards and formats relevant for the rest of the thesis are given in Section 2.3. Finally, Section 2.4 introduces public microarray data repositories and discusses some of their properties.

2.1 Genes and Gene expression

The genetic information in all known living organisms is stored in deoxyribonucleic acid (DNA) molecules. These molecules contain the complete information needed for the normal function and reproduction of the organism. The genomic DNA is divided into locatable functional regions corresponding to a unit of inheritance known as genes. The DNA sequence of a gene consists of coding sequences which determine what the gene does and regulatory sequences which determine how the gene is activated. The process in which the information is copied from genomic DNA to ribonucleic acid (RNA) is known as transcription. Some of the RNA molecules, particularly messenger ribonucleic acid (mRNA) molecules, are further translated into a specific amino acids chain which is then folded into an active protein. The flow of genetic information from DNA to RNA to protein is known as the central dogma of molecular biology (Figure 2.1) [Cri70]. The RNA copy of the gene is referred to as transcript and the expression of all the genes in a given cell in a given time point is called a transcriptome.

The expression levels of genes vary greatly and depend on many internal and external factors. The state of the cell at any time point is dependent

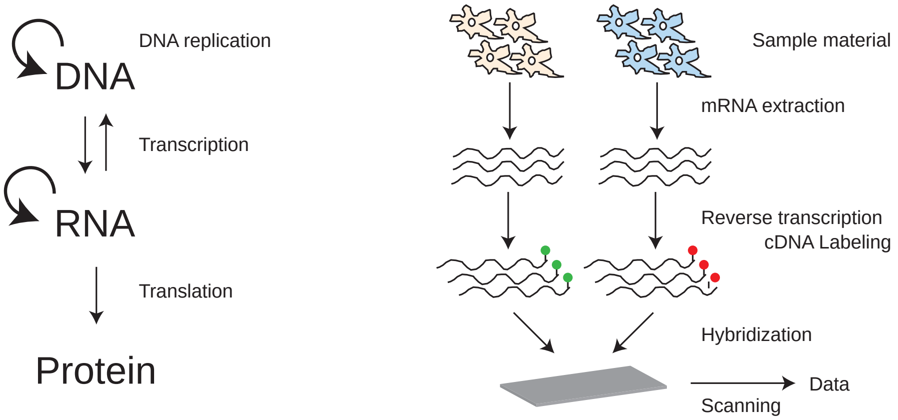

Figure 2.1: Central dogma of molecular biology (left). Steps in gene expression measurement using two colour DNA microarrays (right).

upon, and reflected by, the expression levels of all genes. Studying gene expression therefore provides information about the cells state in general, and about the roles of individual genes.

In the majority of studies, the gene expression is measured for only protein coding genes, i.e. genes from which mRNAs are produced. Studying mRNA expression is biologically relevant since it indirectly gives information on the expression of the proteins which are the key players in a cell. The expression of ribosomal ribonucleic acid (rRNA) and transfer ribonucleic acid (tRNA) molecules is usually ignored as these are expressed in all living cells while micro ribonucleic acid (miRNA) expression is measured by separate dedicated assays. The miRNAs are important for study since they are known to be post-translational regulators of mRNA.

2.2 DNA microarrays

DNA microarrays (or microarrays) are the most commonly used technology for measuring expression levels for thousands of genes in parallel. Alternative low throughput technologies include Northern plotting [AKS77] and RT-PCR [VVF08]. The high throughput alternatives include SAGE (serial analysis of gene expression) [VZVK95], cap analysis gene expression (CAGE) [SKK$^+$03], massively parallel signature sequencing (MPSS)[NH09] and ribonucleic acid sequencing (RNA-Seq)[WGS09], also known as Whole Transcriptome Shotgun Sequencing (WTSS).

The DNA microarray technology, originally derived from Southern blotting technology [Sou75], is used for detection of specific DNA fragments in a larger pool of DNA fragments. In Southern blotting, the known DNA fragments are attached to a substrate and probed with sample DNA fragments to observe which DNA fragment the sample contains. The technology exploits property of nucleic acid sequences to pair with each other in the hybridisation reaction based on the principle of complementarity. In the reaction sample DNA fragments bind to their complements on the substrate. The samples used in the hybridisation, labeled either radioactively or fluorescently, are then detected by the amount of label.

On a DNA microarray, the substrate typically contains an array of thousands of DNA fragments with known sequences. The sequences on the array are designed to complement the mRNA molecules of genes from which expression is measured. The cocktail of mRNA molecules extracted from cells under study are reverse transcribed to complementary deoxyribonucleic acids (cDNAs) and labelled with fluorescent tags. The complementary ribonucleic acid (cRNA) molecules, which are created during the labelling procedure from cDNAs, are then used in a hybridisation reaction where each cRNA finds its DNA complement on the array. The amount of cRNA bound to each DNA fragment (spot/probe) on the array is measured by exposition of the microarray to a light which causes the labels to fluoresce. The microarrays are scanned/photographed and the intensities of the fluorescent signals for each of the spots on the microarray are quantified to numerical values of gene expression by image analysis software.

The weaknesses of the DNA microarray technology are the relative abundance of the measured expression values and limitedness of the results to only these genes for which the DNA fragments on the array exist. Moreover, different microarray platforms (array designs) consist of probes for different lists of genes and/or different DNA sequences for the same genes. These produce results which are not always comparable even when the same biological material is measured.

A variety of techniques exist for DNA microarray manufacture. The sequences of the DNA fragments on the array surface can be spotted or synthesised in situ. The material spotted on the arrays may be oligonucleotides, cDNA or small fragments of polymerase chain reaction (PCR) products that correspond to mRNA. Depending on the microarray type, the location of each DNA fragment on the array may be fixed to a certain location - the case for most current microarray types, or be dynamic for example in microarrays manufactured by Illumina Inc.

2.2.1 Dual channel microarrays

In two-channel (or two-colour) microarrays [SSDB95] [SSB96], cDNA/cRNA from two differently labelled samples are hybridised on microarray at the same time (Figure 2.1 b). The fluorescent labels most commonly used in this type of arrays are Cyanine 5 (Cy5) with fluorescence emission 670 nm and Cyanine 3 (Cy3) with fluorescence emission 570 nm. The microarray is scanned for each wavelength separately and the quantified result for each probe on the array is usually presented as a ratio of fluorophore intensities in one of the channels over the other. In the hybridisation of two-colour microarray, the cDNA molecules from both samples are competing for one set of probes. For each gene/the results therefore show a relative expression difference between the samples. To make individual hybridisations that measure different biomaterial comparable, a common reference in one of the channels is typically used. A low quality sample in one of the channels can seriously bias the obtained results.

In this thesis, the dual channel microarray data was used in Paper I.

2.2.2 Single channel microarrays

Single channel (or one-colour) microarrays [PSS+94] [LDB+96] use one labelled sample per hybridisation. The sensitivity quantified values are absolute for the sample and make the individual hybridisations from the same microarray platform more easy to compare.

Affymetrix

The current market leader in production of single channel oligonucleotide microarray platforms is Affymetrix. The sequences on Affymetrix GeneChip arrays are synthesised in-situ using photolitoghaphic synthesis method developed by the semiconductor industry [FRP+91] [FRH+93] [PSS+94]. The sequences for all spots on the array are synthesised one nucleotide at a time from the bottom of the array. Each transcript on the Affymetrix GeneChip array is presented by eleven 25 base pair long oligonucleotide probe sequences known as a probeset (Figure 2.2). The sequences are designed from the last 500 base pairs of the 3' end of the transcript. Each perfect match (PM) sequence on the array has a mismatch (MM) counterpart that differs from the first by one nucleotide at the 13th position. The nucleotide in the middle of the sequence is subject to purine to pyrimidine (and vice versa) transversion. The purpose of MM probes is to measure non-specific binding and background noise.

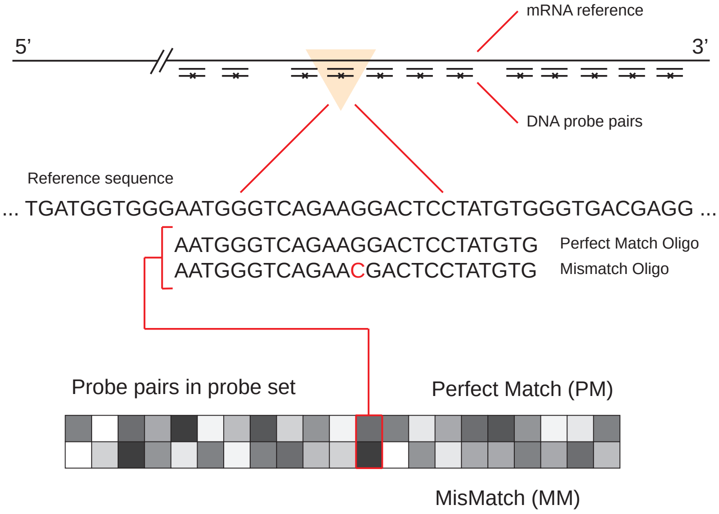

Figure 2.2: Affymetrix Genechip design principles.

Methods used in this thesis for Affymetrix probe level intensity quantification to probe set level intensities of genes included microarray suite version 5 (MAS5) [LMD$^+$02] [HLM02] [IHC$^+$03] in Papers I and II; and robust multi-array average (RMA) [IBC$^+$03] [IHC$^+$03] [BIAS03] in Paper VI. In short, both methods contain steps for background correction, probe level intensity calculation and probe set summarisation. Additionally, RMA uses, before the probe level intensity calculation, a step for arrays wide probe level normalisation; and MAS5 contains a step for scaling summarised intensities to a mean target intensity. Further differences between the two methods lie in techniques used in the listed shared steps and in the usage of only PM probe sets by RMA. The exclusion of MM data in RMA reduces noise as well as loses information and reflect in more precise results and smaller number of false positives. The advantage of MAS5 algorithm which uses both PM and MM probe sets is in larger variance on the lower expression levels, sensitivity to real expression changes and no assumption about equal expression value distributions in different co-analysed biological material.

2.3 Microarray data standards and formats

Microarray community wide standards and formats for public microarray data storage, sharing and exchange are created by the Microarray Gene Expression Database (MGED) society. Over its lifetime, the society has introduced standards for content, format, and semantics known as: minimum information about a microarray experiment (MIAME), MGED Ontology [WPC+06], microarray and gene expression (MAGE), and more recently MAGE-TAB [RRSS+06]. The data standards and formats introduced by the MGED society deal only with the experimental meta-data while the many formats of actual gene expression measurements have not been regulated as these are defined by instrumentation and assay vendors. The experimental meta-data in this thesis is defined as a collection of annotations needed for interpretation of the numerical expression values produced using DNA microarrays. These data include the general description and layout of the experiment, text of relevant laboratory protocols and annotations of measured biological samples and assays.

2.3.1 Formats of gene expression measurements

Diversity in both commercial and academic developed microarray technologies and platforms has prevented the rise of a single common microarray data file format. The data is, however, generally distinguished as either raw or processed/normalised/transformed. The raw microarray data is defined as that produced by microarray image analysis software and contains initial numerical quantifications of the fluorescent intensities measured on microarrays. Processed microarray data is defined as raw data which has been subjected to various data cleaning and smoothing methods which usually include signal background correction, normalisation and transformation to a different scale.

The expression data matrix, or in some cases transformed data matrix or final gene expression data matrix, is a reference to a numerical data matrix of processed microarray data joined across multiple expression measurements. Traditionally the first column and row of the matrix contain references of measured genes and samples.

2.3.2 MIAME

MIAME describes the Minimum Information About a Microarray Experiment needed for unambiguous interpretation of the results of microarray experiment and for potential experiment reproduction [BHQ+01]. The updated MIAME check list contains six elements needed for microarray data based publication:

1. The raw data produced by the microarray image analysis software, such as CEL files for Affymetrix.

2. The final processed data for the set of hybridisations in the experiment. For instance MAS5, RMA etc. normalised data matrices for Affymetrix.

3. The essential sample annotation, including experimental factors and their values.

4. The outline for the experimental design, including sample data relationships.

5. Sufficient annotation of the array design in the form of actual DNA sequences used on the array or database accessions of the sequences where the probes were designed from.

6. Essential experimental and data processing protocols.

MIAME has been criticised for being merely a guideline rather than standard and for leaving too much flexibility for individual interpretation [Bur06]. As a result the amount of information asked by different MIAME compliant public microarray repositories for full MIAME compliance is variable and makes data sharing and comparison across different MIAME compliant databases complex.

2.3.3 MGED Ontology

The primary purpose of MGED Ontology [WPC+06] is to provide standard terms for the annotation of microarray experiments. The terms enable description of how the experiment was performed, provides structure for defining relationships between individual terms and a list of categories for describing used biological material. The MGED Ontology is used in developing software and databases for manipulation and storage of microarray data.

2.3.4 MAGE and MAGE-TAB

The aim of the MAGE standard is to provide a representation of microarray expression data for data exchange between different data systems. The MAGE standard includes the microarray gene expression object model (MAGE-OM) modelled in unified modelling language (UML), the data exchange format MAGE-ML [SMS+02] implemented in extensible markup language (XML) and microarray gene expression software toolkit (MAGEstk) - a collection of software packages that act as converters between MAGE-OM and MAGE-ML.

In practice, MAGE-ML data format has proven to be over-complicated and impractical for laboratories with no dedicated bioinformatics support. A new simpler tab-delimited spreadsheet based format MAGE-TAB to replace the MAGE-ML has been introduced [RRSS$^+$06]. The software tools for MAGE-TAB format microarray data validation, manipulation and conversion presented in Paper IV have become critical for the operation of ArrayExpress (Paper III). MAGE-TAB data format is also used for data import from ArrayExpress to R/Bioconductor in Paper V.

2.4 Public microarray data repositories

The MGED society has promoted all scientific journals to require the submission of microarray data to public repositories as part of the process of publication ([BBC$^+$04]. The databases accepted by the MGED society for public microarray data deposition are ArrayExpress (Paper III and [PKK$^+$09]) at European Bioinformatics Institute (EBI), Gene Expression Omnibus - GEO [BTW$^+$09] in National Center for Biotechnology Information (NCBI), and Center of Information Biology gene Expression database - CIBEX [IIiT$^+$03] hosted by DNA Data bank of Japan (DDBJ).

While the content of corresponding nucleotide databases between EBI, NCBI and DDBJ is exchanged nightly, until recently no official data exchange between central public microarray databases has been agreed. In November 2009, ArrayExpress and GEO entered into a metadata exchange agreement for ultra high throughput sequencing (UHTS) experiments which will appear in both databases regardless of where they were submitted. Moreover, ArrayExpress has started a one sided data import from GEO [PKK$^+$09] and has entered into negotiation with CIBEX over MAGE-TAB format based data exchange. The number of original data submission in CIBEX is marginal to these available in GEO and ArrayExpress.

2.4.1 GEO and ArrayExpress

Usefulness of microarray data lies in its organisation and accessibility. The central units of information in GEO are sample/hybridisation and array platform (array design). Differently from ArrayExpress, the concepts of sample and hybridisation in GEO are the same. GEO human readable sample/hybridisation soft files contain free text sample annotations and links to related raw and processed data files. An experiment SOFT record contains additional experiment level annotations and links to relevant sample and related array design records. The central information units in ArrayExpress are experiment and array design. The data in ArrayExpress is currently stored in a 248 table relational database which database model was automatically generated from MAGE-OM.

Historically, simplicity in the GEO data organisation is best reflected by the fact that nearly all public microarray data based gene expression studies have been explicitly carried out using data from this database. Furthermore, retrieval of microarray data from ArrayExpress in a usable form was difficult before the work invested in Paper III. However, recent improvements in ArrayExpress, such as introduction of MAGE-TAB as a simple data download format and release of Bioconductor package for direct data access to R statistical analysis environment (Paper V), have significantly improved the access to its data.

Chapter 3

Microarray data integration

This chapter discusses the problem of microarray data integration in Paper VI and the sub-problem of re-annotation of experimental meta-data in Papers IV and VI. In this thesis, microarray data integration is defined as usage of data from individual hybridisations, expression profiles or raw data files across the borders of different microarray experiments. Microarray data integration is often seen only as an integration of numerical values. The challenge of integration, however, lies both in ensuring comparability of numerical expression values as well as in consistency of experimental meta-data and in overall comparability of the integrated data. The last two data integration points are regularly overlooked. An important question to ask is which data to use and at which level to integrate. In the integration approach for Paper VI, the data chosen was raw data created on one single-channel microarray platform Affymetrix HG-U133A. The rationale for this approach and clarification for why multi platform data integration was not considered is given in Section 3.1. The focus of Section 3.2 is on integration of experimental meta-data. Section 3.3 discusses the assurance of cross-experiment data quality and comparability and effects that may reduce the reliability of subsequent meta-analysis of the integrated data.

3.1 Challenges of microarray data integration

Microarray data can be integrated at the level of raw or processed expression data, within single or across multiple microarray platforms, or on the level of experimental results with no actual integration of expression measurements. The ideal data for integration is a raw data from one single-channel microarray platform. The raw data is free from experiment specific processing methods, such as normalisation and data transformation which may contribute to cross-experiment incomparabilities. The data from a single channel array platform is free from the problem of potential variable reference channel. Finally, usage of data from only one microarray platform removes the cross-platform data comparability problem. This "ideal" approach is robust [DCD+07] and may require only a routine data quality/comparability check which unfortunately are often not applied. The data for integration in Paper VI were selected according to these principles.

3.1.1 Cross-platform data integration

The data available for any single array platform is, however, limited by the number of available studies and diversity in measured biological material. The amount of data available across different array platforms is always much larger. Moreover, cross-platform data integration has been shown to deliver biologically meaningful results [RRLG03] [JHE+04] [KAO+08]. Nevertheless, the challenges concealed in cross-platform data integration are often not given enough attention and are hence briefly discussed below.

A common start point in cross-platform data integration is the probe mapping between array platforms. The mapping is created by probe annotations or by alignment of probe sequences to a reference genome [RRLG03] [ELS+05] [KAO+08] [PKK+09]. As vendor supplied microarray probe annotations typically originate from the time when particular array was designed, the annotations are often out of date. Therefore, sequence based re-annotation is preferred.

The cross-platform comparability of expression values is influenced by individual probe sequence design and its variations between array platforms. Probe sequence composition influences the final expression value and cross-platform comparability in three different ways. Firstly, the probe sequences, especially in oligonucleotide arrays, are significantly shorter than transcripts of genes against which they are designed. Genes in a cell can be expressed simultaneously as several different transcripts with different length, exon composition, start and end positions [LTB+09]. A selection of probe sequence location impacts to mRNA sub populations of a gene measured by the probe. Secondly, a base pair shift in the genome position for which probe is designed may have a significant effect on the binding affinity of the target mRNA. Thirdly, a small change in a probe sequence composition may change the balance of non-specific binding of other transcripts to the probe. In the design of any particular array, all three of the effects are assessed and minimised. Nonetheless, the size of these effects in cross-platform integration is amplified and hard to predict creating experimental noise generated by the array platforms. The problem in cross-platform data integration is thus minimisation of the noise caused by differences between array designs with at the same time not compromising the biological signal.

Methods used in tackling the effect of platform differences vary. One of these, expression data transformation to rank space is described below. This method was considered in early stages of Paper VI as a universal method for data normalisation and noise reduction.

3.1.2 Integration in rank space

Microarray data integration in rank space uses the assumption that gene expression values in any particular biological condition are monotonic and not influenced by the effects of underlying array platforms. As the order of expression values is globally fixed, the values in the rank space become directly comparable. The facts that different microarrays measure slightly different subsets of genes and in cross-platform comparison, a gene may have a different expression value, become trivial in rank space.

Let X be a set of all genes and Y and Z sets of genes measured by two different array platforms such that $Y \subseteq X$ and $Z \subseteq X$. Let Y_I and Z_I be subsets of genes in Y and Z shared by both array platforms ($Y_I \equiv Z_I$). In monotone transformation of expression values of genes in sets Y and Z, genes with smallest expression values are assigned rank 1, the second smallest 2, and so on. Genes with largest expression value are given ranks m and n equal to the number of elements in Y and Z. In case of monotone transformation of gene expression values in Y_I and Z_I, ranks of any two genes measured by different array platforms under same biological condition should be equal, or in case of some noise, at least close. Similarly, expression profiles of same biological material should also be alike in rank space.

The advantage of data transformation to rank space lies in simplicity and no need for further data normalisation. Even though transformation of expression values to rank space is lossy, it is shown in microarray data to perform well [XTN$^+$05].

The data transformation to rank space could be seen as a method of data normalisation where the ranges of individual datasets are made equal. A similar method for continuous values is quantile normalisation [BIAS03]. In quantile normalisation, the distributions of individual datasets are normalised to an average of all distributions. In this thesis the quantile normalisation was used in Papers II and VI.

3.2 Integration of experimental meta-data

Microarray meta-data is defined as textual information which is needed for understanding the origin and essence of the numerical data obtained by DNA microarrays. Experimental meta-data includes description of the measured biological material – known as sample annotations; experimental protocols that describe the origin and processing of the sample material; and a text for the general outline of the microarray experiment. For experimental meta-data integration, usually only the first two, i.e. sample annotations and experimental protocols used in sample processing, are relevant. The third, the information about the outline and aim of the experiment is usually excluded as the origin of numerical expression values is generally well described already by the first two. In practice also the content of experimental protocols in meta-data integration is ignored, either because the texts of the protocols are not available or due to laboriousness of processing of their content. The effect of experimental protocols on the numerical part of the experimental data, however, can be reduced by integrating data with same experimental protocols. This is usually the case while integrating expression data from single commercial microarray platform which typically use standard sample processing protocols.

The difficulty in sample annotation integration lies in annotation format inconsistencies. Ideally all sample annotations are organised around single well defined format or ontology and use controlled vocabulary. The real world sample annotations, for example from GEO, however, tend to have a large free-text component and loose structure which makes automated sample manipulations and grouping difficult or even impossible. In such cases, sample annotation integration becomes a problem of natural language processing and text mining. The process of sample integration is a problem of sample re-annotation. In previous work, large scale sample re-annotation has been carried out by the help of the lexical sources such as Unified Medical Language System (UMLS) and biomedical ontologies [BK06][FLL+09] while most microarray meta-analysis projects, including [DCD+07][KAO+08][PKK+09] use manual re-annotation.

Sample meta-data integration for Paper VI was carried out semi automatically by means of initial sample re-annotation for microarray experiments retrieved from GEO was carried out by a specially implemented text mining system. The initial annotation was then followed by additional manual curation and polishing. The goal for the text mining was to convert the largely free text GEO annotations to the sample annotation format used by ArrayExpress. Sample annotations in ArrayExpress are presented in (category value) pairs with annotation categories originating from MGED

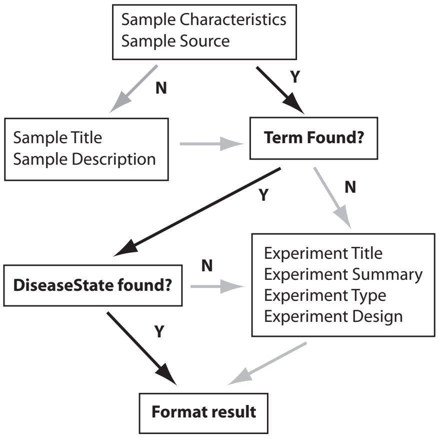

Figure 3.1: Decision tree from the GEO sample meta-data re-annotation pipeline in Paper IV. For each sample category, keywords are first searched from the more informative sections of the annotation text. In case no keywords are found (N), the search is extended to the less specific parts of sample annotations and eventually to the experiment annotations.

ontology where each category has a strict meaning. Category *Organism-Part*, for instance, would refer to anatomical parts of the organism further specified by category *Organism.* The text mining approach in Paper IV and VI use for each ontology category a separate dictionary of keywords and keyword matching regular expressions. The dictionaries of keywords and regular expressions are loaded to a finite state automata which is used to scan the free text annotations and mark all identified keywords by the category to which the particular keyword belongs to. As for each category, the annotation text may contain several different keywords, we implemented a decision tree which first looks for keywords in these parts of annotation text where they are likely to describe the biological sample best (Figure 3.1). In case sample annotations do not contain any keywords, the search is extended to the description of the experiment where the sample belongs to because most of the experiments focus only on small number of similar sample types.

Manual evaluation of the text mining performance showed satisfactory quality for 68% of the assessed samples. The need for perfect sample quality forced us to re-check the quality of all re-annotated samples. The speed of the process of overall re-annotation was, however, significantly boosted by usage of text mining.

Mining and analysis of any large expression data set is more efficient if the relationships and hierarchy between all involved samples are defined. The concept of ontology or ontology like relationship usage for microarray sample annotations, however, seems to be relatively new with [FLL$^+$09] and [KEH$^+$09] being two of few. The need for sample hierarchy was also recognised in Paper VI and due to lack of any appropriate ontology at the time, was created semi-automatically from available sources of biological keywords such as NCI Thesaurus. The appropriate ontology which could be used in further similar integration projects is the recently created Experimental Factor Ontology (EFO) [MRBP08]. EFO is designed for annotating microarray experiment samples and has already become important part of ArrayExpress Atlas [KEH$^+$09]. Usage of these types of ontologies is essential for large scale microarray data integration, meta-analysis, reannotation, grouping and data mining.

3.3 Quality assessment of microarray data

The microarray data quality is often referred to as the quality of expression measurements. The public microarray data repositories promoted by MGED society, on the other hand assess the microarray data quality mainly on the level of meta-data. More explicitly, the data submitted to the repositories is checked manually for MIAME compliance and accepted only on sufficient level of annotation. Differences in data formats and curation practice between the data repositories has lead to a condition where a dataset considered fully MIAME compliant in one of the repositories may not qualify as one in others. ArrayExpress is currently the first and only public microarray database to score the submitted and integrated experiments based on their closeness to full MIAME compliance.

Ideally, the quality of a microarray experiment should be a measure of the quality of meta-data (sample annotations and experimental protocols) as well as the quality of quantified expression measurements. The quality assessment of expression measurements depends on particular array platform and often requires methods specific to particular array platform. The quality of the expression measurements should be checked both within hybridisation and between series of hybridisations that form the experiment or study. The aim for the first is to detect technical artefacts and biases within hybridisation while the second focuses on comparability of individual hybridisations within a study.

Microarray data quality has been assessed by the US Food and Drug Administration initiative Microarray Quality Control (MAQC) and by European Commission Initiative Empowering the Microarray-Based European Research Area to Take a Lead in Development and Exploitation (EMERALD). Even though both initiatives have commonalities, the main focus of MACQC project is to develop standard and quality metrics specific for the use of microarray and next-generation sequencing in clinical practice. The aim of the EMERALD project is more general and focuses on overall microarray data improvement through introduction of best laboratory practices and establishment of disseminate quality metrics. One of the results from the EMERALD project is the creation of arrayQualityMetrics software package [KGH09]. The package is the first to generate universal microarray data quality reports for the expression values created on the majority of the microarray platforms.

3.3.1 Quality assessment of Affymetrix data

The quality measures developed for Affymetrix arrays focus mainly on identification of outliers in a set of arrays. The microarray data quality as a term in this thesis is relative and dependent on the context where the data is analysed. More precisely, an array may be an outlier and low quality in one set of arrays while in the other set it may have a good quality depending on how comparable the computed quality metrics for all of the arrays in these sets are.

The typical first step in the quality analysis of single arrays is to examine the raw probe-level images for spatial artefacts. The analysis is usually carried out on image plots with logarithmically transformed intensities. For large studies with hundreds of arrays, examination of individual arrays may not always be feasible and arrays with spatial artefacts are detected in a set of arrays by other means [BCB$^+$05].

The basic multi-array quality measures for Affymetrix arrays include analysis of probe level data by boxblots, histograms and MA-plots. The boxblot gives a summary of the distribution of probes while the histogram displays the density functions of raw log-scale probe intensities. Arrays with significant differences in these plots may be problematic.

Historically MA-plots are used to visualise intensity dependent ratio of raw microarray data in two channel microarrays. In single-channel arrays, MA-plots are either computed for all array pairs in the set or each array is plotted against a common pseudo array constructed from all arrays in the set. The M and A values are computed for every measured gene as follows

$$M = log_2 R - log_2 G$$

$$A = \frac{1}{2} * (log_2 R + log_2 G)$$

where R and G correspond to either intensities of the gene in different channels of a two channel microarray or intensities of the gene in two different single channel arrays. The MA-plot uses M in y-axis and A in X-axis. The scatter plots are further improved by fitted lowess curves to summarise any non-linear relationships. Arrays with apparent quality problems have oscillating lowess smoother, or different variability in the M values compared to the other arrays.

3.3.2 Affymetrix standard quality metrics

The standard Affymetrix quality metrics used in this thesis include so called Average Background, Scale Factor, Percent Present and computation of $3'/5'$ probe set ratios of RNA quality control genes to measure the RNA degradation. According to Affymetrix guidelines, the Average Background values in a set of arrays are recommended to be comparable while the Scale Factors should lie within 3-fold. The Percent Present values for all arrays should be similar with extremely low values being a possible indicator of poor quality. The $3'/5'$ probe set ratios for RNA degradation are computed for a set of RNA quality control genes, such as β-Actin and GAPDH, each represented by 3 probe sets, one from the $5'$ end, one from the middle and one from the $3'$ end of the transcript. The safe threshold for arrays comparable by $3'/5'$ ratios suggested by Affymetrix is 3.

The $3'/5'$ ratios in standard Affymetrix quality control are measured only for few genes and can be quite variable even in high quality data. Given that probe sets on the Affymetrix array contain 11 and more probe pairs which are spaced equally over the last 600 base pairs of the $3'$ end of the transcript, more global indicators of RNA degradation could be implemented. The additional RNA degradation measures available for the Affymetrix arrays include computations of probe position means and RNA degradation slopes. The suggested average RNA degradation slopes vary between different array types but are recommended not to exceed the average more than factor of 2.

3.3.3 The probe level model derived quality measures

The probe level quality metrics for Affymetrix microarray data are built around the probe level model (PLM) used by RMA normalisation method. The probe level model is a linear model for the background adjusted normalised probe-level data Y_{gij}

$$log(Y_{gij}) = \hat{\theta}_{gi} + \phi_{gj} + \epsilon_{gij}$$

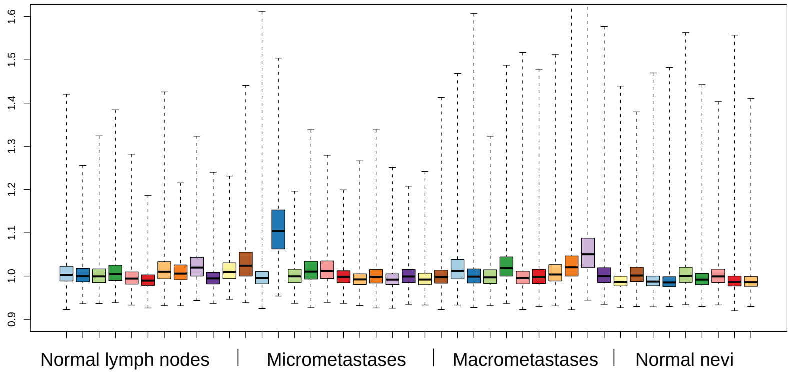

Figure 3.2: NUSE plot of expression data from Affymetrix U133A arrays in Paper II. The third array in a set of micrometastatic samples is an outlier with the box-plot not being comparable with other arrays.

where $\hat{\theta}_{gi}$ represents the log scale estimate of the expression level for gene g on array i, ϕ_{gj} is the effect of the j-th probe representing gene i, and ϵ is the measurement error. The $\hat{\theta}$ in this formula is estimated robustly by median polish and is used for computation of Relative Log Expression (RLE) and Normalised Unscaled Standard Error (NUSE) quality metric plots.

The boxplot of RLE displays relative expression

$$M_{gi} = \hat{\theta}_{gi} - m_g$$

of each gene g on each array i where m_g is the median expression value of gene g across all arrays. The boxes for the arrays on the box plot represent genes which have very little difference in gene expression between the arrays. In many situations the majority of genes between the arrays do not change and the boxplots should have small spread centred around $M = 0$. An array with quality problems may have a box with greater spread or is not centred near 0.

NUSE values for genes g on arrays i are computed as

$$NUSE(\hat{\theta}_{gi}) = \frac{SE(\hat{\theta}_{gi})}{med_i(SE(\hat{\theta}_{gi}))}$$

where SE is the standard error and $med_i(SE(\hat{\theta}_{gi}))$ is the median of standard errors computed for array i. Similarly to RLE values, NUSE values are shown with a box plot for each array. Lower quality arrays are likely to have significantly elevated or more spread boxes (see Figure 3.2). The suggested good single number summary for the method is median NUSE of the array.

Both RLE and NUSE produce quality metrics relative to the arrays used in the computation and are not comparable across different datasets. The advantage of the PLM based quality metrics over the Affymetrix standard methods is that they are directly related to the quality of the expression measures.

3.4 Integration and data quality

The role of quality assessment introduced in the previous section is to identify individual hybridisations with problematic quality and assure overall data comparability. The quality measures are used to evaluate and, if necessary, remove some of the data to avoid low quality and data incomparability driven biases in consecutive data (meta-)analysis. The importance of systematic quality assessment in data integration is even higher if the data is retrieved from different sources and has no warranty for either quality or global level comparability. Surprisingly, data quality assessment prior to integration is scarce. The extreme rareness of pre-integration quality checks may partially be explained by difficulties assuring uniform quality in cross-platform data integration, which compared to single platform approaches have been more popular. We therefore considered the large scale systematic quality assessment of microarray data prior data integration in Paper VI a novelty.

The methods used for quality assessment in Paper VI are summarised in the section above. All included methods are available in R Bioconductor and were applied, when ever possible, for each expression profile separately. Probe level models RLE and NUSE, however, require memory intensive simultaneous processing of the complete data. In computations for Paper VI, we were limited to 128 gigabytes of random access memory which, given the size of the data, was for simultaneous RLE and NUSE assessment insufficient. Hence, a heuristic was used to solve the problem. The maximum number of arrays, given the amount of memory, that we were able to process at one time was around 2000. We hypothesised that a sufficiently large random sample from all arrays should closely approximate RLE and NUSE value distributions of all of the arrays. A similar assumption for probe level model based normalisation has been suggested by [KIL+06] and practiced in [DCD+07]. RLE and NUSE values for the data in Paper VI were therefore computed in 5 randomly assigned batches of 1800 arrays.

An important question in quality assessment is how the quality is de-

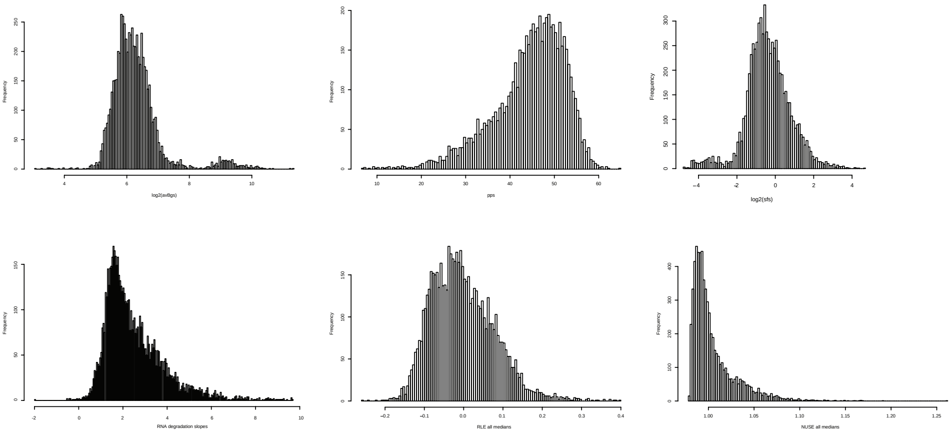

Figure 3.3: Histograms of values obtained for quality metrics in Paper VI. The top row from the left contains histograms for AverageBackground, Percent Present and Scale Factor values while the bottom row from the left contains histograms for RNA degradation slope, RLE median and NUSE median values.

fined. In Paper VI, the quality is firstly defined by ranges of quality metric values suggested by [BCB$^+$05]; and secondly, by selection of ranges for each quality metric such that the number of good quality arrays according to all metrics would be maximised (Figure 3.3). The range cut-offs for each metric were selected manually and the final set of good quality arrays were retrieved by intersection of lists of good quality arrays of each metric.

3.5 Factors proposed to affect the quality of microarray data

Several factors affecting the quality of microarray data have been suggested. The quality of microarrays that use Cy5 label have, for instance, been shown to fluctuate together with atmospheric ozone levels [FCD$^+$03]. The quality of Affymetrix arrays has been demonstrated to correlate with the time measured from the last scanner service [USGR$^+$09]. However, most importantly, the analysis results of integrated data were shown to be influenced by so called lab or batch effect which is expressed by higher correlation between microarrays processed by one lab or in a batch [IWS$^+$05] [JLR07]. The effect was also studied in Paper VI by comparison of similarities be-

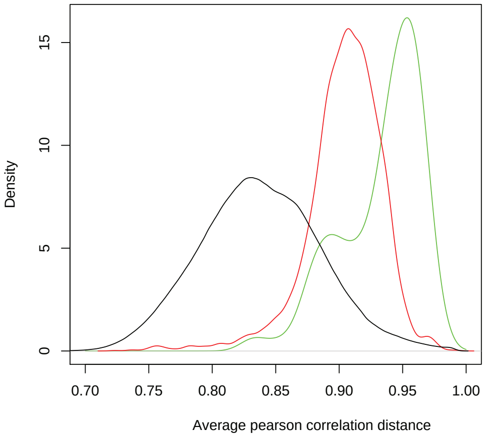

Figure 3.4: Distribution of pairwise correlations between all expression profiles in Paper VI (black), distribution of average similarities between expression profile subgroups from different laboratories within the same biological group (green), distribution of average similarities between expression profile subgroups from different biological groups within the same laboratory (red).

tween expression profiles measured by laboratories that processed at least two different types of biological material and within expression profiles of a same biological material which came from at least two different laboratories (Figure 3.4). It has been demonstrated for example by [IWS+05], that the lab effect in Affymetrix data is much smaller than in two colour cDNA microarray data. The analysis of the data in Paper VI showed that the lab effect may be significant but can be removed at least partially by strict quality control and data comparability. For instance, in Paper VI, the quality assessment lead to almost complete exclusion of initially included Novartis human gene expression atlas data [SWB+04] which prior to global quality assessment clustered together regardless of the diverse content of its biological samples.

Chapter 4

Microarray data analysis

The aim of microarray data analysis is to gain new information about the properties of the biological material and genes which expression is measured. Traditionally this is achieved by grouping of the genes and samples based on expression similarities and identifying genes which best distinguish the groups of interest. This chapter discusses the methods that were used in microarray data analysis in Papers I, II and VI. Section 4.1 introduces a few distance and correlation measures that were used for data grouping and clustering. Section 4.2 gives a brief introduction to different data clustering methods. Section 4.3 focuses on methods used for identification of differentially expressed genes, while section 4.4 gives a very brief introduction to Principal Component Analysis (PCA).

4.1 Distance and correlation

In the gene expression matrix, each sample and gene could be seen as a vector of many dimensions or a point in a higher dimensional space. In linear algebra, the distance d between two vectors p and q of length n is computed as

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

and is known as Euclidean distance. The closer the two points in the Euclidean space, the smaller the distance and vice versa. Euclidean distance is a good measure when the data is standardised by the mean and variance. For other types of data, the correlation-based metrics are likely to perform better.

In Euclidean geometry, the dot product, length and angle of the vectors are related and the angle between two vectors x and y is computed as

$$\theta = arccos(\frac{x \cdot y}{|x||y|}).$$

The cosine correlation distance (or uncentered correlation coefficient) considered in the early microarray studies [ESBB98] equals of the cosine of the angle of the vectors x and y:

$$d_{COR}(x,y) = \frac{x \cdot y}{|x||y|} = \frac{\sum_{i=1}^{m} x_i y_i}{\sqrt{\sum_{i=1}^{m} x_i^2 \sum_{i=1}^{m} y_i^2}}$$

The widely used Pearson correlation coefficient [Bra03]

$$d_{PEAR}(x,y) = \frac{\sum_{i=1}^{m}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{m}(x_i - \bar{x})^2 \sum_{i=1}^{m}(y_i - \bar{y})^2}}$$

could be seen as a more general form of the equation by being mean centered. The cosine correlation would then be a special case of Pearson correlation with $\bar{x}$ and $\bar{y}$ both replaced by zero.

Similarly to other correlation measures, the Pearson correlation is invariant to location and scale transformations and sensitive to linear relationships between variables. However, the Pearson correlation has tendency to be adversely affected by the outliers in which case non-parametric correlation measures such as Spearman rank correlation coefficient would be preferred.

A non-parametric Spearman rank correlation coefficient [Bra03] is summarised as

$$\rho = d_{SPEAR}(x,y) = \frac{\sum_{i=1}^{m}(x_i' - \bar{x'})(y_i' - \bar{y'})}{\sqrt{sum_{i=1}^{m}(x_i' - \bar{x'})^2 \sum_{i=1}^{m}(y_i' - \bar{y'})^2}}$$

where $x_i' = rank(x_i)$ and $y_i' = rank(y_i)$. The measure differs from Pearson correlation mainly by the operation in rank space which makes it invariant to monotone transformations.

Microarray data is in general high dimensional data and suffers from the curse of dimensionality [KKZ09]. In the context of distance and correlation measures, the curse of dimensionality is expressed in less precise distance concept as the number of dimensions grows.

4.2 Clustering methods

Clustering is a method of unsupervised learning with the aim to assign a set of observations into subsets called clusters based on some similarity measure. The clustering algorithms can be either hierarchical or partitional.

4.2.1 Hierarchical clustering

Hierarchical clusters can be built successively either bottom-up (agglomerative) [ESBB98] or top-down (divisive) [ABN$^+$99].

Agglomerative hierarchical clustering starts by assigning observations to separate clusters of one. The algorithm continues by computing distances between all clusters and merging the two closest. Re-computation of distances and merging of the two closest clusters is repeated until either all of the clusters are merged into one, a threshold of maximum distance, or a preferred number of clusters is reached.

As the merged clusters contain more than one observation, different methods exist how the distance or similarity between the groups could be computed. The most well known of these include:

1. The minimum distance between observations of each cluster, known as single linkage.

2. The maximum distance between observations of each cluster, known as complete linkage.

3. The average distance between observations of each cluster, known as average linkage.

In the context of microarray data, single linkage performs well in cases of strong natural classes with irregular shapes. However, the method tends to create chained or sticky clusters and therefore is not recommended for microarray data analysis [Ste03]. The advantage of complete linkage is production of small and compact clusters where the natural clusters in the data are well defined. The method does not work well for fuzzy data. The average linkage is the computationally slowest but is less influenced by random noise and experimental error.

The disadvantages of agglomerative hierarchical clustering include difficulty to define distinct clusters, sensitivity to strong clusters which may affect the rest of the results and unoptimised ordering of the branches of the clustering tree.

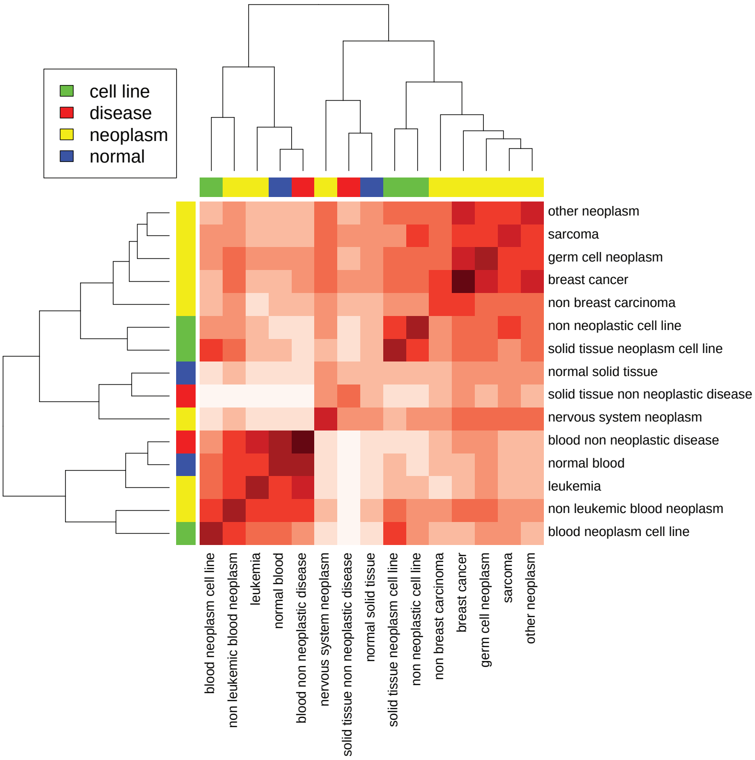

Figure 4.1: Hierarchical clustering of 15 groups of biological samples in Paper VI. The average Pearson correlation distances between groups is visualised by heatmap.

The agglomerative hierarchical clustering is often one of the first steps in the microarray data analysis and a valued method for data visualisation. The results from hierarchical clustering are displayed as dendrograms which illustrate the computed hierarchy between the genes (probes), samples (hybridisations) or both. The clustering dendrograms are frequently combined with data heatmaps. Heatmaps are graphical representations of numerical matrices where the values are transformed to colour intensities (Figure 4.1).

Dendrograms and heatmaps are practical for visualising no more than tens to hundreds of samples or genes. Larger number of tree branches and data points make the interpretation of the results difficult not only because relationships between tree branches and data points in a large tree and heatmap are difficult to grasp but also because of unoptimised ordering of the tree branches. In paper VI, this problem was solved by

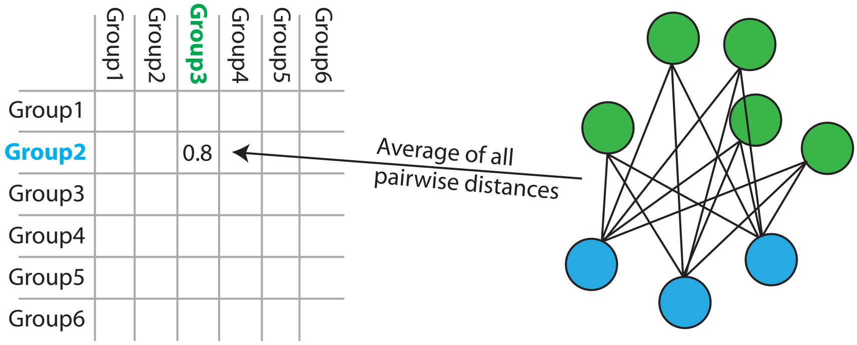

Figure 4.2: Heuristic for hierarchical clustering of large data matrices used in Paper VI. Based on biological annotations of the samples, the data is divided into sample groups which are assumed to be in average more similar to each other than to samples in other sample groups. An input for hierarchical clustering is the similarity matrix of the sample groups where each value is an average of all pairwise correlation distances between particular two groups.

following heuristic. The data (expression measurements from all samples) was first collapsed to groups of samples of same annotation. A similarity distance matrix of sample groups was computed and assigned to hierarchical clustering (see Figure 4.2).

4.2.2 K-means clustering

K-means clustering [Har75] is a partitional clustering method where differently from hierarchical clustering, all of the clusters are determined at once.

K-means clustering groups the data into k independent clusters. The algorithm starts by either randomly or directly assigning the k clusters and cluster centres (centroids) which are averages of all points in the cluster. In the next step all of the points in the data are re-assigned to the nearest cluster centres. Based on the new assignments, the cluster centres are recomputed, and the procedure of assigning all of the points to the nearest centroids and re-computation of the centroids is then repeated until some criterion of convergence is met. A typical criterion of convergence is that the content of the clusters between two runs of centroid re-computation does not change.

The advantage of K-means clustering is its simplicity, the speed of computation and ability to pre-define the number of clusters. The main disadvantages are the difficulty to find good value for k and that the results from any two clustering runs may not be identical as the initial assignment of centroids is random.

K-means clustering with different k was used initially for data exploration in Paper VI but was discarded after similar clusters with better data separation were observed in Principal Component Analysis.

4.3 Differential expression

For microarray studies with measurements from two and more different biological groups, it is natural to study how much the groups differ and which genes between the groups drive the differential expression. This is traditionally done by either Student's t-test or by its derivatives or related alternatives.

4.3.1 Student's t-test

The student t-test is a parametric statistical test used for testing if the means of the values obtained from two normally distributed populations are equal under null hypothesis. The basic t-test equation:

$$t = \frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2 + \sigma_2^2}}$$

uses the means μ and standard deviations σ of groups 1 and 2 and assumes equal group sizes and variances. In case these assumptions are not true, an alternative equation, known as Welch's t-test [Wel47],

$$t = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

could be used. The standard deviations of the groups in Welch's t-test are corrected by the sizes of the groups n_i. Depending on the data properties, other types and variants of t-test exist and could be applied. A non-parametric alternative for t-test of unpaired groups is Mann-Whitney U test [MW47] while Wilcoxon rank-sum test [Wil45] assumes equal sample sizes.

4.3.2 T-test derivatives

For microarray data analysis various modifications and adjustments of the original t-test exist. One of these, Signal to Noise metric (S2N), introduced by Golub and Slonim et al [GST+99] [RRLG03]:

$$d_{S2N} = \frac{\mu_1 - \mu_2}{\sigma_1 + \sigma_2}$$

modifies the denominator of the standard t-test to be a sum of groups standard deviations. Modified denominator is also used by Significance Analysis of Microarrays (SAM)[TTC01]. In SAM, the relative difference between two sample groups is

$$d_{SAM} = \frac{\mu_1 - \mu_2}{s + s_0}$$

where s is the standard deviation and s_0 a small positive constant. The standard deviation is defined as

$$s = \sqrt{\frac{\frac{1}{n_1} + \frac{1}{n_2}}{(n_1 + n_2 - 2)} \sum_m [x_m - \bar{x}_1]^2 + \sum_n [x_n - \bar{x}_2]^2}$$

and the small positive constant s_0 is chosen to minimise the coefficient of variation.

The parametric test used for cases of more than two sample groups is the analysis of variance (ANOVA) [KMC00]. ANOVA is a statistical test used to test the heterogeneity of group means via analysis of the group variances. The test can be seen as generalisation of Student's t-test for more than 2 groups.

S2N and SAM as well as ANOVA were used in the analysis of the data in Paper II.

4.3.3 Permutation test

In Student's t-test, the p-value is obtained from the t distribution. As the underlying distribution of values can not always be assumed, other means for obtaining the p-value exist. Permutation test [Goo00] is a type of non-parametric significance test where the reference distribution is obtained by calculation of all possible test results by re-arrangement of the group labels of observed data points. In case of t-test and its derivatives, the metric distribution is obtained by randomly assigning the data points into two test groups. The p-value is computed as

$$p - value_{perm} = \frac{\#observed}{\#expected}$$

where *#observed* is the number of permutation results where metric value is at least as extreme as in the original test and *#expected* is the number of permutations.

4.3.4 Multiple testing

An outcome of a statistical test is called significant if the result is unlikely to have occurred by chance. In statistical terms it means comparison of the test result p-value with the specified significance level $0 \leq \alpha \leq 0$ and rejection of the null hypothesis in case $\alpha >$p-value. Typically α is chosen as 5% (0.05) or 1% (0.01).

In multiple testing [BY95] where certain statistical test is executed many times over different parts of the dataset, the standard values of significance level α, become problematic. This is because larger number of executed tests for constant α naturally increases the number of expected "significant" results. Several methods for the α correction in multiple testing have been proposed, from which Bonferroni correction [Mil81] is the simplest and most widely used.

In Bonferroni correction, significance level α is assigned to a set of a size of n statistical tests. The significance level β of any single test in a set is computed as $\beta = \alpha/n$.

A more powerful Holm–Bonferroni [Hol79] correction adjusts the significance levels of individual tests depending on how extreme each particular p-value in a set is. The significance for the smallest p-value is adjusted to α/n, the second smallest $\alpha/(n-1)$, and so on according to

$$\beta_i = \frac{\alpha}{n-i}$$

where i is the position of p-value in the extremeness scale of a set.

Other multiple testing correction methods such as Benjamini-Hochberg [BY95] exist.

4.4 Principal component analysis

Principal Component Analysis (PCA) [Jol02] is a non-parametric method for orthogonal linear transformation of data into a new coordinate system. The transformation is carried out such that the greatest variance by any projection of the data is captured by the first coordinate known as the first principal component, the second greatest variance on the second coordinate known as second principal component, and so on.

Principal component analysis can be summarised by equation

$$Y = PX$$

where X is the data matrix of m variables in rows and n measurements in columns and Y is the re-representation of the data X through trans-

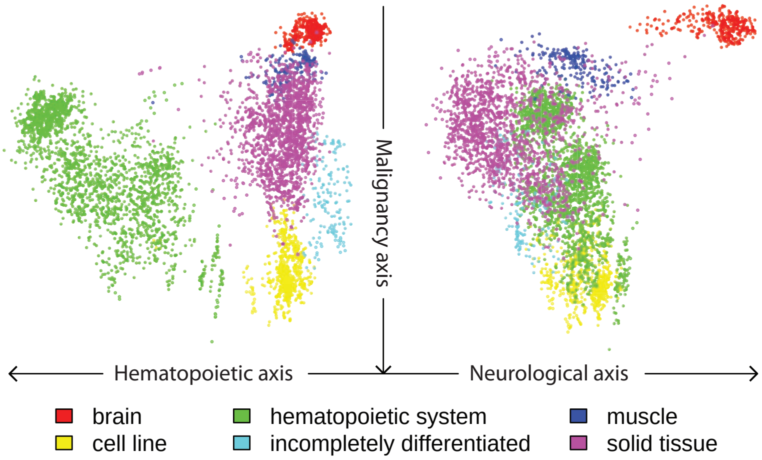

Figure 4.3: Principal component analysis of the data presented in Paper VI. Each dot represents a sample from 5372 samples projected on the plane of first and second (left), and second and third (right) principal components. The first principal component separates hematopoietic cells from solid tissues and was named hematopoietic axis. The second component separates normal and diseased tissues, neoplastic tissues and immortal cell lines and was named the axis of malignancy. The third component separates nervous system related samples from the rest.

formation with matrix P. In the analysis, the orthonormal transformation matrix P is selected such that the covariance matrix C_Y for Y is

$$C_Y \equiv \frac{1}{n-1} Y Y^T.$$

Rows of matrix P are principal components of X. Geometrically, P is a rotation and a stretch of X that results Y.

PCA relies on following assumptions:

- Linearity of the data.
- Statistical importance of mean and covariance as there is no guarantee that the direction of maximum variance will contain good features for discrimination.
- Large variances have important dynamics while smaller variances represent uninteresting components and noise.

Principal component analysis is used for dimensionality reduction and recognition of the few most variable components in the data. As the less variable components tend to be highly correlated and have minimal contribution to the total variance, these components may be dropped with the minimal loss of information.

In context of microarray data, PCA is used both for data analysis and visualisation (Figure 4.3). In studies with large number of measurements the global differences between biological samples are often visualised by the projection of pairs of first few principal components on planes or in three dimensional space [DCD+07][KAO+08][SBD+05][GBL+08].

Chapter 5

Analysis of human melanoma data

Melanoma is a malignant tumour of melanocytes [Ali02] [CFBM69] [Bre70]. It is mainly found on skin, but also in the eye (uveal melanoma) and bowel. Melanoma is more common among Caucasians living in sunny climates and in females. The number of melanoma related death is, however, higher in males [PBFP05]. Melanoma is one of the less common types of skin cancer but responsible for majority of skin cancer related deaths [LMSA06]. The prognosis with regional metastasis is poor and occurrence of distant metastasis refers to a largely incurable disease [LMSA06] [Rus00]. The mechanisms of melanoma metastasation are not clear making early diagnosis vital for patients.

This chapter is an introduction to Papers I and II which contribute to the research of melanoma progression and diagnostics. Section 5.1 provides outline for the sentinel lymph node based melanoma diagnostics. Section 5.2. describes the data preparation for the analysis in Paper II and Section 5.3 will give a rationale for the usage of fold change method in determination of the differentially expressed genes in Paper I.

5.1 Sentinel lymph node based melanoma diagnostics

Since the formal acknowledgment in 1840, it generally still holds true that the chance of survival from melanoma is dependent upon the early removal of the disease by operation [Coo40]. In 1892, Herbert Snow introduced an opportunity to improve the survival of melanoma patients by wide excision and elective lymph node dissection [Sno92]. The essence of this method was to control the lymphatic permeation of melanoma metastases. One hundred years later, in 1992, dissatisfaction with the procedure lead to a development of a new, finer technique, known as intraoperative lymphatic mapping and sentinel lymphadenectomy [MWW+92] [CWM92]. The lymphatic mapping exploits the hypothesis that the dermal lymphatics provide direct connection from the primary melanoma to regional lymph field. The sentinel lymph node, the closest to the primary site of melanoma, is hence the first regional site for melanoma metastasis. Sentinel lymphadenectomy, meaning systematic removal of sentinel lymph nodes, is practiced as a preventive measure to stop disease progression while also providing tissue material for more accurate disease staging.

Despite the controversy about the benefits of sentinel lymphadenectomy criticised for example by [Tho05] [Tho07] and existing clinical alternatives [VvASH+09], histologic sentinel lymph node examination has become a standard of care [Cas99].

5.2 A global search of metastatic melanoma specific marker genes

Histologic examination of sentinel lymph nodes is time consuming and error-prone. Hence, alternative RT-PCR based assays for melanoma cell detection in sentinel lymph nodes have been suggested. The goal in Paper II was to construct such an assay using a list of metastatic melanoma specific genes identified by a transcriptome wide gene expression study. The transcriptome wide coverage in Paper II was achieved by usage of two Affymetrix arrays, HG-133A and HG-U133B, known as HG-U133Set. The material from studied biological material was hybridised on both of the arrays, which in the light of integration issues discussed in Chapter 3, was a source of problems in the analysis.

It is important to note that the data available for the analysis were MAS5 intensities from Affymetrix CHP files. The main questions prior to analysis were: 1) if the data should be somehow pre-processed and 2) if the analysis should be carried out on both arrays separately or the data from two array platforms should be first integrated. Given that the material hybridised to HG-U133A array (A array) had a fixed volume while material hybridised on HG-U133B array (B array) had a variable volume, depending on how much material was left over from the hybridisation to A array, it was hypothesised that observed expression values in B arrays may be volume biased. As the A and B arrays share a set of 100 identical control probe sets, we hypothesised that these could be used in detection and correction of the possible material volume induced expression biases in B arrays. The bias was estimated and corrected by computing a ratio of mean expressions

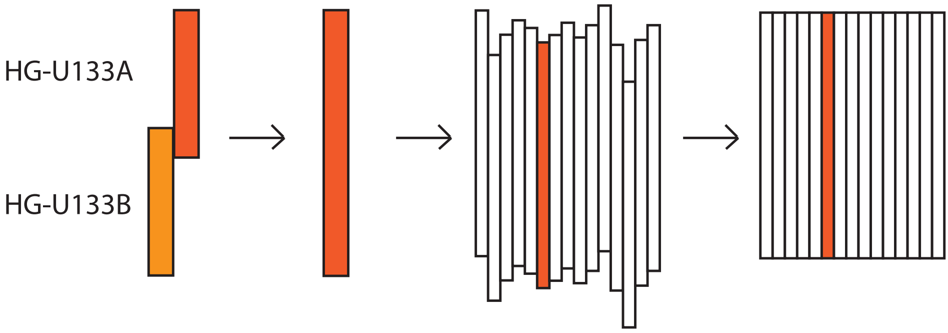

Figure 5.1: Pre-processing of the microarray data in Papers I and II.

of shared probe sets in A and B arrays and multiplying the values in B arrays by the computed factor. The adjusted expression values of B arrays were then merged with the data from A arrays into joint virtual AB arrays and the data from all virtual AB arrays was combined to one data matrix (Figure 5.1).

We also realised that depending on the amount of material hybridised to B arrays, the scales of A arrays and the combined virtual AB arrays may need further adjustment. It is reasonable to assume that the amount of mRNA produced in an average cell or tissue type are comparable. Hence, the adjustment of the scales of individual arrays to the same level is justified. The scales of the AB arrays in the joint data matrix in Paper II were equalised by quantile normalisation for which the method was implemented in C according to [BIAS03].

The methods and measures used for identification of differentially expressed genes for the construction of PCR assay in Paper II have already been covered in Chapter 4. However, the final selection of genes for the PCR assay was performed considering statistical significance as well as fold change and biological significance.

5.3 For the justification of fold change usage

Genes tenascin-C, fibronectin, and procollagen-I in the centre of Paper I were identified in microarray data as differentially expressed using the fold change.

The microarray literature knows two different definitions of fold change.

The fold change F_{ratio} of a gene i is usually referred as

$$F_{ratio} = \frac{\bar{x_i}}{\bar{y_i}}$$

where x_i and y_i are the expression levels of the gene in two groups of samples, while fold change in [CBM+05] and [GLW+06] has also been referred as

$$F_{diff} = \bar{x_i} - \bar{y_i}.$$

Importantly, depending on the definition of the signal, for instance in log space $log(a/b) = log(a) - log(b)$, the fold changes of both measures may be the same. In Papers I and II, the fold was defined as F_{ratio} even though for manuscript preparation, both F_{ratio} and F_{diff} were computed.

The use of fold change among biologists has been criticised and the usage of more sophisticated methods for identification of differentially expressed genes, such as t-test and its derivatives, have been promoted. A problem in using fold change is the difficulty in defining if genes of certain fold are significantly differentially expressed as the widely used two or three fold significance cut-off levels are mere suggestions. However, it has been shown that fold change in selection of differentially expressed genes not only performs well but has very high rate of reproducibility ($> 90\%$) while usage of t-test derived p-values correlates with reproducibility [GLW+06] [STF+05] [SRJ+06] [PLFS+06]. A positive correlation between higher fold change and better reproducibility has been observed, while results from smaller fold changes are less likely to reproduce well. Furthermore, the impact of the normalisation methods on the reproducibility of the gene lists in fold change is minimal compared to p values and global scaling of the data differently from p values does not influence the results. While comparing three methods: t-test, SAM and fold change [GLW+06], the fold change from the three in finding differentially expressed genes is reportedly superior; though concordance based on SAM was clearly improved over that of the simple t-test. Witten and Tibshirani [WT07] have additionally studied the matter, both on simulated as well as on real data, and showed that SAM and fold change clearly outperform the simple t-test, and that the behaviour of two versions of fold change are quite similar. They also point out if absolute changes in expression are relevant, like for instance in Paper I, fold change should be used. Based on this evidence, MAQC suggests fold change assessment together with a non-stringent p-value cutoff [SRJ+06]. We therefore found the usage of fold change in papers I and II justifiable.

Chapter 6

Summary and Conclusions

In this thesis we constructed a global map of human gene expression and studied its structure. We also identified genes which are involved in human melanoma metastasation, carried out genome-wide search of metastatic melanoma specific genes in the context of normal lymph tissue and demonstrated that found genes work well in a clinical assay in detection of melanoma metastasis in sentinel lymph nodes.

The three groups of computational problems assessed in this thesis were problems with DNA microarray data retrieval and formats, problems with microarray data integration, and problems in microarray data analysis and visualisation.

Construction of a global map of human gene expression required a vast amount of gene expression data. The source of these data are public microarray data repositories, such as GEO and ArrayExpress. The data used in this thesis originated from these two databases. Retrieval, processing, re-formatting and cleaning of these data from both of the databases, but especially from ArrayExpress, both computationally as well as manually, was perhaps the most time consuming part of the work invested to this thesis. Even though the access to the data in ArrayExpress has been significantly improved, a large scale microarray data retrieval and local organisation will remain an exercise of scripting and local data organisation even in the future.

The challenge in microarray data integration assessed in this thesis was in assurance of integrity and comparability of both numerical expression values as well as the accompanying microarray meta-data. The general approach taken was new even though most of the applied methods were well known. The consistency of numerical expression measurements was ensured by computation of series of data quality metrics and selection of expression profiles with quality metric values within a range of comparable hand picked metric cut-offs. A possible improvement here would be a selection of ranges of cut-offs for each metric computationally e.g. by dynamic programming as it would assure the maximisation of the the amount of comparable expression profiles globally.

The problem in the microarray meta-data integration was identification of information rich keywords in GEO free text sample annotation records and converting these into MGED Ontology compatible microarray sample annotations. This was solved by creation of custom ontology of MGED Ontology derived sample categories and lists of keywords and regular expressions for each category. The keywords were searched in GEO free text sample and experiment entries, categories were assigned according to identified keywords and keyword category pairs were then made subject of either selection or rejection depending on their context of occurrence. The main weakness of this approach was the incompleteness of the ontology and in some cases lack of descriptiveness of the original annotation text. While for the latter there is no cure, the approach may be improved by usage of now available Experimental Factor Ontology [MHA$^+$10]. This ontology is developed by a team of dedicated ontologists especially for managing microarray sample annotations and is therefore likely to produce better results with less manual curation required in the end.

Microarray data analysis is an active field of research with often more than one method available to solve any one type of problem. Which of the methods to choose is, however, not always clear. This is also reflected in this thesis. For instance, the lists of differentially expressed genes in the included papers were computed by fold change, SAM, and limma [Smy05]; with t-test, S2N and other methods considered and practiced in parallel. In each case, the decision of which method or methods to use use was driven by authors confidence for the deliverance of most accurate results.

The arsenal of traditional microarray data partitioning and visualisation methods is, however, more limited. The methods of choice include hierarchical agglomerative clustering, heatmaps, and K-means clustering, while PCA, Self Organising Maps (SOM) [Koh95] and other methods are used less frequently. In this thesis, partitioning and visual exploration of the data of a human map of gene expression by hierarchical agglomerative clustering proved a challenge. While majority of the microarray studies apply hierarchical clustering from tens to hundreds of samples, our data was captured in a data matrix of 5300 x 22000. A question if usage of hierarchical clustering under such circumstance would be practical and justified was asked. Our experience, however, showed that usage of other methods, such as NeRV [VK07] or Principal Component Analysis alone did not offer the amount of exploratory flexibility and detail that one would have anticipated. Hence the usage of heuristic to collapse the data by another data derived sample ontology and still use the agglomerative hierarchical clustering, now on a matrix of average pairwise correlation distances of ontology defined sample groups. The clustering could have been carried out also differently, for instance by implementation of hierarchical agglomerative clustering algorithm where the inputs are the full data and pre-defined ontology derived sample groups. In such implementation, the clustering would start from pre-defined sample groups and data collapsing prior to clustering would not be necessary.

- - -

In summary, this thesis contributed to the fields of bioinformatics and computational biology by the tools developed; improved data display and access to one of the largest public microarray data repository; and by new techniques and heuristics for large scale microarray data integration and analysis. This thesis also contributed to cancer and systems biology by its new findings reported by included papers. This thesis demonstrated the richness of data accumulated in public microarray data repositories and showed one way of unlocking it. This thesis also demonstrated computational difficulties of working with large quantities of diverse microarray data, and shows how existing data processing and analysis methods can be combined and may need adjustments to reach the goal. The work included points to the need of improved biomedical and specialised ontologies which are curtail for working with large microarray datasets, and limitations of current algorithms which, while working with large integrated microarray datasets, can barely deliver the results even in a large computers not accessible for an average user. Finally, this work exhibits the diversity of skills from general programming and working with relational databases, to statistics, data and text mining, ontology building, good understanding of human anatomy and physiology, cell and molecular biology and wet lab procedures used in microarray data generation which all are needed while working on the field of microarray informatics.

References

[ABN$^+$99] U. Alon, N. Barkai, D. A. Notterman, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc Natl Acad Sci U S A*, 96(12):6745–6750, Jun 1999.

[AKS77] J. C. Alwine, D. J. Kemp, and G. R. Stark. Method for detection of specific rnas in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with dna probes. *Proc Natl Acad Sci U S A*, 74(12):5350–5354, Dec 1977.

[Ali02] M. Alison. *The Cancer Handbook*. Nature Publishing Group, London, 2002.

[BBC$^+$04] C. A. Ball, A. Brazma, H. Causton, et al. Submission of microarray data to public repositories. *PLoS Biol*, 2(9):E317, Sep 2004.

[BCB$^+$05] B. M. Bolstad, F. Collin, J. Brettschneider, et al. Quality assessment of affymetrix genechip data. 2005.

[BHQ$^+$01] A. Brazma, P. Hingamp, J. Quackenbush, et al. Minimum information about a microarray experiment (miame)-toward standards for microarray data. *Nat Genet*, 29(4):365–371, Dec 2001.

[BIAS03] B. M. Bolstad, R. A. Irizarry, M. Astrand, et al. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, Jan 2003.

[BK06] A. J. Butte and I. S. Kohane. Creation and implications of a phenome-genome network. *Nat Biotechnol*, 24(1):55–62, Jan 2006.

[Bra03] A. Brazma. Analysis of gene expression data matrices. In C. H. Causton, Q. John, and B. Alvis, editors, *Microarray Gene Expression Data Analysis: A Beginner's Guide.*. Blackwell Publishing, Oxford, UK, 2003.

[Bre70] A. Breslow. Thickness, cross-sectional areas and depth of invasion in the prognosis of cutaneous melanoma. *Ann Surg*, 172(5):902–908, Nov 1970.

[BTW$^+$09] T. Barrett, D. B. Troup, S. E. Wilhite, et al. Ncbi geo: archive for high-throughput functional genomic data. *Nucleic Acids Res*, 37(Database issue):D885–D890, Jan 2009.

[Bur06] L. D. Burgoon. The need for standards, not guidelines, in biological data reporting and sharing. *Nat Biotechnol*, 24(11):1369–1373, Nov 2006.

[BY95] Y. Benjamini and H. Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57:289–300, 1995.

[Cas99] N. Cascinelli. Who declares lymphatic mapping to be the standard of care for melanoma. *Oncology*, 13(3):288, 1999.

[CBM$^+$05] S. E. Choe, M. Boutros, A. M. Michelson, et al. Preferred analysis methods for affymetrix genechips revealed by a wholly defined control dataset. *Genome Biol*, 6(2):R16, Jan 2005.

[CFBM69] W. H. Clark, L. From, E. A. Bernardino, et al. The histogenesis and biologic behavior of primary human malignant melanomas of the skin. *Cancer Res*, 29(3):705–727, Mar 1969.

[Coo40] S. Cooper. *First lines of theory and practice of surgery.*. London: Longman, Orme, Brown, Green and Longman, 1840.

[Cri70] F. Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–563, Aug 1970.

[CWM92] A. J. Cochran, D. R. Wen, and D. L. Morton. Management of the regional lymph nodes in patients with cutaneous malignant melanoma. *World J Surg*, 16(2):214–221, Mar/Apr 1992.

[DCD+07] A. Day, M. R. J. Carlson, J. Dong, et al. Celsius: a community resource for affymetrix microarray data. *Genome Biol*, 8(6):R112, 2007.

[ELS+05] L. L. Elo, L. Lahti, H. Skottman, et al. Integrating probe-level expression changes across generations of affymetrix arrays. *Nucleic Acids Res*, 33(22):e193, Dec 2005.

[ESBB98] M. B. Eisen, P. T. Spellman, P. O. Brown, et al. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*, 95(25):14863–14868, Dec 1998.

[FCD+03] T. L. Fare, E. M. Coffey, H. Dai, et al. Effects of atmospheric ozone on microarray data quality. *Anal Chem*, 75(17):4672–4675, Sep 2003.

[FLL+09] L. French, S. Lane, T. Law, et al. Application and evaluation of automated semantic annotation of gene expression experiments. *Bioinformatics*, 25(12):1543–1549, Jun 2009.

[FRH+93] S. P. Fodor, R. P. Rava, X. C. Huang, et al. Multiplexed biochemical assays with biological chips. *Nature*, 364(6437):555–556, Aug 1993.

[FRP+91] S. P. Fodor, J. L. Read, M. C. Pirrung, et al. Light-directed, spatially addressable parallel chemical synthesis. *Science*, 251(4995):767–773, Feb 1991.

[GBL+08] E. Grundberg, H. Brändström, K. C. L. Lam, et al. Systematic assessment of the human osteoblast transcriptome in resting and induced primary cells. *Physiol Genomics*, 33(3):301–311, May 2008.

[GLW+06] L. Guo, E. K. Lobenhofer, C. Wang, et al. Rat toxicogenomic study reveals analytical consistency across microarray platforms. *Nat Biotechnol*, 24(9):1162–1169, Sep 2006.

[Goo00] P. Good. *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypothesis*. Springer-Verlag, New York, second edition, 2000.

[GST+99] T. R. Golub, D. K. Slonim, P. Tamayo, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, Oct 1999.

[Har75] J. A. Hartigan. *Clustering Algorithms*. John Wiley & Sons, New York, 1975.

[HLM02] E. Hubbell, W.-M. Liu, and R. Mei. Robust estimators for expression analysis. *Bioinformatics*, 18(12):1585–1592, Dec 2002.

[Hol79] S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 2(6), 1979.

[IBC$^+$03] R. A. Irizarry, B. M. Bolstad, F. Collin, et al. Summaries of affymetrix genechip probe level data. *Nucleic Acids Res*, 31(4):e15, Feb 2003.

[IHC$^+$03] R. A. Irizarry, B. Hobbs, F. Collin, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264, Apr 2003.

[IIiT$^+$03] K. Ikeo, J. Ishi-i, T. Tamura, et al. Cibex: center for information biology gene expression database. *C R Biol*, 326(10-11):1079–1082, Oct/Nov 2003.

[ItHV$^+$08] A. E. Ivliev, P. A. C. 't Hoen, M. P. Villerius, et al. Microarray retriever: a web-based tool for searching and large scale retrieval of public microarray data. *Nucleic Acids Res*, 36(Web Server issue):W327–W331, Jul 2008.

[IWS$^+$05] R. A. Irizarry, D. Warren, F. Spencer, et al. Multiple-laboratory comparison of microarray platforms. *Nat Methods*, 2(5):345–350, May 2005.

[JHE$^+$04] A.-K. Järvinen, S. Hautaniemi, H. Edgren, et al. Are data from different gene expression microarray platforms comparable? *Genomics*, 83(6):1164–1168, Jun 2004.

[JLR07] W. E. Johnson, C. Li, and A. Rabinovic. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127, Jan 2007.

[Jol02] I. T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, New York, second edition, 2002.

[KAO$^+$08] S. Kilpinen, R. Autio, K. Ojala, et al. Systematic bioinformatic analysis of expression levels of 17,330 human genes across 9,783 samples from 175 types of healthy and pathological tissues. *Genome Biol*, 9(9):R139, Sep 2008.

[KEH+09] M. Kapushesky, I. Emam, E. Holloway, et al. Gene expression atlas at the european bioinformatics institute. *Nucleic Acids Res*, Nov 2009.

[KGH09] A. Kauffmann, R. Gentleman, and W. Huber. arrayqualitymetrics–a bioconductor package for quality assessment of microarray data. *Bioinformatics*, 25(3):415–416, Feb 2009.

[KIL+06] S. Katz, R. A. Irizarry, X. Lin, et al. A summarization approach for affymetrix genechip data using a reference training set from a large, biologically diverse database. *BMC Bioinformatics*, 7:464, Oct 2006.

[KKZ09] H.-P. Kriegel, P. Kröger, and A. Zimek. Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *TKDD*, 3(1), 2009.

[KMC00] M. K. Kerr, M. Martin, and G. A. Churchill. Analysis of variance for gene expression microarray data. *J Comput Biol*, 7(6):819–837, 2000.

[Koh95] T. Kohonen. *Self-Organizing Maps*. Springer-Verlag, Berlin, 1995.

[LDB+96] D. J. Lockhart, H. Dong, M. C. Byrne, et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol*, 14(13):1675–1680, Dec 1996.

[LMD+02] W.-m. Liu, R. Mei, X. Di, et al. Analysis of high density expression microarrays with signed-rank call algorithms. *Bioinformatics*, 18(12):1593–1599, Dec 2002.

[LMSA06] R. Lucas, T. McMichael, W. Smith, et al. Solar ultraviolet radiation: Global burden of disease from solar ultraviolet radiation. *Environmental Burden of Disease Series*, (13), 2006.

[LTB+09] K. Q. Lao, F. Tang, C. Barbacioru, et al. mrna-sequencing whole transcriptome analysis of a single cell on the solid system. *J Biomol Tech*, 20(5):266–271, Dec 2009.

[MHA+10] J. Malone, E. Holloway, T. Adamusiak, et al. Modeling sample variables with an experimental factor ontology. *Bioinformatics*, Mar 2010.

[Mil81] R. G. Miller. *Simultaneous statistical inference.* Springer-Verlag, New York, second edition, 1981.

[MRBP08] J. Malone, T. Rayner, Z. Bradley, et al. Developing an application focused experimental factor ontology: embracing the obo community. *Proc. of ISMB 2008 SIG meeting on Bio-ontologies*, 2008.

[MW47] H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60, 1947.

[MWW$^+$92] D. L. Morton, D. R. Wen, J. H. Wong, et al. Technical details of intraoperative lymphatic mapping for early stage melanoma. *Arch Surg*, 127(4):392–399, Apr 1992.

[NH09] V. Nygaard and E. Hovig. Methods for quantitation of gene expression. *Front Biosci*, 14:552–569, Jan 2009.

[PBFP05] D. M. Parkin, F. Bray, J. Ferlay, et al. Global cancer statistics, 2002. *CA Cancer J Clin*, 55(2):74–108, Mar/Apr 2005.

[PKK$^+$09] H. Parkinson, M. Kapushesky, N. Kolesnikov, et al. Arrayexpress update–from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Res*, 37(Database issue):D868–D872, Jan 2009.

[PLFS$^+$06] T. A. Patterson, E. K. Lobenhofer, S. B. Fulmer-Smentek, et al. Performance comparison of one-color and two-color platforms within the microarray quality control (maqc) project. *Nat Biotechnol*, 24(9):1140–1150, Sep 2006.

[PSS$^+$94] A. C. Pease, D. Solas, E. J. Sullivan, et al. Light-generated oligonucleotide arrays for rapid dna sequence analysis. *Proc Natl Acad Sci U S A*, 91(11):5022–5026, May 1994.

[RRLG03] S. Ramaswamy, K. N. Ross, E. S. Lander, et al. A molecular signature of metastasis in primary solid tumors. *Nat Genet*, 33(1):49–54, Jan 2003.

[RRSS$^+$06] T. F. Rayner, P. Rocca-Serra, P. T. Spellman, et al. A simple spreadsheet-based, miame-supportive format for microarray data: Mage-tab. *BMC Bioinformatics*, 7:489, Nov 2006.

[Rus00] D. Rusciano. Differentiation and metastasis in melanoma. *Crit Rev Oncog*, 11(2):147–163, 2000.

[SBD$^+$05] C. G. Son, S. Bilke, S. Davis, et al. Database of mrna gene expression profiles of multiple human organs. *Genome Res*, 15(3):443–450, Mar 2005.

[SKK$^+$03] T. Shiraki, S. Kondo, S. Katayama, et al. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc Natl Acad Sci U S A*, 100(26):15776–15781, Dec 2003.

[SMS$^+$02] P. T. Spellman, M. Miller, J. Stewart, et al. Design and implementation of microarray gene expression markup language (mage-ml). *Genome Biol*, 3(9):RESEARCH0046, Aug 2002.

[Smy05] G. K. Smyth. Limma: linear models for microarray data. In R. Gentleman, V. Carey, S. Dudoit, et al., editors, *Bioinformatics and Computational Biology Solutions using R and Bioconductor*, pages 397–420. Springer, New York, 2005.

[Sno92] H. L. Snow. Abstract of a lecture on melanotic cancerous disease. *The Lancet*, 140(3607):872–874, 1892.

[Sou75] E. M. Southern. Detection of specific sequences among dna fragments separated by gel electrophoresis. *J Mol Biol*, 98(3):503–517, Nov 1975.

[SRJ$^+$06] L. Shi, L. H. Reid, W. D. Jones, et al. The microarray quality control (maqc) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat Biotechnol*, 24(9):1151–1161, Sep 2006.

[SSB96] D. Shalon, S. J. Smith, and P. O. Brown. A dna microarray system for analyzing complex dna samples using two-color fluorescent probe hybridization. *Genome Res*, 6(7):639–645, Jul 1996.

[SSDB95] M. Schena, D. Shalon, R. W. Davis, et al. Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270(5235):467–470, Oct 1995.

[Ste03] D. Stekel. *Microarray bioinformatics*. Cambridge University Press, Cambridge, United Kingdom, 2003.

[STF$^+$05] L. Shi, W. Tong, H. Fang, et al. Cross-platform comparability of microarray technology: intra-platform consistency and appropriate data analysis procedures are essential. *BMC Bioinformatics*, 6 Suppl 2:S12, Jul 2005.

[SWB$^+$04] A. I. Su, T. Wiltshire, S. Batalov, et al. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A*, 101(16):6062–6067, Apr 2004.

[Tho05] J. M. Thomas. Time to re-evaluate sentinel node biopsy in melanoma post-multicenter selective lymphadenectomy trial. *J Clin Oncol*, 23(36):9443–9444, Dec 2005.

[Tho07] J. M. Thomas. Sentinel-node biopsy in melanoma. *N Engl J Med*, 356(4):418; author reply 419–418; author reply 421, Jan 2007.

[TTC01] V. G. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A*, 98(9):5116–5121, Apr 2001.

[USGR$^+$09] G. J. G. Upton, O. Sanchez-Graillet, J. Rowsell, et al. On the causes of outliers in affymetrix genechip data. *Brief Funct Genomic Proteomic*, 8(3):199–212, May 2009.

[VK07] J. Venna and S. Kaski. Nerv: Nonlinear dimensionality reduction as information retrieval. *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics (AISTATS*07)*, Mar 2007.

[VvASH$^+$09] C. A. Voit, A. C. J. van Akkooi, G. Schäfer-Hesterberg, et al. Rotterdam criteria for sentinel node (sn) tumor burden and the accuracy of ultrasound (us)-guided fine-needle aspiration cytology (fnac): can us-guided fnac replace sn staging in patients with melanoma? *J Clin Oncol*, 27(30):4994–5000, Oct 2009.

[VVF08] H. D. VanGuilder, K. E. Vrana, and W. M. Freeman. Twenty-five years of quantitative pcr for gene expression analysis. *Biotechniques*, 44(5):619–626, Apr 2008.

[VZVK95] V. E. Velculescu, L. Zhang, B. Vogelstein, et al. Serial analysis of gene expression. *Science*, 270(5235):484–487, Oct 1995.

[Wel47] B. L. Welch. The generalisation of "studentś" problem when several different population variances are involved. *Biometrica*, 34(1-2):28–35, 1947.

[WGS09] Z. Wang, M. Gerstein, and M. Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, 10(1):57–63, Jan 2009.

[Wil45] F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

[WPC+06] P. L. Whetzel, H. Parkinson, H. C. Causton, et al. The mged ontology: a resource for semantics-based description of microarray experiments. *Bioinformatics*, 22(7):866–873, Apr 2006.

[WT07] D. Witten and R. Tibshirani. A comparison of fold-change and the t-statistic for microarray data analysis, Nov 2007. Http://www-stat.stanford.edu/ tibs/ftp/FCTComparison.pdf.

[XTN+05] L. Xu, A. C. Tan, D. Q. Naiman, et al. Robust prostate cancer marker genes emerge from direct integration of inter-study microarray data. *Bioinformatics*, 21(20):3905–3911, Oct 2005.

Paper I

E Kääriäinen, P Nummela, J Soikkeli, M Yin, M Lukk, T Jahkola, S Virolainen, A Ora, E Ukkonen, O Saksela and E Hölttä

Switch to an invasive growth phase in melanoma is associated with tenascin-C, fibronectin, and procollagen-I forming specific channel structures for invastion

The Journal of Pathology 2006 Oct;210(2):181-91.

Paper II

J Soikkeli[1], M Lukk[1], P Nummela, S Virolainen, T Jahkola, R Katainen, L Harju, E Ukkonen, O Saksela and E Hölttä

Systematic search for the best gene expression markers for melanoma micrometastasis detection

The Journal of Pathology 2007 Oct;213(2):180-9.

[1] These authors contributed equally to this work.

Paper III

H. Parkinson, M. Kapushesky, M. Shojatalab, N. Abeygunawardena, R. Coulson, A. Farne, E. Holloway, N. Kolesnykov, P. Lilja, M. Lukk, R, Mani, T. Rayner, A. Sharma, E. Williams, U. Sarkans and A. Brazma

ArrayExpress–a public database of microarray experiments and gene expression profiles

Nucleic Acids Research 2007 Jan;35(Database issue):D747-50.

Paper IV

Tim F. Rayner, Faisal Ibne Rezwan, Margus Lukk, Xiangqun Zheng Bradley, Anna Farne, Ele Holloway, James Malone, Eleanor Williams and Helen Parkinson

MAGETabulator, a suite of tools to support the microarray data format MAGE-TAB

Bioinformatics 2009 Jan 15;25(2):279-80.

Paper V

Audrey Kauffmann, Tim F. Rayner, Helen Parkinson, Misha Kapushesky, Margus Lukk, Alvis Brazma and Wolfgang Huber

Importing ArrayExpress datasets into R/Bioconductor

Bioinformatics 2009 Aug 15;25(16):2092-4.

Paper VI

Margus Lukk, Misha Kapushesky, Janne Nikkilä, Helen Parkinson, Angela Goncalves, Wolfgang Huber, Esko Ukkonen and Alvis Brazma

A global map of human gene expression

Nature Biotechnology 2010 Apr;28(4):322-324.