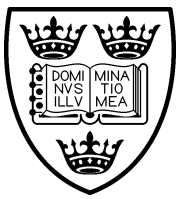
University of Oxford




# Increasing statistical power and generalizability in genomics microarray research

Adaikalavan Ramasamy

Wolfson College

A report submitted for DPhil

17th June 2010

*Department of Statistics, 1 South Parks Road, Oxford*

I certify that this is my own work (except where otherwise indicated).

Candidate .....

Signed .....

Dated .....

## Acknowledgements

I would like to express my deep gratitude towards both my supervisors Professor Doug Altman (Centre for Statistics in Medicine, Oxford) and Professor Chris Holmes (Department of Statistics, University of Oxford) for their guidance, patience and inspiration. I would like to thank Cancer Research UK for kindly sponsoring this postgraduate study.

I would also like to thank Dr. Adrian Mondry (National University Hospital, Singapore) for extensive proof-reading, thorough restructuring and many other helpful suggestion and moral support that has significantly and greatly improved the quality of this thesis and the scientific papers arising from this work.

Special thanks to my past and current supervisors Dr. Lance Miller (Wake Forest University), Professor David Balding (University College London) and Dr. Debbie Jarvis (Imperial College London) who have and are continuing to inspire me to develop into a professional individual. A substantial part of this thesis was done in collaboration with Dr. Francesco Pezzella and Dr. Jianting Hu (Cancer Research UK).

Tracy Edwards deserves a special mention for helping out with the many administrative tasks as do Denise and Brian Leigh for having the patience and heart to act as my life coaches.

My experience in undertaking this PhD is similar to a roller coaster ride with many exhilarating and thrilling ups with some nerve wrecking down moments. Therefore, I and indebted to my numerous friends, the staff and students that make up the Wolfson College community for the love, support and the zaniness during the down cycles. Here are some friends that I am forever grateful for, in alphabetical order: Alexander Antonyuk, Ali Zaidi, Alicia Izharrudin, Ammar Waraich, Anli Wang, Anne Jensen, Asha Titus, Avinash Kolli, Benoit Pujol, Charlotte Mondry, Doreen Tembo, Farooq Waraich, François Szymanski, Gareth Hughes, Gerald Moncayo, Hardave Kharbandra, How Wei Theng, Huan Huang, Kanishka Bhattacharya, JanaLee Cherneski, Kumaran Mande, Linariah Lukman, Lorel Chow, Madelaine Hamilton, Manjulapramila Thimmajanarthanan, Michael Zrenner, Muiz Shariffuddin, Myrtani Pieri, Neil Chisholm, Renee Lee, Satomi Iwakoshi, Satyajit Saste, Syed Rahman, Tahir Mansoori, Tania Oh, Tiki Shabudin, Vidhya Nagarathnam, Wasim Malik, Yeoh Kheng-Wei, YY Teo, Zuzanna Olszewska.

Finally, I also like to thank my saintly mother, my loving father, my motherly sister, supportive brother-in-law, my delightful nephews Ramu and Laxman, and my dear wife Kartika Selvam for the patience and moral support they provided me throughout the years.

# Contents


**1. Introduction** **1**

1.1. Brief introduction to microarrays . . . . . 2

1.1.1. The rise and current state of microarray technology . . . . . 2

1.1.2. Advantages of microarrays . . . . . 3

1.1.3. Applications for gene expression arrays . . . . . 4

1.1.4. Some criticisms of microarray technology . . . . . 5

1.2. The objective of this thesis . . . . . 6

1.2.1. Motivation for sample size calculation . . . . . 7

1.2.2. Motivation for meta-analysis for microarray datasets . . . . . 7

1.3. Biological background for the thesis . . . . . 8

1.3.1. Brief introduction to cancer biology . . . . . 8

1.4. Brief introduction to molecular biology . . . . . 9

1.5. Technical background for the thesis . . . . . 10

1.5.1. Brief introduction to microarray technology . . . . . 10

1.5.2. Probe-level and gene-level identifiers . . . . . 11

1.5.3. Microarray data formats . . . . . 12

1.5.4. Microarray platforms and the associated preprocessing algorithms . . 13

1.6. Summary . . . . . 16

**2. Statistics for two-class comparison in microarray** **17**

2.1. Simple summaries of gene expression . . . . . 18

2.2. Statistical difference . . . . . 19

2.2.1. The multiple hypothesis testing problem . . . . . 20

2.3. Effect size as a measure of biological difference . . . . . 21

2.4. Software packages used . . . . . 23

2.5. Summary . . . . . 23

**3. Sample size for microarrays: Minimum number of biological samples 24**

3.1. Introduction . . . . . 25

3.2. Definition and importance of statistical power . . . . . 25

3.3. Commonly used formulaes for sample size calculation . . . . . 26

3.4. Three simple improvements to the common formula . . . . . 28

3.4.1. Improvement 1: Generalizing to unequal allocation ratio . . . . . 28

3.4.2. Improvement 2: Joint estimation of $\delta$ and $\sigma$ . . . . . 30

3.4.3. Improvement 3: The correction term $z^2_{\alpha/2}/2$ is important for small $\alpha$ . 34

3.5. How to estimate the sample size for microarrays . . . . . 35

3.5.1. Step 1: Calculate the standardized effect Sizes . . . . . 35

3.5.2. Step 2: Choose a cutoff for effect size . . . . . 36

3.5.3. Step 3: Choose a significance level ($\alpha$) . . . . . 37

3.5.4. Step 4: Choosing a power level ($= 1 - \beta$) . . . . . 37

3.5.5. Step 5: Choose the allocation ratio (k) . . . . . 37

3.5.6. Step 6: Calculate the total sample size . . . . . 38

3.5.7. Special case of equal allocation ($k = 1$) . . . . . 39

3.6. Discussion . . . . . 41

3.7. Summary . . . . . 42

**4. Sample size for microarrays: Minimum number of technical replicates 43**

4.1. Motivation . . . . . 44

4.2. Methodology . . . . . 44

4.3. Unequal recruitment cost for the two groups . . . . . 46

4.4. Estimation of the additional parameters . . . . . 46

4.4.1. Estimating $\Delta_m$ for $m > 1$ . . . . . 47

4.4.2. Estimating $R = \sigma^2_E/\sigma^2_B$ . . . . . 47

4.5. Discussion . . . . . 48

4.6. Summary . . . . . 48

**5. Sample size for microarrays: A case study 49**

5.1. Introduction to the clinical problem . . . . . 50

5.2. Choosing the number of technical replicates . . . . . 50

5.3. Choosing the number of biological replicates . . . . . 52
5.4. Summary . . . . . 55

**6. Meta-analysis of microarray datasets: Key issues 56**
6.1. Motivation . . . . . 57
6.2. Introduction . . . . . 57
6.3. Key issues in meta-analysis of microarray studies . . . . . 59
6.3.1. Issue 1: Finding relevant studies . . . . . 59
6.3.2. Issue 2: Data Extraction . . . . . 61
6.3.3. Issue 3: Preparing datasets from different platforms . . . . . 62
6.3.4. Issue 4: Annotating the individual datasets . . . . . 64
6.3.5. Issue 5: Resolving the many-to-many relationship . . . . . 65
6.3.6. Issue 6: Combine the study-specific estimates . . . . . 66
6.3.7. Issue 7: Analyze, present and interpret results . . . . . 67
6.4. Discussion . . . . . 67
6.5. Summary . . . . . 70

**7. Meta-analysis of microarray datasets: Choosing a meta-analysis technique 71**
7.1. Available techniques for meta-analysis of microarrays . . . . . 72
7.1.1. Combining votes . . . . . 72
7.1.2. Combining $p$-values . . . . . 72
7.1.3. Combining ranks and lists . . . . . 73
7.1.4. Combining effect sizes . . . . . 74
7.1.5. The integrative correlation method . . . . . 74
7.2. Qualitative arguments for selecting a technique . . . . . 74
7.3. Quantitative arguments for selecting a technique . . . . . 77
7.3.1. Comparison methodology . . . . . 78
7.3.2. Results . . . . . 79
7.3.3. Discussion on quantitive arguments . . . . . 83
7.4. The inverse variance technique for meta-analysis . . . . . 84
7.4.1. $p$-value calculation for the inverse-variance technique . . . . . 85
7.5. Summary . . . . . 87

**8. Meta-analysis of microarray studies: Two case studies. 88**

8.1. Motivation . . . 89
8.1.1. Meta-analyses of protein folding genes . . . 89
8.1.2. Meta-analyses of solid tumours . . . 90
8.2. Methods . . . 90
8.2.1. General methods . . . 90
8.2.2. Extra information about annotation and mapping . . . 95
8.2.3. Identifying Protein Folding Genes . . . 96
8.2.4. Frequency of identification . . . 97
8.2.5. Gene Set Enrichment Analysis (GSEA) . . . 97
8.2.6. Logistics . . . 99
8.3. Results of Meta-analyses of Solid Tumours . . . 99
8.3.1. Gene-centric view . . . 99
8.3.2. Annotation-centric view . . . 102
8.3.3. Identifying Biological Themes using Annotation Clusters . . . 104
8.4. Results of Meta-analyses of Protein Folding Genes . . . 108
8.5. Summary . . . 110

**9. Exploring data sharing issues in microarray genomics research 111**

9.1. Introduction . . . 112
9.1.1. MIAME requirement on raw microarray data . . . 112
9.1.2. Why is the raw data important? . . . 114
9.2. Collection of studies assessed . . . 115
9.3. Data extraction . . . 116
9.4. Statistical analysis . . . 119
9.5. Results . . . 119
9.5.1. Reporting and sharing of raw data . . . 119
9.5.2. Variables associated with reporting of raw data . . . 121
9.5.3. Variables associated with sharing of raw data when requested . . . 122
9.6. Discussion . . . 123
9.6.1. Limitations of our study . . . 124
9.7. Recommendations . . . 125
9.8. Summary . . . 127

**10. Summary, Discussion and Recommendations** **128**

10.1. Summary of sample size chapters . . . . . 129

10.2. Summary of meta-analysis chapters . . . . . 130

10.3. Noteworthy contribution of this thesis to the field . . . . . 133

**A. Appendix** **134**

A.1. Derivation of Equation 3.3 . . . . . 134

A.2. Proof for points in Section 3.5.5: choosing the allocation ratio, $k$ . . . . . 136

A.3. Derivation of $m_{opt}$ (Equation 4.3) . . . . . 137

A.4. R codes for sample size calculations . . . . . 138

A.5. R codes for meta-analysis of microarray datasets . . . . . 139

A.5.1. Some generic R codes . . . . . 140

A.5.2. R codes to combine effect sizes . . . . . 141

A.6. Major milestones in microarray data reporting requirements . . . . . 146

# List of Figures


1.1. Publication trend of microarray related articles . . . . . 3
1.2. Illustration of Central Dogma of Molecular Biology . . . . . 10
1.3. The flow from data to information in gene expression microarray research . . 12
1.4. Illustration of the normalization procedure . . . . . 15

3.1. The approximate allocation efficiency for the 30 collected studies . . . . . 29
3.2. Agreement in $\Delta$ values across different preprocessing algorithm . . . . . 31
3.3. Choice of preprocessing algorithm affects $S_p$ . . . . . 32
3.4. Choice of preprocessing algorithm does not affect $\Delta$ . . . . . 33
3.5. Graphical depiction of sample size requirements. . . . . 40

7.1. Comparison results from Fisher's method . . . . . 81
7.2. Comparison results from RankProd method . . . . . 82
7.3. Comparison results from the inverse-variance technique . . . . . 82
7.4. Assessing the fit for $p$-value calculation in inverse-variance techniques . . . . 86

8.1. Data identification and acquisition for case study . . . . . 94
8.2. A summary plot for meta-analyses of 26 solid tumours dataset . . . . . 100
8.3. Annotation terms from two top ranking genes from meta-analyses. . . . . 102
8.4. A screen capture of the top few terms enriched in GENELIST. . . . . 103
8.5. A screen capture of the first three annotation clusters from DAVID. . . . . 106
8.6. A summary plot of the meta-analyses of protein folding genes. . . . . 108
8.7. Forest plot of Crystallin, Alpha B (Hs.408767). . . . . 110

9.1. Screen capture of the official MIAME website accessed on $20^{th}$ August 2008. 113
9.2. Overlap in the number of studies in the three sets . . . . . 116
9.3. Reporting and sharing of raw data in 146 microarray-based studies. . . . . 120

9.4. Study sample sizes and journal impact factors by reporting pattern . . . . . 121
9.5. Distribution of publication date by email responses . . . . . 123

A.1. Plot of the multiplier due to allocation ratio, $k$ . . . . . 136

# List of Tables


2.1. An illustrative example of a GEDM. . . . . 18
2.2. Some summary statistics for the GEDM presented in Table 2.1 . . . . . 18
2.3. The value of the correction factor in Hedge's adjusted $g$ . . . . . 22

3.1. Relationship between some statistical concepts related to power . . . . . 26
3.2. Empirical values for $\delta$, $S_p$ and $\Delta$ from Singh *et al.* (2002) . . . . . 31
3.3. Values for selected quantiles of $S_p$ and $\Delta$. . . . . 34
3.4. Value of the correction term as a function of significance level ($\alpha$) . . . . . 35
3.5. Some useful numbers for sample size calculations . . . . . 38
3.6. Minimum total sample size requirements with equal allocation ($k = 1$). . . . 40

4.1. Explanation of the symbols used in this chapter. . . . . 44

5.1. Cost components for microarray studies . . . . . 51
5.2. Effect size estimate by comparison and coverage . . . . . 52

6.1. A checklist for conducting meta-analysis of microarray datasets. . . . . 58
6.2. Useful resources to identify studies for meta-analysis . . . . . 60
6.3. Advantages and disadvantages of meta-analysis of microarray datasets . . . . 69

7.1. Agreement of the vote counting method for testing up-regulated genes. . . . 80
7.2. Agreement of the vote counting method for testing down-regulated genes. . . 80

8.1. Methods for meta-analysis for the two case studies. . . . . 93
8.2. Datasets with FLEO data used for the case study . . . . . 95
8.3. Characteristics of microarrays chips used in the studies for meta-analysis . . 96
8.4. Frequency of UniGene ID identification across datasets. . . . . 97
8.5. List of genes significant at p-value $< 10^{-6}$ for the solid tumours case study . 101

8.6. Enrichment table for a GO term . . . . . 104
8.7. Summary of the significant annotation clusters from DAVID . . . . . 107
8.8. Results the inverse-variance technique for the significant genes . . . . . 109

9.1. Characteristics of the studies in the three sets. . . . . 118
9.2. The frequency and percentage of the "data reported" variable . . . . . 120
9.3. Summary of email responses by commercial affiliation . . . . . 122

# 1. Introduction

*The high-throughput technologies developed during the second half of the 1990s have revolutionized the speed of data accumulation in the life sciences. As a result we have very rich and complex data that holds great promise to solving many complex biological questions.*

*This thesis focuses on statistical issues arising from one such technology that is very widespread and relatively mature: the DNA microarray or gene expression microarray. The aim of this thesis is to contribute to the development of statistics that allow the end users to obtain robust and meaningful results from gene expression microarray for further investigation.*

## 1.1. Brief introduction to microarrays

### 1.1.1. The rise and current state of microarray technology

Microarray technology measures the messenger RNA (mRNA) levels of tens of thousands of genes in tissue samples simultaneously in a high-throughput and cost-effective manner.

Since its introduction over a decade ago (Schena *et al.*, 1995; DeRisi *et al.*, 1996), microarrays have found widespread use in the field of molecular genetics, functional genomics and medicine. The technology has rapidly advanced the miniaturization and need for high-throughput process in many genomic fields (e.g. gene expression, comparitive genomic hybrdisation arrays, single nucleotide polymorphism arrays) and in the proteomics field (e.g. tissue microarrays, multiplex polymerase chain reaction).

Microarray technology is now in the mainstream of many life science laboratories and its popularity is every growing. Figure 1.1, which shows the increase in relevant publications, is an indication of how widely the technology has been adopted.

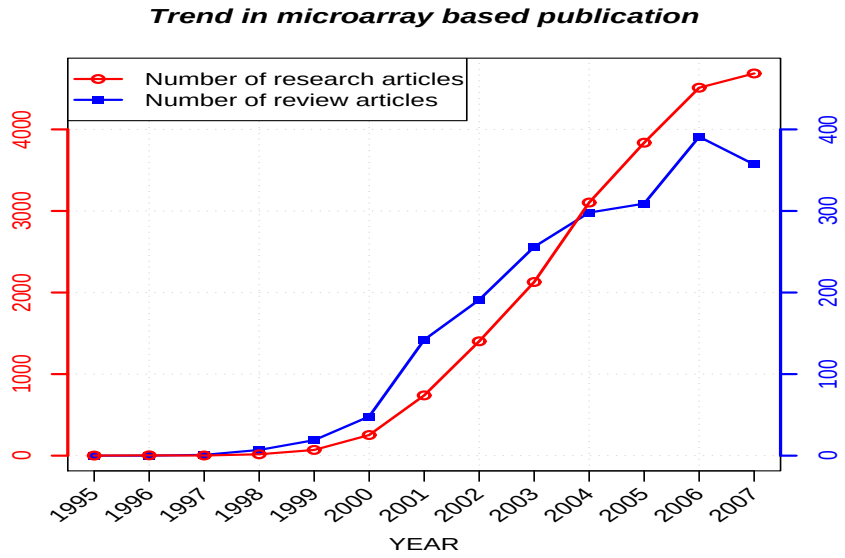



Figure 1.1.: The number of journal and review articles by year available in PubMed as of $20^{th}$ August 2008. The right hand axis corresponds to the number of review articles. We identified the studies using the search terms "gene" and "microarray". For example, the search query to find the number of publications in 1999 would be `gene AND microarray AND ("1999"[PDat]:"1999"[PDat])`.

The technology has been applied to understand underlying biological mechanisms (DeRisi *et al.*, 1996), to discover novel subgroups of diseases (Golub *et al.*, 1999; Perou *et al.*, 1999; Alizadeh *et al.*, 2000), to examine drug response (Staunton *et al.*, 2001; Dan *et al.*, 2002), to classify patients into disease groups (Golub *et al.*, 1999) and to predict disease outcomes (van 't Veer *et al.*, 2002; van de Vijver *et al.*, 2002; Chang *et al.*, 2005). Some of the molecular signatures discovered with microarray technology are now being evaluated in prospective randomized clinical trials (Bogaerts *et al.*, 2006; Paik, 2007).

### 1.1.2. Advantages of microarrays

Microarray technology offers several advantages over other genomic technologies: a comprehensive coverage of the genome, high-throughput experimentation, cost effective solutions, safety and low sample volume requirement.

Current microarray technology allows one to measure mRNA levels of over 55,000 probes which makes it possible to measure entire transcriptomes. For example, a human genome contains of approximately 25,000 genes.

Microarray allows one to measure mRNA levels of tens of thousands of genes simultaneously with relative ease, reducing the experimental variability, cost and time required if one were to quantify each gene individually. The availability of standard protocols for experiments and use has greatly simplified matters.

The chip and reagent cost required for the quantification of a single sample can be as high as £1000 but the cost per gene is extremely cheap, making it an effective tool if one wants to capture a snapshot of the genome. Furthermore, advances in technology and competition among vendors are increasing quality and lowering microarrays costs.

The technology is safe as it does not involve any dangerous chemicals or radiation. It also requires low volumes of biological material. If the starting quantities are too low, as in the use of laser capture microdissection which aims to isolate single cells, one can amplify the amounts of mRNA contained in the samples by using the polymerase chain reaction technique.

### 1.1.3. Applications for gene expression arrays

The areas of application of the microarray technology have grown wider along with the increasing number of publications. Simon *et al.* (2004) classifies microarray technology broadly into three groups of application: class discovery, class comparison and class prediction.

In *class discovery*, one attempts to identify new subgroups of patients. This can be done when one believes that subgroups of patients with different prognostic outcomes or drug responses need to be identified. This is typically achieved via exploratory methods such as clustering and dimension reduction techniques. Such methods can also be used to identify any potentially misclassified patients or to demonstrate microarray's capability to segregate phenotypes based on genotype. The hypothesis generated by such methods has to be biologically meaningful and further validated.

In *class comparison*, one wishes to identify genes that are differentially expressed between

two or more conditions. The accurate classification of patients is not the primary objective here. Genes can be ranked and filtered according to some summary statistics for each gene. The results can be corrected for the multiple hypothesis problem and a threshold can be drawn, or one can simply rank the genes. Typically, some of the high ranking genes are then validated by more precise but low throughput methods, and a hypothesis about their interaction is generated. In a further step, the investigators may try to validate this hypothesis by experimentation or through literature search. The class comparison problem remains at the heart of many microarray studies.

In *class prediction* one attempts to find the best (and often minimal) subset of biomarkers able to identify specific conditions. In this application of DNA microarray technology, genes are ranked by a statistical method and then added sequentially into a classification algorithm along with any other existing complementary (e.g. clinical) information. Next, the minimal number of genes to adequately classify a patient into a group or predict their clinical outcomes is determined. The stable gene products of a selected subset are identified and may be validated in a larger cohort of patients.

There are various types of microarrays including: Gene Expression microarrays (also known as DNA microarrays) to measure gene transcripts; Comparative Genomic Hybrdisation microarrays for detecting copy number changes at the chromosomal level; Single Nucleotide Polymorphism microarray for detecting point mutations and Tissue Microarrays which detect the protein levels in tissues.

This thesis will focus exclusively on gene expression microarrays. Henceforth, in the text *microarray* refers only to gene expression microarrays unless stated otherwise.

### 1.1.4. Some criticisms of microarray technology

Despite their great promise, microarray based studies may report findings that are not reproducible (Ntzani and Ioannidis, 2003) or not robust to the mildest of data perturbations (Michiels *et al.*, 2005; Ein-Dor *et al.*, 2005). Common causes include improper analysis or validation, insufficient control of false positives and inadequate reporting of methods (Jafari and Azuaje, 2006; Dupuy and Simon, 2007). The situation is exacerbated by the small sample sizes relative to large numbers of potential predictors; typically tens of thou-

sands of probes are investigated in only tens or hundreds of biological samples. The low number of samples are often due to a combination of either financial constraints, lack of sample availability or logistical restrictions.

This scenario, where the number of variables is much greater than the number of observations is also known as the "curse of dimensionality" (Bellman, 1961). One clear example this is often seen during the model selection process. In model selection, one searches among subsets of all possible independent variables to find the minimum relevant subset that adequately explains the dependent variable. With tens of thousands of variables, it is possible to find some subset of irrelevant variables that will classify the existing data perfectly, but will perform poorly on new samples.

Peduzzi *et al.* (1996) used simulation studies to show that one needs at least 10 samples per independent variable for robust estimates to use a logistic regression, a far cry from the case of microarray based studies. However, most microarray applications are used as screening tools to filter out the large proportion of genes which are expected not to be differentially expressed and not for modelling.

## 1.2. The objective of this thesis

This thesis aims to help researchers to prioritize and select a robust list of genes for further investigation from microarray studies conducted with class comparison as primary aim. Therefore we present and discuss two inter-related topics. First, we look into the possibility of statistically designing a large single study. Second, we look at the possibility of synthesizing information from different studies which not only increases the statistical power but also the generalizability of the results.

Chapters 3 - 5 are devoted to sample size estimation methodology in microarray studies while Chapters 6 - 8 are devoted to meta-analysis of microarray studies. For each topic, we present a case study and guided steps to help researchers to perform sample size estimation and meta-analysis. Additionally, Chapter 9 discusses the availability of raw data in microarray studies which is important in meta-analysis to avoid biased results and in sample size for parameter estimation.

### 1.2.1. Motivation for sample size calculation

Since small sample sizes frequently constitute a first hurdle to any meaningful data analysis, we address the minimum sample size calculation using the simplest approach in Chapter 3. In Chapter 4, we extend the sample size framework to incorporate the cost and variance structure into the design process for meaningful gene expression microarray studies.

Addressing the sample size question also allows us to have an idea of the statistical power of each of the individual studies. Indirectly, one can use the findings from this chapter to get a feel for the power of a meta-analysis.

### 1.2.2. Motivation for meta-analysis for microarray datasets

While a large and high quality study can provide more precise answers, designing such a study can be financially expensive and time consuming and requires good logistics management and sample availability. Furthermore, it does not address the issue of generalizability across studies (Ferguson, 2004).

What is generalizability? It is the extent to which the results from a particular valid study can be applied to other circumstances, and needs to be assessed before considering widespread practical application. For example, the findings of a study using historical controls from a particular geographical region may not be applicable to newer cohorts of patients or different regions. Other conditions that could cause findings to be study-specific are population structure, technical issues such as sample handling procedures, and data analysis methods.

**In other words, even findings from a sufficiently large, properly analyzed and correctly reported study need to be evaluated against findings from other laboratories.**

An attractive option to solve both the problem of small sample sizes and generalizability is to synthesize information from different existing studies. Statistical techniques which combine results from independent but related studies are called 'meta-analysis'. However, the term meta-analysis is also widely used to describe the whole study process (as we do here), not just the statistical techniques, for which an alternative term is a systematic review.

Another major advantage of meta-analysis is that it makes comprehensive use of already available data. Not only is this relatively inexpensive but it is also time efficient. Furthermore, if we choose to collect and re-analyze the raw data in a consistent manner, we can hope to minimize the influences of different analysis methods and inadequate reporting.

Therefore, the increased statistical power, generalizability, inexpensiveness, ability to assess heterogeneity of the overall estimate and potential for standardization of methods makes meta-analysis an attractive options for microarray datasets.

## 1.3. Biological background for the thesis

### 1.3.1. Brief introduction to cancer biology

Much of the current worldwide research focus in the life sciences is cancer biology. The microarray technology is also very frequently used in this context, and thus this thesis utilizes data from cancer studies for exemplary purposes. Thus a brief introduction to cancer biology and molecular biology become necessary before introducing the microarray technology.

A phenotype is the set of an individual's observable characteristics and can be determined by his/her genetic background, exposure to environmental factors and random variation. Many genetic diseases have a significant genetic component, caused by abnormal expression of one or more genes. Therefore, identifying genes that are expressed abnormally is an important step to help understand the underlying biology, and towards the aims of early detection, predicting diagnosis and prognosis and finally towards prevention and cure.

A cancer is uncontrollable growth of abnormal tissues. Malignancy is defined as the ability of such growth to spread to other tissues. Some cancers arise through accumulation of harmful mutations in the DNA. Mutation is either a change, insertion, deletion or rearrangement of DNA sequence and can be hereditary or non-hereditary. DNA may acquire mutations through copying errors during DNA replication or after exposure to radiation and carcinogens (any agent that promotes cancer, e.g. asbestos, some ingredients in tobbaco).

From the perspective of this thesis, cancer is an ideal biological field of study as, because of its background in abnormal changes of DNA expression, it has been in the focus of researchers

using microarray experiments right from the start.

## 1.4. Brief introduction to molecular biology

Deoxyribonucleic acid (DNA) contains the genetic information for the production of proteins, which are essential to the structure, function and regulation of the body's cell, tissue and organs. DNA can be found in the nucleus of every cell and consists of a pair of molecules held together by hydrogen bonds of complementary nucleotides in a double helix structure.

There are four types of nucleotides or bases Adenine (A), Cytosine (C), Guanine (G) and Thymine (T) in the DNA. Note that the thymine is replaced by Uracil (U) in the single-stranded ribonucleic acid (RNA). Due to the difference in chemical structure of the bases, the hydrogen bonds will form only between A and T and between G and C. This base pairing rule allows us to automatically deduce the sequence of one strand of the DNA from its complementary strand. This is fundamental to understanding how the microarray technology works. Here is a simple graphical illustration of how the base pairing rule works :

```
... A C A A G A T G C C A T T G T C C C ...      (Strand 1)

    | | | | | | | | | | | | | | | | | |

... T G T T C T A C G G T A A C A G G G ...      (Strand 2)
```

DNA does not produce proteins directly but rather through several intermediary products. First DNA is *transcribed* into RNA which contains many coding (exons) and non-coding (introns) regions. The exons are *spliced* and combined to produce mature mRNAs which are then able to transport the genetic information through the nuclear pore into the cyctoplasm. The mRNA nucleotides are read in triplets by the *translational* machinery as each nucleotide triplet codes for a specific amino acid. The string of assembled amino acids (primary structure of a protein) undergoes post-translational modification to fold into its conformational or functional shape in three-dimensions. This flow of genetic information was proposed by Crick (1970) and is commonly called the Central Dogma of Molecular Biology.

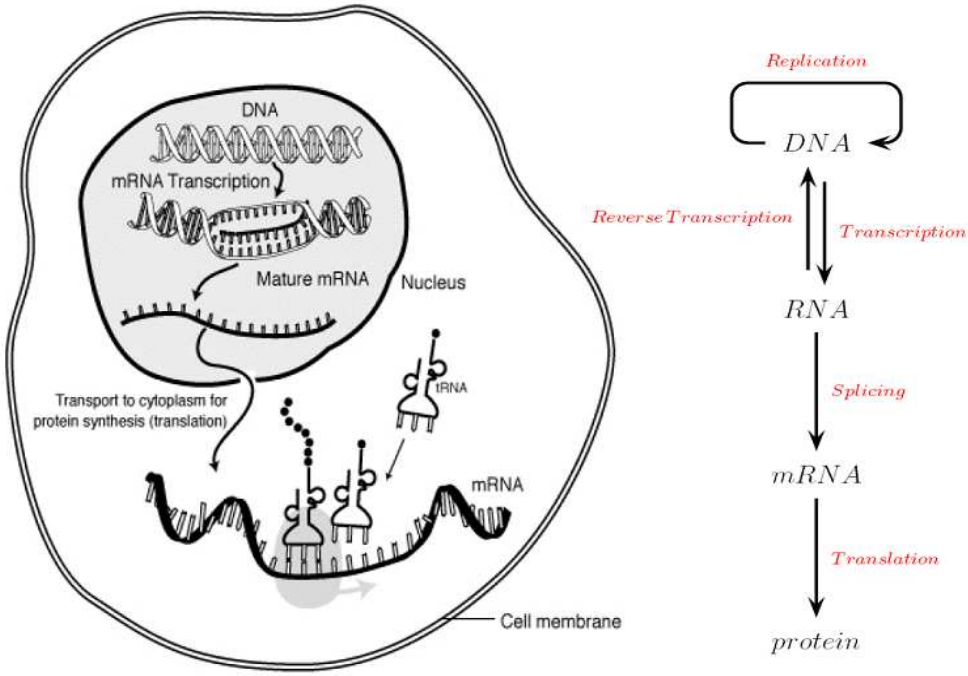



Figure 1.2.: A complex (left) and simplified (right) illustration of the central dogma of molecular biology. The illustration on the left is courtesy of the National Human Genome Research Institute.

## 1.5. Technical background for the thesis

### 1.5.1. Brief introduction to microarray technology

Microarrays are a comprehensive, high throughput and cost effective tool to study the gene expression of a given biological sample. Measuring gene expression is important in trying to understand the stages on individual development, potential genetic component of disease etiology, detection and prevention of diseases, effects of different drugs or treatments, and for establishing the varying susceptibility of genetic subpopulations to each of these (Bloche, 2004; Mondry *et al.*, 2005). DNA microarrays allow us to simultaneously measure the relative transcript abundance of thousands of messenger RNA (mRNA) present in a sample. The defining features of a microarray are its compact physical size and high gene density.

An oversimplified analogy to how microarray quantifies mRNA abundance uses coin sorting

machines and vending machines. Such machines are capable of sorting coins of mixed denomination into different slots (1$p$, 2$p$, 5$p$, 10$p$, 20$p$, 50$p$, £1, £2) by taking advantage of different physical attributes (size, thickness, shape, weight) of different coins. Next, they quantify the number of coins in each slot by either their weight or the height of the stack of coins. Similarly, a microarray first "sorts" the thousands of mRNA in the sample by taking advantage of the complementary base pairing rules. Next it "quantifies" the number of mRNA copies in each slot by measuring the luminescence of fluorescent molecules that were tagged onto all DNA sequences in the sample.

The technical explanation of microarray technology is more detailed. A *probe* is the DNA sequence that represents a gene that is permanently *spotted* (i.e. immobilized) in a regular grid onto a solid support (chemically modified glass or nylon). A DNA microarray contains thousands of different spots and in each spot there are millions of copies of the same probe. DNA sequences from the sample in which one wants to measure the mRNA abundance are called *targets*. Fluorescent molecules, that can be excited by laser, are appended to each target. The *hybridization process* involves applying the prepared sample onto the microarray surface and placing it in a hybridization chamber, where it undergoes a pre-specified cycle of heating and cooling. The heat cycle denatures the hydrogen bonds of the double stranded DNA sequence of both target and probes. The chips are physically rotated in the chamber which allows the single stranded DNA sequence of targets to be attracted to the probes with the best *base complementarity*. During the cooling cycle, the targets will *hybridize* (i.e. anneal by forming hydrogen bonds) with their complimentary probes. Once the hybridization is complete, excess and unhybridized target DNA are washed away. Lasers of specific wavelengths are used to excite the fluorescent molecules and the mRNA abundance is quantified by the fluorescent intensity level. Good *probe design* is very important to ensure that the probes are highly specific to a gene. Otherwise mRNA can *cross-hybridise* with other probes leading to noise in the output.

### 1.5.2. Probe-level and gene-level identifiers

The full-length gene sequence, which could be several hundreds or thousands of bases long, is not used in its entirety for designing a probe (the DNA sequence spotted on the microarray surface) because it is likely to contain non-coding regions and common motifs in the genome

leading to high non-specific binding or noise. Thus, a probe is chosen from a short and highly specific region of the gene to minimize such artifacts. Different design criteria can lead to the creation of many different probes for the same gene.

There are various probe-level nomenclatures, some of which are assigned by the manufacturers (e.g. the IMAGE Clone ID (Lennon *et al.*, 1996) and Affymetrix ID) and some which are annotated collection of genetic sequence (e.g. GenBank (Benson *et al.*, 2008)). The genes are identified using gene nomenclatures such as UniGene Wheeler *et al.* (2003), RefSeq (Pruitt *et al.*, 2007) or EntrezGene ID Entrez.

How and why one maps from the probe-level identifiers to gene-level identifiers is addressed in Section 6.3.4.

### 1.5.3. Microarray data formats

Figure 1.3 shows the four types of data arising from microarray analysis.

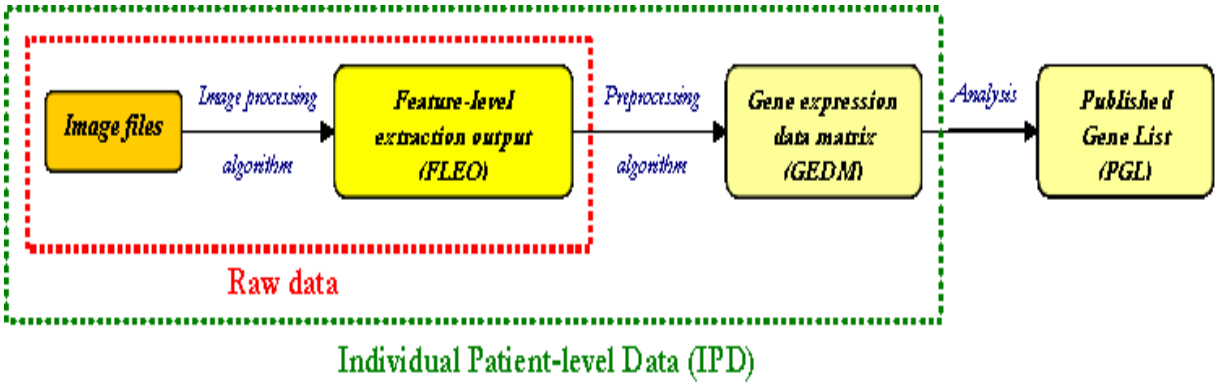



Figure 1.3.: The flow from data to information to biological knowledge in gene expression microarray research.

The intensities of the probes of a hybridized chip are optically scanned and stored as an image file. This data is then quantified using image analysis software and stored as the feature level extraction output (FLEO). Each row in a FLEO file usually represents a probe and its pixel-level intensity summaries of the foreground signal, background signal and quality measures. Examples of FLEO files generated include probe results (CEL) files for the Affymetrix platform and GenePix Results (GPR) files for most two-channel technology platforms. Each

array type is associated with a library file that contains, for every probe in a FLEO file, its location on the array, biological annotation and sometimes the sequence data. Examples of the annotation library include the Chip Description File (CDF) for Affymetrix and the GenePix Array List (GAL) file for most two-channel technology platforms.

The FLEO data is then converted using preprocessing algorithms into the gene expression data matrix (GEDM). Preprocessing algorithms try to minimize systematic noise and summarize the various intensities per probe into a single value. The GEDM is typically represented as a matrix with rows representing probes and columns representing samples and is the starting point for many statistical analysis.

A published gene list (PGL) represents the genes which are declared as differently expressed in a given study. These form the basis for biological interpretation and discussion. PGLs are often presented in the main or supplementary text of microarray based studies.

Every image file, every FLEO file and every column in a GEDM file can be regarded as individual patient-level data (IPD) since it represents the information for a sample which represents an individual.

### 1.5.4. Microarray platforms and the associated preprocessing algorithms

Some laboratories with sufficient research funding own robotic machines to make their own custom-made cDNA chips. However this requires a high startup cost, good probe design skills and strict quality control. The alternative is to purchase pre-fabricated microarray chips from commercial companies. These companies usually provide the consumables and protocols to help hybridize the biological sample onto the chip as well as machines and softwares to optically scan and quantify the hybridized chips.

Unfortunately, different laboratories and commercial companies use different sets of probes and designs which may require distinctly different preprocessing algorithms. Pre-processing algorithms try to reduce some of the systematic errors, quantify and summarize the expression values.

We focus our attention on the two most widespread and popular microarray platforms: the

Affymetrix platform and a set of platforms that could be generically classified as "two-channel technology" platforms.

### Two-channel technology platforms

Arrays from most two-channel technology platforms typically use long sequence probes, print-head spotting and competitive two-sample hybridization to control for printing variability. The sample of interest and a control sample are stained with dyes that fluorescence at different wavelengths (see Section 1.5.1). After hybridization and scanning, the images are gridded and spotted, manually or using automated software, to account for non-uniformity in spatial positioning and spot intensity. This allows image analysis software to estimate the foreground (also known as the signal) and background intensity for the two channels/dyes. The superimposition of these two intensities can be translated into a false color representation.

Figure 1.4 illustrates the distribution of the ratio before and after transformations from a typical study. The left most panel shows the boxplot of the signal-to-background ratio from ten different arrays and each boxplot summarizes information for tens of thousand of genes from an array. Since we expect half the genes to be under-expressed by chance, about 50% of the data will be between 0 and 1 (red horizontal line) and the remaining 50% of the data to be above 1. A log transformation helps to visualize the data better because it represents the reciprocals symmetrically. The distribution of the log ratio shows that the signals from some arrays are systematically low or high for all genes and also the spread of the log ratios are uneven. This could be due to the different starting RNA quantities, strengths in dye concentration, scanning intensity or other systematic bias. A simple approach of removing them is to center the data from each array to have mean zero and unit variance.

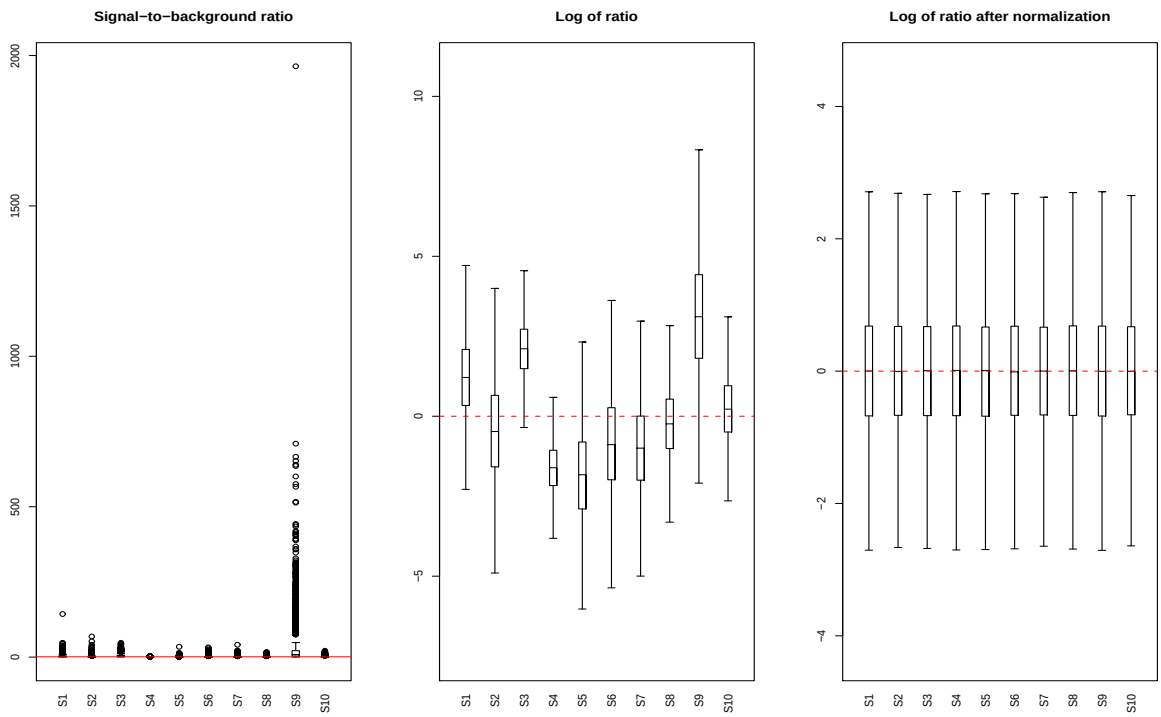



Figure 1.4.: Illustration of the normalization procedure. The distribution of the ratio before any transformation (left), after log transformation (middle) and after standardizing each array to have zero mean and unit variance (right).

The approach described above is a simplistic one. A more sophisticated and better approach would be to calculate the ratio of the background adjusted signal for the two channels in each array using a LOESS normalization (Smyth and Speed, 2003; Yang *et al.*, 2002), a local regression method, to correct for the trend of the ratios depending on the local signal intensity. The LOESS normalization could be stratified by the print-head tip. The print head contains many print-tips and minor non-uniformity among these can result in different spotting intensity. i.e. different amounts of biological material being deposited. However, print-head information from published study may be difficult to obtain as there are many two-channel array designs.

The most common transformation is to take log base 2. This is a simple and by far the popular approach in microarray though other options such as variance stabilizing techniques are also available (Huber *et al.*, 2002; Lin *et al.*, 2008). Finally, we set the distribution of log of gene expression ratios in an array to have mean (or median) zero and unit variance (or median absolute deviation), to account for between-array variability. Now, we can assume

that the expression values for any given gene is Gaussian distributed between arrays and thus use many of the statistical tool already available for analyses of univariate variables (e.g. $t$-test, ANOVA).

### Affymetrix platform

Arrays from the Affymetrix platform are characterized by short sequence probes, photolithographic printing, high density, multiple probes per probeset, and perfect match and mismatch probes to control for unspecific hybridization. Preprocessing for Affymetrix arrays involves automatic background noise correction within an array, normalization between arrays and calculating a probeset summary.

For Affymetrix data there are four popular expression measures:

- MAS 5.0 : Microarray Affymetrix Suite 5.0 (Affymetrix, 2001)
- DCHIP : Li-Wong's Model Based Expression Index (Li and Wong, 2001)
- RMA : Robust Multichip Analysis (Irizarry *et al.*, 2003a)
- GC-RMA : Robust Multichip Analysis with GC content correction (Wu *et al.*, 2004)

Bolstad *et al.* (2003) shows that RMA and dChip are more precise, especially for genes with low transcript abundance, but perhaps slightly less accurate than MAS 5.0. Wu and Irizarry (2004) compares GC-RMA to RMA. Whatever their relative merits, all four expression measures are popular and affects parameter estimation (see Section 3.4.2)

## 1.6. Summary

This introductory chapter sets the background for the subsequent chapters. We have reviewed the basics of microarray technology and it's application for molecular biology and cancer. We also introduced the two topics that this thesis intends to cover: sample size calculations and meta-analysis of microarray datasets.

# 2. Statistics for two-class comparison in microarray

*This brief chapter introduces some of the basic terminology and statistics used for the sample size estimation and meta-analysis chapters.*

## 2.1. Simple summaries of gene expression

The gene expression data matrix (GEDM) is the starting point for many statistical analysis. It is often structured as a matrix with rows representing probes (which in turn represents genes) and columns representing the samples. Here, we assume that the GEDM is sufficiently preprocessed and transformed so that the gene expression values are normally distributed. This should be checked by visual inspection (e.g. using boxplots as in Figure 1.4). Further we assume there are two clinically well defined groups, say, tumors (A) and normals (B).

Table 2.1 illustrates a GEDM with 5 tumor samples and 3 normal samples shown only for the first five probes. Missing values which may arise as a result of poor hybridization or artifacts is coded as "NA" in the table below.

| | $Tumor1$ | $Tumor2$ | $Tumor3$ | $Tumor4$ | $Tumor5$ | $Normal1$ | $Normal2$ | $Normal3$ |
|---|---|---|---|---|---|---|---|---|
| $probe1$ | -0.63 | -0.82 | 1.51 | -0.04 | 0.92 | -0.06 | 1.36 | -0.41 |
| $probe2$ | 0.18 | NA | 0.39 | -0.02 | 0.78 | -0.16 | -0.10 | -0.39 |
| $probe3$ | -0.84 | 0.74 | -0.62 | 0.94 | 0.07 | 1.53 | 3.39 | 2.94 |
| $probe4$ | 1.60 | 0.58 | -2.21 | 0.82 | -1.99 | -0.48 | -0.05 | NA |
| $probe5$ | 0.33 | -0.31 | 1.12 | NA | 0.62 | -2.58 | -4.38 | -2.24 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |

Table 2.1.: An illustrative example of a GEDM.

The expression profile of a given gene can be succinctly summarized by the group means ($m_A$ and $m_B$), within-group standard deviations ($s_A$ and $s_B$) and group size ($n_A$ and $n_B$). Table 2.2 presents these summary statistics along with few other statistics that will be explained later in this chapter.

| | $n_A$ | $m_A$ | $s_A$ | $n_B$ | $m_B$ | $s_B$ | $\delta$ | $S_p$ | $\Delta$ | $g$ | $var(g)$ | $t$ | d.f. | $p$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $probe1$ | 5 | 0.19 | 1.00 | 3 | 0.30 | 0.94 | -0.11 | 0.98 | -0.11 | -0.10 | 0.73 | -0.15 | 6 | 0.884 |
| $probe2$ | 4 | 0.33 | 0.34 | 3 | -0.22 | 0.15 | 0.55 | 0.28 | 1.95 | 1.64 | 1.01 | 2.55 | 5 | 0.051 |
| $probe3$ | 5 | 0.06 | 0.79 | 3 | 2.62 | 0.97 | -2.56 | 0.86 | -2.99 | -2.60 | 1.17 | -4.10 | 6 | 0.006 |
| $probe4$ | 5 | -0.24 | 1.74 | 2 | -0.26 | 0.30 | 0.02 | 1.56 | 0.02 | 0.01 | 0.84 | 0.02 | 5 | 0.985 |
| $probe5$ | 4 | 0.44 | 0.60 | 3 | -3.07 | 1.15 | 3.51 | 0.86 | 4.07 | 3.43 | 1.58 | 5.33 | 5 | 0.003 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |

Table 2.2.: Some summary statistics for the GEDM presented in Table 2.1

## 2.2. Statistical difference

The difference in population group means, $\delta$, is also known as *log fold change* (since the raw expression values are log-transformed to achieve normality). This parameter is calculated as the difference in observed group means, $\delta = m_A - m_B$.

If we assume the variability of the expression profiles for the two groups are the same, then the best estimate of population standard deviation, $\sigma$, is the pooled within-group standard deviations (Cohen, 1988, page 44) described in the equation below:

$$S_p^2 = \frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2} \tag{2.1}$$

The Student's $t$-test (Student, 1908) can be used to test the null hypothesis that the two (unpaired) group means are equal. It assumes the population variation for the two groups is equal. It is calculated as

$$t = \frac{m_A - m_B}{S_p\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}} \tag{2.2}$$

and follows a $t$-distribution with $n_A + n_B - 2$ degrees of freedom. Another alternative to use is the Welch t-test (Welch, 1947) which does not make the equal variance assumption. It replaces the denominator for Equation 2.2 with $\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}$ and follows a $t$-distribution with the degrees of freedom approximated by the Welch-Satterthwaite equation (Welch, 1947; Satterthwaite, 1946) instead.

In microarray research, where several thousand of genes are studied simultaneously, one could utilize information from other genes to better estimate gene-specific variances. Some of the test statistics developed along this line include the significance analysis of microarrays, SAM (Tusher *et al.*, 2001); empirical Bayes analysis of microarrays, EBAM (Efron *et al.*, 2001; Efron and Tibshirani, 2002) and linear models for microarray data, LIMMA (Smyth, 2005). These modify the standard $t$-test (and the corresponding degrees of freedom) by adding a constant to the denominator of Equation 2.2, effectively shrinking the test statistics towards zero.

These test statistics can be converted into $p$-values which can be adjusted for the multiple hypothesis testing problem (see next subsection). $p$-value is defined as the probability of obtaining a test statistic as extreme or more extreme than the observed statistic, *when*

*the null hypothesis is true.* The $p$-value can be calculated from the $t$-distribution with the appropriate degrees of freedom or empirically using permutation testing.

The permutation method in microarray research often involves calculating the test statistics several tens of thousands of times, each time randomly assigning class labels to arrays, to generate the empirical distribution for the null hypothesis. Then one simply estimates the $p$-value as the proportion of test statistics under the null hypothesis which are as large or larger than the observed statistics.

### 2.2.1. The multiple hypothesis testing problem

A *Type I error* or *false positive* occurs when one incorrectly rejects the null hypothesis. In other words, one incorrectly declares a gene as significantly differentially expressed when there is no evidence for it.

In microarray research, one often tests the null hypothesis, i.e. that an individual gene is not differentially expressed between two conditions, for tens of thousands of genes simultaneously. Therefore using a traditional level for Type I error of 0.05 for selecting genes from an array with 10,000 genes can lead to 500 genes being declared as significantly differentially expressed by chance alone. The false positives could overwhelm the true positives and thus the list of genes selected as differentially expressed could be meaningless.

In order to reduce the number of false positive, one often specifies a much smaller per-gene Type I error rate. The so called *multiple hypothesis problem* in microarray studies has fueled many developments in the statistical methodology (Dudoit *et al.*, 2003; Lee and Whitmore, 2002). Two popular methods to overcome the problem are the Bonferroni correction (Bonferroni, 1936) and False Discovery Rate (Benjamini and Hochberg, 1995).

The Bonferroni approach seeks to control the family-wise error rate which is the probability of rejecting at least one hypotheses (i.e. gene is not statistically significant) in the family of hypotheses. This approach treats the hypotheses as independent and ignores the dependence between genes. The Bonferroni corrected $p$-value is calculated as

$$P_{BONFERRONI} = pmin(n_g \times \tilde{p}, 1.0) \tag{2.3}$$

where $\tilde{p}$ is the vector containing the nominal $p$-values for all probes and $n_g$ is the total number

of hypothesis/probes tested. *pmin* calculates the parallel minimum between two vectors (i.e. the element-by-element minimum). Mathematically, $pmin(\tilde{x}, \tilde{y}) = \{min(x_i, y_i);\ i = 1, ..., n\}$, provided $\tilde{x}$ and $\tilde{y}$ have exactly $n$ elements.

Benjamini and Hochberg (1995) approach aims to control the False Discovery Rate (FDR), which is the expected proportion of false positives in the list of genes declared as significantly differentially expressed. For example, if we declare 300 genes as differently expressed with FDR adjusted $p$-value of 0.05 or lower, then we expect (on average) 15 of these genes to be false. The FDR procedure is less conservative than the Bonferroni method and thus more appropriate for microarray data analyses as we are often willing to tolerate a higher number of false positives which can be eliminated through further investigation rather than missing out on any truly important genes (Devlin *et al.*, 2003; Nakagawa, 2004).

$$P_{FDR} = cummin\left( \left[ \frac{p_{[i]} \times n_g}{n_g + 1 - i}\ ;\ i = 1, ..., n_g \right] \right) \tag{2.4}$$

where $p_{[i]}$ represents the $i^{th}$ element of the ordered nominal $p$-value (from largest to smallest). This means that the output from Equation 2.4 needs to be re-sorted into the original order. *cummin* calculates the cumulative minimum of a vector sequentially. Mathematically, if $\tilde{z} = cummin(\tilde{x})$ then $\{z_i = min(x_i, z_{i-1})\ ;\ i = 2, ..., n\}$ and $z_1 = x_1$.

## 2.3. Effect size as a measure of biological difference

In the preceding sections of this chapter, we discussed how to find genes with statistically significant difference. However, a statistically significant difference does not guarantee the difference will be of practical importance. Note that since statistical significance depends on the sample size, even a biologically small difference could be translated as a statistically significant difference if sample sizes are large enough. Conversely, a large biological difference may not be statistically significant if the sample sizes are very small.

The practical difference or *effect size* measures the strength of a gene to discriminate between the two classes. As we show in subsequent chapters, the effect size plays a crucial role in both the sample size estimation and as the common currency to summarize findings from different studies (i.e. a meta-analysis).

There are several ways of defining sample size for a gene expression profiling which is a continuous measure in a two-class comparison. Firstly, one could simply measure it as the unstandardized difference in the group means, $\delta = m_A - m_B$. The other alternative is to use **Cohen's d** (Cohen, 1988) which is the standardized log fold change, a unitless measure. Note that we write Cohen's d as $\Delta$ instead of **d** in this thesis for typographical clarity.

$$\Delta = \frac{\delta}{\sigma} \tag{2.5}$$

In Section 3.4.2, we show that the distribution of estimated $\delta$ across genes depends on the microarray platform and preprocessing algorithm used while $\Delta$ is more robust. Thus this humble transformation plays a pivotal role in sample size estimation for microarray research.

Hedges and Olkin (1985) showed that Cohen's d overestimates the effect size for studies with small samples sizes. They proposed a small correction factor to give the unbiased estimate, which is known as the **Hedge's adjusted *g*** is given below. Please see pages 5 - 6 of Hedges (1981) (also accessible from http://www.jstor.org/stable/1164588) for more details on the Hedge's approximation term.

$$g = \frac{\delta}{\sigma} \times \left(1 - \frac{3}{4(n_A + n_B) - 9}\right) \tag{2.6}$$

For large sample sizes, the difference between Hedge's adjusted $g$ and Cohen's $d$, which is quantified by the correction factor, $\lambda = \frac{g}{\Delta} = 1 - \frac{3}{4(n_A+n_B)-9}$, is small as shown in the table below.

| $n_A + n_B$ | 5 | 10 | 15 | 20 | 25 | 50 | 100 |
|---|---|---|---|---|---|---|---|
| $\lambda$ | 0.73 | 0.90 | 0.94 | 0.96 | 0.97 | 0.98 | 0.99 |

Table 2.3.: The value of the correction factor in Hedge's adjusted $g$ for various sample sizes

The variance of Hedge's g, $v = var(g)$, is calculated as

$$v = \frac{n_A + n_B}{n_A n_B} + \frac{g^2}{2(n_A + n_B - 3.94)} \tag{2.7}$$

## 2.4. Software packages used

The R software (R Development Core Team, 2004) is an open source and free statistical software developed and maintained by experts around the world. There are many additional packages which are contributed by leading researchers resulting in an extensible software. A large subset of the R packages focusing on genetics and molecular biology were compiled and/or developed under the umbrella of BioConductor project (Gentleman and Carey, 2005). Both R and BioConductor have very active public mailing lists as well as few more special interest group mailing lists. The table below gives the URL to download the softwares, additional packages and main mailing lists for each project.

| | |
|---|---|
| R software and packages | http://www.cran.r-project.org/ |
| R mailing list | https://stat.ethz.ch/mailman/listinfo/r-help/ |
| BioConductor software and packges | http://www.bioconductor.org/ |
| BioConductor mailing list | https://stat.ethz.ch/mailman/listinfo/bioconductor/ |

We used R version 2.7.0 and BioConductor version 2.2.0 throughout this thesis.

## 2.5. Summary

This brief chapter covers the basic statistical terminology and formulae that we will use in the subsequent chapters.

# 3. Sample size for microarrays: Minimum number of biological samples

*One of the rationale for meta-analysis is the lack of statistical power in individual studies. Indeed, it is generally accepted that the sample size in microarray studies in inadequate. In this chapter we formally assess the the statistical power of individual microarray studies, before proceeding to combine them in a meta-analysis.*

*During the course of this research, we noticed a lack of clarity on how to calculate sample sizes in practise. Thus, in this chapter, we also provide a step-by-step guide for estimating the minimal sample size, which is equivalent to calculating the statistical power.*

*Materials from this chapter are being prepared for submission as:*

Ramasamy A, Mondry A, Holmes CC, Altman DG (2009)
Three simple extensions to the commonly used univariate, $z$-test based formula for sample size estimation in two-class microarray designs (in preparation).

Ramasamy A, Mondry A, Holmes CC, Altman DG (2009)
Step-by-step guide for sample size calculation for a two-class microarray experiment (in preparation).

## 3.1. Introduction

In this chapter we consider the minimum number of independent *biological replicates* or *subjects* required for a two-class comparison. By contrast, the term *technical replicates* refers to the number of measurements (i.e. microarray hybridization) performed for each subject.

First, we define and explain the importance of statistical power. Next we discuss the commonly used algorithm for two-class comparisons, its shortcomings and three simple adjustments for its improvement. Then we describe how to estimate the parameters and therefore how to estimate sample sizes. We also provide the R codes for generic sample size and power estimation in Appendix A.4 and a look-up table for special case of equal allocation in Table 3.6

The 24 Affymetrix comparisons from studies used for meta-analysis in Chapter 8 and listed in Table 8.2 is used for empirical demonstration of concepts presented here and for parameter estimation in sample size calculation. We use the notations introduced in Section 2.

In Chapter 4, we extend this work to consider both the biological replicates and the number of *technical replicates* that minimizes the total cost of the study. This chapter also contains a case study to demonstrate how to robustly estimate parameters from multiple studies. Other types of replication (e.g pooling DNA, arrays with replicate spots) are not considered in this thesis.

## 3.2. Definition and importance of statistical power

Suppose we have a method to test individuals for the presence of a disease. The test is correct when it either yields a positive result for a person with disease or a negative result for a person without the disease. *Type I error*, at the rate $\alpha$, occurs when the test incorrectly yields a positive result when the disease is absent while *Type II error*, at the rate $\beta$, occurs when the test fails to detect the presence of a disease. Since Type I error is typically considered more serious, it is usual to calculate the sample size by varying $\beta$ for a fixed maximum acceptable $\alpha$. The relationship between the test result and true disease status can be summarized in

the table below.

| | | Test Result: Negative | Test Result: Positive |
|---|---|---|---|
| True Nature | Non-disease | True Negative | False Positive<br>Type I error ($\alpha$) |
| | Disease | False Negative<br>Type II error ($\beta$) | True Positive<br>Power ($1-\beta$) |

Table 3.1.: Relationship between some statistical concepts related to power

Statistical power ($= 1 - \beta$) is the ability to correctly identify the disease when it is present. This is an important concept in planning experiments. An underpowered design will not be able to identify the changes even though it may be present and therefore we may miss important discoveries. An overpowered design may be an inefficient use of resources, which are often obtained from public funding, and can be put to better use such as validating the findings instead.

## 3.3. Commonly used formulaes for sample size calculation

The most common and simplest formula to use is the traditional univariate sample size calculation based on $t$-test or $z$-test procedures with a very stringent Type I error rate to account for the multiple hypothesis problem. This method is popular both because of its simplicity and because many researchers are already familiar with controlling for the multiple hypothesis problem in a similar manner in other experimental settings. There are several

variants in use that we discuss next.

**Sample size based on non-central $t$-test.** This is the most accurate representation of the derivation of sample size but has been applied by only a couple of researchers in the context of microarrays (Wei *et al.*, 2004; Yang *et al.*, 2002). The total sample size, $n = n_A + n_B$, is given by

$$n_A + n_B = 4 \left( \frac{t_{n-2,ncp,\alpha/2} + t_{n-2,ncp,\beta}}{\delta/\sigma} \right)^2 \tag{3.1}$$

where $n_A$ and $n_B$ are the sample sizes for the two groups and $\delta$, represents the difference in population group means and $\sigma$ is the standard deviation of the distribution. $\delta$ is estimated as the difference in observed group means, $m_A - m_B$ and $\sigma$ as the pooled standard deviation (see Equation 2.1). And $t_{n-2,ncp,i}$ is the $i^{th}$ quantile of a $t$-distribution with $n-2$ degrees of freedom and non-centrality parameter, $ncp = \frac{\delta}{\sigma\sqrt{1/n_A + 1/n_B}}$.

**Sample size based on central $t$-test.** Equation 3.1 is computationally and theoretically the most intractable since it requires iterative solution since the degree of freedom and non-centrality parameter are both functions of sample size. A common simplification is to ignore the non-centrality parameter which leads to the use the central t-distribution (McShane *et al.*, 2003; Wang and Chen, 2004; Zhang and Gant, 2004). However, the formula still involves degrees of freedom which remain a function of sample size and thus requiring iterative solution.

**Sample size based on $z$-test.** A more useful simplification is to make use of the asymptotic convergence of a $t$-distribution to $z$-distribution (a standard gaussian distribution) to get a non-iterative closed form solution. Articles proposing this method in the context of microarrays include Dobbin and Simon (2002); Cui and Churchill (2002); Wenisch (2002); Black and Doerge (2002); Simon *et al.* (2002); Dobbin *et al.* (2003); Yang *et al.* (2003); Wit and McClure (2004); Simon *et al.* (2004); Wei *et al.* (2004); Dobbin and Simon (2005) and Seo *et al.* (2006). The total sample size based on the z-test is given as:

$$n_A + n_B = 4 \left( \frac{z_{\alpha/2} + z_\beta}{\delta/\sigma} \right)^2 \tag{3.2}$$

where $z_i$ is the $i^{th}$ quantile of a standard Gaussian distribution. Equation 3.2 is quite intuitive and shows that we need a larger sample size to detect a smaller difference in population means ($\delta$) or when the variability of measurement ($\sigma^2$) is higher.

While Equation 3.2 is intuitive and easy to implement and thus the most common. However,

it is has few drawbacks. First, the solution from Equation 3.2 is consistently smaller sample size than Equation 3.1. Second, it assumes an equal allocation of sample sizes between the two groups $n_A = n_B$. And finally, we show that estimating $\delta$ and $\sigma$ separately can lead to inconsistent sample size estimates.

## 3.4. Three simple improvements to overcome the shortcoming of the sample size based on $z$-test

We propose the following alternative formulation of Equation 3.2 (see Appendix A.1 for derivation) that leads to a more accurate, generalized and robust solution in the context of microarray research.

$$n_A + n_B = \frac{(k+1)^2}{k}\left(\frac{z_{\alpha/2} + z_\beta}{\Delta}\right)^2 + \frac{z_{\alpha/2}^2}{2} \tag{3.3}$$

where $k = n_A/n_B$ is the allocation ratio and $\Delta = \delta/\sigma$ is the Cohen's d measure (Equation 2.5).

In brief, we propose three modifications in Equation 3.3. First, we generalize the formula to incorporate an allocation ratio, $k$. Note that the multiplicative factor of $(k+1)^2/k$ reduces to 4 when $k = 1$. Secondly, we replace $\delta$ and $\sigma$ with $\Delta = \delta/\sigma$. Finally, we include an additive correction term of $z_{\alpha/2}^2/2$, which is an adjustment term for using the normal approximation instead of a $t$-distribution.

The proposed changes may look cosmetic but have important implications in the context of microarrays. Next, we discuss why each of the modification is important.

### 3.4.1. Improvement 1: Generalizing to unequal allocation ratio

It is easy to show that the statistical power for a *fixed total sample size* is maximized with equal allocation of samples between the two classes i.e. $k = 1$. This can be shown by differentiating Equation 3.3 with respect to $k$ or graphically (see Appendix A.2). This is perhaps why the most common formulation does not account for unequal allocation ratio.

However, when samples from one of the groups may be prohibitively expensive or difficult to obtain, it is possible to increase the power, albeit inefficiently, by sampling more from the more common group. Such scenarios arise when one compares pathological samples from a low risk group (e.g. lung cancer from non-smokers) to those from a high risk group (e.g. lung cancer from smokers) or a disease with low prevalence or when there are logistical and ethical difficulties. Thus, we need to allow for unequal allocation ratio as we do in Equation 3.3.

With equal allocation ratio, $(k+1)^2/k$ reduces to 4. If we define the *allocation efficiency* of an existing design as the sample size relative to a hypothetical design that uses the same total sample size but with $k = 1$, then it can be approximated as $\frac{4k}{(k+1)^2}$.

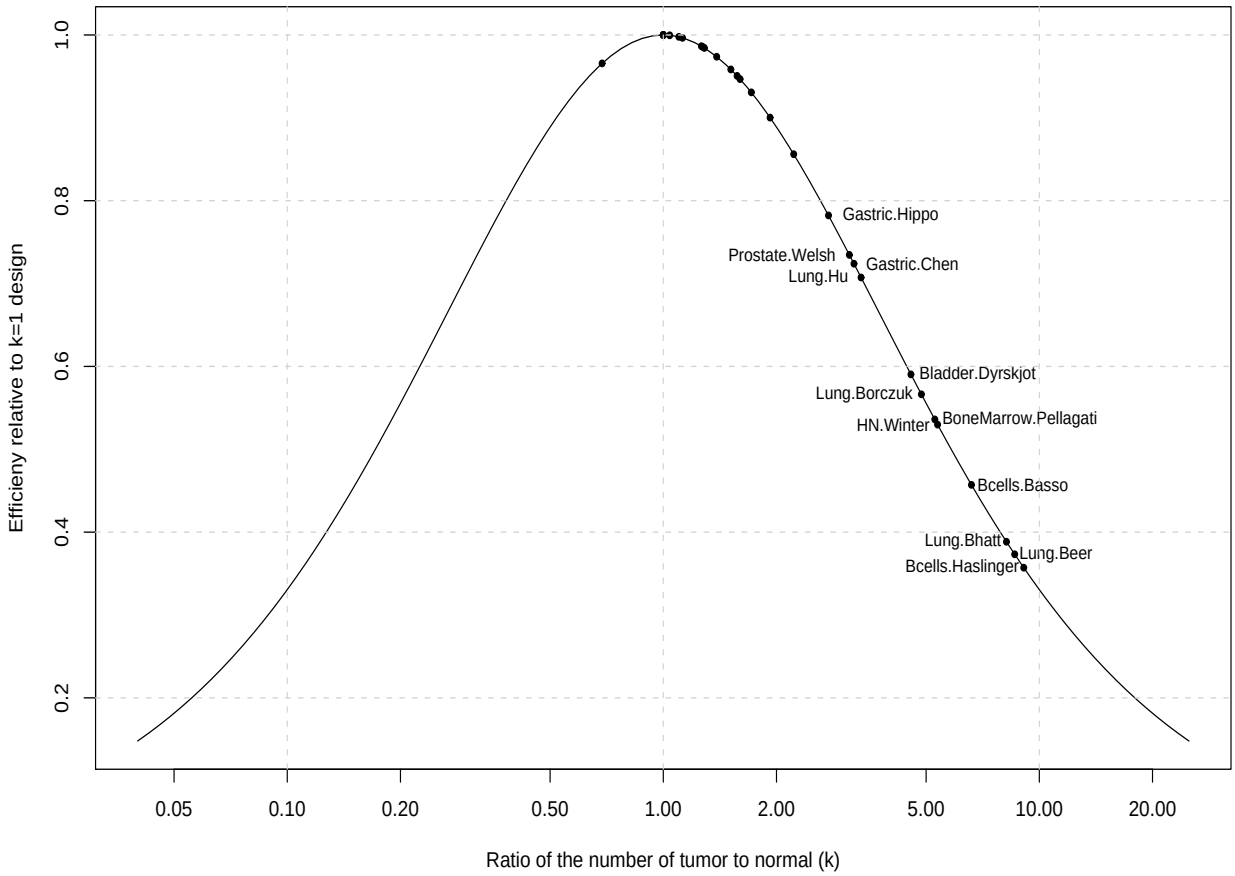



Figure 3.1.: Plot showing the approximate allocation efficiency of current design relative to studies with equal allocation between tumors and normals (k=1) for the studies used in Chapter 8. The study labels for the 12 studies with efficiency ratio less than 80% is shown.

Figure 3.1 depicts the approximate allocation efficiency for the studies used in Chapter 8. For example, Basso *et al.* (2005) which compares 10 normals with 66 tumors ($n_A + n_B = 76$, $k = 6.6$) has an allocation efficiency of 0.457 compared to a study that compares 38 normals to 38 tumors ($k = 1$). Or to put it another way, the researchers could have achieve the same statistical power by only 34 samples ($\approx 0.457 \times 76$) if they compared 17 normals to 17 tumors. However, we acknowledge that the tumor-normal comparison may not have been the primary objectives of the collected studies here.

### 3.4.2. Improvement 2: Joint estimation of $\delta$ and $\sigma$

Equation 3.2 suggests that one is able to estimate the sample size by fixing $\delta$ and estimating $\sigma$ separately. Indeed, the majority of authors who propose the use of traditional sample size formula for microarray studies suggest using a two-fold change (i.e. $\delta = 1$ on a $\log_2$ scale) and employ some "typical value" of the variance ($\sigma$) from published data. Some of these authors (e.g. Wei *et al.* (2004)) then suggest using a quantile (e.g. $25^{th}, 50^{th}, 75^{th}, 90^{th}$) of the distribution of the empirical standard deviation across genes as the "typical value" of $\sigma$.

However, in practice we found that the distribution of estimated $\delta$ and $S_p$ for one-dye technologies varies greatly by preprocessing algorithm. Therefore fixing $\delta = 1$ and then estimating a typical value of $S_p$ as suggested above can be misleading. We hypothesized that the output from the various preprocessing algorithms are on different measurement scales and standardizing the log fold change relative to its variability would provide for a more coherent framework for sample size calculations irrespective of the preprocessing method used. In other words, we propose to jointly estimate $\delta$ and $\sigma$ as $\Delta = \delta/\sigma$ which is a common approach in many areas of statistics to remove any effect due to scales of measurement.

The first line of evidence is summarized in Table 3.2 which presents the selected quantiles from one of the published studies that we are going to introduce later. It shows that the distribution of $\delta$ and $S_p$ across 12,625 genes do vary hugely between preprocessing algorithms. However the values of $\Delta = \delta/\sigma$ is much more consistent across the four preprocessing algorithm as can be seen from the last four rows of this table and from Figure 3.2.

| | | Quantiles | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Quantity** | **Algorithm** | **0%** | **1%** | **5%** | **10%** | **20%** | **50%** | **80%** | **90%** | **95%** | **99%** | **100%** |
| $\delta$ | MAS 5.0 | -2.12 | -0.92 | -0.66 | -0.55 | -0.42 | -0.17 | 0.21 | 0.48 | 0.72 | 1.24 | 2.77 |
| | dChip | -1905.53 | -174.79 | -79.87 | -51.61 | -29.16 | -7.96 | 8.53 | 37.01 | 88.48 | 502.27 | 3045.53 |
| | RMA | -1.24 | -0.33 | -0.20 | -0.17 | -0.13 | -0.06 | 0.07 | 0.22 | 0.41 | 0.94 | 2.28 |
| | GCRMA | -2.47 | -0.50 | -0.22 | -0.15 | -0.11 | -0.05 | 0.08 | 0.30 | 0.60 | 1.24 | 3.73 |
| $S_p$ | MAS 5.0 | 0.24 | 0.41 | 0.58 | 0.69 | 0.84 | 1.12 | 1.38 | 1.52 | 1.63 | 1.90 | 2.90 |
| | dChip | 6.67 | 10.64 | 14.32 | 17.74 | 24.3 | 50.23 | 120.04 | 195.82 | 315.24 | 1047.11 | 4318.36 |
| | RMA | 0.10 | 0.14 | 0.16 | 0.18 | 0.20 | 0.28 | 0.42 | 0.60 | 0.85 | 1.42 | 2.38 |
| | GCRMA | 0.02 | 0.07 | 0.11 | 0.15 | 0.19 | 0.29 | 0.54 | 0.88 | 1.24 | 1.99 | 3.58 |
| $\Delta$ | MAS 5.0 | -1.97 | -0.86 | -0.58 | -0.49 | -0.39 | -0.17 | 0.22 | 0.46 | 0.63 | 0.91 | 2.50 |
| | dChip | -1.78 | -0.93 | -0.63 | -0.53 | -0.44 | -0.24 | 0.20 | 0.44 | 0.62 | 0.91 | 1.98 |
| | RMA | -1.76 | -1.03 | -0.66 | -0.54 | -0.44 | -0.23 | 0.23 | 0.49 | 0.65 | 0.92 | 2.33 |
| | GCRMA | -1.71 | -0.90 | -0.56 | -0.45 | -0.37 | -0.26 | 0.18 | 0.45 | 0.63 | 0.93 | 2.47 |

Table 3.2.: Empirical values for $\delta$, $S_p$ and $\Delta$ from Singh *et al.* (2002)

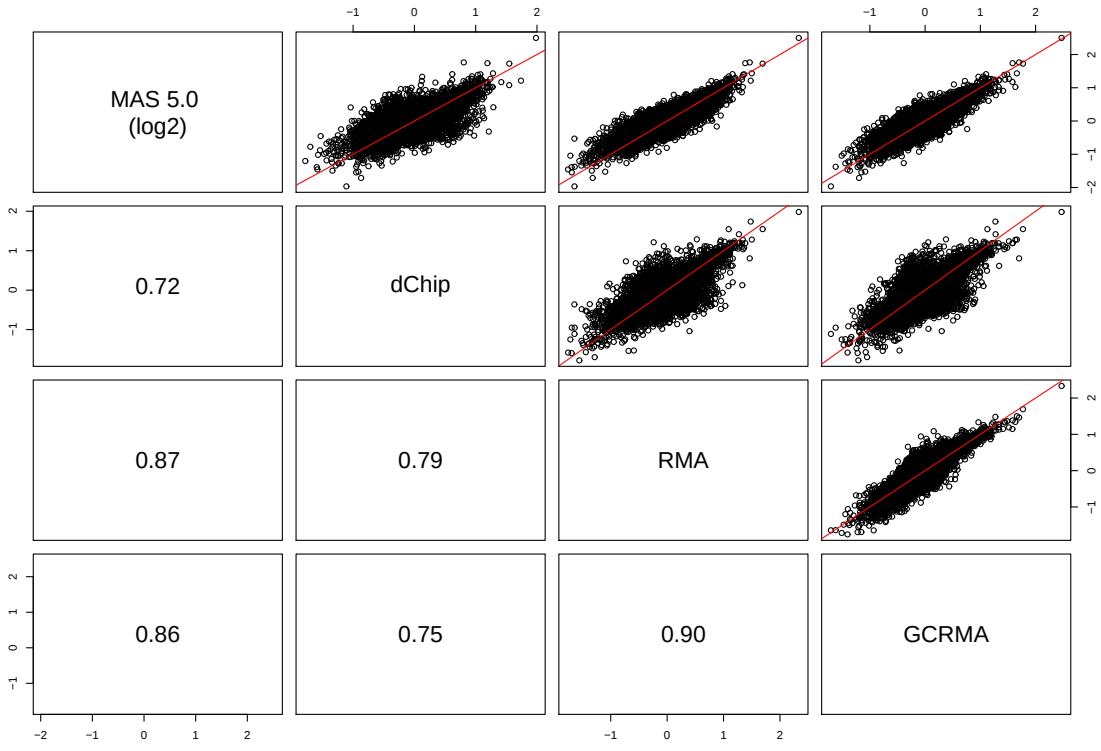



Figure 3.2.: Illustration of the agreement in $\Delta$ values of genes across different preprocessing algorithm in Singh *et al.* (2002). The pairwise linear correlation is displayed in the lower panels.

To generalize this finding, we use the data from 24 Affymetrix two-class comparisons used in Chapter 8 under different preprocessing algorithms. See Section 1.5.4 for more information on preprocessing algorithms. Again, we can see that the distribution of $S_p$ (see Figure 3.3) is strongly affected by the different preprocessing algorithms used in Affymetrix platform; while the distribution of $\Delta$ (see Figure 3.4) is much more consistent within any given study confirming our hypothesis.

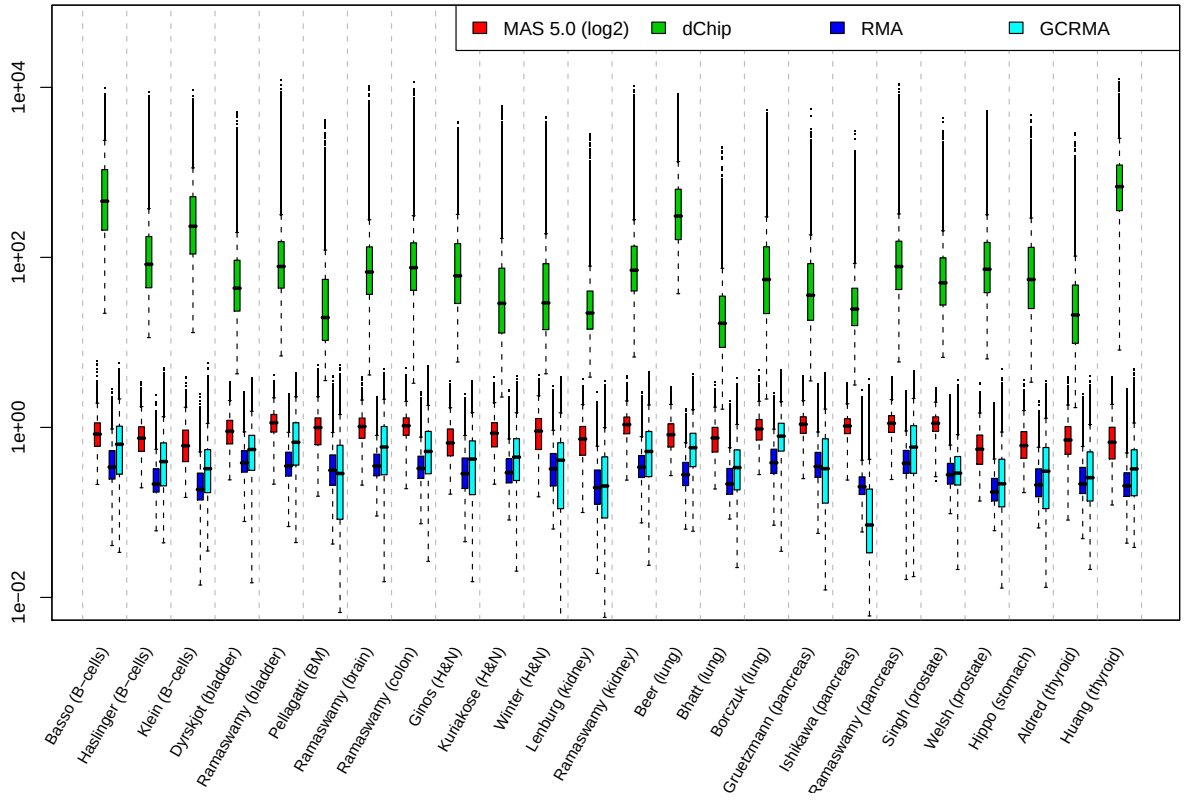



Figure 3.3.: Boxplot of pooled standard deviation $S_p$ (calculated using Equation 2.1) for all genes with each combination of 24 two-class comparison and 4 preprocessing algorithms. The boxplots are grouped first by study then by different preprocessing algorithm. The y-axis is on a $\log_{10}$ scale to aid visualization.

While, there is still a small amount of variability in $\Delta$ between the studies, the figures above and Table 3.3 suggest that it is possible to obtain general estimates from these 24 comparisons for future sample size calculations. For example, the median of the $90^{th}$ quantile of $|\Delta|$ across the 24 datasets vary between 0.82 and 0.86 which would yield sample sizes between 98 - 108 for $\alpha = 0.001, \beta = 0.20, k = 1$. On the other hand, fixing $\delta = 1$ and choosing, say, the $90^{th}$ quantile of $S_p$ for the MAS 5.0, RMA, and GCRMA algorithms (which are 1.44, 0.76,

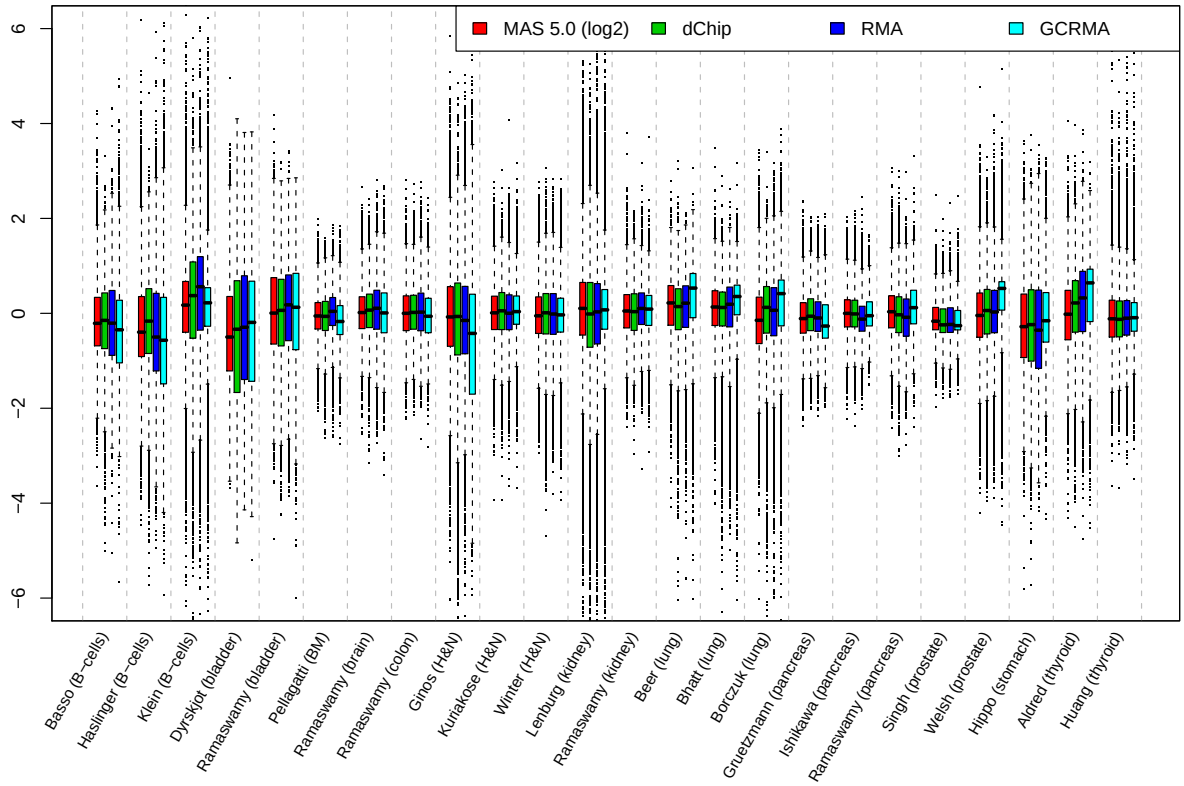



Figure 3.4.: Boxplot of $\Delta$ for all genes with each combination of study and preprocessing algorithm. The figure shows that the distribution of $\Delta$ is more similar between preprocessing algorithm within each study.

1.00 respectively) would yield sample sizes of 146, 46 and 74 for $\alpha = 0.001, \beta = 0.20, k = 1$. The median of the $90^{th}$ quantile of the dChip algorithm is so large that it would require an infinitely large sample size.

By contrasts, we expect the problem of differing measurement scale to be of less significance in the two-channel platforms. This is because the expression values from two-channel platforms are calculated as a ratio relative to the two dyes/samples, thereby automatically removing any effect due to the measurement scale. Affymetrix arrays (and other one-channel platform) use the absolute signal values.

To our knowledge, only Seo *et al.* (2006) has investigated the effect of choice of preprocessing algorithm on sample size estimates. In brief, they found large differences in sample size and power based on probe set algorithm selection with RMA having higher sensitivity at low signal noises as previously shown by Bolstad *et al.* (2003). However, they did not provide

| | Median of the quantiles of $S_p$ | | | | | Median of the quantiles of $\|\Delta\|$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 80 | 85 | 90 | 95 | 99 | 80 | 85 | 90 | 95 | 99 |
| MAS5.0 | 1.24 | 1.32 | 1.44 | 1.62 | 1.95 | 0.48 | 0.65 | 0.84 | 1.09 | 1.69 |
| dChip | 162.05 | 202.56 | 279.46 | 457.95 | 1640.46 | 0.55 | 0.68 | 0.82 | 1.03 | 1.52 |
| RMA | 0.58 | 0.66 | 0.76 | 0.97 | 1.46 | 0.58 | 0.73 | 0.86 | 1.11 | 1.65 |
| GCRMA | 0.76 | 0.86 | 1.00 | 1.27 | 2.12 | 0.56 | 0.69 | 0.85 | 1.08 | 1.63 |

Table 3.3.: The median (across 24 studies) value of selected quantiles of the $S_p$ distribution (left columns) and the $\Delta$ distribution (right columns) for each of the four preprocessing algorithms. The median of the quantiles from $\Delta$ distribution shows greater consistency across preprocessing algorithms.

a solution that leads to a coherent sample size estimation under different preprocessing algorithms as we have done here.

### 3.4.3. Improvement 3: The correction term $z^2_{\alpha/2}/2$ is important for small $\alpha$

The exact sample size for a two-class comparison should be calculated using an iterative procedure involving non-central $t$-distribution. Using the asymptotic convergence of a $t$-distribution to normal distribution, we can get a non-iterative approximate solution (Equation 3.2). Guenther (1975) and Guenther (1981) show that the addition of a correction term of $z^2_{\alpha/2}/2$ provides a surprising good approximation, often yielding the exact sample size given by the iterative non-central $t$-distribution.

Why has the correction factor received so little attention? To answer this, we need to inspect Table 3.4 which shows the size of the correction factor (i.e. the difference in outputs between Equation 3.3 and Equation 3.2) for a varying levels of $\alpha$ when all other parameters are fixed.

| $\alpha$ | 0.05 | 0.01 | 0.001 | 0.0001 | 0.00001 |
|---|---|---|---|---|---|
| $\frac{z^2_{\alpha/2}}{2}$ | 1.92 | 3.32 | 5.41 | 7.57 | 9.76 |

Table 3.4.: Value of the correction term as a function of significance level ($\alpha$)

We can see that when $\alpha = 0.05$, as in a conventional study, the correction term adds only 2 additional samples to the total sample size. Further, an increase of two additional samples may represent only a tiny fraction of the total sample size, which may explain why it has not received more attention.

However in the context of microarray, one would typically choose $\alpha$ between 0.0001 and 0.00001, where it represents $8 - 10$ additional samples. Further since the total sample size for many microarray experiments is small to begin with, this correction term could represent a significant proportional increase over the total. Therefore ignoring the correction term as Equation 3.2 does, is inappropriate in the context of microarray experiments.

## 3.5. How to estimate the sample size for microarrays

For reasons mentioned in the earlier section, we prefer to use Equation 3.3. In this section, we provide a practical step-by-step guide to assist researchers in using this formula to calculate the sample sizes for their study. We start with the estimation of the parameters used in Equation 3.3 for which we assume the researchers have access to either a pilot data or other similar published study (see Section 6.3.1 for some guidance on finding suitable studies).

### 3.5.1. Step 1: Calculate the standardized effect Sizes

First, we need to estimate the distribution of standardized effect sizes for all genes. One can calculate the effect size using Equation 2.5, $\Delta = \delta/\sigma$. Here where the numerator $\delta = \mu_A - \mu_B$ is estimated as the difference in sample group means $m_A - m_B$. The square of the denominator $\sigma^2$ can be estimated as a whole by the pooled (within group between subject) variance from Equation 2.1.

However if the pilot data or published data is very small, say $n_A + n_B < 30$, we strongly suggest using Hedge's adjusted $g$ as defined in Equation 2.6 instead of $\Delta$ to account for the bias in effect size estimation due to the small sample size of pilot study.

How large is the difference in sample size estimated using $\Delta$ and and the same size estimated using $g$? We can answer this by making use of two simple points that have already been discussed. First, we know that $\lambda = g/\Delta = 1 - \frac{3}{4(n_A+n_B)-9}$ where $n_A, n_B$ here refers to the sample size of existing pilot study (not the planned study for which sample size is to be estimated). Second, we know from Equation 3.3 that the sample size estimate is inversely proportional to the effect size. Thus if we denote sample size estimated using $g$ as $SS_g$ and the sample size estimated using $\Delta$ as $SS_\Delta$, then we have $SS_g \propto \frac{1}{g^2} = \frac{1}{(\lambda\Delta)^2} = \frac{1}{\lambda^2} SS_\Delta$. Using the values of $\lambda$ tabulated in Table 2.3, the sample size using $\Delta$ would need to be inflated by approximately 23% if the pilot study had only 10 samples in total. The inflation factor drops to 13%, 8%, 6% if the pilot data has 15, 20, 25 samples in total respectively.

### 3.5.2. Step 2: Choose a cutoff for effect size

Reformulating the sample size in terms of standardized fold change now poses the question of which value of effect size to choose. We like to answer this by first pointing out that there is no absolute reason to choose a two-fold change or a $p$-value cutoff of 0.05. These values were simply used out of convention as a threshold beyond which an observed difference is considered interesting or believable.

Step 1 produces a distribution for $\Delta$ or $g$ across the genes. If we assume that majority of the important genes are to be found in, say, the top 10% of the ranked genes, then we power the study to detect at least the $90^{th}$ quantile of $\Delta$ or $g$. This information is best elicited from the domain expert of the study. If the chosen value is based on other similar studies, we must be careful to account for the differences in experimental quality, number of probesets and redundancy among probesets. We should use the absolute values of the effect size if we are equally interested in finding the up- and down-regulated of genes (i.e. two-sided testing).

### 3.5.3. Step 3: Choose a significance level ($\alpha$)

As described in Section 2.2.1, we need to select a more stringent value of $\alpha$ to account for multiple testing. The two popular approaches are Bonferroni correction (Bonferroni, 1936) and False Discovery Rate (Storey, 2002), which have been condensed into one formulae by Yang *et al.* (2003)

$$\alpha = \alpha_F \ \max(1, n_0)/n_g \tag{3.4}$$

where $\alpha_F$ is the overall family-wise error rate that we wish to control. $n_g$ is the total number of hypothesis/genes being tested and $n_0$ is the number of genes expected to be differentially expressed.

The value of $n_0$ can be guessed by the domain expert, or derived from models using data from pilot or published studies. When the experimenter is unable to specify $n_0$, one can use $\alpha = \alpha_F/n_g$, which corresponds to the conservative Bonferroni solution.

### 3.5.4. Step 4: Choosing a power level $(= 1 - \beta)$

Power represents the probability of correctly identifying differentially expressed genes. Power is an important concept because an overpowered study makes inefficient use of resources and an underpowered study would not sufficiently address the question of interest and results could be misleading. There is no scientific or empirical justification for choosing a power level but typically a power level of 80 to 95% (or $\beta = 0.20$ to 0.05) is used. As we shall show later, there is complete interchangeability between $\alpha$ and $\beta$.

### 3.5.5. Step 5: Choose the allocation ratio (k)

If there are availability and financial constraints on obtaining samples from one of the groups, then one may need to choose an unequal allocation. When choosing the allocation ratio, the following three points needs to be kept in mind. Let $P_{n_A:n_B}$ denote the power of a design with sample size $n_A$ and $n_B$ for the two groups respectively with fixed values for $\alpha, \beta$ and $\Delta$.

1. For any fixed total sample size, power is maximized with equal allocation ($k = 1$).
   E.g. $P_{25:25} > P_{(50-i):i}$ for $i \neq 25$ and $0 < i < 50$. See Appendix A.2 for proof.

2. When the sample size from one group is fixed (i.e. the rare group), one can increase the power of the design by increasing the sample numbers from the more common group (i.e. the group for which samples can be obtained more easily).
   E.g. $P_{(25+i):25} > P_{25:25}$ for $i > 0$.

3. When $k \neq 1$, increasing the number of subjects from the smaller group will always yield a bigger gain in power than increasing subjects from the larger group.
   E.g. $P_{30:26} > P_{31:25} > P_{30:25}$.

### 3.5.6. Step 6: Calculate the total sample size

Having estimated $\Delta$ and chosen a value for $\alpha$, $\beta$ and $k$, we can now calculate the total sample size according to Equation 3.3. Table 3.5 gives some useful numbers to help calculate the sample size by hand. Finally, we estimate the size of the two groups as $\frac{k}{k+1}$ and $\frac{1}{k+1}$ of the final total sample size.

| | Value of the $(z_{\alpha/2} + z_{\beta})^2$ term | | | | | $\frac{z^2_{\alpha/2}}{2}$ |
|---|---|---|---|---|---|---|
| | $\beta = 0.20$ | $\beta = 0.15$ | $\beta = 0.10$ | $\beta = 0.05$ | $\beta = 0.01$ | |
| $\alpha = 0.05$ | 7.8 | 9.0 | 10.5 | 13.0 | 18.4 | 1.9 |
| $\alpha = 0.01$ | 11.7 | 13.0 | 14.9 | 17.8 | 24.0 | 3.3 |
| $\alpha = 0.001$ | 17.1 | 18.7 | 20.9 | 24.4 | 31.5 | 5.4 |
| $\alpha = 0.0001$ | 22.4 | 24.3 | 26.8 | 30.6 | 38.7 | 7.6 |
| $\alpha = 0.00001$ | 27.7 | 29.7 | 32.5 | 36.7 | 45.5 | 9.8 |

Table 3.5.: Values of the $(z_{\alpha/2}+z_{\beta})^2$ term (left columns) and the correction factor $\frac{z^2_{\alpha/2}}{2}$ (rightmost column) that are useful for sample size calculations.

Here is a simple example of calculating sample sizes. Assume we chose $\alpha = 0.00001$, $\beta = 0.20$, $k = 2$ and $\Delta = 1.2$. From Table 3.5, we know that $(z_{\alpha/2} + z_{\beta})^2 = 27.7$ and we can apply

Equation 3.3 as follows:

$$\begin{aligned} n_A + n_B &= \frac{(k+1)^2}{k}\left(\frac{z_{\alpha/2} + z_\beta}{\Delta}\right)^2 + \frac{z_{\alpha/2}^2}{2} \\ &= \frac{(2+1)^2}{2} \times \frac{27.7}{1.2^2} + 9.8 \\ &= \frac{4.5 \times 27.2}{1.44} + 9.8 \\ &= 86.6 + 9.8 \\ &= 96.4 \end{aligned}$$

and thus we have a total sample size of 96. Finally, we allocate 32 ($= 96 \times \frac{1}{3}$) samples to one group and 64 ($= 96 \times \frac{2}{3}$) samples to the other group.

Alternatively, if one has access to a computer with R software installed (R Development Core Team, 2004), they can also use the function `find.ss` given in Appendix A.4 to obtain the same results.

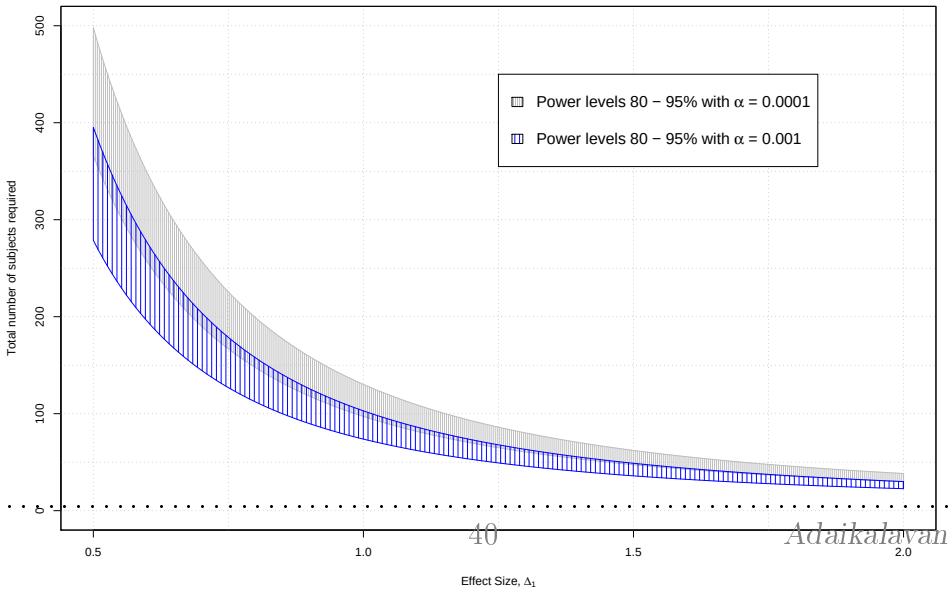
```
> find.ss( Delta=1.2, sig.level=0.00001, power=0.80, k=2)
$‘Delta=1.2‘
            power=0.8
alpha=1e-05      96.1
```

### 3.5.7. Special case of equal allocation ($k = 1$)

As discussed earlier, when the total sample size is fixed an equal allocation ($k = 1$) maximizes statistical power. Table 3.6 and Figure 3.5 give the total sample size in this case. We chose to tabulate and plot for values of $\Delta$ between 0.5 and 2.0 because these values cover the $80^{th}$ and $99^{th}$ quantiles averaged across the 24 comparisons (see Table 3.3).

| | $\alpha = 0.001$ | | | | $\alpha = 0.0001$ | | | | $\alpha = 0.00001$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Power Level | | | | Power Level | | | | Power Level | | | |
| $\Delta$ | 80% | 85% | 90% | 95% | 80% | 85% | 90% | 95% | 80% | 85% | 90% | 95% |
| 2.0 | 22 | 24 | 26 | 30 | 30 | 32 | 34 | 38 | 38 | 40 | 42 | 46 |
| 1.9 | 24 | 26 | 28 | 32 | 32 | 34 | 38 | 42 | 40 | 42 | 46 | 50 |
| 1.8 | 26 | 28 | 32 | 36 | 36 | 38 | 40 | 46 | 44 | 46 | 50 | 56 |
| 1.7 | 30 | 32 | 34 | 40 | 38 | 42 | 44 | 50 | 48 | 52 | 54 | 60 |
| 1.6 | 32 | 34 | 38 | 44 | 42 | 46 | 50 | 56 | 54 | 56 | 60 | 68 |
| 1.5 | 36 | 40 | 42 | 50 | 48 | 52 | 56 | 62 | 60 | 62 | 68 | 76 |
| 1.4 | 40 | 44 | 48 | 56 | 54 | 58 | 62 | 70 | 66 | 70 | 76 | 86 |
| 1.3 | 46 | 50 | 56 | 64 | 60 | 66 | 72 | 80 | 76 | 80 | 86 | 98 |
| 1.2 | 54 | 58 | 64 | 74 | 70 | 76 | 82 | 92 | 86 | 92 | 100 | 112 |
| 1.1 | 62 | 68 | 74 | 86 | 82 | 88 | 96 | 110 | 102 | 108 | 118 | 132 |
| 1.0 | 74 | 80 | 90 | 104 | 98 | 104 | 114 | 130 | 120 | 130 | 140 | 158 |
| 0.9 | 90 | 98 | 108 | 126 | 118 | 128 | 140 | 160 | 146 | 156 | 170 | 192 |
| 0.8 | 112 | 122 | 136 | 158 | 148 | 160 | 176 | 200 | 182 | 196 | 214 | 240 |
| 0.7 | 146 | 158 | 176 | 204 | 190 | 206 | 226 | 258 | 236 | 252 | 276 | 310 |
| 0.6 | 196 | 214 | 238 | 276 | 256 | 278 | 306 | 348 | 318 | 340 | 370 | 418 |
| 0.5 | 278 | 306 | 340 | 396 | 366 | 396 | 436 | 498 | 452 | 486 | 530 | 598 |

Table 3.6.: Minimum total sample size requirements rounded up to the nearest even integer with equal allocation ($k = 1$) between groups.



We can also use Figure 3.5 to determine the influence of each parameter in the sample size calculations. We can see that the sample size is most sensitive to the Type I error rate ($\alpha$) and the power level, particularly at smaller values of $\Delta$. The required sample size can vary from 74 to 158 for $\Delta = 1.0$ and 62 to 132 for $\Delta = 1.1$ depending on the choice of $\alpha$ and power.

Note that there is complete interchangeability between $\alpha$ and $\beta$. This is not too surprising from Equation 3.3. For example, if we decided to power the study to detect at least the top 5% ranked genes (which corresponds to $|\Delta| = 1.085$ from Table 3.3), with a power level of 95%, $\alpha = 0.001$, $k = 1$ requires 88 subjects and this is equivalent to choosing a more stringent error rate of $\alpha = 0.0001$, but with a reduced power level of 85% requires 90 subjects.

## 3.6. Discussion

Here we have explicitly focused on the most commonly used sample size formula for a two-class microarray design. However, its frequent use does not mean it is the most appropriate method under all circumstances.

One can approach the sample size estimation with greater emphasis on false discovery rate or on the correlation structure between genes (Lee and Whitmore, 2002; Jung *et al.*, 2005; Tsai *et al.*, 2005; Pan *et al.*, 2002; Zien *et al.*, 2003; Muller *et al.*, 2004; Pawitan *et al.*, 2005). For example, Pan *et al.* (2002) calculate the power in a multivariate setting by modelling the gene expression statistic as a mixture of normal distributions. Zien *et al.* (2003) fit a complex model to the raw gene expression matrix and estimate the biological and technical variation from this model but it is however unclear what format the data input should be in this model and most researchers would prefer to use standard preprocessing algorithm. Muller *et al.* (2004) finds the sample size that minimizes the loss function which is a weighted average of the False Discovery Rate and False Negative Rate and their use of a Bayesian approach allows the incorporation of prior knowledge. Alternatively, one can try ranking based (Matsui *et al.*, 2008; Matsui and Oura, 2009). Pavlidis *et al.* (2003) provide an empirical study that uses a random sub-sampling approach to find the minimum sample size that does not compromise stability and reproducibility of results when compared to the results from using all available samples. Sample size calculations have also been proposed for more than two-groups designs

(Pounds and Cheng, 2005, 2009) and for other purposes such as estimating the minimum number of replicate probes needed to be spotted on an array (Black and Doerge, 2002) and for classification of samples (Fu *et al.*, 2005; Hwang *et al.*, 2002; Mukherjee *et al.*, 2003).

While all these approaches have immense theoretical importance, they are only rarely used in practice. This fact explains our focus on the univariate formula for sample size estimation. The suggested alterations as per Equation 3.3 increase the generalizability of this formula while maintaining its ease of use.

## 3.7. Summary

While reviewing these articles using the traditional sample size formula, we noticed several issues that were not fully addressed.

First, most of the authors implicitly assume an equal allocation ratio. This poses a problem when samples from one group are difficult or expensive to obtain. For such a scenario, we show how it is possible to increase the statistical power (though less efficiently) by increasing samples from the more common group. We generalized the commonly used formula to allow for an unequal allocation ratio.

Next, we noticed that many authors estimate the variation in gene expression levels while fixing the minimum fold change (difference in group means). Using empirical evidence, we show that this leads to a situation where the sample size numbers are highly dependent on the type of preprocessing algorithm due to different scales of measurement. Thus we develop the idea of standardizing the fold change relative to its variability and empirically show that this allows for a more consistent sample size framework independent of the choice of preprocessing algorithm.

Third, we show that approximating the $t$-distribution with normal distribution consistently underestimates the sample size required, especially for very small values of $\alpha$. We propose a simple addition term that makes the approximation very accurate.

Finally, we provide a step-by-step guide for sample size calculations with empirical evidence.

# 4. Sample size for microarrays: Minimum number of technical replicates through cost analysis

*In Chapter 3, we calculated the minimum number of subjects (i.e. biologically independent replicates) required assuming one measurement per subject (i.e. technical replicate). In this Chapter, we move on from this assumption and show how one can determine the optimal number of measurement per subject.*

## 4.1. Motivation

The standard sample size calculation (see Equation 3.3) does not have any provision for calculating the number of technical replicates required. This reflects the common assumption that one technical replicate per biological replicate is optimal - an assumption that is rarely tested. Increasing the number of technical replicates could reduce the number patients required while maintaining the same level of statistical power. In this Chapter, we show that the number of technical replicates is directly related to the relative cost structure and relative variance components of DNA microarray study design.

## 4.2. Methodology

First recall Equation 3.3 as represented below, for which we will now introduce new parameters and refine the definitions of some of the existing parameters as listed in Table 4.1.

$$n_A + n_B = \frac{(k+1)^2}{k}\left(\frac{z_{\alpha/2} + z_\beta}{\Delta_m}\right)^2 + \frac{z_{\alpha/2}^2}{2} \tag{4.1}$$

| Symbol | Meaning | Relationships |
|---|---|---|
| $n_i$ | Number of *biological replicates* in group $i$ $(i = A, B)$ | $n_B = kn_A$ |
| $\alpha$ | Type I error rate | |
| $\beta$ | Type II error (or 1 - power) | |
| $z_i$ | $i^{th}$ quantile of a standard normal distribution | |
| $\Delta_m$ | The standardized effect size | $\Delta_m = \delta/\sigma_m$ |
| $m$ | Number of technical replicates per biological subject | |
| $\sigma_m^2$ | Variability of the difference in group means with $m$ technical replicates per biological subject | $\sigma_m^2 = Var(\delta)$ |
| $\sigma_B^2$ | The biological variance due to different subjects | $\sigma_m^2 = \sigma_B^2 + \frac{\sigma_E^2}{m}$ |
| $\sigma_E^2$ | The technical variance due to measurement errors | |
| $R$ | Ratio of technical variation to biological variation | $R = \sigma_E^2/\sigma_B^2$ |
| $C_1$ | The cost to recruit a single subject | |
| $C_2$ | The cost to perform a single microarray measurement | |

Table 4.1.: Explanation of the symbols used in this chapter.

Equation 4.1 does not yet provide a direct way to calculate the number of technical replicates required. It does not take into account the different variability and cost components either. For example, if the technical variability was high or if measurements were cheap to perform then one could potentially be better off with a design that has fewer biological replicates but uses several technical replicates each.

We will now demonstrate how the number of technical replicates can be calculated. First, we define $C_1$ as the recruitment cost per subject and $C_2$ as the cost of making a single measurement. In the context of microarrays, $C_1$ typically represents the cost of treatment and RNA extraction. Additionally in the case of animal models, $C_1$ could also include the cost of purchasing the animals and/or breeding as well as housing costs. $C_2$ represents the cost of cRNA production, array cost, reagents and hybridization costs.

If we have recruited $n_A$ and $n_B = kn_A$ subjects and measured each subject $m$ times ($m = 1, 2, \ldots$), then the total cost is given by

$$\begin{aligned} \text{Total Cost} &= (n_A \times C_1 + n_A \times m \times C_2) + (n_B \times C_1 + n_B \times m \times C_2) \\ &= (n_A + n_B)(C_1 + mC_2) \end{aligned} \tag{4.2}$$

Now we will have to find the best combination of $\{n_A, m\}$ that minimizes the total cost while being subject to a fixed power level. Using the Lagrangian method to minimize Equation 4.2 subject to a fixed value of $\beta$ in Equation 4.1 gives (see Appendix A.3 for derivation):

$$\begin{aligned} m_{opt} &= -fR + \sqrt{f^2R^2 + (1-2f)\frac{C_1}{C_2}R} \\ \text{where } f &= \frac{0.25 \times z^2_{\alpha/2}}{n_A + n_B} \end{aligned} \tag{4.3}$$

Equation 4.3 and Equation 4.1 need to be solved simultaneously using an iterative procedure. However, note that $f$ represents half the correction factor, needed to make the approximation of the $t$-distribution to Gaussian accurate (see Section 3.4.3), divided by the total sample size. For example, from Table 3.5 we know that $0.25 \times z^2_{\alpha/2}$ equals to 2.7 when $\alpha = 0.001$ and equals to 4.9 when $\alpha = 0.00001$. Thus, if we are willing to assume that the total sample size is large relative to the correction factor $z^2_{\alpha/2}/2$ (i.e. $f \approx 0$), then we can simplify Equation 4.3 to :

$$m_{opt} \approx \sqrt{\frac{C_1}{C_2} \times \frac{\sigma_E^2}{\sigma_B^2}} \tag{4.4}$$

Equation 4.4 allows for the intuitive interpretation that when the recruitment cost ($C_1$) and/or technical variation ($\sigma_E^2$) is relatively high, it might be more efficient to increase the number of technical replicates. We can also use this approximate solution as the starting value for Equation 4.3. Note that we set $m = 1$ when $m_{opt} < 1$. Similarly, when $m_{opt}$ suggests a number that is beyond the number of aliquots that is physically possible, then we set it to the maximum possible.

## 4.3. Unequal recruitment cost for the two groups

We could envisage a situation where the recruitment costs could differ between two groups. For example, in some animal experiments researchers compare common wild type mice with gene-knockout or mutant mice which are more expensive to obtain.

If we distinguish the recruitment cost for group A and group B as $C_1^A$ and $C_1^B$ respectively, then the solution is the same as Equation 4.3 and Equation 4.4, except that $C_1$ is replaced by its weighted recruitment cost ($\hat{C}_1$)

$$\begin{aligned} \hat{C}_1 &= \gamma C_1^A + (1-\gamma) C_1^B \\ \text{where } \gamma &= \frac{n_A}{n_A + n_B} = \frac{1}{1+k} \end{aligned} \tag{4.5}$$

We can further extend the model to increasingly complex situations such as unequal replication between groups, unequal replication within samples and unequal measurement costs. However, we are content to have covered the most practical designs as proof of principle.

## 4.4. Estimation of the additional parameters

In addition to the parameters discussed in Section 3.5, we need to estimate the cost components and two more parameters. We discuss the cost component estimation when we present the case study in Section 5.

### 4.4.1. Estimating $\Delta_m$ for $m > 1$

When Equation 4.4 or Equation 4.5 suggests $m > 1$, we need to estimate $\Delta_m$. $\Delta_m$ can be calculated either directly from a pilot study or existing data with the appropriate technical replicates. However, obtaining sufficiently high quality studies with technical replicates, especially those based on an Affymetrix platform, can prove challenging. If no previous data allows direct calculation of $\Delta_m$, we can try to estimate it indirectly from $\Delta_1$ from studies with similar biological background. We can use the relationship $\sigma_m^2 = \sigma_B^2 + \sigma_E^2/m$ to obtain the following relationship, necessitating the need to estimate $R = \sigma_E^2/\sigma_B^2$.

$$\Delta_m = \frac{1+R}{1+R/m}\,\Delta_1 \tag{4.6}$$

Note that $\Delta_m$ approaches $(1+R)\,\Delta_1$, or equivalently $\sigma_m^2$ approaches $\sigma_B^2$, as $m$ increases. This implies that there is a lower bound to the number of subjects required to achieve a specified power level even with a very large number of measurements being made on each subject.

### 4.4.2. Estimating the ratio of technical to biological variation, $R = \sigma_E^2/\sigma_B^2$

This quantity is required for the calculation of $m_{opt}$ in Equation 4.3 - 4.4 or if we wish to estimate $\Delta_m$ indirectly from $\Delta_1$. The variance components can be determined by maximum likelihood methods (Pinheiro and Bates, 2000) or by an analysis of variance (Fisher, 1990). Again, we obtain a distribution of values for $R$ and would need to select a suitable value. We suggest looking at the distribution of $R$ conditional on the value of $\Delta_m$ exceeding the chosen threshold.

Intuitively, one would expect $\sigma_E^2 < \sigma_B^2$ (or equivalently $R < 1$) for genes that are differentially expressed and there is some support for this assumption. Dobbin and Simon (2002) report that $R < 1$ for 73% of the genes in a cDNA experiment on breast cancer cell lines, while Yang *et al.* (2003) calculated the sample size using various values of $R$ that are less than 1.

## 4.5. Discussion

Lee *et al.* (2000) consider the accuracy of the results to detect the presence of "spiked-in genes" (genes with known concentration) using 1, 2 or 3 technical replicates. They conclude with that at least 3 technical replicates are required per sample. We disagree with this recommendation when the interest of investigation lies in finding differentially expressed genes. The authors have not accounted for fixed resources or the influence of increasing biological samples instead. Our concerns are shared by Dobbin *et al.* (2003), who show that in the case with limited resources, replication at population level (i.e. increasing number of subjects) reduces both biological and technical variability, while replication at sample level (i.e. increasing technical replicates) only reduces the latter.

Finally, the univariate sample size approach has also been utilized for other objectives. For example Black and Doerge (2002) calculate the minimum number of times a spot needs to be duplicated on array. Wit and McClure (2004) consider the number of individual RNA's to pool and number of pools required. This would allow us to reduce the number of arrays required but only if the individuals in each pool were homogenous in their gene expression profile, an assumption that we do not wish to make here.

## 4.6. Summary

In this chapter, we derived the optimal number of technical replicates and shown it to be directly proportional to the ratio of recruitment to measurement cost and to the ratio of technical to biological variation. We also considered the case of unequal recruitment cost. Finally, we provided a section on estimating the additional parameters required.

# 5. Sample size for microarrays: A case study

*In this chapter, we use a real case study to solidify the concepts presented in Chapters 3 and 4. The case study arose from a collaboration with Dr. Andrea Pellagatti, Dr. Jacqueline Boultwood and Professor James S. Wainscoat of Leukemia Research Foundation, Molecular Haematology Unit, John Radcliffe Hospital, Oxford.*

## 5.1. Introduction to the clinical problem

Myelodysplastic Syndrome (MDS) is a fatal disease of the bone marrow that causes death either through bone marrow failure, or through transformation to acute myleoid leukemia (AML). There are various histopathological MDS subtypes, each with its own prognosis and likelihood of progressing to AML. The MDS subtype (Bennett *et al.*, 1976) depends on the percentage of undifferentiated progenitor cells (blasts), cytogenetic abnormality, chromosomal abnormality and the blood cell count. The median survival times varies from 1 to 5 years depending on MDS subtype. The experimental objective is to characterize the subtypes of MDS by gene expression microarray.

The core of this study involved the following groups:

- normal bone marrow extracted from patients undergoing hip replacement operation.
- RA : refractory anemia ($< 5\%$ blasts)
- RAEB : refractory anemia with excessive blast counts (5 - 20% blasts). RAEB is a more severe form of MDS than RA.

The main comparisons of interest are between normal vs. RA and RA vs. RAEB. Data from a pilot study with 11 normal, 13 RA and 15 RAEB samples using the Affymetrix HGU-133A 2.0 array which contains $n_g = 54,675$ probesets was available. An additional practical constraint in this study is that the number of bone marrow samples were limited.

## 5.2. Choosing the number of technical replicates

We plan to use the guide presented in Section 4.4 to estimate the optimal number of technical replicates ($m_{opt}$). Once this is done, we can proceed as detailed in Section 3.5, indirectly estimating the effect size $\Delta_m$ from $\Delta_1$ using Equation 4.6.

The estimation of $m_{opt}$ requires the ratio of recruitment to measurement costs ($C_1/C_2$) and the ratio of technical to biological variation ($R = \sigma_E^2/\sigma_B^2$). We will first estimate the $C_1/C_2$ ratio. Estimates of the cost components for this project are tabulated in Table 5.1 and show that the $C_1/C_2$ ratio varies between 0.15 - 0.25. The current price of cRNA

production, microarray and reagents vary over a possible range to allow for the possibility of bulk purchase discounts. Costs reflect typical market prices at the time of writing (2007).

| **Recruitment cost per sample,** $C_1$ | **Cost (£)** |
|---|---|
| Sample extraction | 40 - 70 |
| RNA extraction | 30 |
| Quantification using spec | 20 |
| Quality assessment using bio-analyser | 30 |
| Total, $C_1$ | 120 - 150 |

| **Measurement costs,** $C_2$ | **Cost (£)** |
|---|---|
| cRNA production | 250 - 300 |
| Microarray | 200 - 300 |
| Reagents | 150 - 200 |
| Total, $C_2$ | 600 - 800 |

Table 5.1.: Cost component estimates shows that the $C_1/C_2$ ratio varies between 0.15 - 0.25.

Note that when calculating the final cost, we may need to add the fixed costs such as labor, hardware and software. Further we need to allow for the fact that some recruited and hybridized samples may fail the quality assessment. The approximate total cost for this project using the average prices is £ $(n_1 + n_2) \times (135 + 700m)$ plus the fixed costs.

The second step is to estimate $R = \sigma_E^2/\sigma_B^2$. Unfortunately, we did not identify sufficient numbers of high quality studies using technical replicates. We proceed following Dobbin *et al.* (2003) suggestion to assume either $R = 0.5$ for a high quality experiment or $R = 2.0$ for a low quality experiment.

Since the recruitment costs for two groups to be compared in this case study are approximately the same, we use Equation 4.4 to calculate $m_{opt}$. This value of $m_{opt}$ is largest when $C_1/C_2$ and $R$ are maximum and thus the maximum value of $m_{opt}$ is $\sqrt{0.25 \times 2} = 0.7071$

which is less than one. In fact, the technical variance needs to be at least four times larger than the biological variation (i.e. $R > 4$) before the use of any replicates for this study is to be considered. Therefore we conclude that choosing $m = 1$ is the best choice in this case study.

## 5.3. Choosing the number of biological replicates

From the results of the pilot data, we estimate at least several hundred probes are differentially expressed for each comparison. Therefore, we decided to calculate the power of the planned study to be sufficient for either 1000 or 2000 probes.

The main comparisons of interest are between normal vs. RA and RA vs. RAEB. A pilot study looked at 11 normal, 13 RA and 15 RAEB samples using the Affymetrix HGU-133A 2.0 array which contains $n_g = 54,675$ probesets. An additional practical constraint in this study is that the number of bone marrow samples were limited.

Since the pilot data contains less than 30 samples for the comparison of any two groups, we calculated the effect size using $g$ rather instead of *Delta* (see Section 3.5.1).

| | **Probe Coverage** | |
|---|---|---|
| | top 1000 probes (*low*) | top 2000 probes (*high*) |
| normal vs. RA | 1.032 | 0.883 |
| RA vs. RAEB | 0.930 | 0.804 |

Table 5.2.: The magnitude of effect size estimated using Hedge's adjusted $g$ value by comparison and coverage type.

Note that the values for the comparison of RA vs. RAEB are always lower than for the comparison of normal vs. RA. This indicates that the differences between RA and RAEB are more subtle and thus require larger sample sizes to detect. Therefore, we first calculate the required sample size for the comparison RA vs. RAEB under the following conditions:

- Low coverage (1000 probesets) or high coverage (2000 probesets) which corresponds to an effect size of 0.930 and 0.804 respectively.
- The pilot data indicates that there are at least $n_0 = 100$ interesting genes, and we can calculate the corresponding $\alpha$ from Equation 3.4. We also check the results under the more stringent condition of when $n_0 = 10$.
- Three levels of power : 85%, 90%, 95%.
- Equal allocation ($k = 1$) as samples from both groups are equally available.

This can be calculated using the R codes given in Appendix A.4 as follows

```
>  find.ss( Delta=c(0.804, 0.930), sig.level=c(10,100)/54675,
            power=c(0.85, 0.90, 0.95), k=1 )


$`Delta=0.804`
                            power=0.85 power=0.9 power=0.95
alpha=0.000182898948331047      148.3     163.2      186.6
alpha=0.00182898948331047       111.6     124.6      145.2


$`Delta=0.93`
                            power=0.85 power=0.9 power=0.95
alpha=0.000182898948331047      112.6     123.7      141.2
alpha=0.00182898948331047        84.7      94.4      109.8
```

The results show that a study with $\alpha = 0.00183$, 95% power for 2000 probesets requires 146 samples, which is almost equivalent to a study with $\alpha = 0.000183$, 95% power for 1000 probesets which requires 142 samples. We choose the former setting as it leads to the more conservative estimate of the two. As we have chosen an equal allocation, this means 69 samples from RA and 69 samples from RAEB respectively.

Let us now look at the sample size required for the normal vs. RA comparison. Comparing

73 normals with 73 RA using $\alpha = 0.00183$ for a study powered for 1000 genes (i.e. effect size of 0.883) shows that the second comparison is overpowered. This is useful information as the availability of normal bone marrow samples is clinically more limited than RA.

```
> find.power( Delta=0.883, sig.level=0.00183, n1=73, n2=73 )

[1] 0.9833153
```

We can reduce the number of samples from the normal group since the number of samples from RA is fixed (by the RA vs. RAEB comparison). By incrementally decreasing the sample size for the normal groups, we find out that we only require 52 subjects from the normal group.

```
> pow <- find.power( Delta=0.883, sig.level=0.00183, n1=73:1, n2=73 )

> cbind(n1=73:1, power=pow)

       n1     power

  [1,] 73 0.9833153

  [2,] 72 0.9825220

  [3,] 71 0.9816823

  [4,] 70 0.9807931

  [5,] 69 0.9798510

  [6,] 68 0.9788524

  [7,] 67 0.9777934

  [8,] 66 0.9766697

  [9,] 65 0.9754769

 [10,] 64 0.9742100

 [11,] 63 0.9728638

 [12,] 62 0.9714326

 [13,] 61 0.9699102
```

```
[14,] 60 0.9682901

[15,] 59 0.9665650

[16,] 58 0.9647274

[17,] 57 0.9627687

[18,] 56 0.9606802

[19,] 55 0.9584519

[20,] 54 0.9560734

[21,] 53 0.9535333

[22,] 52 0.9508194

[23,] 51 0.9479183

[24,] 50 0.9448156

[25,] 49 0.9414959

...
```

Therefore we require 52 normal, 73 RA and 73 RAEB. From previous experience, we anticipate about 5% of the samples will be unusable due to poor quality or failed hybridizations. Accounting for this means we need to collect DNA from approximately 55 normal samples, 77 RA samples and 77 RAEB samples.

## 5.4. Summary

In this chapter, we demonstrated sample size calculations for a real case study using materials developed from the previous two chapters. We also observed that one technical replicate is optimal in this case study and needs further testing under a wider setting.

# 6. Meta-analysis of microarray datasets: Key issues

*We previously stated that meta-analysis of microarray datasets is an attractive option to increase statistical power and generalizability while being inexpensive and allowing for standardization of methods. In this chapter, we discuss seven key issues that must be addressed for any good quality meta-analysis.*

*We give practical guidance to assist those conducting or reviewing such a meta-analysis. The approaches presented here can be adapted to other areas of high-throughput biological data analysis.*

*Materials from this chapter has now been published as:*

Ramasamy A, Mondry A, Holmes CC, Altman DG (2008)
Key issues in conducting a meta-analysis of gene expression microarray datasets.
PLoS Medicine 5(9): e184. http://dx.doi.org/10.1371/journal.pmed.0050184

## 6.1. Motivation

The advantages of meta-analysis of gene expression microarray datasets have not gone unnoticed by researchers in various fields (Rhodes *et al.*, 2002; Lee *et al.*, 2004; Pilarsky *et al.*, 2004; Rhodes *et al.*, 2004a; Wang *et al.*, 2004; Grützmann *et al.*, 2005; Mehra *et al.*, 2005; Bianchi *et al.*, 2007; Kim *et al.*, 2007; Silva *et al.*, 2007). Several meta-analysis techniques have been proposed in the context of microarrays (Rhodes *et al.*, 2002; Choi *et al.*, 2003; Smid *et al.*, 2003; Stuart *et al.*, 2003; Choi *et al.*, 2004; Rhodes *et al.*, 2004a; Parmigiani *et al.*, 2005; Warnat *et al.*, 2005; Yang *et al.*, 2005; Aggarwal *et al.*, 2006; DeConde *et al.*, 2006; Hong *et al.*, 2006; Wang *et al.*, 2006; Zintzaras and Ioannidis, 2008). However, no discussion of key issues or comprehensive framework exists on how to carry out a meta-analysis of microarray datasets.

## 6.2. Introduction

There is a considerable literature to guide the whole review process, including statistical methods, for clinical trials and epidemiological studies (Sutton *et al.*, 2000; Deeks *et al.*, 2001; Whitehead, 2002). As yet however, there is little guidance for conducting a meta-analysis of microarray datasets.

Therefore, in this chapter, we disentangle this complex topic and identify seven distinct key issues specific to meta-analysis of microarray datasets, each comprising several steps. We discuss in detail the first five key issues in this chapter which are related to data acquisition and curation. Chapter 7 looks at the sixth issue of choosing a meta-analysis technique for two-class comparison. The seventh issue of analyzing, presenting and interpreting is discussed using a case study of meta-analysis of 30 datasets in Chapter 8.

We provide a practical checklist, shown in Table 6.1, that should enable the reader to make informed decisions on how to conduct a meta-analysis, and to understand better the underlying concepts that make this approach so attractive for analysis of microarray datasets.

| STEP | *Identify suitable microarray studies (Issue 1)* |
|---|---|
| 1 | Formulate objectives and a review protocol. |
| 2 | Define inclusion-exclusion criteria and suitable keywords. |
| 3 | Perform literature search using the keywords on the websites listed in Table 6.2(a) |
| 4 | Search public microarray repositories listed in Table 6.2(b) and 6.2(c). |
| 5 | Contact collaborators and experts in the field to help find published and unpublished data. |
| 6 | Search the reference section of retrieved studies for other relevant studies. |
| 7 | Check the selected study against inclusion-exclusion criteria. |

| STEP | *Extract the data from studies (Issue 2)* |
|---|---|
| 8 | Scan the literature to identify feature-level extraction output (FLEO) (e.g. CEL, GPR files). |
| 9 | If the main text does not contain a link to FLEO data, search the repositories and group/lab's webpages. If unsuccessful, write to the authors. |
| 10 | If multiple publications use overlapping data, identify the most comprehensive one. Combine any training and validation dataset together. |

| STEP | *Prepare the individual datasets (Issue 3)* |
|---|---|
| 11 | Identify and remove any arrays with poor quality. |
| 12 | Preprocess the FLEO data into GEDM. |
| 13 | Check for batch effects among arrays, especially in large studies. |
| 14 | (optional) Filter out any probes with poor spot quality in the arrays. |
| 15 | Aggregate any technical replicates. |
| 16 | Check the processed expression values from multiple platforms are compatible. |

| STEP | *Annotate the individual datasets (Issue 4)* |
|---|---|
| 17 | Identify the probe sequence or the most sequence-specific probe annotation information. |
| 18 | Either a) cluster the probe sequences or b) map the most sequence-specific probe annotation to a gene-level identifier. Use the same mapping build for all datasets. |

| STEP | *Resolve the many-to-many relationship between probes and genes (Issue 5)* |
|---|---|
| 19 | Discard any probe that does not map to any GeneID$^\sharp$. |
| 20 | For every GeneID within a study, calculate the study-specific estimate(s). |
| 21 | If a probe maps to multiple GeneIDs within a study, "expand" it by replacing it with a new record for each GeneID with the same study-specific estimate(s) or expression profile. |
| 22 | For GeneIDs with multiple records within a study, "summarize" them by either selecting one of the records or by aggregating them. |

| STEP | *Combine the study-specific estimates (Issue 6)* |
|---|---|
| 23 | For every GeneID, identify the studies which provide usable information. Optionally, discard any GeneID that is not found in at least a pre-specified number of studies. |
| 24 | For every GeneID, combine the study-specific estimates across the studies using a meta-analytic technique. Record the resulting summary statistic(s). |
| 25 | Calculate the nominal $p$-value for every GeneID and adjust for multiple testing. |

| STEP | *Analyze, present and interpret results (Issue 7)* |
|---|---|
| 26 | Examine the sensitivity of results to individual studies with a leave one out analysis and by varying the selections made (e.g. type of data available). |
| 27 | Analyze findings using computational tools (e.g. gene set enrichment analysis). |
| 28 | Present the summary statistics graphically (e.g. forest plot) for genes of interest. |
| 29 | If possible, validate using an alternative technology and/or different samples. |
| 30 | Consider strength of evidence, limitations and generalizability of current findings. |

Table 6.1.: A checklist for conducting meta-analysis of microarray datasets. $^\sharp$ refers to either the sequence cluster or gene-level identifier used in STEP 18. See text for further details.

## 6.3. Key issues in meta-analysis of microarray studies

### 6.3.1. Issue 1: Finding relevant studies

The first step in any research project is to clearly define the objectives [STEP 1 in Table 6.1]. Meta-analysis approach to microarray datasets has been used for the following purposes:

- identify genes which are differentially expressed between two biologically different groups (Rhodes *et al.*, 2002; Choi *et al.*, 2003; Smid *et al.*, 2003; Choi *et al.*, 2004; Rhodes *et al.*, 2004a; Parmigiani *et al.*, 2005; Yang *et al.*, 2005; DeConde *et al.*, 2006; Hong *et al.*, 2006; Zintzaras and Ioannidis, 2008)
- increase the robustness of cross-platform classification (Warnat *et al.*, 2005)
- identify overlaps between samples from heterologous datasets (Smid *et al.*, 2003)
- to investigate co-expression of genes or to reconstruct gene networks (Stuart *et al.*, 2003; Aggarwal *et al.*, 2006; Wang *et al.*, 2006).

Designing a review protocol can further help clarify the research objectives and methods and to minimize bias from unplanned data-driven analysis. We suggest developing the review protocol by outlining the solutions to the steps in the checklist shown in Table 6.1. For example, [STEP 7] (Check the selected study against inclusion-exclusion criteria) might be expanded in a review protocol as follows: "Two reviewers will check the eligibility of the identified studies, with disagreements resolved by a third reviewer. A log of excluded studies, with reasons for exclusions, will be maintained." The protocol can be turned into a useful project management tool by incorporating time line and division of labor.

The inclusion-exclusion criteria [STEP 2] are eligibility criteria for studies that will help achieve the stated objectives. These could be biological (e.g. specific disease, type of outcome, type of tissues) or technical (e.g. density of array, minimum number of arrays). The retrieved articles must be evaluated as to whether they meet the inclusion criteria.

Once the inclusion-exclusion criteria have been defined, one needs to perform a comprehensive literature search [STEP 3] to identify suitable studies, usually based on appropriate keywords for automated queries. We recommend searching all the major online repositories of abstracts listed in Table 6.2(a) to maximize data acquisition. Review articles and directly

contacting researchers in relevant fields [STEP 5] may help to identify both work potentially missed by the automated search, and ongoing research efforts with possibly unpublished data. One should also look at the reference lists of relevant articles [STEP 6] to identify further suitable studies.

In the case of microarrays, one should also search public microarray data repositories [STEP 4] recommended by the Minimum Information About a Microarray Experiment (MIAME) requirements (Brazma *et al.*, 2001; Ball *et al.*, 2004) which is listed in Table 6.2(b); as well as a few more specialized repositories listed in Table 6.2(c).

(a) Online repositories of abstracts

| **Database** | **Website URL** | **Reference** |
|---|---|---|
| PubMed | http://www.pubmed.gov/ | |
| Google Scholar | http://scholar.google.com/ | |
| Web of Science♯ | http://wos.mimas.ac.uk/ | |
| SCOPUS♯ | http://www.scopus.com/ | |

(b) Microarray repositories recommended by MIAME for mandatory data deposition

| | | |
|---|---|---|
| Array Express | http://www.ebi.ac.uk/arrayexpress/ | Brazma *et al.* (2003) |
| CIBEX | http://cibex.nig.ac.jp/ | Ikeo *et al.* (2003) |
| Gene Expression Omnibus | http://www.ncbi.nlm.nih.gov/geo/ | Edgar *et al.* (2002) |

(c) Other useful sites for data identification

| | | |
|---|---|---|
| ONCOMINE | http://www.oncomine.org/ | Rhodes *et al.* (2004b) |
| Stanford Microarray Database | http://smd.stanford.edu/ | Demeter *et al.* (2007) |

Table 6.2.: Useful internet resources to identify studies for meta-analysis of microarray studies. ♯ sites which requires paid subscription.

Having identified potentially eligible studies from abstracts, one needs to retrieve the articles, where available, and confirm eligibility [STEP 7]. This process may best be done by at least two people.

### 6.3.2. Issue 2: Data Extraction

Before we consider how to extract the data, we need to first decide what type of data to extract. This partially depends on the choice of meta-analysis technique (Issue 6), but the underlying principles will be discussed here. First, recall the four types of data formats introduced in Section 1.5.3: image files, feature level extraction output (FLEO), gene expression data matrix (GEDM) and published gene list (PGL).

PGLs are often presented in the main or supplementary text of microarray based studies and thus easy to obtain. Unfortunately such PGLs are of limited use for meta-analysis since they represent only a subset of the genes actually studied, and information from many genes will be completely absent. Furthermore, PGLs depend heavily on the preprocessing algorithm, the analysis method, the significance threshold and annotation builds used in the original study, all of which usually differ between studies (Suárez-Fariñas *et al.*, 2005). Thus Individual Patient-level data (IPD), which for microarrays represents the measurement for every probe in every hybridization, are far more useful. Ioannidis *et al.* (2002) discusses further the advantages of a meta-analysis using IPD vs PGLs. We consider now which of the following three IPD formats to choose as the data input for meta-analysis of microarray datasets.

Since GEDM represents the gene expression summary for every probe and sample, it is ideally suited as input for meta-analysis. *Published GEDMs*, however, are unsuitable for meta-analysis because they depend on the choice of the preprocessing algorithms used, which may produce gene expression on different scales of measurement (e.g. see Figure 3.3) and thus not directly comparable with other studies.

In order to eliminate bias due to specific algorithms used in the original studies and to allow consistent handling of all datasets, we recommend obtaining the FLEO files `[STEP 8]`, such as CEL and GPR files, and converting them to GEDM in a consistent manner (see Issue 3). FLEO files are likely to be available, especially for newer studies, as the widely supported MIAME requirements (Ball *et al.*, 2004) ask authors to make the FLEO data available in public microarray repositories.

Image files at present are neither routinely deposited in public microarray repositories nor technologically uniform enough to be used as input for meta-analysis.

If the main text and supplementary information do not state the location to the FLEO data, then one should try searching public microarray repositories or the research group's webpage before contacting the authors [STEP 9]. If multiple publications use overlapping sets of data, one should identify and use the most comprehensive dataset available [STEP 10], and combine any datasets that were split for algorithm training and validation purposes.

### 6.3.3. Issue 3: Preparing datasets from different platforms

FLEO data has to be converted into GEDM, which can then be used as input for the meta-analysis. The same preprocessing algorithm should be used for multiple studies conducted on the same platform. To combine studies from different platforms, which may have different designs and thus have different options of preprocessing algorithms, it is desirable to try to identify comparable preprocessing algorithms. There are many microarray platforms but we again focus mainly on the Affymetrix and two-channel technology platforms (see Section 1.5.4 for details).

Before the preprocessing step, one may wish to first identify and remove any arrays that are of poor quality [STEP 11]. Visual inspection of the image files can identify physical artifacts and the array report files may indicate sample RNA degradation problems. One can also assess these information directly from FLEO files instead (Larsson and Sandberg, 2006). For two-channel technology platforms, one could investigate the correlation of the two channels using the M-A plots. By comparing the location and spread of gene expressions for all arrays or using dimensional reduction techniques, one could identify outliers. There are many comprehensive, free and open-source packages in BioConductor (Gentleman and Carey, 2005) for quality assessment including arrayMagic (Buness *et al.*, 2005) for the two-channel technology platform and simpleaffy (Wilson and Miller, 2005), and affyPLM (Bolstad, 2006) for the Affymetrix platform.

Next, all good quality arrays should be preprocessed consistently to remove any systematic differences [STEP 12]. This is an important stage, since preprocessing directly affects the gene expression measurements, and thus all subsequent steps. In practice, researchers are likely to combine datasets from multiple platforms and there are very few preprocessing algorithms which can be applied universally, such as the Variance Stabilizing Normalization

(Huber *et al.*, 2002), which accounts for the dependence between variance and mean of the output expression measure. By contrast, it is more common to use different preprocessing algorithms for each platform, as carried out in several large high quality studies specifically designed to assess cross-platform comparability (Larkin *et al.*, 2005; Irizarry *et al.*, 2005; Bammler *et al.*, 2005; M. A. Q. C. Consortium *et al.*, 2006). Unfortunately, there is currently no consensus on which preprocessing algorithm(s) produce comparable expression measurements across different platforms.

One may also want to check and correct for any batch effects `[STEP 13]`, especially in large studies. Batch effects could arise whenever the number of arrays are logistically too large to be handled in a single batch. Some of the contributing factors include multiple technicians in a project, the order in which the chips were manufactured or hybridized, sample handling and ambient conditions. This will be hard to investigate if the original authors did not provide information on factors that may lead to batch effects. Unsupervised visualization Benito *et al.* (2004) can help to identify any grouping caused by experimental factors.

One needs to decide whether to use all available probes on the array, or a filtered set of probes `[STEP 14]`. In a single study analysis, it is common to filter out probes that have visible defects (e.g. using quality flags), probeset calls (e.g. using the absent/present calls from MAS 5.0 preprocessing algorithm) or probes that show little variation (e.g. using the minimum coefficient of variation) in single study analysis. However, we feel that such filtering is not beneficial in a meta-analyses context, as we do anticipate the same set of probes to consistently fail across the different studies.

Fifth, one needs to deal with multiple technical replicates (i.e. multiple measurements from the same biological subject which include replicated hybridization and dye-swaps design) if relevant `[STEP 15]`. These should not be treated as independent observations. One approach is to select one of the replicates at random. Alternatively, one can average the replicates. If we assume that all technical replicates have similar array quality, then a simple average or median can be used. A more sophisticated approach would be to treat the technical replicates as nested values within patients.

Finally, one could check that the processed expression values from multiple platforms are comparable `[STEP 16]`. Microarray platform manufacturers typically include housekeeping genes or negative controls, which are genes expected to be transcribed at a constant level,

that may be used for this purpose. Additionally, one may use a visualization technique such as multidimensional scaling (Kruskal and Wish, 1978; Venables and Ripley, 2002) to inspect for any clustering of arrays by studies.

### 6.3.4. Issue 4: Annotating the individual datasets

In Section 1.5.2, we made the distinction between probes and genes and the associated nomenclatures. It is very difficult to combine information for each probe, especially when different chip designs are used, as the highly specific nature of probe identifiers leads to a poor overlap in probes. Therefore one needs to combine information at a gene level. This necessitates the need to identify which probes represent the same gene within and across the studies included in the meta-analysis.

One option is to cluster the probes based on the sequence data `[STEP 17a]` using the BLAST algorithm (Altschul *et al.*, 1990), for example, by using the Ensembl browser (Birney *et al.*, 2004) `[STEP 18a]`. It has been shown that sequence-matched datasets can increase cross-platform concordance (Morris *et al.*, 2005). Such methods can also accommodate Affymetrix probeset redefinitions (Carter *et al.*, 2005), which better addresses the problem of alternative splicing. However, the probe sequence may not be available for all platforms and the clustering of probe sequences could be computer intensive for very large numbers of probes.

Alternatively, one can map the probe-level identifiers such as IMAGE Clone ID, Affymetrix ID or GenBank Accession Number to a gene-level identifier such as UniGene, RefSeq or EntrezGene ID. For example, UniGene (Wheeler *et al.*, 2003), which is an experimental system for automatically partitioning sequences into non-redundant gene-oriented clusters, is a popular choice to unify the different datasets. For example, UniGene Build #211 (release date $12^{th}$ March 2008) reduces nearly 7 million human sequences to 124,181 clusters.

To translate probe-level identifiers to gene-level identifiers, one can use either the annotation packages in BioConductor (Gentleman and Carey, 2005) or web tools such as SOURCE (Diehn *et al.*, 2003) and RESOURCERER (Tsai *et al.*, 2001) `[STEP 18b]`. We suggest using IMAGE Clone ID or Affymetrix ID first, if available, as they are more sequence-specific `[STEP 17b]` . The same mapping build, ideally the most recent, should be used for all datasets to avoid inconsistencies between releases (Noth *et al.*, 2005; Perez-Iratxeta and Andrade, 2005).

### 6.3.5. Issue 5: Resolving the many-to-many relationship between probe and gene-level identifiers

In this section, we will refer to either the sequence cluster ID or the gene-level identifier used to annotate the datasets as described in the previous issue simply as the ***GeneID***.

Many probes can map to the same GeneID because of the clustering nature of the UniGene, RefSeq and BLAST systems involved or because the microarray chip used contains duplicately spotted probes. On the other hand, a probe may map to more than one GeneID if the probe sequence is not specific enough. Sometimes, a probe has insufficient information to be mapped to any GeneID and we recommend to omit these from further analysis `[STEP 19]`. Inconsistencies between various annotation databases or releases (Noth *et al.*, 2005; Perez-Iratxeta and Andrade, 2005) and software (Zeeberg *et al.*, 2004) complicate the matter further. The case study presented in Chapter 8 contains 537,686 probes. 47,154 (or 8.7%) of these probes could not be mapped to any UniGene ID while 29,774 (or 6.1%) of the remaining probes map to more than one UniGene ID.

This "many-to-many" relationship can fragment the available information for meta-analysis. For example, a probe could map to GeneID X in half of the datasets but to both GeneIDs X and Y in the remaining datasets. Software which performs automated meta-analysis on several thousand genes will treat such probes as two separate gene entities and will fail to fully combine the information for GeneID X from all studies.

A simple approach is to use only the probes with one-to-one mapping for further analysis, but this means losing information, and so is not recommended. In the example above, potentially half of the information for GeneID X (i.e. probes mapping to both X and Y) will be ignored. Further, it would be particularly biased against probes that consistently map to multiple keys across studies.

Therefore, when relevant, we recommend replacing probes with multiple GeneIDs by a new record for each GeneID `[STEP 21]`. This greedy approach of "expanding" the probes with multiple GeneIDs ensures the software utilizes all possible information.

On the other hand, how should one deal with multiple probes that map to the same GeneID *within a given study*? Grützmann *et al.* (2005) treated these as independent observations in

the meta-analysis but we recommend summarizing them [STEP 22] into a single representative value per GeneID within a study.

Several options are available to summarise information in this situation. First, one could select a probe at random but again this means losing information. Simply averaging the expression profiles before proceeding is not desirable either as different probe sequences have different binding affinity, giving rise to the problem of different measurement scales. Thus, it is preferable to work with standardized measures such as the $p$-value or effect size. When working with standardized measures, one could select the most extreme value as it is least likely to occur by chance. For example Rhodes *et al.* (2002) used the smallest $p$-value of the probes that corresponded to each GeneID. A more sophisticated approach, when working with effect size, is to take a weighted average by meta-analysis.

The MicroArray Quality Control (MAQC) project (M. A. Q. C. Consortium *et al.*, 2006) describes another alternative to resolve the many-to-many mapping. For a probe that maps to multiple RefSeq IDs, the authors selected the RefSeq ID that was annotated by Taqman assays and secondarily one that was present in the majority of platforms. Next, if many probes mapped to a given RefSeq ID, they chose the one closest to the 3' end of the gene.

After resolving for the many-to-many relationship by expanding and summarizing probes, we are left with a summary statistic per GeneID per study. In the next step, we proceed with meta-analyzing the summary statistic for each GeneID in turn across the studies.

### 6.3.6. Issue 6: Combine the study-specific estimates

Having identified and curated the individual datasets, the next step is to combine them in a meta-analysis [STEPS 23 – 25]. However, this necessitates the researcher to choose a meta-analysis technique, an issue complicated by the availability of numerous methods with little empirical comparison to assessing them. In Chapter 7, we briefly introduce the available techniques and discuss the various strengths and weaknesses of each one.

### 6.3.7. Issue 7: Analyze, present and interpret results

After meta-analyzing the studies, we typically end up with several summary estimates per gene. Unlike a meta-analyses of clinical trials or epidemiological studies, we have the summary estimates from tens of thousands of genes and *apriori* we typically expect only a small proportion of these genes to be interesting. Thus, we face the challenge of choosing a handful of genes from a large pool for further investigation. The exact approach of doing this depends largely on the meta-analyses technique used since it dictates the type of summary estimates outputted.

In most cases, we can estimate a statistical significance (e.g. summary estimate divided by the standard error in the case of inverse-variance technique) or the rank of each gene. We can then utilize a "volcano plot" (Cui and Churchill, 2003) to summarize the results for all the genes in a single graph. A volcano plot, commonly used in genomics and other high-throughput biological data analyses, is a scatter plot of the statistical significance (e.g. log p-values or ranks) versus the biological significance (e.g. $\delta$ or $\Delta$) to enable easy identification of genes that can be investigated further. A volcano plot can be successfully used for most of the meta-analyses techniques.

Having selected a handful of interesting genes, we can then proceed to investigate them in more detail as in a conventional meta-analysis. These may include checking that the overall pooled estimate is not influenced largely by one study alone or investigating the between study heterogeneity.

In Chapter 8, we demonstrate how one could analyze, present and interpret the results [STEPS 26 – 30] from a meta-analysis using a real case study. We extend the graphics used in traditional meta-analysis for microarray datasets where one typically meta-analyzes thousands of genes making visualization difficult.

## 6.4. Discussion

Meta-analysis of microarray datasets shares many features with meta-analysis in other health care research. Perhaps the main differences are the large numbers of variables involved and technical complexities of integrating data across multiple platforms. Furthermore, most

microarray studies are not prospectively planned and often do not have detailed protocols, but rather tend to make use of existing samples. Table 6.3 gives an overview of the advantages and disadvantages of various aspects of meta-analysis of microarray datasets. We discuss some of these points below.

We argue that working with FLEO files allows for better standardization of information and the incorporation of data from unpublished studies but it also requires significant effort to acquire and manage the datasets. This is further hampered by data sharing issues as discussed in Ventura (2005); Larsson and Sandberg (2006); Piwowar *et al.* (2007); Ioannidis *et al.* (2007). We too explore the issue of raw data sharing in Chapter 9.

A major concern associated with meta-analysis in many clinical and epidemiological studies is the problem of publication bias which is a consequence of selectively publishing statistically significant and favorable results (Dickersin *et al.*, 1992; Egger and Smith, 1998). On the surface, we do not expect to find a publication bias at a gene-level in a given study because of the discovery-driven and high-density nature of microarrays. However, anecdotal evidence based on sales figures (John P. Ioannidis, personal communication) suggests that data from only 10% of all the Affymetrix chips sold are published. The possibility of massive publication bias in microarray research needs further investigation.

Another major concern, especially in epidemiological studies, is adjusting for covariates (a variable that is possibly predictive of outcome) such as sex and age. Adjusting for covariates in a single study analyses not only helps us address any imbalance in baseline covariates between the two groups but also reduces the variability in the effect estimate for the genes. However, researchers in microarray tend to rely heavily on unadjusted methods such as $t$-test instead. In the context of meta-analyses, collecting and adjusting for covariates becomes even more desirable to further account for the differences in the baseline characteristics between studies and sampling biases that may exist. Unfortunately, clinical covariates of interest are rarely made available in practice even when the FLEO or GEDM is available. Requesting such information is hampered by the fact that it does not constitute a "minimum" requirement by MIAME (Brazma *et al.*, 2001) and concerns over patient confidentiality.

Furthermore, within a single-study microarray analysis the particular choice of down-stream statistical analysis may lead to different results depending on the objective of the study (Mondry *et al.*, 2008; Loh and Mondry, 2008). It is unclear to what extent this problem

| Advantages | Disadvantages |
|---|---|
| *Combining independent but related studies* | |
| • Increases statistical power<br>• Increases generalizability of results<br>• Financially inexpensive as it uses existing studies | • Assumes there are sufficient numbers of suitable studies available<br>• Study selection issues such as study diversity and data quality needs to be addressed<br>• Time and effort to acquire and manage data can be significant<br>• Potential publication bias might limit meta-analysis |
| *Individual Patient-level Data (IPD) vs. Published Gene Lists (PGLs)* | |
| • IPD permits re-analysis of individual studies for reproducibility and further novel analysis<br>• IPD avoids selective reporting of genes | • Harder to acquire IPD data compared to gene lists |
| *Feature-Level Extraction Output (FLEO) data vs. Gene Expression Data Matrix (GEDM)* | |
| • FLEO data allows us to standardize preprocessing algorithms and analysis methods | • More time and effort required to acquire and manage FLEO relative to GEDM<br>• In practice, some researchers may withhold access to FLEO and thereby introduce a possible bias |
| *Inverse-variance technique vs. other techniques for combining data* | |
| • Interpretable results with standard error to construct confidence intervals<br>• Treats rarely-studied and frequently-studied genes equally<br>• Weights the contribution from different studies by its precision<br>• Good ability to rank results when applied on small number of studies | • A large study in the collection may influence overall results of meta-analysis<br>• Ignores correlation between genes (and so do most techniques).<br>• IPD may not be available for all studies. |
| *Use of forest plots* | |
| • Visualize contributions from each individual studies<br>• Assess heterogeneity of result across datasets | • Only possible to plot the forest plots for a small number of genes<br>• Descriptive in nature |

Table 6.3.: Advantages and disadvantages of various aspects of meta-analysis of microarray datasets

affects meta-analysis of microarray, even with coherently preprocessed datasets.

The key issues presented here have to be addressed if we are to exploit fully the wealth of the information available from microarray datasets. These issues are fundamental, so the ideas presented here can be adapted to other areas of high-throughput biological data analysis. We hope this discussion helps to form the basis for further debate and research into biological data integration.

## 6.5. Summary

In this chapter, we have formulated and explored seven key issues encountered in conducting a meta-analysis of microarray datasets. We considered the available solutions and made some practical recommendations, which is neatly captured in a practical checklist, shown in Table 6.1.

Here we have discussed Issues $1-5$ which are related to the data identification, acquisition and curation in greater detail. First, we showed how to obtain suitable datasets by searching the published literature and public microarray repositories. Second, we proposed that using the FLEO data allows for better standardization of information. Third, we outlined the issues involved in preparing datasets from multiple platforms. Fourth, we discussed how to match the different datasets using gene-level identifiers. Fifth, we explained how to resolve the problems caused by the many-to-many relationship between the probes and genes by "expanding" probes with multiple GeneIDs and then "summarising" any probes that correspond to a GeneID within a study. We discuss Issues 6 and 7 in Chapters 7 and 8 respectively.

# 7. Meta-analysis of microarray datasets: Choosing a meta-analysis technique

*Numerous techniques for meta-analysis of microarray datasets have been proposed but their relative merits have not been compared. We briefly introduce these techniques and present a series of questions to help researchers choose a technique. In this chapter, we assume that the dataset has been suitably prepared for meta-analysis according to the steps discussed in the previous chapter.*

*The codes to implement some of the meta-analysis techniques presented here is being packaged into an R package and will be released as a BioConductor package:*

Ramasamy A (2009)
metaGEM: an R package for meta-analysis of gene expression microarray studies.
A beta version of this open source package (along with the vignette) is available at
http://spiral.imperial.ac.uk/handle/10044/1/4217

## 7.1. Available techniques for meta-analysis of microarrays

The choice of an appropriate meta-analysis technique depends on the type of outcome measure (e.g. difference in group means, odds ratio) that we wish to combine. In this chapter, we focus on a fundamental application of microarrays: the *two-class comparison* where the objective is to identify differentially expressed genes between two well-known conditions. There are four generic ways of combining information in such a situation.

### 7.1.1. Combining votes

Combining votes is also known as vote counting (Bushman *et al.*, 1994) as one simply counts the number of studies a given gene was declared as significant at a particular significance threshold. The significance measure used can be the test statistics (e.g. fold change, $t$-test) or the corresponding $p$-value which could be either adjusted or unadjusted for multiple testing. For small numbers of studies, the results could be visualized using a Venn diagram (Venn, 1880).

Vote counting has been used in an ad hoc manner in microarray studies, often presented as a Venn diagram, as little computation is required. It is perhaps best formally described by Rhodes *et al.* (2004a) who also suggest calculating the null distribution of votes using permutation testing. Alternatively, one could calculate the significance of the overlaps using the normal approximation to binomial as described in Smid *et al.* (2003). Yang *et al.* (2005) extend both of these methods into the concept of meta-analysis pattern matches.

### 7.1.2. Combining $p$-values

Fisher's sum of logs method (Fisher, 1932) simply adds up the logarithm of $p$-values across studies for a given gene. Rhodes *et al.* (2002) first used this method in the context of microarrays to combine four studies comparing prostate cancer and normal prostate.

Note that the $p$-value used must be obtained from the one-sided test. This sum of logs can

be compared with those obtained from a $\chi^2$ distribution with $2N$ degrees of freedom, where $N$ is the number of studies combined. Alternatively, permutation testing could be used to generate an empirical distribution of the null hypothesis.

### 7.1.3. Combining ranks and lists

These methods try to aggregate the rank of the test statistics or $p$-values across studies. Unlike vote counting, which simply dichotomizes the significance values, it is able to take the ordering of the significance of genes into account. Since it uses ranks instead of the actual test statistics, it is perhaps more robust than Fisher's method when the distribution of $p$-values differ greatly from study to study. Three such methods have been proposed in the context of microarrays.

Hong *et al.* (2006) extend the RankProd (Breitling and Herzyk, 2005) concept for meta-analysis. Briefly, this technique first calculates the differences in gene expression values between every sample in one class with every sample in the other class. So if a study contains $n_A$ tumours and $n_B$ normal tissues, then we end up with $n_A \times n_B$ columns of differences. This process is repeated for every study and the column of differences from all studies are merged using GeneID, creating missing values as necessary. Next, the differences are ranked within the columns and the geometric average is calculated for each gene.

DeConde *et al.* (2006) use three different approaches to aggregate the rankings of, say, the top 100 genes (the 100 most significantly up-regulated or down-regulated genes) from different studies. Two of these algorithms use Markov Chains to convert the pairwise preference between the gene lists to a stationary distribution while the third algorithm is based on an order-statistics model.

Zintzaras and Ioannidis (2008) proposed METa-analysis of RAnked DISCovery datasets (METRADISC) method which is based on the average of the standardized rank and has the advantage of incorporating the between-study heterogeneity (sum of squared deviations from the average). The null distributions for the average rank and heterogeneity are then estimated using non-parametric Monte Carlo permutation testing and matched for pattern of occurrence in studies.

### 7.1.4. Combining effect sizes

The inverse-variance technique (Cochran, 1937; Fleiss, 1993) consists of two general steps. In the first step, we calculate the effect size and its variance for every study. In the second step, we combine the study-specific effect sizes across the different studies into a weighted average. As the name suggests, the study weights are inversely proportional to the variance of the study-specific estimates. More details are given in Section 7.4.

This technique was first used in the context of microarrays on a gene by gene basis by Choi *et al.* (2003) and then subsequently applied by other researchers (Choi *et al.*, 2004; Grützmann *et al.*, 2005).

### 7.1.5. The integrative correlation method

The integrative correlation method proposed by Parmigiani *et al.* (2005) is, strictly speaking, not a meta-analysis technique for combining information but it could be used to select only the "reproducible" genes for meta-analysis. Unlike other techniques described here, it has the advantage of incorporating the dependence structure between genes.

The integrative correlation works as follows. First, the correlation profile of gene G is calculated as the correlation between gene G and every other gene in a study. Next, the correlation of correlation profiles of gene G in every pair of studies is computed, and if the average exceeds a certain threshold, the gene is called reproducible.

## 7.2. Qualitative arguments for selecting a technique

In view of the the various statistical options available for meta-analysis, one is left with the task of choosing the most suitable technique. We present a series of questions that could help a meta-analyst make an informed choice.

1. **What is the required minimum data format for each technique?**
   RankProd, inverse-variance technique and the integrative correlation absolutely require the IPD data, which are less readily available than PGL data. Vote counting,

Fisher's method, METRADISC and algorithms proposed in DeConde *et al.* (2006) could potentially work with the PGL data but may not be able to do so in practice. For example, most publications report the $p$-values or the corresponding rank from two-sided hypothesis testing while these techniques require the values from one-sided testing. Using $p$-values from two-sided testing means ignoring the directionality of the significance and may lead one to spuriously select genes that are discordant in the direction of gene regulation between the studies.

2. **Which set of genes are used if we had the IPD data for all studies?**
   Different chips could contain different sets of genes, leading to potentially a poor overlap. For example, there is only 30.5% overlap in UniGene IDs between Affymetrix chips version HU-6800 and HGU-133 plus 2.0 (a newer and more comprehensive chip). Since microarrays is often used as a hypothesis generating tool, one should prefer techniques that captures information from as many genes as possible.

   The integrative correlation technique only considers genes which are present in all studies and is thus limited by the least gene-diverse platform and/or large numbers of platforms with little overlap. Vote counting and rank aggregation techniques (using PGL data) only consider the genes declared significant in the original studies. Thus, not only do these techniques require one to subjectively select a significance threshold, they also completely ignore genes that fall below this selected threshold. It is interesting to note, that the ranking in an individual study depends on which other genes are included in the chip and thus can influence the rank aggregation method. By contrast, Fisher's method, the rank aggregation techniques (using IPD data) and the inverse-variance technique consider information from all available genes.

3. **How does each technique treat frequently-studied and rarely-studied genes?**
   This question is related to the previous questions. Newer microarrays chips have more comprehensive sets of genes compared to older chips. Thus some genes will be studied more frequently across the studies than others. For example, Affymetrix version HGU-133 plus 2.0 contains almost all of the genes (as identified by UniGene IDs) available in Affymetrix version HU-6800, plus a further additional 13,624 genes, which accounts for nearly 70% of the union of the two versions. Ideally, we prefer a method that treats a frequently-studied and a rarely-studied gene equally.

Integrative correlation only considers genes that are present in all studies ignoring those missed out in even one of the studies. Vote counting and rank aggregation techniques (using PGL data) use the genes declared as significant in the original studies, they ignore the frequency of the genes. For example, a gene found significant in 4 studies and *not significant* in 16 studies will be favored over a gene found significant in 3 studies but *absent* in the other 17 studies. METRADISC accounts for this by matching each gene to the null distribution of genes that have the same absent/present patterns. The test statistic for Fisher's method is based on an unstandardized sum, but it can address this problem by comparing it to a $\chi^2$ distribution where the degree of freedom is determined by the number of studies. The inverse-variance technique and RankProd can address this problem in a more direct way as it yields a weighted average of the effect sizes.

4. **What is the ranking ability of each technique if the number of available studies was small**

   In practice, only a small number of studies, say $3-5$, may be available for meta-analysis. A ranked list from meta-analysis can help researchers to prioritize genes for further testing and validation. We expect vote counting to produce very granular results while the other techniques should produce a much finer result that can be used to ranks genes better.

5. **What is the computational requirement for each technique?**

   The question of computational requirements becomes important especially if one wants to estimate the null distribution using permutation techniques. To answer this, we use the meta-analysis of 30 datasets presented in Chapter 8. The computing time (for analysis starting from GEDMs) for vote counting, Fisher's method, Rank Product and the inverse-variance model took approximately 1, 1, 10, 3 minutes respectively on a Linux-based server with 4 GB of RAM memory. We used R version 2.7.0 (R Development Core Team, 2004)without any code optimization in C or C++ programming. Further, if we started from the PGL data format, we would require additional time and effort to extract the list of significant genes from the publication in a standardized format.

6. **What other arbitrary decisions are required for each technique?**

Some techniques require some arbitrary decisions to be made, this adds extra complexity. The vote counting, Fisher's method and rank aggregation techniques require $p$-values (or the corresponding ranks) which necessitate choosing a test statistic. It is not clear which test statistics is the "best" in the context of microarrays. Mondry et al (2008) shows the top-50 genes selected by five different test statistics (t-test, SAM, eBayes, LIMMA, Wilcoxon) can vary significantly. Additionally, vote counting requires one to select a $p$-value threshold.

While the inverse-variance technique does not require one to select a test statistics, the meta-analyst may be required to decide between a fixed effect and random effects model. There are tests of heterogeneity to formally compare between these two models. However, we argue that one should always use the random effects model to combine information across studies if there is *apriori* evidence that the studies are not homogenous.

## 7.3. Quantitative arguments for selecting a technique

In the previous section, we outlined several questions that can help us choose a technique. Answering these questions gives us an idea of the relative strengths and weaknesses of the various techniques. However, it is still possible for a technique with fewer desirable properties to outperform one with more desirable properties. Therefore in this section, we try to get an idea of the empirical performances.

Benchmarking the various techniques is not straightforward as different techniques output results on non-comparable scales. For example, the vote counting returns the number of favorable votes (a non-negative integer with limited support), Fisher's method returns twice the negative sum of the log $p$-values (a non-negative real value with unlimited support), rank aggregation returns the aggregated rank (a non-negative rational number) and inverse variance returns the summary effect size (a real value with unlimited support). Because of this apples-and-oranges problem and the lack of a dataset with known truths, we cannot directly compare the performance of one technique with another.

However, we can compare the ability of each meta-analysis technique to approximate its large

sample size outputs to get an initial indication. In other words, can a technique reproduce the results of a large study by meta-analyzing subsets of the study?

### 7.3.1. Comparison methodology

In brief, our comparison methodology can be summarized as follows for each technique.

1. Select a large enough dataset which has no missing values.
2. For every gene, calculate the *observed summary statistic(s)*.
3. Randomly split the whole dataset into $N$ subsets, by sampling arrays without replacement into each of the $N$ set, while maintaining the ratio of the two classes constant and similar to the combined dataset.
4. Calculate the summary statistic(s) for each of the $N$ sets as in Step 2.
5. Combine the set-specific summary statistic(s) across the $N$ sets into a *pooled summary statistic(s)*.
6. Compare the pooled summary statistic(s) from Step 5 with the observed summary statistic(s) from Step 2.

The summary statistic(s) used in Step 2 and Step 4 depends on the meta-analysis. For the inverse-variance, the summary statistics are the effect size (Equation 2.6) and its variance (Equation 2.7). For the RankProd technique, it is the geometric average of the ranks of differences as described above. For Fisher's method, we need to first calculate the $p$-value from, say, a two-sample $t$-test with equal variance assumption. Then the summary statistic is negative twice of the sum of log $p$-values. For vote counting, the summary statistic is the number of studies a gene was found significant at certain $p$-value threshold.

Note that for vote counting, Fisher's method and RankProduct we considered the testing the null hypothesis that there are no significantly up-regulated and the null hypothesis that there are no significantly down-regulated genes separately. Therefore we calculated two one-sided $p$-values and corresponding ranks and did not adjust for the multiple hypothesis testing problem. For vote counting, we considered a gene significant if the $p$-value was less than 0.05.

For inverse-variance technique, we used the fixed effect model as we expect no heterogeneity between the five sets. However, in a real meta-analysis, where we are certain that the datasets are generated from different labs or different platforms or the subjects are from different very populations, we strongly recommend using the random effects model instead.

For rank aggregation, we used the R implementation for RankProd (Hong *et al.*, 2006). We coded all other techniques in R and is available in Appendix A.5. We plan on releasing these codes to the community as the R package `metaGEM` in the future.

There are several advantages to using a single dataset as we have done here. First it avoids the need to find a realistic simulation or a dataset that can act as a gold standard. Second, the variation between different sets is smaller than using data from different sources of origin. Third, all the sets contain the same probeset, which we treat here as having a unique GeneID each, so we can avoid the many-to-many relationship between probes described in Section 6.3.5.

### 7.3.2. Results

We present the results by applying the methodology on Singh *et al.* (2002) which contains the gene expression for 12625 probesets (treated as independent genes here) for 50 normal prostate tissues and 52 cancerous prostate tissues. We chose $N = 5$ sets and randomly sampled (without replacement) 10 normal tissues and $10 - 11$ cancers into each set. The entire procedure was repeated few times with similar conclusions. Thus we show the results for one such random sampling.

**Vote counting**

We calculated the one sided $p$-value (from equal variance t-test) for up- and down-regulated genes in the combined data and each of the set. The decision rule when analyzing the combined dataset is to simply declare genes as significant if the $p$-value is below the threshold $p_{comb}$. The decision rule when analyzing the $N$ sets of data is to declare genes as significant if it is found significant at the $p$-value threshold $p_x$ in at least $x$ of the $N$ studies. In the latter case, if we let X be the random variable "number of studies with significant results", then $X \sim Binomial(N, p_x)$. Thus, in order to make a valid comparison, one needs fix two of three parameters $x, p_x, p_{comb}$ and estimate the third parameter by solving

$\sum_{x=x}^{N} \binom{N}{x} p_x^x (1-p_x)^{N-x} = p_{comb}$.

In our simulation, selecting genes which are significant at $p_x = 0.05$ in at least 3 studies (i.e. more than half the studies) gives $p_{comb} = 0.001158125$. The cross-tabulation between the genes declared significant in the combined dataset and individual sets is given in Tables 7.1-7.2. The last two columns depicts the decision rules for classifying genes as significant. Using these to calculate the misclassification rate gives 2.6% and 4.1% for up-regulated and down-regulated genes. One could choose to weight the false negatives much more than false positives instead, but doing so adds yet another arbitrary decision.

| | x = Number of sets in which p $<$ 0.05 | | | | | | Decision Rule | |
|---|---|---|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | 4 | 5 | x $<$ 3 | x $>=$ 3/5 |
| $p \geq 0.001158125$ in combined | 8804 | 2406 | 569 | 117 | 1 | 0 | 11779 | 118 |
| $p < 0.001158125$ in combined | 0 | 24 | 185 | 360 | 144 | 15 | 209 | 519 |

Table 7.1.: Agreement of the vote counting method for testing up-regulated genes.

| | x = Number of sets in which p $<$ 0.05 | | | | | | Decision Rule | |
|---|---|---|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | 4 | 5 | x $<$ 3 | x $>=$ 3/5 |
| $p \geq 0.001158125$ in combined | 6321 | 3439 | 1813 | 249 | 5 | 0 | 11573 | 254 |
| $p < 0.001158125$ in combined | 1 | 29 | 232 | 340 | 155 | 41 | 262 | 536 |

Table 7.2.: Agreement of the vote counting method for testing down-regulated genes.

**Fisher's method**

Figure 7.1 shows the agreement of the test statistics and $p$-values for Fisher's sum of logs method. Notice that the meta-analyzed test statistic is greater than the test statistics from the combined dataset which is counter intuitive at first sight, given that the combined dataset should be statistically more powerful. This is because the test statistics on the combined dataset is $\chi^2_2$ distributed while the meta-analyzed test statistics is $\chi^2_{10}$ distributed and thus not comparable. By converting the test statistics into $p$-values, which standardizes for the degrees of freedom, we see an improved agreement (bottom row of Figure 7.1).

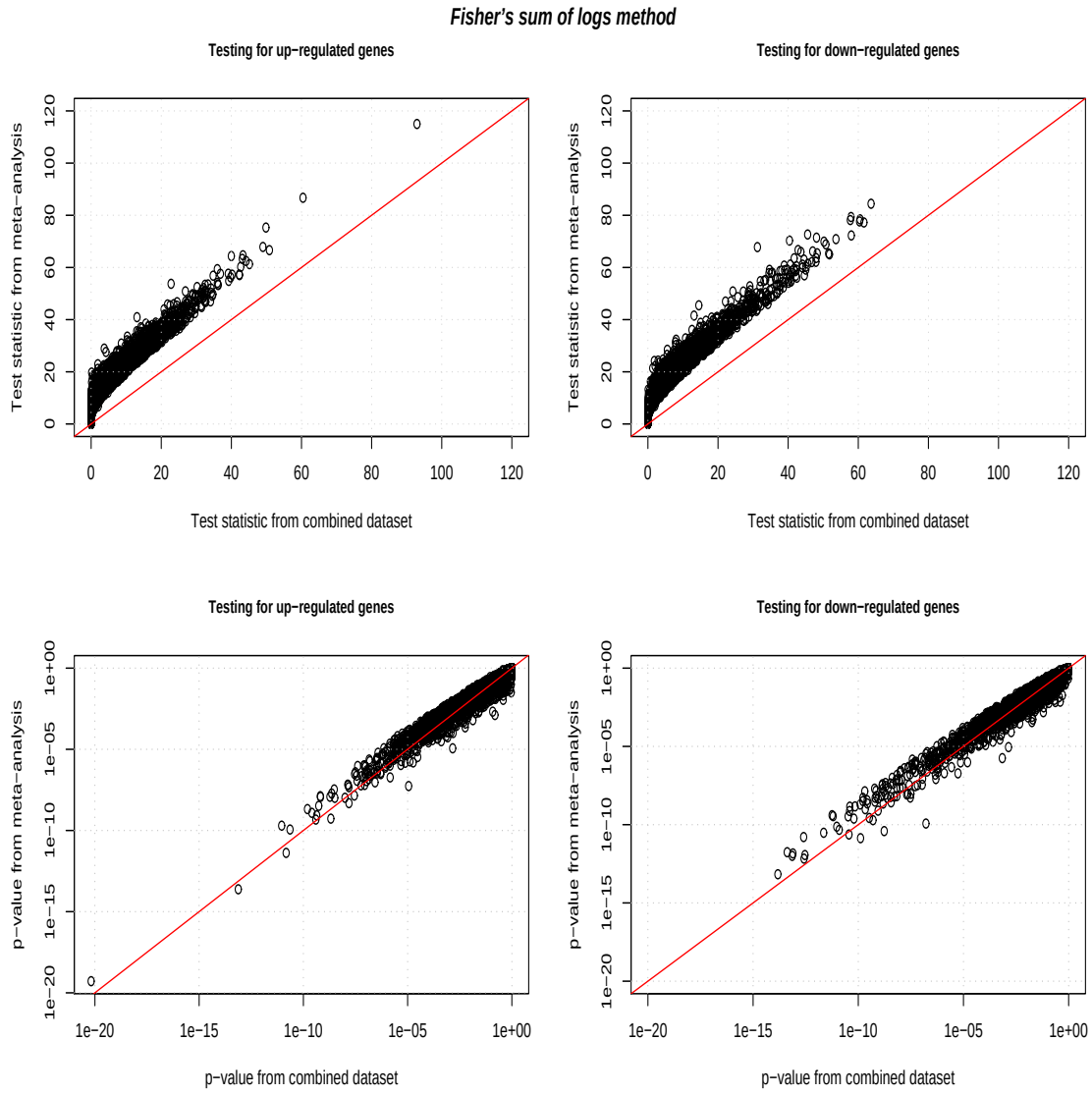



Figure 7.1.: Top row plots the raw Fisher's methods' test statistics (negative twice the sum of log $p$-values) obtained from the combined dataset and the meta-analysis of 5 subsets. The bottom row plots the corresponding $p$-value of the test statistic from comparing the statistics from combined dataset with $\chi^2_2$ and statistics from meta-anlysis with $\chi^2_{10}$. The left and right column shows the results from testing the null hypothesis that there is no significantly up-regulated genes and down-regulated genes. The red line is the line of identity.

### Rank Product

We calculated and plotted the RankProduct statistics in Figure 7.2. We did not not calculate the $p$-values for this method, since it is very computationally intensive and time consuming. However, we expect a good agreement in $p$-values given the good agreement of the statistics here.

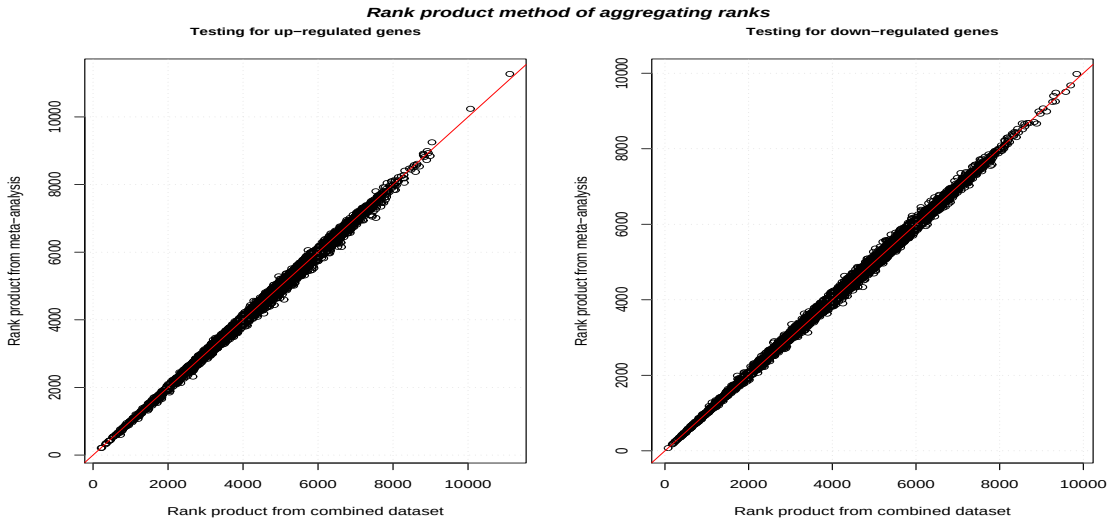



Figure 7.2.: Plot of the test statistics from RankProd (Hong *et al.*, 2006) method for the combined dataset and meta-analysis of 5 subsets for testing of up-regulated genes (left) and down-regulated genes (right).

### Fixed Effect Inverse-Variance

Figure 7.3 shows a good agreement both at the scale of effect sizes and the $p$-value of the effect sizes. Note that we do not need to separate the graphs for up-regulated and down-regulated genes as we are not combining significance measures. If the effect sizes for a gene from two studies are discordant, then they should largely cancel each other out.

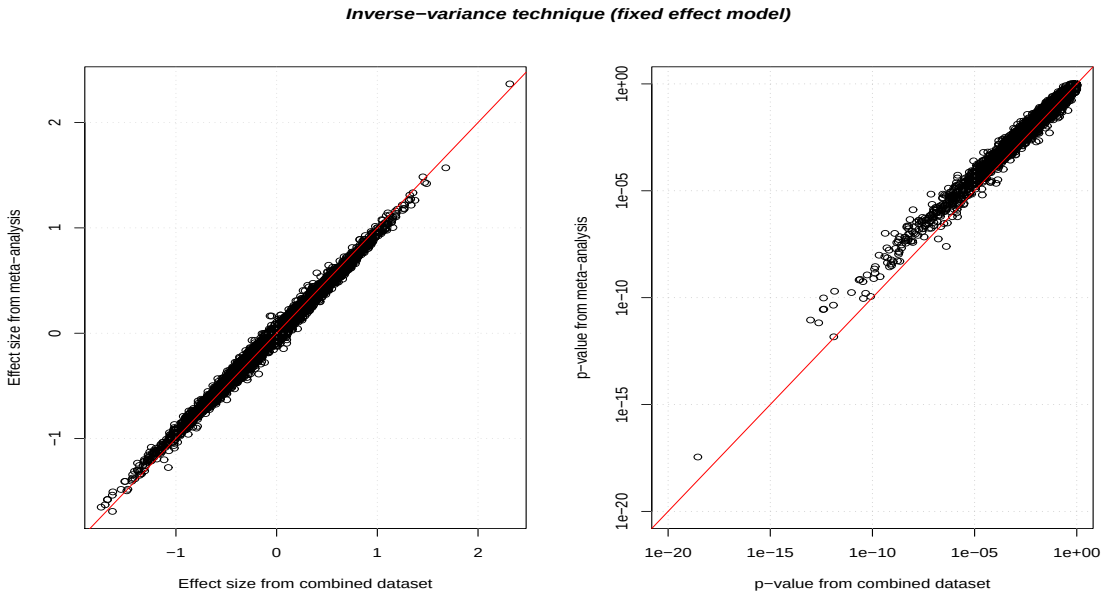



Figure 7.3.: Plot of the effect size (left) and $p$-value (right) between the combined dataset and meta-analysis of 5 subsets using fixed effect inverse-variance technique.

### 7.3.3. Discussion on quantitive arguments

Even though the methodology is applied to a somewhat artificial situation, it shows some of the strengths and weakness of the various techniques. We caution that the results need to be repeated with other datasets and under other conditions (e.g. using sets of different sizes).

From visual inspection of the Figures 7.1 - 7.3, the inverse-variance and Rank Product can approximate both the test statistics and $p$-values of the combined datasets very well. However, a lack of theoretical derivation for Rank Product means that computing the $p$-value has to be done using permutations which is computationally intensive. Fisher's method is unable to approximate the test statistics but does well in approximating the $p$-values. The vote counting method performs very poorly and requires one to choose an arbitrary cutoff.

We believe that combining the effect sizes using an inverse-variance model is the most comprehensive approach for meta-analysis of two-class gene expression microarrays. In addition to the characteristics discussed above, the inverse-variance technique has the following decisive advantages:

1. The output, which is the pooled effect size of differential expression and its standard error, is a biologically interpretable discrimination measure.
2. It is the only method that automatically weights the contribution of each study by its precision, which is related to the study sample size.
3. The contributions of individual studies and the amount of heterogeneity in the overall summary result can be visualized using a forest plot (Lewis and Clarke, 2001)
4. It uses the effect size, a standardized unit, from each study which facilitates the combining of signals from Affymetrix (or other one-channel platforms) and expression ratios from two-channel technology platforms.
5. Unlike other techniques discussed here, we do not need to test the null hypothesis for up-regulated genes and down-regulated genes separately.
6. The only other decision to make is whether to use a random effects model or fixed effect model in situations where the different studies arise from homogenous sources. However, such circumstances are rare and one can formally test the heterogeneity

between studies to guide the decision.

In the next section, we look at the mathematical derivation for the inverse-variance in detail.

## 7.4. The inverse variance technique for meta-analysis

The inverse-variance technique (Deeks *et al.*, 2001; Sutton *et al.*, 2000) allows one to combine the effect sizes from different studies, which estimates the underlying effected size as the *weighted average* of the study-specific effect sizes from $N$ different studies. This is represented mathematically as follows:

$$\hat{g} = \frac{\sum_{i=1}^{N} w_i g_i}{\sum_{i=1}^{N} w_i} \tag{7.1}$$

and its variance as

$$Var(\hat{g}) = \frac{1}{\sum_{i=1}^{N} w_i} \tag{7.2}$$

where $g_i$ is the effect size described by Hedge's adjusted $g$ (see Equation 2.6) of the $i^{th}$ study. $w_i$ represents the weights of $i^{th}$ study and is inversely proportional to the study-specific variances, $v_i$ (see Equation 2.7).

From Equation 2.7, it is clear that the weights are heavily related to sample sizes. It is the difference in the definitions of $w_i$ that distinguishes a random effects model from a fixed effect model.

In the **fixed effect model**, we assume the underlying true effect size is the same for all studies. The weights are reciprocals of the study-specific variance of effect size. In other words, studies with lower variance (i.e. higher precision) are weighted higher.

$$w_i^F = \frac{1}{v_i} \tag{7.3}$$

In the **random effects model**, we assume that the underlying true effect size varies across studies with variance $\tau^2$. Therefore we moderate the study-specific variances with the between-study variance.

$$w_i^R = \frac{1}{\tau^2 + v_i} \tag{7.4}$$

The $Q$ statistic can be used to test the null hypothesis $\tau^2 = 0$ and thus to choose between a fixed effect and random effects model. It is calculated as the weighted sum of squares of the

deviation under the fixed effect model and is distributed as $\chi^2_{N-1}$ under the null hypothesis.

$$Q = \sum_{i=1}^{N} w_i^F (\ g_i - \hat{g}^F\ )^2 \tag{7.5}$$

In the context of microarrays, we obtain a $Q$ statistics (or the corresponding $p$-value) for each gene. Choi *et al.* (2003) propose the use of the random effects model if we reject the null hypothesis for majority of the genes. However, if there are good reasons to assume heterogeneity because of biological and technical details of the studies included, as we do in our case study, then one *must* use the random effects model regardless of the $Q$ statistics value.

The between-study variance ($\tau^2$) is calculated using the weighted least squares approach (Deeks *et al.*, 2001; Sutton *et al.*, 2000).

$$\tau^2 = \max\left(0,\ \frac{Q-(N-1)}{U}\right) \tag{7.6}$$

$$\text{where } U = \frac{S_w^2 - \sum_{i=1}^{N}(w_i^F)^2}{S_w} \tag{7.7}$$

$$\text{where } S_w = \sum_{i=1}^{N} w_i^F \tag{7.8}$$

We expect variation between the microarray studies included in the meta-analysis of gene expression data, due to both the biological differences (e.g. tumor types, different patient populations) and technical differences (e.g. different platforms, laboratories). Therefore, we use the random effects model to combine information across different studies. However, when combining information from within the same study (see Section 6.3.5), one can expect a reasonable level of homogeneity and therefore use the fixed effects model instead.

### 7.4.1. $p$-value calculation for the inverse-variance technique

The $p$-value can be defined as the probability of observing a test statistic that is as extreme or more extreme than the observed value when the null hypothesis is true. A small $p$-value indicates evidence for rejecting the null hypothesis. The null hypothesis here is that the effect size is zero with the alternative hypothesis being that the the effect size is non-zero.

The test statistics used is $z = \hat{g}/SE(\hat{g})$ and there are several ways of deriving the $p$-value for this statistics. First, one can use permutation to empirically calculate the distribution

of $z$ under the null hypothesis. This involves randomly switching group labels within each set and repeating the entire meta-analysis several hundreds of thousands of times, each time calculating and storing the $z$ values for all genes. Next, for each gene, we calculate how many of the (absolute) $z$ values under permutation exceed the observed value.

The permutation method is computationally intensive and time consuming. An more common alternative is to assume the test statistics $z$ is Gaussian distributed. Indeed, it can be shown that this is asymptotically true. The $p$-value is then the cumulative probability of the observed statistics according to the Gaussian distribution.

However, Follmann and Proschan (1999) showed that using the Gaussian assumption for $N \leq 30$, which is often the case in practice, suffers from serious Type I error inflation. They propose using the $t$-distribution with $N - 1$ degrees of freedom instead.

We investigated if the findings from Follmann and Proschan (1999) holds true in the context of meta-analysis of microarray datasets. Therefore, we empirically calculated the $p$-value under 100,000 permutations using the setup and data described in Section 7.3. This $p$-value derived from the Gaussian distribution and $t$-distribution with 4 degrees of freedom was calculated and compared with those based on permutation in Figure 7.4. This figure shows that the Gaussian assumption is sufficiently adequate here. Note that both figures show a saturation at $p = 10^{-5} = 1/100,000$ which is the limit of number of permutations chosen.

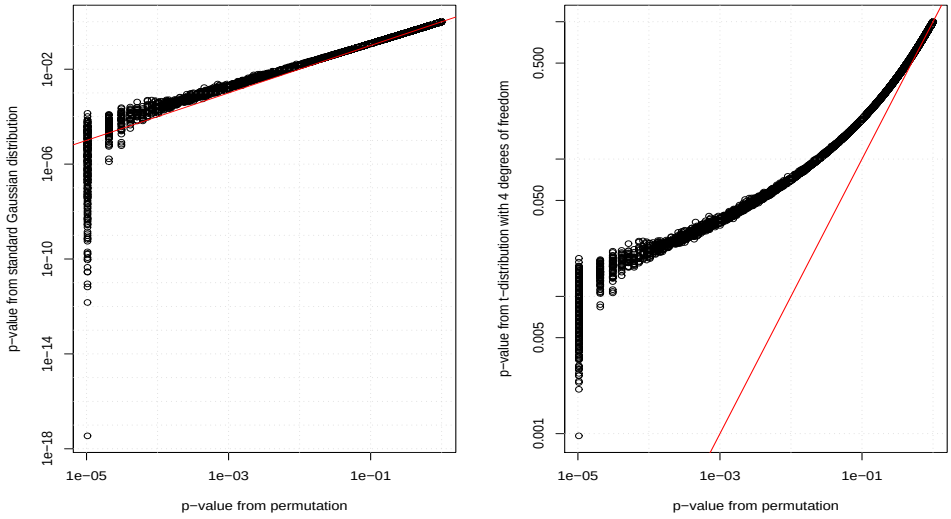



Figure 7.4.: Comparing the $p$-value from Gaussian distribution (left) and $t$-distribution with 4 degrees of freedom with the $p$-value obtained from 100,000 permutations. The red line is the line of identity.

## 7.5. Summary

We started this chapter by briefly discussing the available techniques that can be classified into four categories. Next, we discussed the strengths of each technique quantitatively by asking a series of questions. We tried to empirically quantify the ability of each technique but this did not yield a clear answer as desired. Finally, we chose and described the inverse-variance technique which we believe is the most comprehensive technique for meta-analysis of microarray datasets.

# 8. Meta-analysis of microarray studies: Two case studies.

*In the previous chapter, we discussed the key issues in conducting a meta-analysis. In this chapter we describe two related case studies. The first case study is presented as a proof-of-concept to show that meta-analysis of microarray datasets can reveal useful information. We focus on solid tumors only and analyze all available genes. The second case study, which motivated this section, is focused on apriori protein folding genes and materials from this case study is being prepared as the following pair of articles:*

Ramasamy A, Hu J, Harris AL, Gatter KC, Pezzella F, Altman DG (2009) Meta-analysis of protein folding genes using individuals patient-level microarray data identifies Crystallin alpha B (CRYAB) as a consistently down regulated protein folding gene in tumors (in preparation).

Hu J, Ramasamy A, Tan EY, Inche A, Turley H, Campo L, North P, Couvelard A, Cesario A, Altman DG, Harris AL, La Thangue NB, Gatter KC, Pezzella F (2009) Loss of Expression of Protein-folding Genes in Tumours: a New Oncogenic Mechanism (in preparation).

## 8.1. Motivation

### 8.1.1. Meta-analyses of protein folding genes

Chaperone protein folding genes (PFGs) are an ancient and evolutionarily conserved class of proteins that guide the folding of other proteins. Typical examples of PFG are heat-shock proteins. Increased levels of PFGs are regarded as common in tumours and play an important role in oncogenesis. More specifically, under cellular stress such as those present in tumour cells, many proteins get unfolded. And thus, it is well described (Ma and Hendershot, 2004; Mosser and Morimoto, 2004; Calderwood *et al.*, 2006) that the expression levels of PFGs increases in tumors as a response to stress in order to restore the normal protein-folding environment.

Dr. Francesco Pezzella and Dr. Jianting Hu, our collaborators from Cancer Research UK, observed that some PFGs had instead a lower level of expression in two of their DNA microarray studies: a) Hu *et al.* (2005) which identified 243 genes discriminating non-small cell lung carcinomas from normal lung tissues and b) Couvelard *et al.* (2005) which identified 304 and 215 genes discriminating group 1 and group 2 endocrine pancreatic tumours from normal endocrine pancreatic tissues. Within these clusters of discriminatory genes, they identified several protein folding genes were down regulated. Interestingly, some of the down regulated genes coded for chaperones to well-known tumour suppressor genes.

Thus, they hypothesized that some PFGs could truly be down-regulated in tumours instead. This may represent a new major oncogenic mechanism in which the tumour suppressor activity is lost because an otherwise normal protein cannot fold properly, following reduced availability of chaperone proteins.

To investigate this hypothesis, we proposed carrying out a meta-analysis of PFGs expression levels using available DNA microarray. Since DNA microarrays is a comprehensive technology, we can obtain the expression levels from many PFGs simultaneously from each study. This also allows us meta-analyze the co-expression of PFG expression profiles across different studies in the future.

### 8.1.2. Meta-analyses of solid tumours

This case study is presented as a proof-of-concept that meta-analysis of microarray studies can reveal useful information. Adopting an agnostic approach, we meta-analysis all available genes and check if the resulting gene lists and associated pathways make biological sense.

We decided to focus only on solid tumours (e.g. lung, pancreas, prostate) because our clinical colleagues informed us that the published literature and biology of solid tumours is very different from non-solid tumours (e.g. leukemia) and also since bulk of our data are solid tumours. While this heterogeneity between solid and non-solid tumours is compatible with the objective of the meta-analyses of protein folding genes where generalizability across tumour types is important, it is an unnecessarily complication for this proof-of-concept case study. We do not attempt to interpret the results in detail here as it is beyond the scope of this thesis.

## 8.2. Methods

### 8.2.1. General methods

We concisely describe the methods in Table 8.1 using the same steps and number ordering as in Table 6.1. Table 8.2 lists the characteristics of the included studies and Figure 8.1 shows the data acquisition process.

1 *Objective:* Integrate information from multiple studies using DNA microarray technology for the following two aims.

a) To identify protein folding genes (PFGs) that are consistently down-regulated across solid and non-solid tumours.

b) To identify genes that are either consistently up-regulated or down-regulated (and associated pathways) in solid tumours.

2 *Inclusion-exclusion criteria:* We included any dataset that satisfied **all** of the following:

- has data based on human tissues.
- contains $\geq 7$ primary tumor tissues. Metastatic tumors and cell lines are excluded.
- contains $\geq 7$ normal tissues from corresponding origin. Benign tumors are excluded.
- uses a high density genome-wide array. Specialized or boutique arrays are excluded.
- the feature-level extraction output (FLEO) files must be freely available.

An extra exclusion criteria for the meta-analysis of solid tumours is that we exclude non-solid tumours like those from B-cell and bone marrow tissues.

$3-10$ See Figure 8.1 for data retrieval of 26 studies. Ramaswamy *et al.* (2001) had 5 sets of cancer-normal tissues that satisfied our criteria and thus we have 30 datasets (Table 8.2) for meta-analysis of protein folding genes and 26 datasets for meta-analysis of solid tumours.

11 *Array quality check:* Not performed as this information was not available for all studies.

12 *Preprocessing FLEO files:* Arrays from the Affymetrix platform were RMA preprocessed (Irizarry *et al.*, 2003b) and arrays from two-channel technology were LOESS normalized (Smyth and Speed, 2003; Yang *et al.*, 2002) and then standardized to have the same variability across arrays within studies. We omitted any arrays with more than 25% missing values. Gene expression values for all studies are on a log base 2 scale.

13 & 14 *Batch effect and spot quality check:* Not performed as the required information was not available for all studies.

15 *Aggregate technical replicates:* Bhattacharjee *et al.* (2001) and Winter *et al.* (2007) had technical replicates, which we averaged using a simple mean.

16 *Compatibility check:* Not performed.

17 & 18 *Annotation matching:* For the Affymetrix studies, we used the annotation packages in BioConductor 2.2.0 (built on 26th March 2008) to obtain the UniGene IDs. For the two-channel technology arrays, we mapped the Clone IDs or secondarily the GenBank Accession to UniGene IDs using the web tool SOURCE Diehn *et al.* (2003) on 26th April 2008.

19 *Discard non-identifiable probes:* Only 615,407 or 88% of the combined 700,307 probes could be mapped to UniGene ID. The remaining 12% were discarded from analysis.

20 *Calculate study-specific estimates:* For every probe and for every study, we calculated effect size as Hedge's adjusted $g$ and its variance (see Section 7.4).

21 *Expanding probes with multiple UniGene IDs:* As described in Section 6.3.5.

22 *Summarizing multiple probes per UniGene ID within a study:* We used the fixed effect inverse-variance model as we expected a reasonable level of homogeneity of the information from probes within a study.

23 *Discard poorly represented probes:* We restricted to genes that were present in at least five studies. Additionally for the meta-analysis of protein folding genes, we restricted the analysis to approximately 1.2% of the genes from Step 19 which were identified as protein folding gene using Gene Ontology (Section 8.2.3). See Table 8.3 for further details.

24 *Combine study-specific estimates:* We used the random effects inverse-variance technique for each UniGene ID in turn to obtain the pooled effect size and its standard error. The random effects model differs from the fixed effect model in that it allows for and incorporates the between-study heterogeneity into study weights. We expect significant between-study heterogeneity as the studies combined are both biologically (e.g. different tumors) and technically diverse (e.g. different platforms, laboratories).

| | |
|---|---|
| 25 | The $z$-statistics (ratio of the pooled effect size to its standard error) for every UniGene ID was compared to a standard normal distribution to obtain the nominal $p$-value. We adjusted for multiple testing using the False Discovery Rate (FDR) (Storey, 2002). |

Table 8.1.: Methods for meta-analysis for the two case studies.

The decision to require at least 7 samples in each group in Step 2 is arbitrary, in order to avoid inclusion of very small or highly imbalanced studies. Furthermore, as discussed Section 6.3.2, we decided to collect the raw data from the identified studies in order to minimize the influence of preprocessing algorithm, the analysis method, the significance threshold and annotation builds used in the original studies, all of which usually differ between studies (Suárez-Fariñas *et al.*, 2005).

We preprocessed all available CEL files even if we analyzed only a subset of the whole data. For example, we preprocessed all 360 CEL files in Yeoh *et al.* (2002), even though we only required 98 of them for our purpose. This ensures that the model-based algorithms (RMA, dChip and GC-RMA) have as much information as possible and thus gives better overall estimates. All resulting datasets (after transformation) were visually checked for normality within each array.

Note that Huang *et al.* (2003); Kuriakose *et al.* (2004) had matched samples (tumor and normal tissues from the same patients), while Hippo *et al.* (2002); Borczuk *et al.* (2003); Lenburg *et al.* (2003); Lapointe *et al.* (2004); Winter *et al.* (2007) had a mixture of matched and unmatched samples. Tumor-normal pairs make up only 16% of the data in our case study. The statistical power in paired studies is expected to be maximized if one uses matched/paired test statistics. However, to simplify the analytic approach, we treated these matched and partially matched studies as unmatched. This should yield a slightly conservative results due to reduced statistical power.

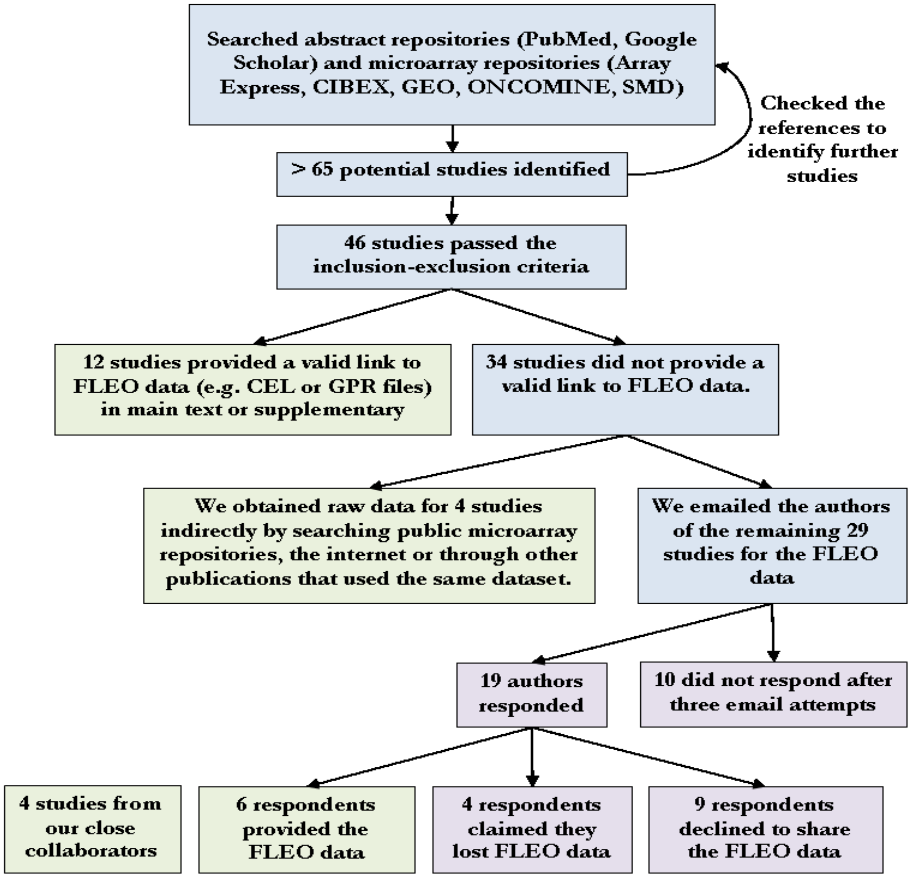



Figure 8.1.: Data identification and acquisition (Steps $3-10$ in Table 8.1). In total, the 26 studies (corresponds to the green boxes) obtained either through our collaborators, from study co-authors, indirectly from other means or directly from the publication itself are included in the meta-analysis.

| Study Information | | | | # Samples | |
|---|---|---|---|---|---|
| **Citation** | **Tissue** | **PMID** | **Platform / Chip type** | **Normal** | **Tumor** |
| Basso *et al.* (2005) | B-cells | 15778709 | HGU-95Av2 | 10 | 66 |
| Haslinger *et al.* (2004) | B-cells | 15459216 | HGU-95Av2 | 11 | 100 |
| Klein *et al.* (2001) | B-cells | 11733577 | HGU-95Av2 | 20 | 32 |
| Dyrskjot *et al.* (2004) | Bladder | 15173019 | HGU-133A | 9 | 41 |
| Ramaswamy *et al.* (2001) | Bladder | 11742071 | HU-6800 + HU-35kSubA | 7 | 11 |
| Pellagatti *et al.* (2006) | Bone Marrow | 16527891 | HGU-133A plus 2.0 | 11 | 58 |
| Ramaswamy *et al.* (2001) | Brain | 11742071 | HU-6800 + HU-35kSubA | 9 | 20 |
| Ramaswamy *et al.* (2001) | Colon | 11742071 | HU-6800 + HU-35kSubA | 11 | 11 |
| Chen *et al.* (2003) | Gastric | 12925757 | cDNA | 28 | 90 |
| Hippo *et al.* (2002) | Gastric | 11782383 | HU-6800 | 8 | 22 |
| Ginos *et al.* (2004) | Head & Neck | 14729608 | HGU-133A | 13 | 25 |
| Kuriakose *et al.* (2004) | Head & Neck | 15170515 | HGU-95Av2 | 17 | 17 |
| Winter *et al.* (2007) | Head & Neck | 17409455 | HGU-133A plus 2.0 | 11 | 59 |
| Lenburg *et al.* (2003) | Kidney | 14641932 | HGU-133A + HGU-133B | 8 | 9 |
| Ramaswamy *et al.* (2001) | Kidney | 11742071 | HU-6800 + HU-35kSubA | 11 | 11 |
| Chen *et al.* (2002) | Liver | 12058060 | cDNA | 75 | 104 |
| Beer *et al.* (2002) | Lung | 12118244 | HU-6800 | 10 | 86 |
| Bhattacharjee *et al.* (2001) | Lung | 11707567 | HGU-95Av2 | 17 | 139 |
| Borczuk *et al.* (2003) | Lung | 14578194 | HGU-95A | 7 | 34 |
| Hu *et al.* (2005) | Lung | 15592519 | cDNA | 14 | 47 |
| Jones *et al.* (2004) | Lung | 15016488 | cDNA | 19 | 24 |
| Couvelard *et al.* (2005) | Pancreas | 15910598 | cDNA | 7 | 12 |
| Grützmann *et al.* (2004) | Pancreas | 15548371 | HGU-133A + HGU-133B | 11 | 14 |
| Ishikawa *et al.* (2005) | Pancreas | 16053509 | HGU-133A + HGU-133B | 16 | 11 |
| Ramaswamy *et al.* (2001) | Pancreas | 11742071 | HU-6800 + HU-35kSubA | 10 | 11 |
| Lapointe *et al.* (2004) | Prostate | 14711987 | cDNA | 41 | 62 |
| Singh *et al.* (2002) | Prostate | 12086878 | HGU-95Av2 | 50 | 52 |
| Welsh *et al.* (2001) | Prostate | 11507037 | HGU-95A | 8 | 25 |
| Aldred *et al.* (2003) | Thyroid | 12776192 | HGU-95Av2 | 7 | 9 |
| Huang *et al.* (2001) | Thyroid | 11752453 | HGU-95A | 8 | 8 |
| | | | | 484 | 1210 |

Table 8.2.: Datasets with FLEO data used for the case study. In total there are 30 datasets from 26 published studies.

### 8.2.2. Extra information about annotation and mapping

Table 8.3 shows the annotation information of the various chip designs used across the 26 studies listed in Table 8.2. For example, Chen *et al.* (2002) used a cDNA chip with 43,330 probes of which 5,539 probes cannot be mapped to any UniGene ID. 34,952 of the remaining 37,800 probes mapped to exactly one UniGene. After expanding the probes that mapped to multiple UniGene IDs (see Step 21 of Table 6.1), there were 40,666 records which corresponds to 28,593 unique UniGenes. Across the 30 datasets, we have 700,307 probes that map into

36,965 unique UniGeneIDs.

| Chip design | # probes | # mapped | # one2one | # records | # UniGene IDs | Studies |
|---|---|---|---|---|---|---|
| cDNA | 43330 | 37800 (299) | 34952 (262) | 40666 (299) | 28593 (147) | Chen *et al.* (2002) |
| cDNA | 42037 | 36479 (268) | 33966 (232) | 39007 (268) | 27648 (155) | Chen *et al.* (2003) |
| cDNA | 40368 | 33454 (243) | 33454 (243) | 33454 (243) | 25267 (152) | Jones *et al.* (2004) |
| cDNA | 45205 | 39470 (322) | 36590 (285) | 42368 (322) | 29274 (180) | Lapointe *et al.* (2004) |
| cDNA | 9910 | 8379 (131) | 8379 (131) | 8379 (131) | 5131 (83) | Couvelard *et al.* (2005) |
| cDNA | 9929 | 8378 (131) | 8378 (131) | 8378 (131) | 5130 (83) | Hu *et al.* (2005) |
| HU-6800 | 7129 | 6912 (98) | 6195 (87) | 7786 (105) | 6338 (96) | Beer *et al.* (2002)<br>Hippo *et al.* (2002) |
| HU-6800 + HU-35kSubA | 16063 | 15086 (197) | 13519 (183) | 16969 (204) | 11375 (152) | Ramaswamy *et al.* (2001) *(5 subsets)* |
| HGU-95A | 12626 | 12054 (183) | 10851 (166) | 13466 (190) | 9783 (147) | Huang *et al.* (2001)<br>Klein *et al.* (2001)<br>Welsh *et al.* (2001)<br>Haslinger *et al.* (2004)<br>Basso *et al.* (2005) |
| HGU-95Av2 | 12625 | 12097 (185) | 10893 (168) | 13506 (192) | 9796 (148) | Bhattacharjee *et al.* (2001)<br>Singh *et al.* (2002)<br>Aldred *et al.* (2003)<br>Borczuk *et al.* (2003)<br>Kuriakose *et al.* (2004) |
| HGU-133A | 22283 | 21291 (315) | 19026 (278) | 23984 (333) | 13899 (175) | Dyrskjot *et al.* (2004)<br>Ginos *et al.* (2004) |
| HGU-133A + HGU-133B | 44928 | 34968 (451) | 31238 (401) | 39371 (475) | 18856 (202) | Lenburg *et al.* (2003)<br>Grützmann *et al.* (2004)<br>Ishikawa *et al.* (2005) |
| HGU-133A plus 2.0 | 54675 | 46976 (515) | 41985 (466) | 52834 (538) | 20927 (207) | Pellagatti *et al.* (2006)<br>Winter *et al.* (2007) |
| | 700307 | 615407 (7428) | 560160 (6729) | 679278 (8147) | 36965 (270) | all 30 datasets |

Table 8.3.: Characteristics of microarrays chips used in the studies listed in Table 8.2. For each design, the number of probes spotted (# probes), the number of probes with any UniGene ID (# mapped), the number of probes that map to exactly one UniGene ID (# one2one), the number of records after expanding probes as described in Step 21 of Table 6.1 (# records), the number of UniGene IDs after summarizing probes as described in Step 22 (# UniGeneID). The numbers in the parenthesis refer to only those involved with protein folding genes.

### 8.2.3. Identifying Protein Folding Genes

Since we are interested in protein folding genes only, we restricted the analysis to genes mapping to Gene Ontology (http://geneontology.org/) identifier GO:0006457 which refers to the biological process "protein folding". Thus any gene that matched this id was considered as Protein Folding Gene. As we show below, using this criteria, about 0.7% of the UniGene

IDs represents PFG.

### 8.2.4. Frequency of identification

Table 8.4 shows the frequency of identification of the 36,965 UniGene IDs across the 30 datasets. Only 2236 or 6% of the 36965 UniGene IDs are present in all 30 datasets.

| **# studies** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
|---|---|---|---|---|---|---|---|---|---|---|
| **all genes** | 820 | 3817 | 8379 | 4394 | 1552 | 249 | 1227 | 1091 | 1058 | 817 |
| **PFG only** | 2 | 5 | 1 | 0 | 6 | 0 | 3 | 3 | 2 | 4 |
| **# studies** | **11** | **12** | **13** | **14** | **15** | **16** | **17** | **18** | **19** | **20** |
| **all genes** | 692 | 367 | 456 | 865 | 409 | 811 | 390 | 181 | 124 | 357 |
| **PFG only** | 9 | 0 | 7 | 3 | 2 | 3 | 1 | 1 | 1 | 1 |
| **# studies** | **21** | **22** | **23** | **24** | **25** | **26** | **27** | **28** | **29** | **30** |
| **all genes** | 414 | 386 | 355 | 975 | 420 | 923 | 828 | 1824 | 548 | 2236 |
| **PFG only** | 4 | 2 | 5 | 7 | 6 | 10 | 4 | 21 | 13 | 20 |

Table 8.4.: Frequency of UniGene ID identification across datasets.

Notice that 12,773 or about 35% of the 36,965 UniGene IDs are present in $3-4$ of the 30 studies. Upon closer inspection, majority of these arise from just 4 studies, (Chen *et al.*, 2002, 2003; Jones *et al.*, 2004; Lapointe *et al.*, 2004), that happens to be of cDNA design. This is also reflected in the higher values of column # UniGene ID in Table 8.3 for these four studies. Although we cannot explain these phenomena, it is possible that most of these UniGenes were excluded by the probe designers in the Affymetrix company.

### 8.2.5. Gene Set Enrichment Analysis (GSEA)

In gene set enrichment analysis, we look for enrichment of biological themes among the list of genes declared as differentially expressed. This has several advantageous over a gene by gene inspection. Subramanian *et al.* (2005) states the following problems that may be overcome by GSEA approach

> (i) After correcting for multiple hypotheses testing, no individual gene may meet the threshold for statistical significance, because the relevant biological differences are modest relative to the noise inherent to the microarray technology.

(ii) Alternatively, one may be left with a long list of statistically significant genes without any unifying biological theme. Interpretation can be daunting and ad hoc, being dependent on a biologist's area of expertise.

(iii) Single-gene analysis may miss important effects on pathways. Cellular processes often affect sets of genes acting in concert. An increase of 20% in all genes encoding members of a metabolic pathway may dramatically alter the flux through the pathway and may be more important than a 20-fold increase in a single gene.

(iv) When different groups study the same biological system, the list of statistically significant genes from the two studies may show distressingly little overlap.

GSEA begins by assigning the annotation terms that are associated with each gene. This information may be obtained from many database including the Gene Ontology which we mentioned in Section 8.2.3. The Gene Ontology project consists of three controlled vocabularies to describe the biological processes, cellular components and molecular functions of a gene.

Next, we count the number of genes associated with a given term in the list of genes that we have declared as differentially expressed. Similarly, we count the number of genes associated with this term in the background/population (i.e. all genes that would be eligible to be tested for differential expression). If there is a significantly higher proportion of genes that is associated in the gene list than the background, then we can declared that this term is enriched. There are many variety of this test which can formally test enrichment using Chi-square test, Fisher's exact test or the hypergeometric distribution.

One of the popular tools for GSEA is a freely available bioinformatics webtool called DAVID (the Database for Annotation, Visualization and Integrated Discovery) (Dennis *et al.*, 2003; Huang *et al.*, 2009) and can be found at http://david.abcc.ncifcrf.gov.

The databases used in DAVID to identify annotation terms are: Online Mendelian Inheritance in Man database, Functional Categories (COG_ONTOLOGY, SP_PIR_KEYWORDS, UP_SEQ_FEATURE), Gene Ontology (GOTERM_BP_FAT, GOTERM_CC_FAT, GOTERM_MF_FAT), Pathways (BBID, BIOCARTA, KEGG_PATHWAY) and Protein Domains (INTERPRO, PIR_SUPERFAMILY, SMART).

### 8.2.6. Logistics

Conducting such a meta-analysis may be inexpensive as it makes use of existing data. However, it can be labor intensive to identify and acquire them, which is complicated by data sharing issues. As shown in Figure 8.1, only 12 or 26%of the 46 identified studies provided valid links to FLEO data in the main text or supplementary of the article. The data for the remaining 34 studies required additional effort to locate.

Performing such a meta-analysis requires good communication skills, planning skills, persistence and patience as data often has to be requested from coauthors of the remaining studies. Contacting coauthors of 29 studies, with over 400 email exchanges, yielded the raw data for only 6 additional studies (20% success rate). At least 9 study coauthors or 31% clearly refused to share their raw data. We explore and report these findings in a larger set of studies in the next chapter.

Such a meta-analysis also has a substantial computational requirement to organize and curate the datasets. The raw data for the 30 datasets take up approximately 21 gigabytes of hard disk space to store. We then had to preprocess each datasets and extract the summary statistics for each study. The preprocessing of some of the larger studies required more that 4 gigabyte of random-access memory which is not possible in a Windows based machine and requires large scale Unix or Linux architecture.

## 8.3. Results of Meta-analyses of Solid Tumours

### 8.3.1. Gene-centric view

For this case study, we select 26 studies of solid tumours which comprises a total of 954 tumours and 432 normals. As described above, we applied meta-analyses using random effects inverse variance across the studies. The pooled effect size and the 95% confidence interval for all genes can be visualized in a summary plot (Figure 8.2). This shows that ACTL6A (Actin-like 6A) which is strongly differentially expressed.

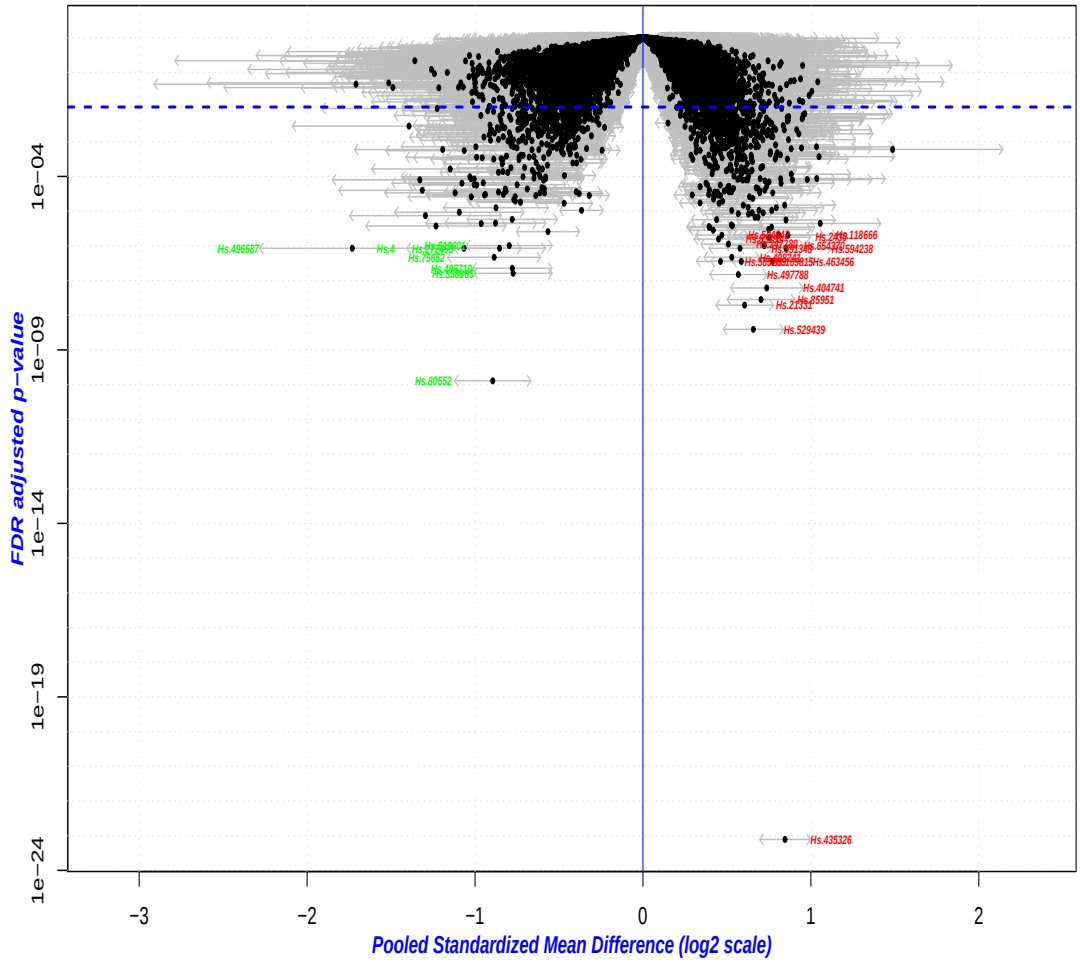



Figure 8.2.: A summary plot of the pooled effect size (black dots) for meta-analyses of 26 solid tumours dataset and its 95% confidence interval (gray bars) sorted by the false discovery rate.

At the False Discovery rate of 0.01, there are 906 significant genes that is present in at least five studies. Thus we can expect 1% of the list or 9.06 genes to be false positives. Due to space restrictions, we list only the genes with p-value $< 10^{-6}$ in Table 8.5.

| UniGene ID | $N$ | log OR | SE | $\tau^2$ | $P_{FDR}$ | Symbol | Name |
|---|---|---|---|---|---|---|---|
| Hs.435326 | 22 | 0.846 | 0.077 | 0.018 | 4.0E-24 | ACTL6A | Actin-like 6A |
| Hs.80552 | 24 | -0.895 | 0.116 | 0.240 | 1.9E-11 | DPT | Dermatopontin |
| Hs.529439 | 8 | 0.657 | 0.091 | 0.000 | 6.5E-10 | ZBTB41 | Zinc finger and BTB domain containing 41 |
| Hs.21331 | 15 | 0.605 | 0.087 | 0.037 | 3.6E-09 | ZWILCH | Zwilch, kinetochore associated, homolog (Drosophila) |
| Hs.85951 | 24 | 0.703 | 0.103 | 0.139 | 6.0E-09 | XPOT | Exportin, tRNA (nuclear export receptor for tRNAs) |
| Hs.404741 | 22 | 0.737 | 0.110 | 0.152 | 1.5E-08 | NFE2L3 | Nuclear factor (erythroid-derived 2)-like 3 |
| Hs.497788 | 24 | 0.568 | 0.087 | 0.091 | 3.9E-08 | EPRS | Glutamyl-prolyl-tRNA synthetase |
| Hs.558865 | 24 | -0.773 | 0.119 | 0.212 | 4.8E-08 | CES1 | Carboxylesterase 1 (monocyte/macrophage serine esterase 1) |
| Hs.495710 | 26 | -0.778 | 0.121 | 0.285 | 6.9E-08 | GPM6B | Glycoprotein M6B |
| Hs.585209 | 22 | 0.462 | 0.073 | 0.031 | 1.2E-07 | C16orf88 | Chromosome 16 open reading frame 88 |
| Hs.169815 | 13 | 0.585 | 0.092 | 0.042 | 1.2E-07 | FBXO45 | F-box protein 45 |
| Hs.463456 | 26 | 0.771 | 0.122 | 0.340 | 1.3E-07 | NME1 | Non-metastatic cells 1, protein (NM23A) expressed in |
| Hs.408241 | 24 | 0.528 | 0.084 | 0.064 | 1.8E-07 | NUPL2 | Nucleoporin like 2 |
| Hs.75652 | 24 | -0.886 | 0.141 | 0.353 | 1.8E-07 | GSTM5 | Glutathione S-transferase mu 5 |
| Hs.496587 | 12 | -1.731 | 0.282 | 0.783 | 3.4E-07 | CHRDL1 | Chordin-like 1 |
| Hs.4 | 26 | -1.066 | 0.174 | 0.647 | 3.5E-07 | ADH1B | Alcohol dehydrogenase 1B (class I), beta polypeptide |
| Hs.272493 | 25 | -0.854 | 0.139 | 0.340 | 3.5E-07 | CCL15 | Chemokine (C-C motif) ligand 15 |
| Hs.594238 | 24 | 0.851 | 0.139 | 0.340 | 3.5E-07 | KPNA2 | Karyopherin alpha 2 (RAG cohort 1, importin alpha 1) |
| Hs.591343 | 13 | 0.578 | 0.094 | 0.052 | 3.6E-07 | MTHFD1L | Methylenetetrahydrofolate dehydrogenase (NADP+ dependent) 1-like |
| Hs.519601 | 24 | -0.796 | 0.131 | 0.307 | 4.5E-07 | ID4 | Inhibitor of DNA binding 4, dominant negative helix-loop-helix protein |
| Hs.654370 | 26 | 0.722 | 0.119 | 0.229 | 4.6E-07 | FAP | Fibroblast activation protein, alpha |
| Hs.247280 | 22 | 0.509 | 0.084 | 0.086 | 5.2E-07 | RBCK1 | RanBP-type and C3HC4-type zinc finger containing 1 |
| Hs.83293 | 8 | 0.450 | 0.075 | 0.014 | 7.3E-07 | ANKIB1 | Ankyrin repeat and IBR domain containing 1 |
| Hs.2430 | 25 | 0.751 | 0.126 | 0.282 | 8.1E-07 | VPS72 | Vacuolar protein sorting 72 homolog (S. cerevisiae) |
| Hs.118666 | 26 | 0.863 | 0.145 | 0.421 | 9.7E-07 | FLAD1 | FAD1 flavin adenine dinucleotide synthetase homolog (S. cerevisiae) |
| Hs.584842 | 26 | 0.469 | 0.079 | 0.077 | 9.8E-07 | SART3 | Squamous cell carcinoma antigen recognized by T cells 3 |

Table 8.5.: List of genes significant at p-value $< 10^{-6}$ for the solid tumours case study. For each UniGene ID, the number of studies the gene was present in ($N$), the summary of the pooled log odds ratio, standard error, between-study heterogeneity measure ($\tau^2$), FDR adjusted $p$-value, the gene symbol and name are shown.

## 8.3.2. Annotation-centric view

We entered the UniGene ID for the 906 genes into DAVID and it recognized 850 or 94% of these as the GENELIST. We entered the UniGene ID for 18,057 genes that were present in at least five studies as the BACKGROUND gene list.

For each gene, DAVID attaches the annotation terms found from various databases (including Gene Ontology). Figure 8.3 shows the available annotation terms for the two most significant genes here.

| Hs.435326 | actin-like 6A | Related Genes | Homo sapiens |
|---|---|---|---|
| BIOCARTA | Control of Gene Expression by Vitamin D Receptor, | | |
| GOTERM_BP_FAT | chromatin organization, chromatin remodeling, transcription, protein amino acid acetylation, chromatin modification, covalent chromatin modification, histone modification, histone acetylation, regulation of growth, protein amino acid acylation, histone H4 acetylation, histone H2A acetylation, regulation of transcription, chromosome organization, | | |
| GOTERM_CC_FAT | histone acetyltransferase complex, nucleoplasm, plasma membrane, SWI/SNF complex, chromatin remodeling complex, membrane-enclosed lumen, nuclear lumen, NuA4 histone acetyltransferase complex, H4/H2A histone acetyltransferase complex, organelle lumen, nucleoplasm part, intracellular organelle lumen, SWI/SNF-type complex, | | |
| GOTERM_MF_FAT | nucleotide binding, nucleoside binding, purine nucleoside binding, chromatin binding, ATP binding, purine nucleotide binding, adenyl nucleotide binding, ribonucleotide binding, purine ribonucleotide binding, adenyl ribonucleotide binding, | | |
| INTERPRO | Peptidase aspartic, active site, Actin/actin-like, Actin, conserved site, | | |
| SMART | ACTIN, | | |
| SP_PIR_KEYWORDS | acetylation, activator, alternative splicing, chromatin regulator, complete proteome, direct protein sequencing, growth regulation, neurogenesis, nucleus, phosphoprotein, Transcription, transcription regulation, | | |
| UP_SEQ_FEATURE | chain:Actin-like protein 6A, modified residue, sequence conflict, splice variant, | | |
| **Hs.80552** | **dermatopontin** | **Related Genes** | **Homo sapiens** |
| GOTERM_BP_FAT | cell adhesion, negative regulation of cell proliferation, biological adhesion, extracellular matrix organization, collagen fibril organization, regulation of cell proliferation, extracellular structure organization, | | |
| GOTERM_CC_FAT | extracellular region, proteinaceous extracellular matrix, extracellular matrix, extracellular region part, | | |
| OMIM_DISEASE | Genetic correlates of longevity and selected age-related phenotypes, | | |
| SP_PIR_KEYWORDS | cell adhesion, complete proteome, disulfide bond, extracellular matrix, polymorphism, Pyrrolidone carboxylic acid, repeat, Secreted, signal, sulfation, | | |
| UP_SEQ_FEATURE | chain:Dermatopontin, disulfide bond, modified residue, region of interest:2 X 53-55 AA tandem repeats, region of interest:3 X 6 AA repeats of D-R-[EQ]-W-[NQK]- [FY], repeat:1-1, repeat:1-2, repeat:2-1, repeat:2-2, repeat:3-3, sequence conflict, sequence variant, signal peptide, | | |

Figure 8.3.: The annotation terms from various sources for two top ranking genes from the meta-analyses of solid tumours. Using multiple databases increases the chances of finding annotation terms associated with each gene.

Viewing the results gene-by-gene to identify common annotation terms is time consuming and subject to human errors. Therefore a common way is to sort terms by their enrichment (i.e. over-representation in the GENELIST). We performed the gene enrichment analyses in DAVID and the screen capture of the top few terms that are enriched in the GENELIST

relative to BACKGROUND is given in Figure 8.4.

| Sublist | Category | Term | RT | Genes | Count | LT | PH | PT | % | P-Value | Fold Enrichment |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ☐ | GOTERM_BP_FAT | mitotic cell cycle | RT | | 61 | 631 | 338 | 10922 | 7.3 | 2.5E-15 | 3.1 |
| ☐ | GOTERM_CC_FAT | organelle lumen | RT | | 159 | 576 | 1580 | 10126 | 19.2 | 4.7E-14 | 1.8 |
| ☐ | GOTERM_CC_FAT | intracellular organelle lumen | RT | | 156 | 576 | 1541 | 10126 | 18.8 | 5.5E-14 | 1.8 |
| ☐ | GOTERM_CC_FAT | membrane-enclosed lumen | RT | | 160 | 576 | 1614 | 10126 | 19.3 | 1.4E-13 | 1.7 |
| ☐ | GOTERM_CC_FAT | nuclear lumen | RT | | 132 | 576 | 1250 | 10126 | 15.9 | 4.9E-13 | 1.9 |
| ☐ | SP_PIR_KEYWORDS | nucleus | RT | | 278 | 794 | 3553 | 14797 | 33.5 | 7.5E-13 | 1.5 |
| ☐ | SP_PIR_KEYWORDS | acetylation | RT | | 202 | 794 | 2343 | 14797 | 24.3 | 8.7E-13 | 1.6 |
| ☐ | SP_PIR_KEYWORDS | mitosis | RT | | 35 | 794 | 166 | 14797 | 4.2 | 9.8E-12 | 3.9 |
| ☐ | GOTERM_BP_FAT | cell cycle process | RT | | 70 | 631 | 509 | 10922 | 8.4 | 1.5E-11 | 2.4 |
| ☐ | GOTERM_BP_FAT | M phase | RT | | 49 | 631 | 291 | 10922 | 5.9 | 2.7E-11 | 2.9 |
| ☐ | GOTERM_BP_FAT | nuclear division | RT | | 39 | 631 | 197 | 10922 | 4.7 | 3.1E-11 | 3.4 |
| ☐ | GOTERM_BP_FAT | mitosis | RT | | 39 | 631 | 197 | 10922 | 4.7 | 3.1E-11 | 3.4 |
| ☐ | SP_PIR_KEYWORDS | cell division | RT | | 42 | 794 | 241 | 14797 | 5.1 | 3.8E-11 | 3.2 |
| ☐ | GOTERM_BP_FAT | M phase of mitotic cell cycle | RT | | 39 | 631 | 201 | 10922 | 4.7 | 5.8E-11 | 3.4 |
| ☐ | GOTERM_BP_FAT | cell cycle phase | RT | | 56 | 631 | 370 | 10922 | 6.7 | 6.2E-11 | 2.6 |
| ☐ | GOTERM_BP_FAT | cell cycle | RT | | 84 | 631 | 694 | 10922 | 10.1 | 8.7E-11 | 2.1 |
| ☐ | GOTERM_BP_FAT | organelle fission | RT | | 39 | 631 | 205 | 10922 | 4.7 | 1.1E-10 | 3.3 |
| ☐ | SP_PIR_KEYWORDS | cell cycle | RT | | 57 | 794 | 414 | 14797 | 6.9 | 1.3E-10 | 2.6 |
| ☐ | SP_PIR_KEYWORDS | phosphoprotein | RT | | 428 | 794 | 6355 | 14797 | 51.6 | 1.5E-10 | 1.3 |
| ☐ | GOTERM_BP_FAT | anaphase-promoting complex-dependent proteasomal ubiquitin-dependent protein catabolic process | RT | | 19 | 631 | 61 | 10922 | 2.3 | 5.0E-9 | 5.4 |
| ☐ | GOTERM_BP_FAT | cell division | RT | | 42 | 631 | 265 | 10922 | 5.1 | 5.7E-9 | 2.7 |
| ☐ | GOTERM_BP_FAT | regulation of ubiquitin-protein ligase activity during mitotic cell cycle | RT | | 19 | 631 | 65 | 10922 | 2.3 | 1.5E-8 | 5.1 |
| ☐ | GOTERM_BP_FAT | DNA metabolic process | RT | | 57 | 631 | 444 | 10922 | 6.9 | 2.0E-8 | 2.2 |
| ☐ | GOTERM_BP_FAT | negative regulation of ubiquitin-protein ligase activity during mitotic cell cycle | RT | | 18 | 631 | 60 | 10922 | 2.2 | 2.6E-8 | 5.2 |
| ☐ | GOTERM_CC_FAT | nucleoplasm | RT | | 82 | 576 | 773 | 10126 | 9.9 | 3.2E-8 | 1.9 |

Figure 8.4.: A screen capture of the top few terms enriched in GENELIST. Count represents the number of genes in the GENELIST that is associated at least once with the term. LT = List Total, PH = Population Hit, PT = Population Total

This shows that term "mitotic cell cycle" from the GOTERM_BP_FAT (Gene Ontology Biological Process database) is the most enriched in the list. Let us focus on this one term to understand how the $p$-value was derived. We can see that in general, only 631 genes

from the GENELIST and 10,922 genes from the BACKGROUND had at least a term in the GOTERM_BP_FAT database. Of these, 61 or 9.7% in the GENELIST and 338 or 3.1% in the BACKGROUND was associated with this term. In other words, this term is enriched 3.1 folds (=9.7%/3.1%). Here is another way of representing these numbers.

| | Associated | Not associated | Total |
|---|---|---|---|
| GENELIST | 61 (=count) | 570 | 631 (=LT) |
| Not in GENELIST | 277 | 10,014 | 10,291 |
| BACKGROUND (i.e. population) | 338 (=PH) | 10,584 | 10,922 (=PT) |

Table 8.6.: Number of genes that are associated and not associated with the Gene Ontology term "Mitotic Cell Cycle"

The p-value of 2.5E-15 shown in Figure 8.4 is based on the EASE Score which is a modified Fisher's Exact Test. The modification is to penalize the count of positive agreement by subtracting 1 unit. Therefore the p-value can be calculated in R as follows.

```
## p-value from EASE score
fisher.test( matrix( c(61-1, 570, 277+1, 10014), nc=2, byrow=T ) )$p.value
[1] 2.516541e-15

## p-value from Fisher's Exact Test for comparison only
fisher.test( matrix( c(61, 570, 277, 10014), nc=2, byrow=T ) )$p.value
[1] 6.984502e-16
```

### 8.3.3. Identifying Biological Themes using Annotation Clusters

We can see from Figure 8.4 that there is a great deal of redundancy in annotation terms that are enriched. There are obvious examples such as when the term "mitosis" appears in

GOTERM_BP_FAT and SP_PIR_KEYWORDS. A less obvious example is when annotation terms are highly related with each other as in "cell cycle process", "cell cycle phase", "cell cycle".

The "Functional Annotation Clustering" tool in DAVID uses fuzzy clustering to reduce these redundancies, allowing the users to identify biological themes or modules. Using this tool at medium classification stringency (similarity threshold=0.50), we identified 240 annotation clusters of which 18 were significant for enrichment at p-value $< 0.01$ (see Table 8.7 for a summary) and majority of these clusters fit in with our current knowledge of cancer.

Figure 8.5 gives a more detailed view of the first three clusters. Please note that the Enrichment Score for a cluster is the geometric mean of the p-values of individual terms contained within, expressed as the negative log base 10.

240 Cluster(s) Download File

| | Annotation Cluster 1 | Enrichment Score: 10.7 | | | Count | P_Value |
|---|---|---|---|---|---|---|
| | GOTERM_CC_FAT | organelle lumen | RT | | 159 | 4.7E-14 |
| | GOTERM_CC_FAT | intracellular organelle lumen | RT | | 156 | 5.5E-14 |
| | GOTERM_CC_FAT | membrane-enclosed lumen | RT | | 160 | 1.4E-13 |
| | GOTERM_CC_FAT | nuclear lumen | RT | | 132 | 4.9E-13 |
| | GOTERM_CC_FAT | nucleoplasm | RT | | 82 | 3.2E-8 |
| | GOTERM_CC_FAT | nucleolus | RT | | 60 | 1.1E-5 |
| | **Annotation Cluster 2** | **Enrichment Score: 10.54** | | | **Count** | **P_Value** |
| | GOTERM_BP_FAT | mitotic cell cycle | RT | | 61 | 2.5E-15 |
| | SP_PIR_KEYWORDS | mitosis | RT | | 35 | 9.8E-12 |
| | GOTERM_BP_FAT | cell cycle process | RT | | 70 | 1.5E-11 |
| | GOTERM_BP_FAT | M phase | RT | | 49 | 2.7E-11 |
| | GOTERM_BP_FAT | mitosis | RT | | 39 | 3.1E-11 |
| | GOTERM_BP_FAT | nuclear division | RT | | 39 | 3.1E-11 |
| | SP_PIR_KEYWORDS | cell division | RT | | 42 | 3.8E-11 |
| | GOTERM_BP_FAT | M phase of mitotic cell cycle | RT | | 39 | 5.8E-11 |
| | GOTERM_BP_FAT | cell cycle phase | RT | | 56 | 6.2E-11 |
| | GOTERM_BP_FAT | cell cycle | RT | | 84 | 8.7E-11 |
| | GOTERM_BP_FAT | organelle fission | RT | | 39 | 1.1E-10 |
| | SP_PIR_KEYWORDS | cell cycle | RT | | 57 | 1.3E-10 |
| | GOTERM_BP_FAT | cell division | RT | | 42 | 5.7E-9 |
| | **Annotation Cluster 3** | **Enrichment Score: 5.96** | | | **Count** | **P_Value** |
| | GOTERM_BP_FAT | DNA metabolic process | RT | | 57 | 2.0E-8 |
| | SP_PIR_KEYWORDS | dna replication | RT | | 20 | 4.6E-8 |
| | GOTERM_BP_FAT | DNA replication | RT | | 27 | 7.0E-6 |
| | KEGG_PATHWAY | DNA replication | RT | | 10 | 2.2E-4 |

Figure 8.5.: A screen capture of the first three annotation clusters from DAVID. Count represents the number of genes associated with the term and the p-values are from the EASE scores (modified Fisher's Exact test).

| Annotation Cluster | Enrichment p-value | # unique genes | # Terms | Example of annotation terms |
|---|---|---|---|---|
| 1 | 2.0E-11 | 161 | 6 | lumen (intracellular organelle, membrane-enclosed, nuclear lumen, organelle); nucleolus; nucleoplasm; |
| 2 | 4.8E-11 | 94 | 13 | cell cycle; cell division; mitotic cell cycle; nuclear division; organelle fission |
| 3 | 1.1E-06 | 59 | 4 | DNA metabolic process; DNA replication |
| 4 | 1.1E-05 | 76 | 6 | DNA damage; DNA repair; cellular response to stress |
| 5 | 7.8E-05 | 122 | 53 | regulation (catalytic activity, cellular protein metabolic process, ligase activity, molecular function, protein metabolic process, protein modification process, protein ubiquitination); proteolysis |
| 6 | 1.1E-04 | 49 | 10 | centromere; chromatin; chromosome; kinetochore |
| 7 | 1.5E-04 | 37 | 3 | cell cycle checkpoint |
| 8 | 2.7E-04 | 168 | 3 | non-membrane-bounded organelle |
| 9 | 9.8E-04 | 140 | 12 | ATP binding; nucleotide binding (adenyl, purine) |
| 10 | 1.2E-03 | 26 | 12 | damaged or mismatch DNA binding; mismatch repair |
| 11 | 3.5E-03 | 89 | 28 | RNA processing and binding; mRNA processing and binding; Spliceosome |
| 12 | 4.7E-03 | 22 | 10 | Chaperone; protein folding |
| 13 | 5.0E-03 | 59 | 8 | macromolecular protein complex assembly |
| 14 | 6.6E-03 | 40 | 6 | heparin binding; glycosaminoglycan binding |
| 15 | 8.2E-03 | 27 | 3 | DNA-dependent transcription |
| 16 | 8.9E-03 | 35 | 8 | ubl conjugation |
| 17 | 9.1E-03 | 3 | 5 | Myelin proteolipid protein PLP |
| 18 | 9.3E-03 | 21 | 6 | ribosome biogenesis |

Table 8.7.: Summary of the 18 annotation clusters from DAVID that is significant at p-value $< 0.01$. For each cluster, we give the Enrichment p-value ($\log_{10}$ of Enrichment score), number of unique genes involved in all terms within the cluster, number of terms involved and some examples.

## 8.4. Results of Meta-analyses of Protein Folding Genes

Using the Gene Ontology criteria, 138 UniGene IDs were identified as protein folding and in at least five of the studies. The frequency of identification is given in Table 8.4.

The pooled effect size and the 95% confidence interval for all 138 genes can be visualized in a summary plot (Figure 8.6). Table 8.8 shows the numerical output for the eleven genes that are significant at a false discovery rate of 1%. At this rate, we can expect 1% of the list or 0.11 gene to be false positives.

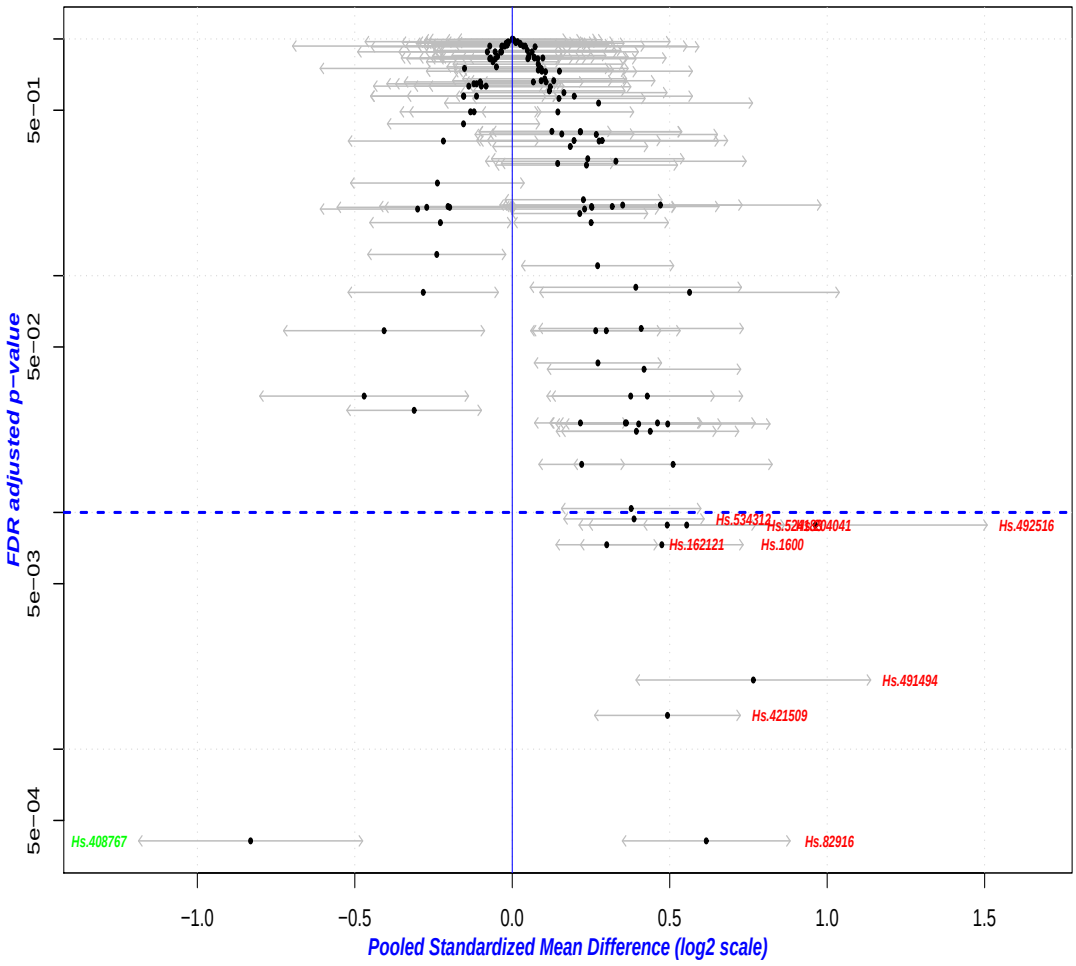



Figure 8.6.: A summary plot of the pooled effect size (black dots) and its 95% confidence interval (gray bars) of the meta-analyses of protein folding genes across 30 datasets sorted by the false discovery rate.

Figure 8.6 shows that while there are many up-regulated genes as theorized in the literature, there is one gene - Crystallin, alpha B (Hs.408767 or CRYAB) - appear to be significantly down-regulated. This gene was studied in all 30 studies and the pooled effect size is $2^{-0.831} = 0.56$ with a 95% confidence interval (0.44, 0.72). Thus, we could expect the expression of CRYAB to be reduced by half in cancer tissues compared to normal tissues.

| **UniGene ID** | $N$ | **log OR** | **SE** | $\tau^2$ | $P_{FDR}$ | **Symbol** |
|---|---|---|---|---|---|---|
| Hs.408767 | 30 | -0.831 | 0.182 | 0.866 | $2.9 \times 10^{-4}$ | CRYAB |
| Hs.82916 | 28 | 0.616 | 0.136 | 0.400 | $2.9 \times 10^{-4}$ | CCT6A |
| Hs.421509 | 30 | 0.493 | 0.118 | 0.312 | $1.1 \times 10^{-3}$ | CCT4 |
| Hs.491494 | 25 | 0.766 | 0.190 | 0.738 | $1.4 \times 10^{-3}$ | CCT3 |
| Hs.162121 | 30 | 0.300 | 0.082 | 0.111 | $5.6 \times 10^{-3}$ | COPA |
| Hs.1600 | 28 | 0.475 | 0.132 | 0.364 | $5.8 \times 10^{-3}$ | CCT5 |
| Hs.204041 | 25 | 0.554 | 0.158 | 0.476 | $7.6 \times 10^{-3}$ | AHSA1 |
| Hs.524183 | 29 | 0.492 | 0.143 | 0.466 | $7.7 \times 10^{-3}$ | FKBP4 |
| Hs.492516 | 9 | 0.963 | 0.279 | 0.555 | $7.7 \times 10^{-3}$ | PFDN2 |
| Hs.534312 | 28 | 0.387 | 0.114 | 0.279 | $8.3 \times 10^{-3}$ | TOR1A |
| Hs.356331 | 27 | 0.378 | 0.113 | 0.244 | $9.3 \times 10^{-3}$ | PPIA |

Table 8.8.: The output from the inverse-variance technique for the six genes which are significant at a false discovery rate of 1%. For each UniGene ID, the number of studies the gene was present in ($N$), the summary of the pooled log odds ratio, standard error, between-study heterogeneity measure ($\tau^2$), FDR adjusted $p$-value and gene symbol are shown.

After having identified the gene(s) of interest, we proceed as in a traditional meta-analysis and visualize the contribution of individual studies using forest plots. For example, Figure 8.7 shows the forest plot for Hs.408767 (Crystallin, alpha B).

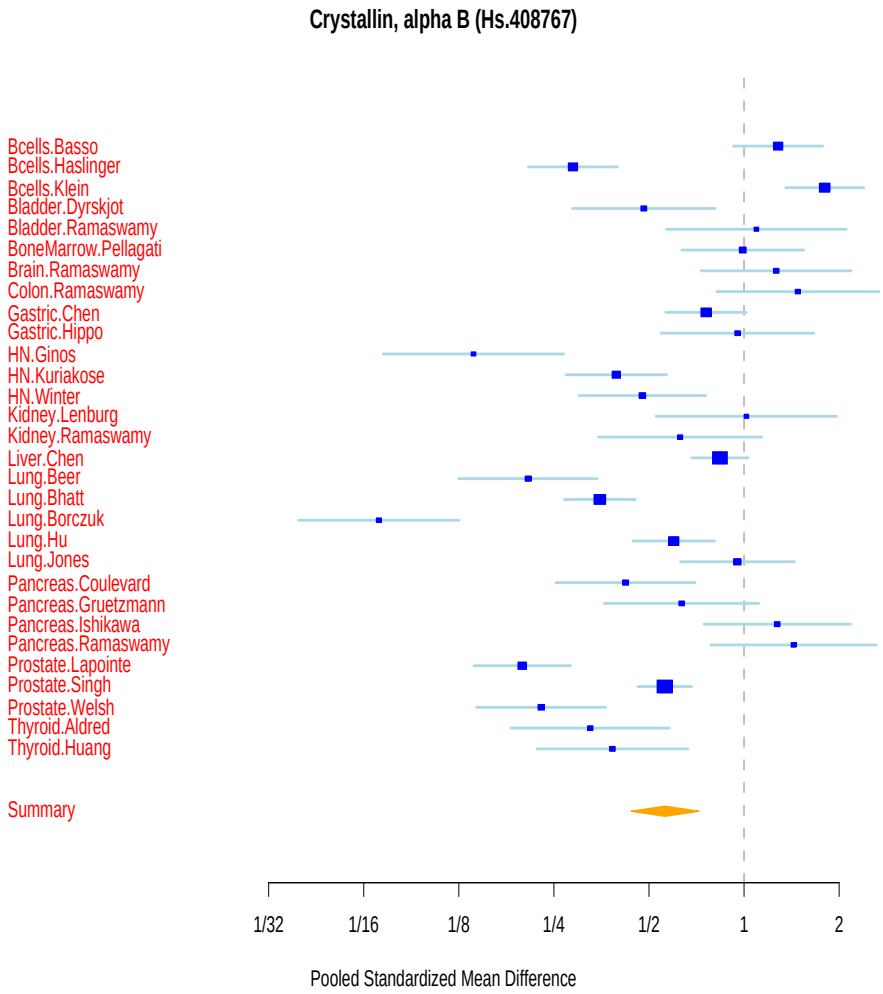



Figure 8.7.: Forest plot of Crystallin, Alpha B (Hs.408767).

## 8.5. Summary

In this chapter, we performed a meta-analysis of original individual patient DNA microarray data. The first case study is a meta-analyses of 26 solid tumours where we showed that we are able to identify genes and biological themes that are meaningful. In the second case study and motivating example of 30 datasets using only the protein folding genes, we showed that while the up-regulation of these genes in cancer is common as we anticipated, at least one protein folding gene (crystallin, alpha B) is consistently down-regulated. This suggests further lines of investigation which we are pursuing now.

# 9. Exploring data sharing issues in microarray genomics research

*Raw data availability is critically important for good science and scientific progress. In particular it allows researchers to validate claims made, carry out additional analysis, plan new studies and, importantly for meta-analysis, enable more precise conclusions to be made. In this article we explore the raw data availability for three collections of published studies and investigate the statistical association of study and journal characteristics with these patterns. Finally, we make several recommendations to remedy this situation, and therefore improve the quality of research using microarray technology.*

*Materials from this chapter has been submitted and is under peer review process as:*

Ramasamy A, Holmes CC, Altman DG (2009).
Failure to share data in microarray research.
PLoS ONE (under review).

## 9.1. Introduction

Recently we attempted to acquire raw data for several published microarray studies. We expected this to be relatively straightforward since the widely adopted Minimum Information About a Microarray Experiment (MIAME) requirements Brazma *et al.* (2001) recommends authors to make raw data publicly available and many leading journals require authors to follow MIAME requirements. However, we were surprised to find that only a quarter of the articles we considered reported a valid link to raw data. We decided to investigate this phenomenon more deeply, incorporating information on two other sets of studies compiled by other researchers (Piwowar *et al.*, 2007; Ioannidis *et al.*, 2007).

### 9.1.1. Minimum Information About a Microarray Experiment (MIAME) requirement on raw microarray data

Shortly after the introduction of DNA expression microarrays Schena *et al.* (1995), several researchers and industrial partners recognized the importance of making microarray data available. The Microarray Gene Expression Data (MGED) Society (http://www.mged.org/) was founded in November 1999 as a grass-roots organization that tasked itself to addressing this issue further. In December 2001, they published data-reporting requirements called the Minimum Information About a Microarray Experiment (MIAME) (Brazma *et al.*, 2001) outlining the information that researchers should make available when reporting microarray experiments in publications. In October 2002, these requirements were summarized into the MIAME checklist and made available online (Ball *et al.*, 2002). Many leading scientific journals including Cell, Lancet, Nature and Science quickly adopted and endorsed the MIAME requirements. The adoption of MIAME requirements was considered so successful that similar "minimum information" requirements for other high-throughput biological technologies modeled after MIAME have followed (Taylor *et al.*, 2007).

In September 2004, Ball *et al.* (2004) updated the MIAME requirements. There were two major changes. First, they recognized Array Express (Brazma *et al.*, 2003), CIBEX (Ikeo *et al.*, 2003) and Gene Expression Omnibus (GEO) (Edgar *et al.*, 2002) as MIAME-compliant repositories. Second, they proposed the compulsory submission of microarray data into

one of these three MIAME-compliant public repositories. Appendix A.6 summarizes the major milestones in microarray data reporting requirements and public repositories.

The MIAME requirements cover many aspects including descriptions of the experimental design, array design, samples, hybridization procedures, measurement data and normalization. All of these aspects are important, but here we focus only on measurement data. Access to original measurement data is important to allow validation and secondary analysis, but we found such access difficult to obtain in practice.

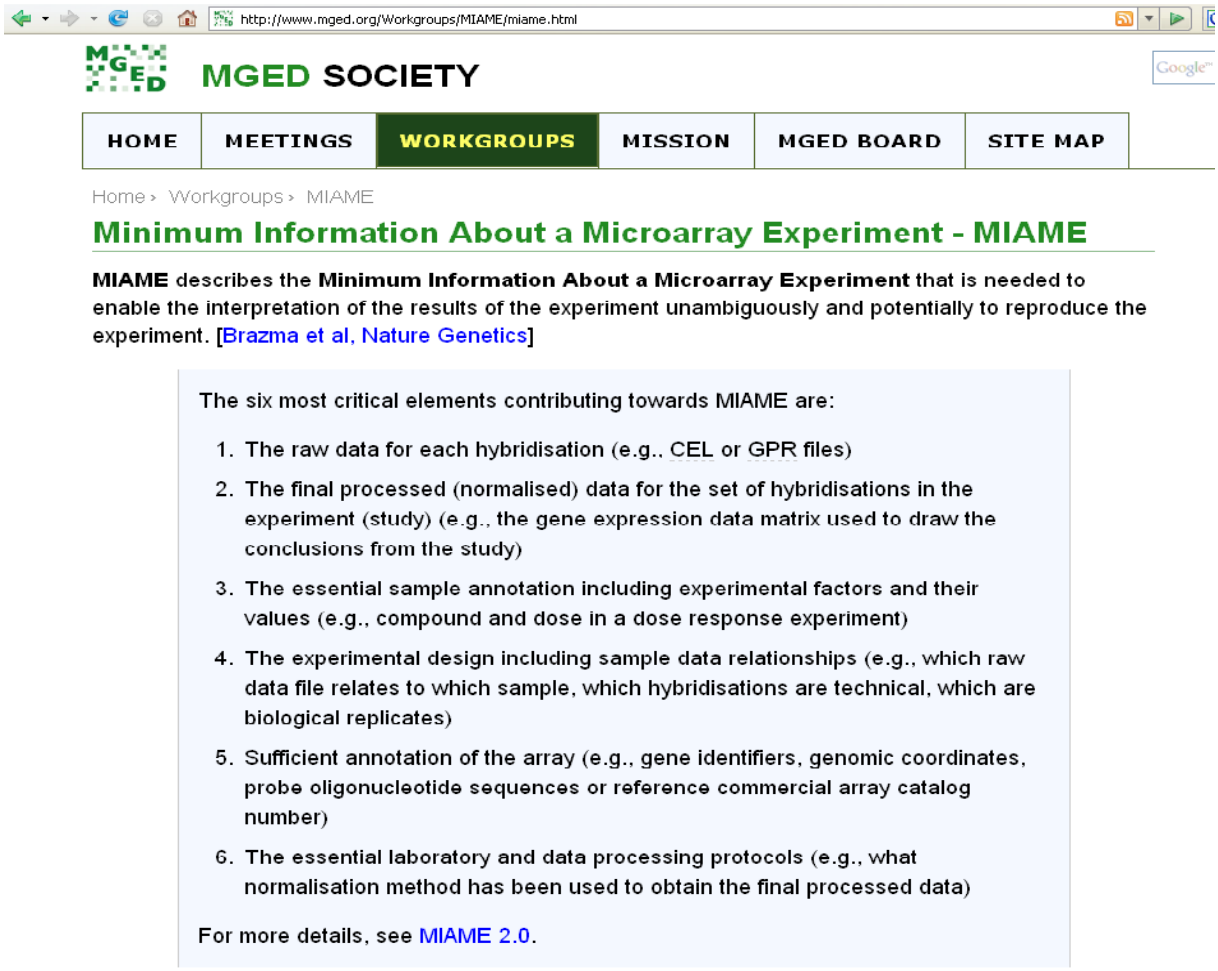



Figure 9.1.: Screen capture of the official MIAME website accessed on $20^{th}$ August 2008.

.

Figure 9.1 clearly shows that MIAME requires **both** the FLEO and GEDM to be made publicly available . Providing only the GEDM data is insufficient as it is severely limited by the choice of preprocessing algorithm used and it does not allow for standardization of information across multiple studies to the same level as FLEO data can.

Note that MIAME working group does not yet have a consensus on whether the provision of images should also be required. The image data is many times larger than FLEO data and few image processing algorithms are freely available. The usefulness of such data is thus limited and image files are not routinely deposited. Almost all the studies investigated in this article that had provided FLEO data also provided the GEDM but only a few also provided the image files. In this article, we consider the provision of FLEO data as a necessary condition of compliance with MIAME requirements to provide raw data.

### 9.1.2. Why is the raw data important?

Access to raw data is important to the general community for the following reasons:

- Funding bodies and other researchers may wish to validate a claim made in a scientific article.
- One can carry out additional analysis that was not reported in the original publication.
- Since preprocessing algorithms are being actively developed and tested, one can easily update existing results by preprocessing the FLEO data using a new algorithm or test the robustness of the results under different algorithms.
- One can use existing relevant datasets to plan for new studies, for example, to estimate parameters for sample size calculations as we have done in Chapters 3-5.
- If the FLEO data are available for several biologically similar studies, then one can combine the information from them, as we have done in Chapters 6-8. The results from such a meta-analysis should yield a more precise and robust estimate of gene differential.

The importance of the raw data is even widely acknowledged in the community. For example, a survey of 313 authors and reviewers in Physiological Genomics (survey response rate of 29%) found that 92% of the survey respondents believed that depositing microarray data was of significant importance and cited many examples where they found such data to be useful for their own research Ventura (2005).

Further, providing the raw data is not only advantageous for the scientific community in general but also to the study authors themselves. Piwowar *et al.* (2007) found that microarray-

based studies that made their individual patient-level data (IPD) publicly available had 69% more citations independently of journal impact factor, date of publication and author country.

## 9.2. Collection of studies assessed

We used three sets of collection of studies to assess for reporting of raw data. The first set is our own collection of 63 cancer related articles which was collected to prepare for the previously described chapters in this project and few other projects. We have excluded four of the 67 studies for which we obtained the raw data from our close collaborators.

The second set is a list of 85 studies compiled by Ntzani and Ioannidis (2003) and used by Piwowar *et al.* (2007) to investigate factors that influence citation. We could not use the information from Supplementary Table S1 of Piwowar *et al.* (2007) directly as the definition of raw data, trial size and location of raw data was inconsistent between Piwowar *et al.* (2007) and our study. To elaborate further, Piwowar *et al.* (2007) defined "raw data" to include both FLEO and GEDM (i.e. IPD) while we define it to be only the FLEO data. Next, Ntzani and Ioannidis (2003), and subsequently Piwowar *et al.* (2007), defined the trial size as the number of biological cases only and excluded any controls or technical replicates. Finally, Piwowar *et al.* (2007) did not clearly distinguish if the link was obtained from the article, by searching the research group's website or from the public microarray repositories. There were ten instances of these different definitions.

The third set comes from Ioannidis *et al.* (2007) who investigated the public availability of raw data for 24 studies across two non-overlapping time periods. After confirming the information in Table 5 of the main text of Ioannidis *et al.* (2007), we incorporated it.

In total, there were 146 unique studies published in 44 different journals. Although all three sets of studies were mostly in cancer, there was little overlap, with 121 studies (82.8%) included in only one set (see Figure 9.2).

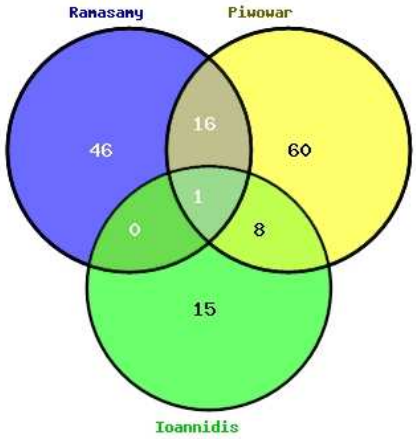



Figure 9.2.: Venn diagram showing the overlap in the number of studies in the three sets.

Diagram created using the VENNY website (Oliveros, 2007).

## 9.3. Data extraction

For each published article, we searched for a link to the raw data in the main text and (if available) the supplementary material. If a web link was reported, we accessed it to check if it was valid and determined the most comprehensive type of information available there. This assessment of web links was carried out in the fourth week of July 2007.

If a study reported the location of the FLEO data, we did not necessarily check for the presence of GEDM as well. Thus "data reported" had four possible values, arranged in decreasing preference: FLEO data, GEDM only (but not FLEO), invalid link and no link to any kind of IPD.

Additionally, we recorded the following study and journal characteristics:

1. **The number of months since publication from January 1999**. Occasionally, articles are published online several weeks or even months before they appear in print. We took the earliest date the article was publicly available.

2. **The sample size of the study.** If a valid link to raw data was identified, we took this to be the number of FLEO files available via the link. There were some minor inconsistencies between the number of cases found in repositories and those reported which we do not discuss here. If no valid link to FLEO data was reported, we used the

total number of microarray hybridizations reported. This information was not always clearly reported in publications and in a few cases we approximated the numbers from visual representations presented in the paper (e.g. number of samples shown in heatmaps).

3. **The 2006 Journal Impact Factor (JIF06)** from the ISI Journal Citation Reports 2006 online at the ISI Web of Knowledge (http://www.isiknowledge.com/) for each of the 44 unique journals that had published the articles in the superset. We tried a log transformation of this variable but it did not change the conclusions, and therefore we report the results from fitting it on the natural scale.

4. Whether one of the co-author(s) of the study was **affiliated with a commercial company**. We did not classify the following as commercial companies: The Emmes Corporation, Mitsubishi Chemical Safety Institute and IBM Watson Research Center.

5. Whether one of the co-authors(s) was **based in the USA, Europe** (including the United Kingdom) or Japan. In the combined superset 93 articles had been co-authored by at least a researcher based in America, 55 by those based in Europe and 22 by those based in Japan. Because an article can have co-authors from multiple regions we used three binary indicator variables.

   We did not specifically consider co-authors based in the following countries because of low frequency as indicated in parenthesis: China including Hong Kong (7), Canada (5), Singapore (3), Australia (2), South Korea (2), Barbados (1), Brazil (1), India (1), Israel (1) and Thailand (1).

For each of the studies we double checked the available information (more specifically the Supplementary Table S1 of Piwowar *et al.* (2007) and Table 5 of Ioannidis *et al.* (2007) and corrected any inconsistencies. We also extracted some additional study and journal characteristics that were not considered by Piwowar *et al.* (2007) and Ioannidis *et al.* (2007).

Table 9.1 gives the characteristics of the studies in each set and in the superset.

| | Our set | Piwowar *et al.* (2007) | Ioannidis *et al.* (2007) | Combined Superset |
|---|---|---|---|---|
| **Number of studies in set** | 63 | 85 | 24 | 146 |
| **Time period covered** | 6/1999 - 10/2006 | 8/1999 - 3/2003 | 6/2002 - 2/2003, 1/2006 - 6/2006 | 6/1999 - 10/2006 |
| **Median of study sample sizes (range)** | 55 (6 - 389) | 34 (10 - 389) | 61 (19 - 240) | 43 (6 - 389) |
| **Median of 2006 Journal** | 10.11 | 12.72 | 14.86 | 10.94 |
| **Impact Factor (range)** | (2.01 - 30.03) | (1.54 - 51.30) | (4.16 - 51.30) | (1.54 - 51.30) |
| **Articles with co-author(s) from USA** | 44 (69.8%) | 56 (65.9%) | 14 (58.3%) | 93 (63.7%) |
| **Articles with co-author(s) from Europe (including UK)** | 19 (30.2%) | 29 (34.1%) | 13 (54.2%) | 55 (37.6%) |
| **Articles with co-author(s) from Japan** | 7 (11.1%) | 14 (16.4%) | 4 (16.6%) | 22 (15.1%) |
| **Articles with co-author(s) with a commercial affiliation** | 16 (25.4%) | 18 (21.2%) | 5 (20.8%) | 33 (22.6%) |

Table 9.1.: Characteristics of the studies in the three sets.

Next, for our own set of studies we attempted to obtain the raw data for all the studies that did not report a valid location of FLEO in the main text or supplementary materials ($n = 47$). We proceeded as in Section 6.3.2.

Some authors responded, some of whom provided the FLEO data. Some authors stated that they had misplaced or lost the raw data and others declined access to raw data. We created an "email response" variable with the following four possible values, arranged in decreasing preference: positive (i.e. provided the FLEO data), lost FLEO, no response, denied access.

## 9.4. Statistical analysis

After presenting descriptive data on reporting and sharing patterns, we explored variables that might be associated with these patterns. We performed univariate logistic regressions for each of the seven variables that we assessed - number of months since January 1999, the size of the study, the 2006 Journal Impact Factor, whether one of the co-authors had a commercial affiliation, if co-author(s) was based in the USA, if co-author(s) was based in Europe, if co-author(s) was based in Japan. We consider any $p$-value of 0.05 or below as indicating a significant association.

After identifying the significantly associated variable, we present tabular or graphical summary of the results. All analysis and graphs were performed in the R-project software (R Development Core Team, 2004).

## 9.5. Results

### 9.5.1. Reporting and sharing of raw data

Figure 9.3 visually describes sequentially the proportion and type of data that was available directly from the articles, from searching the web and upon direct contact with the authors. For seven of the 47 studies in our own set, we obtained the data indirectly. We emailed the authors of the remaining 40 studies and only 30 (75%) responded. Of the responders, 14 provided the FLEO data, six had lost it and 10 denied access.

Four respondents were willing to share their data in return for co-authorship in manuscripts or a share in any commercial profit or intellectual property. This attitude puts at a disadvantage those who have already shared their raw data freely and thus we classified such responses as denials.

Table 9.2 describes the type of data reported in the article for each of the three sets separately and for the combined superset. While there are some differences in the results for each set, the underlying message is clear. The reporting of raw data in microarray studies is very poor, with the large majority of the articles making no attempt to provide any kind of usable data (i.e. IPD).

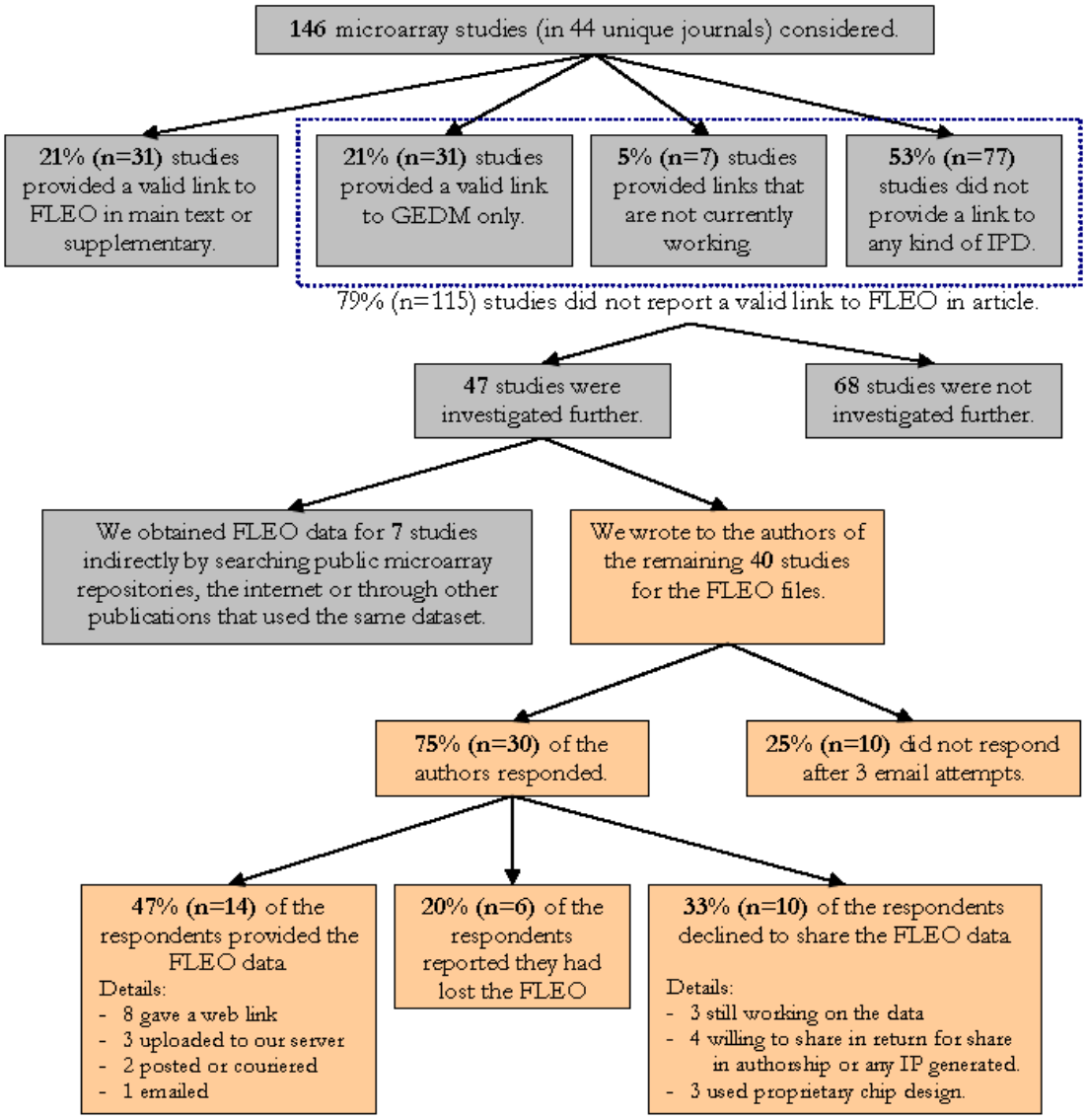



Figure 9.3.: Reporting and sharing of raw data in 146 microarray-based studies.

| | FLEO | GEDM only | Invalid link | No link to any IPD |
|---|---|---|---|---|
| **Our set** ($n = 63$) | 16 (25.4%) | 17 (27.0%) | 5 (7.9%) | 25 (39.7%) |
| **Piwowar *et al.* (2007)** ($n = 85$) | 14 (16.5%) | 14 (16.5%) | 3 (3.5%) | 54 (63.5%) |
| **Ioannidis *et al.* (2007)** ($n = 24$) | 8 (33.3%) | 8 (33.3%) | 0 (0%) | 8 (33.3%) |
| **Combined superset** ($n = 146$) | 31 (21.2%) | 31 (21.2%) | 7 (4.8%) | 77 (52.7%) |

Table 9.2.: The frequency and percentage of the "data reported" variable

Among the articles that provided a valid link to some kind of usable data ($n = 62$), exactly half in the superset provided only the GEDM but not the FLEO ($n = 31$). This finding

is in agreement with Larsson and Sandberg (2006) who found that, on average, only 48% of studies in GEO and ArrayExpress were submitted with FLEO. This suggests that some researchers could be incorrectly associating GEDM as the raw data, perhaps because it is often the starting point for many visualization and statistical analysis.

### 9.5.2. Variables associated with reporting of raw data

To identify variables associated with reporting of raw data, we omitted the seven studies that provided links that were not working. We used univariate logistic regression with the response being the indicator variable if location to FLEO data was reported in main text or supplementary of article ($n = 31$) or not ($n = 108$).

We found that the probability of reporting the location to the raw data increased with larger sample size ($p = 0.0004$) and rising impact factor ($p = 0.039$; see Figure 9.4) but not with time ($p = 0.11$); however, these three variables are inter-related.

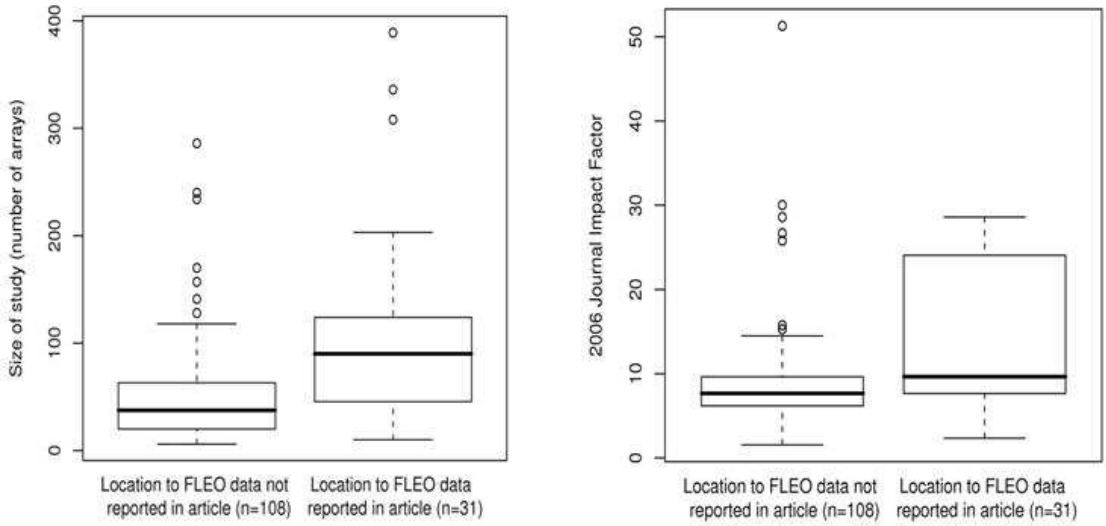



Figure 9.4.: Figure on the left shows the distribution of sample sizes for those who reported a valid link to FLEO data and those who did not. Similarly the figure on the right shows the distribution of the 2006 Journal Impact Factors for these two groups. This is a box-and-whisker plot, where the box represents the interquartile range (where 50% of the observations lie) and the whiskers extend 1.5 times the interquartile range beyond the box. The thick central line in each box represents the group medians.

### 9.5.3. Variables associated with sharing of raw data when its location is not reported in the article

For the analysis of data sharing, we again used univariate logistic regression with the response being the indicator variable if the respondent provided the raw data ($n = 14$) or declined access ($n = 10$). Only a small number of studies ($n = 24$) were involved in this analysis so the results should be taken as only suggestive.

We found that the probability of sharing the raw data, when it is not reported in the article, decreased if one of the co-authors had a commercial affiliation ($p = 0.006$) and with more recent studies ($p = 0.048$). Nine out of the ten articles for which access was denied were published after October 2004.

Table 9.3 summarizes the email responses by commercial affiliation while Figure 9.5 summarizes by date of publication. Please note that these show the data include all the category of email responses, even though we only used two categories (respondents who provided and respondents who declined FLEO data) to identify statistically significant associations.

| | **Provided the FLEO data** | **Lost the FLEO data** | **Did not respond** | **Denied Access** |
|---|---|---|---|---|
| **Articles without co-author(s) with commercial affiliation (n=28)** | 13 (46.4%) | 4 (14.3%) | 8 (28.6%) | 3 (10.7%) |
| **Articles with co-author(s) with commercial affiliation (n=12)** | 1 (8.3%) | 2 (16.7%) | 2 (16.7%) | 7 (58.3%) |

Table 9.3.: Summary of email responses for 40 articles that did not report a link to FLEO data in the main text or supplementary. The percentages are of the row total.

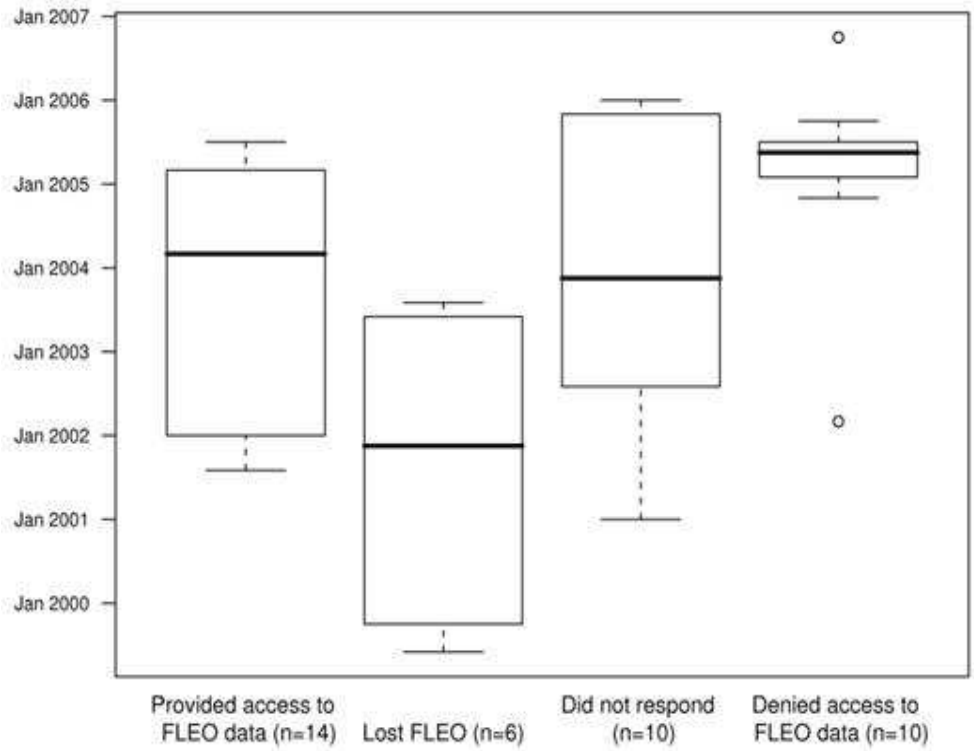



Figure 9.5.: Distribution of the date of publication by email responses for the 40 articles that did not report a link to FLEO data in the main text or supplementary.

## 9.6. Discussion

Our main findings can be summarized as follows

- Reporting of location to the raw data used in microarray-based articles was very poor. Less than a quarter of the articles in the combined superset provided a valid link to raw data. Further, more than half did not attempt to provide the location for any kind of data.
- The probability of reporting the location of the raw data improved with larger sample size studies and higher journal impact factors.
- If the location of raw data was not reported in the article, then the authors were less likely to share it when requested if one of the co-authors had a commercial affiliation.

These findings should serve as a useful warning for those working with other "omics" technologies as the number of studies and volume of data in these field increases. If such poor reporting of data is observed even for the best established high-throughput biological technology, then others should learn from this and try to prevent similar problems arising in related fields.

### 9.6.1. Limitations of our study

First, we studied the current availability of raw data which does not necessarily reflect the availability of the raw data around the time of publication of the study. The latter may be assessed by using the internet archives such as the WayBackMachine (http://www.archive.org/web/web.php) but we were only interested in current availability. We anticipate that over time some authors may have retrospectively made the raw data available either due to increasing awareness of publicly available raw data, faster internet bandwidth, and availability of public microarray repositories or requests from other researchers. Indeed, after our email request, at least one group uploaded the raw data to a website that previous only had GEDM data.

Second, we are measuring association not causation. For example, we have identified that better reporting is associated with rising journal impact factor. However, we are not able to say whether authors who report the location raw data prefer to publish in high impact journals or if high impact journals are better at enforcing the reporting of raw data.

Third, we are using the affiliation of co-authors reported in the article to determine commercial affiliation and where they are based. It is common for researchers to have multiple affiliations and the reported affiliation may not be the only one. We did not assess the conflict of interest information reported in the articles. Further, few researchers co-authored multiple papers in the combined superset with consistent attitudes towards making raw data publicly available. We did not take account of this.

Finally, we have not addressed the usability of the available raw data. Larsson and Sandberg (2006) found that, on average, only 77% of all available raw data in GEO and ArrayExpress had high RNA integrity and high hybridization sensitivity to allow for future integrative research. However, they did not investigate what proportion of the published studies had

submitted their data to GEO or ArrayExpress. Thus if we intersect their finding with our finding that, on average, only 21% of the studies have reported the location to the raw data in the article, we can expect less than a sixth (= 0.21 x 0.77) of all available hybridizations from all microarray studies will be available for future integrative research, without further accounting for the unknown number of studies that never lead to a published article.

## 9.7. Recommendations

The problem of poor reporting of raw data has not gone unnoticed by other researchers in this field. Following Ventura (2005); Piwowar *et al.* (2007); Ioannidis *et al.* (2007) we have also shown that reporting and sharing of raw data for microarray studies is very poor. Serious steps have to be taken now in order to prevent the complete loss of data in the future.

Array Express and GEO now offer a free service to ensure that the data submitted to them are MIAME-compliant (Brazma *et al.*, 2006; Edgar and Barrett, 2006). We strongly encourage journal editors to take advantage of this new service and to continue supporting the efforts of the MGED society.

Our recommendations for authors and journal editors are given below.

**Recommendations for authors**

(a) Previously published studies

- If you have previously published an article using microarray result but have not made the FLEO data available, please make them available in the public microarray repositories. Using a personal or institutional site for this purpose reduces the chances of the data being found.
- Inform the journal of the location of the FLEO data so the link can be tagged to the article.

(b) New studies

- Deposit the FLEO files in one of the recommended public microarray repositories Ball *et al.* (2004): ArrayExpress (Brazma *et al.*, 2003), GEO (Edgar *et al.*, 2002) and CIBEX (Ikeo *et al.*, 2003)
- Report the full URL to the FLEO data, preferably (also) in the abstract.
- Follow all other MIAME requirements. See the MIAME website for details http://www.mged.org/Workgroups/MIAME/miame.html
- Deposit the image files (if possible) or store them safely in a centralized location with additional backups if necessary. They may become useful in the future.

**Recommendations for journal editors**

- Require authors to follow the MIAME requirements.
- Clarify instructions to authors to include the following:
    1. both the FLEO and GEDM data must be deposited in accordance with the MIAME requirements
    2. the data must be deposited into either ArrayExpress, GEO or CIBEX, as recommended by MIAME (Ball *et al.*, 2004)
    3. the full URL to the location must be stated in the main text, and preferably also in the abstract. Check the URL is correct and working.
- Check the submission of both the FLEO and GEDM data into public repositories along with adherence to other MIAME criteria before final acceptance of article. Take advantage of the new services offered by Array Express (Brazma *et al.*, 2006) and GEO (Edgar and Barrett, 2006) if needed.
- Carry out regular MIAME compliance checks on published articles. Contact authors of previously published articles to deposit the FLEO and GEDM data if they have not done so already and electronically link to main text or supplementary of article.

## 9.8. Summary

The results across three different sets of studies were similar. Of the 146 articles in the superset, only 31 articles (21%) provided a valid link to the raw data in the main text or supplementary of the article. Another 38 articles (26%) either provided a valid link to summary data or gave an invalid link. Strikingly, 77 articles (53%) did not even attempt to provide a link to any kind data. In other words, we could not identify the raw data for 115 of the 146 articles considered (79%) from the main text or (if available) the supplementary of the published articles.

# 10. Summary, Discussion and Recommendations

*This thesis set out with the aim of helping researchers select a robust list of genes from microarray studies for further investigation. To achieve this objective we discuss two inter-related topics: sample size calculation and meta-analysis of microarray data. In this concluding chapter, we summarize the work presented so far, discuss the strengths and weakness of the thesis and recommend some future directions to resolve unanswered questions.*

The introductory chapter of this thesis provides the background for the readers. We discussed the history, growth and importance of microarray to molecular biology as well as some criticisms for it. We also gave a very brief background to cancer biology and microarray technology, including different data formats, platforms available and types of preprocessing algorithms.

The second chapter provided an introduction to the statistical concepts and common formulae for the sample size and meta-analysis chapters. It also introduced most of the notation used in subsequent chapters.

Chapter 9 discussed the issue of raw data availability, which affects both the sample size and meta-analysis chapters. In this chapter, we revealed evidence for poor raw data sharing despite requirements by journal editors and the impetus of the wider community. We also explored the factors that are statistically associated with data sharing patterns. Finally, we made some recommendations to help improve the sharing of raw data. The limitations of our research is addressed in the Discussion section of this chapter.

## 10.1. Summary of sample size chapters with discussion of strengths and future work

Chapter 3 introduced the most commonly used sample size formula in gene expression microarray studies, its shortcomings and potential amelioration. These include the provision for unequal allocation between two groups, why one should estimate the fold change and variability jointly and how to minimize the difference in approximating the $t$-distribution by a Gaussian distribution. We provided a step-by-step guide and a look-up table for sample size estimation.

Chapter 4 aimed to identify the optimal number of measurements per subject (i.e. technical replicates). We calculated this by minimizing the total cost, which can be divided into recruitment and measurement cost, while subject to a fixed power level using Lagrangian method. The optimal number of measurements was found to be inversely proportional to the square root of recruitment-to-measurement cost and technical-to-biological variation.

Chapter 5 demonstrated the calculation of sample sizes for the case study of MDS experiments. The R codes to calculate the sample size is given in Appendix A.4.

Future works on the sample size strand includes:

1. Calculating the sample size using methods that allow for the **correlation structure of the genes** (Pan *et al.*, 2002; Zien *et al.*, 2003) on the dataset presented here. It would be useful to see if these more sophisticated methods give similar answers to our simple approach here.

2. Similarly, to calculate the sample size using methods that use Bayesian methods or **loss functions** (Muller *et al.*, 2004) on the datasets presented here.

3. **Investigate if one technical replicate is optimal** under more generic situation. To do this, we need to first identify suitable empirical data to estimate the technical-to-biological variation.

## 10.2. Summary of meta-analysis chapters with discussion of strengths and future work

Chapter 6 identified seven key issues in conducting a meta-analysis of gene expression microarray datasets and we discussed in detail the first five key issues which are related to data identification, acquisition and curation. We argued that combining independent studies not only increases statistical power through increased sample size but also allows us to assess generalizability of findings. We also provided a step-by-step checklist for performing meta-analysis in Table 6.1.

Chapter 7 addressed the choice of meta-analysis technique. We briefly introduced the available techniques that can be loosely classified into one of four categories. We then presented a series of qualitative questions that can help researchers select a meta-analysis technique. Then we attempted to quantify the performance of each of the techniques. Finally, we chose the inverse-variance technique as the most comprehensive method and presented the details of this technique.

Chapter 8 demonstrated the inverse-variance technique by applying it to a case study of protein folding genes. Here we identified, acquired and curated 30 different tumour-normal sets containing over 1700 arrays to test the hypothesis that there are some protein folding genes which are truly down-regulated in many cancer tissues. We also highlighted a few logistical issues which were integral to this study. The results show that Crystallin, Alpha B (Hs.408767) is significantly down-regulated in cancer studies and this is a novel biological finding.

Future works on the meta-analyses strand includes:

1. Develop and test a way to **meta-analyze on pathway level**. This concept is attractive for two related reasons. First, we may be able to boost the signal from genes with modest difference. Second, depending on cancer type and location, we may hypothesize that only a subset of genes from the same biological pathway may be activated to perform identical role, thus meta-analyzing at gene-level would miss the contribution of pathways. We have already seen in Section 8.3.2 that running a gene set enrichment analyses (GSEA) on the gene-level results reveals well known biological pathways. A rather simplistic approach to try would be to run a GSEA on each study individually and store the resulting annotation clusters and enrichment $p$-value; then proceed to merge on the annotation clusters and meta-analyze the enrichment $p$-values across studies.

2. **Cluster the datasets on sequence-level data**. This would involve first collecting the sequence data for the probes considered here and then creating a pipeline for clustering them in an efficient manner. There is evidence that sequence-matched datasets increases cross-platform concordance (Morris *et al.*, 2005). Furthermore, having the sequence data extends the usefulness and shelf-life of these datasets when new and interesting questions arise (e.g. merging at exon-level rather than gene-level could address the alternative splicing problem).

3. Investigate if **adjusting for clinical variables** changes the conclusion. Again, the first step would be to collect the clinical variables from various researchers and we anticipate this to be a challenging task. Adjusting for covariates can address imbalance in baseline covariates between the two groups and also reduce variability in effect estimates of the gene.

4. Develop a better way to **access compatibility between datasets that are pre-processed using different algorithms**.

5. Develop a better way to **benchmark the various meta-analyses techniques**. Findings from this work could also have implications for single marker studies, clinical trials and epidemiological studies.

6. Assess the problem of **publication bias**. Prof. John Ioannidis believes that data from only 10% of all the Affymetrix chips are published. This is an intriguing gap given the high cost of purchasing the chips (and associated equipment and reagents) and the low failure rate of these chips.

7. Try to increase awareness of the importance of depositing the **raw data in the public repository**.

8. Extend the area of application to other fields. Meta-analyses technique is already being used in genome-wide association studies where large scale collaboration is often the norm.

## 10.3. Noteworthy contribution of this thesis to the field

1. Showed that joint estimation of fold change and variability is important to allow for consistent sample size calculations independent of pre-processing algorithm (Chapter 3.4.2).
2. Developed a guide to sample size calculations and parameter estimation (Chapter 3.5).
3. Demonstrated that one technical replicate is probably sufficient for Affymetrix data (Chapter 4 and 5).
4. Identified seven key issues in meta-analysis of microarray datasets (Chapter 6-8).
5. Produced a checklist for conducting a meta-analysis of microarray datasets (Table 6.1).
6. Made first attempts to formally compare different meta-analysis techniques for microarray datasets (Chapter 7).
7. Contributed the R package `metaGEM` to the community to perform meta-analysis of microarray datasets (see Appendix A.5 or http://hdl.handle.net/10044/1/4217).

# A. Appendix

## A.1. Derivation of Equation 3.3

Equation 3.3 is derived by combining the following sources of information:

- The standard sample size assuming equal allocation ratio and Normal approximation of the non-central $t$-distribution (also presented as Equation 3.2 here)

$$n_A + n_B = 4 \left( \frac{z_{\alpha/2} + z_\beta}{\delta/\sigma} \right)^2$$

- The common rule of thumb that says a design with allocation ratio $k$ requires $\frac{(k+1)^2}{4k}$ times the sample size of a design with equal allocation ratio e.g. (e.g. see Altman, 1991, pages 459 - 460). Combining with the point above, we can rewrite it as

$$n_A + n_B = \frac{(k+1)^2}{k} \left( \frac{z_{\alpha/2} + z_\beta}{\Delta} \right)^2$$

- Guenther (1981) who showed that using the Normal distribution to approximate the non-central $t$-distribution leads to a quantifiable and constant difference in the sample size. We rewrite Equation 3.1. from Guenther (1981) for a two-sided hypothesis testing problem as

$$n_A = 2 \left( \frac{z_{\alpha/2} + z_\beta}{\Delta} \right)^2 + \frac{z_{\alpha/2}^2}{4}$$

However, there is some ambiguity as what the correction factor should be when we calculate the total sample size $n_A + n_B$. It could either be a constant or vary according to the allocation ratio $k$. I used the following R codes to calculate the error in sample sizes when approximating the non-central t-test based formula with the z-test based formula.

```
find.power <- function(Delta, sig.level, n1, n2, k=1){

  if( missing(n2) ){
    if( k < 1 ) stop("k is defined as k >= 1")
    n2 <- k*n1
  }

  nu  <- n1 + n2 - 2
  qu  <- qt(sig.level/2, nu, lower = FALSE)
  ncp <- sqrt( (n1 * n2) / (n1 + n2) ) * Delta
  pow <- pt(qu, nu, ncp=ncp, lower = FALSE) + pt(-qu, nu, ncp=ncp, lower = TRUE)
  return(pow)
}

approximation.error <- function(sig.level, power, Delta, k){

    ## Find the exact solution using recursive non-central t-test
    fn       <- function(n1)
                    find.power(n1=n1, Delta=Delta, sig.level=sig.level, k=k) - power
    nA.exact <- uniroot( fn, lower=2, upper=1000000 )$root
    n.exact  <- (1+k)*nA.exact

    ## Use normal theory approximation without correction
    num <- ( qnorm( 1-sig.level/2 ) + qnorm( power ) )^2 * (k+1)/k
    nA.approx <- num/Delta^2
    n.approx  <- (1+k)*nA.approx

    return( n.exact - n.approx )  # the underestimation
}

alphas <- c(0.05, 0.01, 0.001, 0.0001, 0.00001)

errs  <- matrix( nr=5, nc=length(alphas) )
rownames(errs) <- paste("alpha=", alphas, sep="")
colnames(errs) <- paste("k=", c(1,2,5,10,100), sep="")

for( i in 1:5 ){
 kkk <- c(1,2,5,10,100)[i]

 for(j in 1:length(alphas)){
   errs[j, i] <- approximation.error( k=kkk, sig.level = alphas[j],
                                      power = 0.80, Delta = 1.0 )
 }
}

round( cbind( errs, NA,
              CorrectionFactor=0.5*qnorm(1 - alphas/2)^2 ), digits=2 )
```

which gives the following values:

| | Errors in approximation | | | | | |
|---|---|---|---|---|---|---|
| $\alpha$ | $k=1$ | $k=2$ | $k=5$ | $k=10$ | $k=100$ | $0.5z^2_{\alpha/2}$ |
| 0.05 | 2.03 | 2.02 | 1.98 | 1.96 | 1.93 | 1.92 |
| 0.01 | 3.42 | 3.41 | 3.37 | 3.35 | 3.32 | 3.32 |
| 0.001 | 5.48 | 5.47 | 5.45 | 5.44 | 5.42 | 5.41 |
| 0.0001 | 7.58 | 7.58 | 7.58 | 7.58 | 7.57 | 7.57 |
| 0.00001 | 9.72 | 9.72 | 9.73 | 9.74 | 9.75 | 9.76 |

The last column shows that the approximation errors do not vary much by the allocation ratio and is very close to the $0.5z^2_{\alpha/2}$ value (last column). Thus, we obtain the Equation 3.3.

## A.2. Proof for points in Section 3.5.5: choosing the allocation ratio, $k$

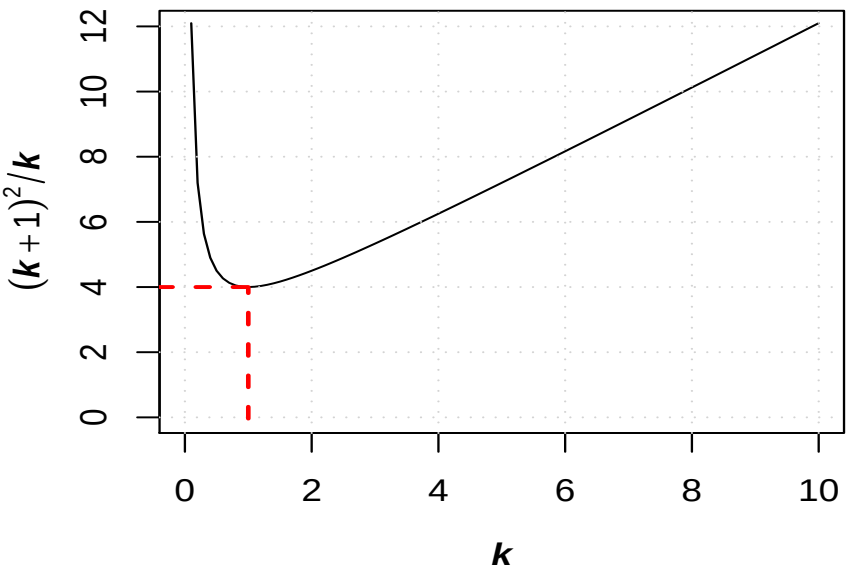



Figure A.1.: Plot of the multiplier factor in Equation 3.3 due to the allocation ratio ($k$) vs. the allocation ratio ($k$).

From Equation 3.3 and if we let $n = n_A + n_B$

$$\begin{aligned} n &= \frac{(k+1)^2}{k}\left(\frac{z_{\alpha/2}+z_\beta}{\Delta}\right)^2 + \frac{z_{\alpha/2}^2}{2} \\ n &= \left(k+2+\frac{1}{k}\right)\left(\frac{z_{\alpha/2}+z_\beta}{\Delta}\right)^2 + \frac{z_{\alpha/2}^2}{2} \\ \frac{\mathrm{d}n}{\mathrm{d}k} &= \left(1-\frac{1}{k^2}\right)\left(\frac{z_{\alpha/2}+z_\beta}{\Delta}\right)^2 \\ \frac{\mathrm{d}^2n}{\mathrm{d}k^2} &= \frac{2}{k^3}\left(\frac{z_{\alpha/2}+z_\beta}{\Delta}\right)^2 \end{aligned}$$

Setting $\frac{\mathrm{d}n}{\mathrm{d}k} = 0$ gives $k = 1$ (since $k > 0$ by definition) and evaluating the second derivative show that this value of $k$ minimizes $n$.

## A.3. Derivation of $m_{opt}$ (Equation 4.3)

First let us rewrite Equation 4.1 with $n = n_A + n_B$ and $\psi = \frac{z_{\alpha/2}^2}{2}$

$$n = \frac{(k+1)^2}{k}\left(\frac{z_{\alpha/2}+z_\beta}{\Delta_m}\right)^2 + \psi$$

If we rewrite the above using $\Delta_m = \frac{\delta}{\sigma_m}$ and $\sigma_m^2 = \sigma_B^2 + \frac{\sigma_E^2}{m}$, we obtain

$$\begin{aligned} n - \psi &= \frac{(k+1)^2}{k}\left(\frac{z_{\alpha/2}+z_\beta}{\delta}\right)^2\left(\sigma_B^2 + \frac{\sigma_E^2}{m}\right) \\ \frac{1}{n-\psi}\left(\sigma_B^2 + \frac{\sigma_E^2}{m}\right) &= \frac{k}{(k+1)^2}\left(\frac{\delta}{z_{\alpha/2}+z_\beta}\right)^2 \end{aligned}$$

Since $k, \alpha, \beta, \delta$ are predetermined values, the right hand side of the equation above is a constant and we set this to $A$.

$$\frac{1}{n-\psi}\left(\sigma_B^2 + \frac{\sigma_E^2}{m}\right) = A \tag{A.1}$$

Recall that the total cost given Equation 4.2 can be rewritten as $n(C_1 + mC_2)$. The Lagragian

operator to minimize this cost subject to Equation A.1 is given by:

$$\begin{aligned} L(n,m,\lambda) &= n(C_1+mC_2)+\lambda\left[\frac{1}{n-\psi}\left(\sigma_B^2+\frac{\sigma_E^2}{m}\right)-A\right] \\ \frac{\partial L}{\partial n} &= C_1+mC_2-\frac{\lambda}{(n-\psi)^2}\left(\sigma_B^2+\frac{\sigma_E^2}{m}\right) \\ \frac{\partial L}{\partial m} &= nC_2-\frac{\lambda}{n-\psi}\frac{\sigma_E^2}{m} \end{aligned}$$

$$\frac{\partial L}{\partial n}=0 \Longrightarrow \frac{\lambda}{n-\psi} = \frac{(n-\psi)(C_1+mC_2)}{\left(\sigma_B^2+\frac{\sigma_E^2}{m}\right)} \tag{A.2}$$

$$\frac{\partial L}{\partial m}=0 \Longrightarrow \frac{\lambda}{n-\psi} = \frac{nmC_2}{\frac{\sigma_E^2}{m}} \tag{A.3}$$

Equating Equation A.2 to Equation A.3 gives

$$m^2+m\frac{\psi}{n}R-\left(1-\frac{\psi}{n}\right)\frac{C_1}{C_2}R=0 \tag{A.4}$$

where $R=\sigma_E^2/\sigma_B^2$.

Note that Equation A.4 is quadratic in $m$. The quadratic formula states that the solution to $ax^2+bx+c=0$ is given by $x=\frac{-b\pm\sqrt{b^2-4ac}}{2a}$. Applying the quadratic formula in our case gives

$$m=\frac{-\psi R}{2n}\pm\sqrt{\left(\frac{\psi R}{2n}\right)^2-\left(1-\frac{\psi}{n}\right)\frac{C_1}{C_2}R}$$

Since $\psi>0, R>0, n>0$, then in order for m $>0$, we choose the upper root of the solution.

Further if we let $f=\frac{\psi}{2n}=\frac{\frac{z_{\alpha/2}^2}{2}}{2(n_A+n_B)}=\frac{0.25\times z_{\alpha/2}^2}{n_A+n_B}$, we get Equation 4.3:

$$m_{opt}=-fR+\sqrt{f^2R^2+(1-2f)\frac{C_1}{C_2}R}$$

## A.4. R codes for sample size calculations

```
find.power <- function(Delta, sig.level, n1, n2, k=1){

  if( missing(n2) ){
    if( k < 1 ) stop("k is defined as k >= 1")
    n2 <- k*n1
  }

  nu  <- n1 + n2 - 2
  qu  <- qt(sig.level/2, nu, lower = FALSE)
```

```
  ncp <- sqrt( (n1 * n2) / (n1 + n2) ) * Delta
  pow <- pt(qu, nu, ncp=ncp, lower = FALSE) + pt(-qu, nu, ncp=ncp, lower = TRUE)
  return(pow)
}

find.ss.internal <- function(sig.level, power, Delta, k, exact=TRUE){

  if(exact){ ## Find the exact solution using recursive t-test
    fn  <- function(n1)
       find.power(n1=n1, Delta=Delta, sig.level=sig.level, k=k) - power
    nA  <- uniroot( fn, lower=2, upper=1000000 )$root

  } else {   ## Use normal theory approximation to above

    mult <- ( (k+1)^2 / k )
    num  <- mult * ( qnorm( 1-sig.level/2 ) + qnorm( power ) )^2
    cf   <- 0.5*qnorm(1 - sig.level/2)^2
    nA.plus.nB <- num/Delta^2 + cf
    nA  <- nA.plus.nB/(k+1)
  }
  return( c( nA=nA, nB=k*nA) )
}

find.ss <- function(Delta, sig.level, power, k=1, digits=1, exact=TRUE){
  ## Returns total sample size with allocation ratio of 1:k (k>1)

  names(sig.level) <- paste("alpha=", sig.level, sep="")
  names(power)     <- paste("power=", power, sep="")
  names(Delta)     <- paste("Delta=", Delta, sep="")

  wrapper <- function(x, y, my.fun, ...) {
    sapply(seq(along = x), FUN = function(i)
           sum( find.ss.internal(sig.level=x[i], power=y[i], ...) ) )
  }

  out <- lapply( Delta, function(x)
                 outer(sig.level, power, FUN=wrapper, Delta=x, k=k, exact=exact) )
  out <- lapply(out, round, digits=digits)
  return(out)
}
```

## A.5. R codes for meta-analysis of microarray datasets

These codes are now available as an R package. To download it and to read the Vignette, please go to http://spiral.imperial.ac.uk/handle/10044/1/4217

### A.5.1. Some generic R codes

```
library(rmeta)
library(multtest)

se <- function(x) sd(x)/length(x)

cleanNA <- function(x) return( x[!is.na(x) & is.finite(x) ] )

gsub.formula <- function(pattern, replacement, x, ...)
                  as.formula( gsub(pattern, replacement, as.expression(x), ...) )

install.packages.BioC <- function(x, ...){
   if( !exists("getBioC") ) source("http://www.bioconductor.org/getBioC.R")
   getBioC(x, ...)
}

blow <- function(x, scale=0.10)
             c(mini = ( min(x) - scale*abs(min(x)) ),
               maxi = ( max(x) + scale*abs(max(x)) ) )
               ## superseeded by extendrange()

expand.df <- function(df, key.name="keys", keys.sep = ","){

   ## every row in df may have multiple identities
   ## these identities are defined in keys, with individual keys separated by keys.sep
   ## this function expands every row that has 1:n mapping to n x 1:1 rows

   key.name <- as.character(key.name)
    skey <- strsplit(df[ , key.name], split = keys.sep)
    df   <- df[rep(1:nrow(df), sapply(skey, length)), ]
    df[ , key.name] <- unlist(skey)
    return(df)
}

multimerge <- function(mylist){
    ## iterative version of merge using all=T option
    ## it assumes rownames is unique and is the common key

    unames <- unique( unlist( lapply( mylist, rownames ) ) )
    n      <- length(unames)

    out <- lapply( mylist, function(df){
        tmp <- matrix( nr=n, nc=ncol(df),
                       dimnames=list( unames, colnames(df) ) )
        tmp[ rownames(df), ] <- as.matrix(df)
        return(tmp)
    })
```

```
    bigout <- do.call( cbind, out )
    colnames(bigout) <- paste(rep( names(mylist), sapply(mylist, ncol) ),
                              sapply(mylist, colnames), sep="_")
    return(bigout)
}

pairwise.apply <- function(x, FUN, ...){
   ## output of FUN must be scalar

   n <- nrow(x)
   r <- rownames(x)
   output <- matrix(NA, nc=n, nr=n, dimnames=list(r, r))


   for(i in 1:n){
       for(j in 1:n){
           if(i >= j) next()
           output[i, j] <- FUN( x[i,], x[j,], ... )
       }
   }
   return(output)
}
```

## A.5.2. R codes to combine effect sizes

```
##########################################
### Calculate and combine within study ###
##########################################

getstats <- function(v, g){
    ## function to calculate basic statistics for two-class comparison for a gene

    stopifnot( identical( length(v), length(g) ) )

    x <- cleanNA( v[ which(g==1) ] )
    y <- cleanNA( v[ which(g==0) ] )

    n1 <- length(x); n2 <- length(y)
    if( n1 < 2 | n2 < 2 )
        return( c(n1=NA, m1=NA, sd1=NA, n2=NA, m2=NA, sd2=NA,
                  diff=NA, pooled.sd=NA, g=NA, se.g=NA) )

    m1   <- mean(x); m2 <- mean(y)
    diff <- m1 - m2

    sd1  <- sd(x);  sd2 <- sd(y)
```

```
    sp   <- sqrt( ( (n1-1)*sd1^2 + (n2-1)*sd2^2 )/( n1 + n2 - 2 ) )

    cf   <- 1 - 3/( 4*(n1 + n2) - 9 )
    g    <- cf * diff/sp
    se.g <- sqrt( (n1+n2)/(n1*n2) + 0.5*g^2 /(n1+n2-3.94) )

    return( c(n1=n1, m1=m1, sd1=sd1, n2=n2, m2=m2, sd2=sd2,
              diff=diff, pooled.sd=sp, g=g, se.g=se.g) )

}

getstats2 <- function(Tmat, Nmat){
    mat <- cbind(Tmat, Nmat)
    grp <- rep( 1:0, c(ncol(Tmat), ncol(Nmat)) )
    out <- t( apply(mat, 1, getstats, g=grp) )
    return(out)
}

summ.eff.within.study <- function(effects, option="fixed.iv"){
    ## summarizing multiple probes that map to the same UniGene within a study
    stopifnot( all( c("g", "se.g", "keys") %in% colnames(effects) ) )
    effects <- effects[ , c("g", "se.g", "keys")]
    effects$keys  <- as.character(effects$keys)

    if( length( grep(",", effects$keys) ) > 0 )
        stop("Multiple keys detected. Please expand geneID first")

    if(nrow(effects)==1){ ## hack when only effect size is available per study
        rownames(effects) <- effects$keys
        effects <- effects[ , c("g", "se.g")]
        return(effects)
    }

    ## Deal with singletons first ##
    singles.keys <- names(which(table(effects$keys) == 1))
    singles.ind <- which(effects$keys %in% singles.keys)
    out <- effects[ singles.ind, ]
    rownames(out) <- out$keys;  out$keys <- NULL

    ## Next, deal with multiple keys within a study ##
    multis <- effects[ -singles.ind, ]
    multis$abs.z <- abs( multis$g/multis$se.g ) # used if extreme option is choosen

    if(nrow(multis) > 0){  # skip if no multiple ID found

        tmp <- split(multis, multis$keys)

        if (option == "fixed.iv") {
```

```
            out2 <- sapply(tmp, function(m) {
              x <- unlist(meta.summaries(m$g, m$se.g, method = "fixed")
              x[c("summary", "se.summary")])
          })
          out2 <- t(out2)
          colnames(out2) <- c("g", "se.g")
        }

        if (option == "extreme") {
            out2 <- lapply(tmp, function(mat) mat[which.max(mat$abs.z), ])
            out2 <- do.call(rbind, out2)
            out2 <- out2[, c("g", "se.g")]
         }
         out <- rbind(out, out2)
   }

    out <- out[sort(rownames(out)), ]
    return(out)
}


##############################
### Combine across studies ###
##############################

pool.inverseVar <- function( g, se.g, method ){

   stopifnot( identical( rownames(g), rownames(se.g) ) )
   out <- matrix( nr=nrow(g), nc=5,
                  dimnames=list( rownames(g), c("n.studies", "summary", "se.summary",
                                                "tau2", "p.value") ) )

   for(j in 1:nrow(g)){
       e  <- cleanNA(    g[j, ] )
       se <- cleanNA( se.g[j, ] )
       n  <- length(e)

       if(n==1){
           summ <- e;   se.summ <- se;   tau2 <- NA
       } else {
           fit <- meta.summaries(e, se, method = method)
           summ <- fit$summary
           se.summ <- fit$se.summary
           tau2 <- ifelse( method=="fixed", NA, fit$tau2 )
           rm(fit)
       }

       pval     <- 2*pnorm( abs(summ/se.summ), lower.tail=FALSE )
```

```
        out[j, ] <- c(n, summ, se.summ, tau2, pval)
        rm(e, se, n, summ, se.summ, tau2, pval)
    }
    return(out)
}


combine.effect.sizes <- function (list.of.effects, between.method="random",
                                  within.method="fixed.iv", everything=TRUE){

    if( is.null(names(list.of.effects)) )
        names(list.of.effects) <- paste("data", 1:length(list.of.effects), sep="")

    study.effects <- lapply(list.of.effects, function(effects) {

        effects <- data.frame(effects)
        effects$keys <- as.character(effects$keys)

    ## remove probes that cannot be mapped or have insufficient info
    ## observations to calculate effect size
        bad <- which( is.na(effects$g) | is.na(effects$keys) | effects$keys=="NA" )
        effects <- effects[ setdiff(1:nrow(effects), bad), ]

        ## expand probes that maps to multiple keys
        effects <- expand.df( effects )

        ## summarize multiple probes within a study
        effects <- summ.eff.within.study(effects, option = within.method)
    })


    tmp <- multimerge(study.effects)
    g    <- tmp[, paste(names(study.effects), "_g", sep = ""), drop=FALSE]
    se.g <- tmp[, paste(names(study.effects), "_se.g", sep = ""), drop=FALSE]

    pooled.estimates <- data.frame( pool.inverseVar(g, se.g, method=between.method ) )

    if (everything) {
        return(list(g=g, se.g=se.g, pooled.estimates=pooled.estimates))
    } else {
        return(pooled.estimates)
    }
}



####################
```

```
### Visualization ###
#####################

plot.sumsum <- function (db, cex = 0.5, K=NULL, signif.thr=1, labels=NULL, ...) {

    db <- db$pooled.estimates
    p.adj <- p.adjust(db[, "p.value"], method = "fdr")
    summ <- db[, "summary"]
    se.summ <- db[, "se.summary"]
    if( is.null(labels) ) labels <- rownames(db)
    LCL <- summ - qnorm(0.975) * se.summ
    UCL <- summ + qnorm(0.975) * se.summ
    yrange <- blow(c(LCL, UCL))

    plot(yrange, range(p.adj), type = "n", log = "y", ann = F )
    title(ylab = list("FDR adjusted p-value", font = 4, col = 4),
          xlab = list("Pooled Standardized Mean Difference (log2 scale)",
          font = 4, col = 4), line = 2)
    grid()
    arrows(LCL, p.adj, UCL, p.adj, code = 3, length = 0.05, angle = 45, col = 8)
    points(summ, p.adj, pch = 20, cex = 0.75, col = 1)
    abline(v = 0, col = 4, lwd = 1)

    if(is.null(K)){
        neg <- which(p.adj < signif.thr & summ < 0)
        pos <- which(p.adj < signif.thr & summ > 0)
        abline(h = signif.thr, col = 4, lwd = 2, lty = 2)
    } else {
        s <- split( p.adj, sign(summ) )
        r <- sapply( s, rank )
        rr <- unsplit( r, sign(summ) )

        pos <- which( summ > 0 & rr < K )
        neg <- which( summ < 0 & rr < K )
    }

    if (length(neg) > 0)
        text(LCL[neg] - 0.125, p.adj[neg], labels = labels[neg],
             col = "green", cex = cex, font = 4)
    if (length(pos) > 0)
        text(UCL[pos] + 0.125, p.adj[pos], labels = labels[pos],
             col = "red", cex = cex, font = 4)
}

forest.plot <- function(db, key, main=key, ...){
    stopifnot(identical(rownames(db$g), rownames(db$se.g)))
    stopifnot(identical(rownames(db$g), rownames(db$pooled.estimates)))
```

```
    g    <- cleanNA( db$g[ key, ] )
    g    <- g[ sort(names(g)) ]

    se.g <- cleanNA( db$se.g[ key, ] )
    se.g <- se.g[ sort(names(se.g)) ]

    pool    <- db$pooled.estimates[ key, "summary" ]
    se.pool <- db$pooled.estimates[ key, "se.summary" ]

    metaplot( g, se.g, labels=names(g),
             summn=pool, sumse=se.pool, sumnn=1/se.pool^2,
             xlab="Standardized Mean Difference (log2 scale)", ylab="", main=main,
             colors=meta.colors(box="blue", lines="lightblue",
             zero="grey", summary="orange", text="red"), ... )
}
```

## A.6. Major milestones in microarray data reporting requirements and public repositories.

| Date | Event | Reference | Cumulative number of articles |
|---|---|---|---|
| **Oct 1995** | First seminal publication using microarray technology | Schena *et al.* (1995) | 1 |
| **Nov 1999** | Microarray Gene Expression Data (MGED) Society founded | | 75 |
| **Jul 2000** | Gene Expression Omnibus (GEO) starts accepting submission from the public. | Edgar *et al.* (2002) | 255 |
| **Dec 2001** | MGED publishes the MIAME requirements. | Brazma *et al.* (2001) | 1555 |
| **Feb 2002** | ArrayExpress starts accepting public submissions from the public (capable of FLEO data). | Brazma *et al.* (2003) | 1944 |
| **Oct 2002** | MIAME checklist published in response to user requests. | Ball *et al.* (2002) | 3153 |
| **Oct 2002** | Major scientific journals (including Cell, Lancet, Nature and Science) endorse MIAME requirements. | See journal's editorials | |
| **Oct 2003** | CIBEX starts accepting submissions from the public (capable of FLEO data). | Ikeo *et al.* (2003) | 5885 |
| **Apr 2004** | GEO starts accepting FLEO data files. | Barrett *et al.* (2005) | 7724 |
| **Sep 2004** | MGED requires mandatory submission of microarray data into a MIAME-compliant public repository.<br>MGED recognises ArrayExpress, GEO and CIBEX as MIAME-compliant repositories. | Ball *et al.* (2004) | 9238 |
| **Nov 2006** | ArrayExpress offers new service for reviewers/editors to check a submission for compliance to critical elements of MIAME | Brazma *et al.* (2006) | 18429 |
| **Dec 2006** | GEO offers new service for reviewers/editors to check a submission for compliance to critical elements of MIAME | Edgar and Barrett (2006) | 18748 |

# Bibliography


Affymetrix (2001) Statistical algorithm reference guide. Tech. rep., Affymetrix.

Aggarwal, A., Guo, D. L., Hoshida, Y., Yuen, S. T., Chu, K.-M., So, S., Boussioutas, A., Chen, X., Bowtell, D., Aburatani, H., Leung, S. Y. and Tan, P. (2006) Topological and functional discovery in a gene coexpression meta-network of gastric cancer. *Cancer Res*, **66**, 232–241. URL `http://dx.doi.org/10.1158/0008-5472.CAN-05-2232`.

Aldred, M. A., Morrison, C., Gimm, O., Hoang-Vu, C., Krause, U., Dralle, H., Jhiang, S. and Eng, C. (2003) Peroxisome proliferator-activated receptor gamma is frequently downregulated in a diversity of sporadic nonmedullary thyroid carcinomas. *Oncogene*, **22**, 3412–3416. URL `http://dx.doi.org/10.1038/sj.onc.1206400`.

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Marti, G. E., Moore, T., Hudson, J., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, J. O., Warnke, R., Levy, R., Wilson, W., Grever, M. R., Byrd, J. C., Botstein, D., Brown, P. O. and Staudt, L. M. (2000) Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503–511. URL `http://dx.doi.org/10.1038/35000501`.

Altman, D. G. (1991) *Practical Statistics for Medical Research*. Chapman and Hall.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–10.

Ball, C., Brazma, A., Causton, H., Chervitz, S., Edgar, R., Hingamp, P., Matese, J., Parkinson, H., Quackenbush, J., Ringwald, M., Sansone, S., Sherlock, G., Spellman, P., Stoeckert, C., Tateno, Y., Taylor, R., White, J. and Winegarden, N. (2004) Submission of microarray data to public repositories. *PLoS Biology*, **2**, e317.

Ball, C. A., Sherlock, G., Parkinson, H., Rocca-Sera, P., Brooksbank, C., Causton, H. C., Cavalieri, D., Gaasterland, T., Hingamp, P., Holstege, F., Ringwald, M., Spellman, P., Stoeckert, C. J., Stewart, J. E., Taylor, R., Brazma, A., Quackenbush, J. and Society, M. G. E. D. M. (2002) Standards for microarray data. *Science*, **298**, 539.

Bammler, T., Beyer, R. P., Bhattacharya, S., Boorman, G. A., Boyles, A., Bradford, B. U., Bumgarner, R. E., Bushel, P. R., Chaturvedi, K., Choi, D., Cunningham, M. L., Deng, S., Dressman, H. K., Fannin, R. D., Farin, F. M., Freedman, J. H., Fry, R. C., Harper, A., Humble, M. C., Hurban, P., Kavanagh, T. J., Kaufmann, W. K., Kerr, K. F., Jing, L., Lapidus, J. A., Lasarev, M. R., Li, J., Li, Y.-J., Lobenhofer, E. K., Lu, X., Malek, R. L., Milton, S., Nagalla, S. R., O'malley, J. P., Palmer, V. S., Pattee, P., Paules, R. S., Perou, C. M., Phillips, K., Qin, L.-X., Qiu, Y., Quigley, S. D., Rodland, M., Rusyn, I., Samson, L. D., Schwartz, D. A., Shi, Y., Shin, J.-L., Sieber, S. O., Slifer, S., Speer, M. C., Spencer, P. S., Sproles, D. I., Swenberg, J. A., Suk, W. A., Sullivan, R. C., Tian, R., Tennant, R. W., Todd, S. A., Tucker, C. J., Houten, B. V., Weis, B. K., Xuan, S., Zarbl, H. and of the Toxicogenomics Research Consortium, M. (2005) Standardizing global gene expression analysis between laboratories and across platforms. *Nat Methods*, **2**, 351–356.

Barrett, T., Suzek, T. O., Troup, D. B., Wilhite, S. E., Ngau, W.-C., Ledoux, P., Rudnev, D., Lash, A. E., Fujibuchi, W. and Edgar, R. (2005) Ncbi geo: mining millions of expression profiles–database and tools. *Nucleic Acids Res*, **33**, D562–D566. URL `http://dx.doi.org/10.1093/nar/gki022`.

Basso, K., Margolin, A., Stolovitzky, G., Klein, U., Dalla-Favera, R. and Califano, A. (2005) Reverse engineering of regulatory networks in human B cells. *Nat. Genet.*, **37**, 382–90.

Beer, D. G., Kardia, S. L. R., Huang, C.-C., Giordano, T. J., Levin, A. M., Misek, D. E., Lin, L., Chen, G., Gharib, T. G., Thomas, D. G., Lizyness, M. L., Kuick, R., Hayasaka, S., Taylor, J. M. G., Iannettoni, M. D., Orringer, M. B. and Hanash, S. (2002) Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat Med*, **8**, 816–824. URL `http://dx.doi.org/10.1038/nm733`.

Bellman, R. (1961) *Adaptive Control Processes: A guided tour*. Princeton, New Jersey: Princeton University Press.

Benito, M., Parker, J., Du, Q., Wu, J., Xiang, D., Perou, C. M. and Marron, J. S. (2004)

Adjustment of systematic microarray data biases. *Bioinformatics*, **20**, 105–114.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, **57**, 289–300.

Bennett, J., Catovsky, D., Daniel, M., Flandrin, G., Galton, D., Gralnick, H. and Sultan, C. (1976) Proposals for the classification of the acute leukaemias. french-american-british (fab) co-operative group. *Br J Haematol.*, **33**, 451–8.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. and Wheeler, D. L. (2008) Genbank. *Nucleic Acids Res*, **36**, D25–D30. URL `http://www.ncbi.nlm.nih.gov/`.

Bhattacharjee, A., Richards, W. G., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M., Loda, M., Weber, G., Mark, E. J., Lander, E. S., Wong, W., Johnson, B. E., Golub, T. R., Sugarbaker, D. J. and Meyerson, M. (2001) Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc Natl Acad Sci U S A*, **98**, 13790–13795. URL `http://dx.doi.org/10.1073/pnas.191502998`.

Bianchi, F., Nuciforo, P., Vecchi, M., Bernard, L., Tizzoni, L., Marchetti, A., Buttitta, F., Felicioni, L., Nicassio, F. and Fiore, P. P. D. (2007) Survival prediction of stage i lung adenocarcinomas by expression of 10 genes. *J Clin Invest*, **117**, 3436–3444. URL `http://dx.doi.org/10.1172/JCI32007`.

Birney, E., Andrews, T. D., Bevan, P., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cuff, J., Curwen, V., Cutts, T., Down, T., Eyras, E., Fernandez-Suarez, X. M., Gane, P., Gibbins, B., Gilbert, J., Hammond, M., Hotz, H.-R., Iyer, V., Jekosch, K., Kahari, A., Kasprzyk, A., Keefe, D., Keenan, S., Lehvaslaiho, H., McVicker, G., Melsopp, C., Meidl, P., Mongin, E., Pettett, R., Potter, S., Proctor, G., Rae, M., Searle, S., Slater, G., Smedley, D., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Storey, R., Ureta-Vidal, A., Woodwark, K. C., Cameron, G., Durbin, R., Cox, A., Hubbard, T. and Clamp, M. (2004) An overview of ensembl. *Genome Res*, **14**, 925–928. URL `http://dx.doi.org/10.1101/gr.1860604`.

Black, M. and Doerge, R. (2002) Calculation of the minimum number of replicate spots

required for detection of significant gene expression fold change in microarray experiments. *Bioinformatics*, **18**, 1609–16.

Bloche, M. (2004) Race-based therapeutics. *New England Journal of Medicine*, **351**, 2035–7.

Bogaerts, J., Cardoso, F., Buyse, M., Braga, S., Loi, S., Harrison, J. A., Bines, J., Mook, S., Decker, N., Ravdin, P., Therasse, P., Rutgers, E., van 't Veer, L. J., Piccart, M. and consortium, T. R. A. N. S. B. I. G. (2006) Gene signature evaluation as a prognostic tool: challenges in the design of the mindact trial. *Nat Clin Pract Oncol*, **3**, 540–551.

Bolstad, B. (2006) *affyPLM: Methods for fitting probe-level models.* URL `http://bmbolstad.com`. R package version 1.10.0.

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

Bonferroni, C. E. (1936) Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, **8**, 3–62.

Borczuk, A. C., Gorenstein, L., Walter, K. L., Assaad, A. A., Wang, L. and Powell, C. A. (2003) Non-small-cell lung cancer molecular signatures recapitulate lung developmental pathways. *Am J Pathol*, **163**, 1949–1960.

Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C., Causton, H., Gaasterland, T., Glenisson, P., Holstege, F., Kim, I., Markowitz, V., Matese, J., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J. and Vingron, M. (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nature Genetics*, **29**, 365–71.

Brazma, A., Parkinson, H. and ArrayExpress team, E. M. B. L-E. B. I (2006) Arrayexpress service for reviewers/editors of dna microarray papers. *Nat Biotechnol*, **24**, 1321–1322.

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

Breitling, R. and Herzyk, P. (2005) Rank-based methods as a non-parametric alternative of the t-statistic for the analysis of biological microarray data. *J Bioinform Comput Biol*, **3**, 1171–1189.

Buness, A., Huber, W., Steiner, K., Sültmann, H. and Poustka, A. (2005) arraymagic: two-colour cdna microarray quality control and preprocessing. *Bioinformatics*, **21**, 554–556. URL http://dx.doi.org/10.1093/bioinformatics/bti052.

Bushman, B. J., Cooper, H. and Hedges, L. V. (1994) *The Handbook of Research Synthesis.*, chap. Vote counting methods in meta-analysis., 193–214. New York, NY, USA: Russell Sage Foundation Publications.

Calderwood, S. K., Khaleque, M. A., Sawyer, D. B. and Ciocca, D. R. (2006) Heat shock proteins in cancer: chaperones of tumorigenesis. *Trends Biochem Sci*, **31**, 164–172. URL http://dx.doi.org/10.1016/j.tibs.2006.01.006.

Carter, S. L., Eklund, A. C., Mecham, B. H., Kohane, I. S. and Szallasi, Z. (2005) Redefinition of affymetrix probe sets by sequence overlap with cdna microarray probes reduces cross-platform inconsistencies in cancer-associated gene expression measurements. *BMC Bioinformatics*, **6**, 107. URL http://dx.doi.org/10.1186/1471-2105-6-107.

Chang, H. Y., Nuyten, D. S. A., Sneddon, J. B., Hastie, T., Tibshirani, R., Sørlie, T., Dai, H., He, Y. D., van't Veer, L. J., Bartelink, H., van de Rijn, M., Brown, P. O. and van de Vijver, M. J. (2005) Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. *Proc Natl Acad Sci U S A*, **102**, 3738–3743. URL http://dx.doi.org/10.1073/pnas.0409462102.

Chen, X., Cheung, S. T., So, S., Fan, S. T., Barry, C., Higgins, J., Lai, K.-M., Ji, J., Dudoit, S., Ng, I. O. L., Rijn, M. V. D., Botstein, D. and Brown, P. O. (2002) Gene expression patterns in human liver cancers. *Mol Biol Cell*, **13**, 1929–1939. URL http://dx.doi.org/10.1091/mbc.02-02-0023.

Chen, X., Leung, S. Y., Yuen, S. T., Chu, K.-M., Ji, J., Li, R., Chan, A. S. Y., Law, S., Troyanskaya, O. G., Wong, J., So, S., Botstein, D. and Brown, P. O. (2003) Variation in gene expression patterns in human gastric cancers. *Mol Biol Cell*, **14**, 3208–3215. URL http://dx.doi.org/10.1091/mbc.E02-12-0833.

Choi, J., Choi, J., Kim, D., Choi, D., Kim, B., Lee, K., Yeom, Y., Yoo, H., Yoo, O. and Kim,

S. (2004) Integrative analysis of multiple gene expression profiles applied to liver cancer study. *FEBS Lett*, **565**, 93–100.

Choi, J., Yu, U., Kim, S. and Yoo, O. (2003) Combining multiple microarray studies and modeling inter-study variation. *Bioinformatics.*, **19 Suppl 1**, i84–90.

Cochran, W. (1937) Problems arising in the analysis of a series of similar experiments. *Journal of the Royal Statistical Society*, **Supplement**, 102–118. URL `http://www.jstor.org/view/14666162/di993107/99p0051c/`.

Cohen, J. (1988) *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum, 2nd edn.

Couvelard, A., O'Toole, D., Leek, R., Turley, H., Sauvanet, A., Degott, C., Ruszniewski, P., Belghiti, J., Harris, A. L., Gatter, K. and Pezzella, F. (2005) Expression of hypoxia-inducible factors is correlated with the presence of a fibrotic focus and angiogenesis in pancreatic ductal adenocarcinomas. *Histopathology*, **46**, 668–676. URL `http://dx.doi.org/10.1111/j.1365-2559.2005.02160.x`.

Crick, F. (1970) Central dogma of molecular biology. *Nature*, **227**, 561–563.

Cui, X. and Churchill, G. (2002) How many mice and how many arrays? Replication in mouse cDNA microarray experiments. URL `http://www.jax.org/staff/churchill/labsite/pubs/camda$_$cui.pdf`. The Jackson Labotary.

Cui, X. and Churchill, G. A. (2003) Statistical tests for differential expression in cdna microarray experiments. *Genome Biol*, **4**, 210.

Dan, S., Tsunoda, T., Kitahara, O., Yanagawa, R., Zembutsu, H., Katagiri, T., Yamazaki, K., Nakamura, Y. and Yamori, T. (2002) An integrated database of chemosensitivity to 55 anticancer drugs and gene expression profiles of 39 human cancer cell lines. *Cancer Res*, **62**, 1139–1147.

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

Deeks, J., Altman, D. and Bradburn, M. (2001) *Systematic Reviews in Health Care: Meta-Analysis in Context*, chap. 15, 285–312. BMJ Publishing Group, 2nd edn.

Demeter, J., Beauheim, C., Gollub, J., Hernandez-Boussard, T., Jin, H., Maier, D., Matese, J., Nitzberg, M., Wymore, F., Zachariah, Z., Brown, P., Sherlock, G. and Ball, C. (2007) The Stanford Microarray Database: implementation of new analysis tools and open source release of software. *Nucleic Acids Research*, **35**, D766–770.

Dennis, G., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C. and Lempicki, R. A. (2003) David: Database for annotation, visualization, and integrated discovery. *Genome Biol*, **4**, P3.

DeRisi, J., Penland, L., Brown, P. O., Bittner, M. L., Meltzer, P. S., Ray, M., Chen, Y., Su, Y. A. and Trent, J. M. (1996) Use of a cdna microarray to analyse gene expression patterns in human cancer. *Nat Genet*, **14**, 457–460. URL `http://dx.doi.org/10.1038/ng1296-457`.

Devlin, B., Roeder, K. and Wasserman, L. (2003) False discovery or missed discovery? *Heredity*, **91**, 537–538. URL `http://dx.doi.org/10.1038/sj.hdy.6800370`.

Dickersin, K., Min, Y. I. and Meinert, C. L. (1992) Factors influencing publication of research results. follow-up of applications submitted to two institutional review boards. *JAMA*, **267**, 374–378.

Diehn, M., Sherlock, G., Binkley, G., Jin, H., Matese, J., Hernandez-Boussard, T., Rees, C., Cherry, J., Botstein, D., Brown, P. and Alizadeh, A. (2003) SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data. *Nucleic Acids Res.*, **31**, 219–23. URL `http://source.stanford.edu/`.

Dobbin, K., Shih, J. and Simon, R. (2003) Questions and answers on design of dual-label microarrays for identifying differentially expressed genes. *J Natl Cancer Inst*, **95**, 1362–9.

Dobbin, K. and Simon, R. (2002) Comparison of microarray designs for class comparison and class discovery. *Bioinformatics*, **18**, 1438–45.

Dobbin, K. and Simon, R. (2005) Sample size determination in microarray experiments for class comparison and prognostic classification. *Biostatistics*, **6**, 27–38.

Dudoit, S., Shaffer, S. and Boldrick, J. (2003) Multiple hypothesis testing in microarray experiments. *Statistical Science*, **18**, 71–103.

Dupuy, A. and Simon, R. (2007) Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. *J Natl Cancer Inst*, **99**, 147–157.

Dyrskjot, L., Kruhoffer, M., Thykjaer, T., Marcussen, N., Jensen, J., Moller, K. and Orntoft, T. (2004) Gene expression in the urinary bladder: a common carcinoma in situ gene expression signature exists disregarding histopathological classification. *Cancer Res*, **64**, 4040–8.

Edgar, R. and Barrett, T. (2006) Ncbi geo standards and services for microarray data. *Nat Biotechnol*, **24**, 1471–1472. URL `http://dx.doi.org/10.1038/nbt1206-1471`.

Edgar, R., Domrachev, M. and Lash, A. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, **30**, 207–10.

Efron, B. and Tibshirani, R. (2002) Empirical bayes methods and false discovery rates for microarrays. *Genet Epidemiol*, **23**, 70–86. URL `http://dx.doi.org/10.1002/gepi.1124`.

Efron, B., Tibshirani, R., Storey, J. D. and Tusher, V. (2001) Empirical bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, **96**, 1151–1160.

Egger, M. and Smith, G. D. (1998) Bias in location and selection of studies. *BMJ*, **316**, 61–66.

Ein-Dor, L., Kela, I., Getz, G., Givol, D. and Domany, E. (2005) Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*, **21**, 171–178. URL `http://dx.doi.org/10.1093/bioinformatics/bth469`.

Ferguson, L. (2004) External validity, generalizability, and knowledge utilization. *J Nurs Scholarsh*, **36**, 16–22.

Fisher, R. (1932) *Statistical methods for research workers*. Oliver and Boyd (London), 4 edn.

Fisher, R. (1990) *Statistical Methods, Experimental Design and Scientific Inference*. Oxford University Press. Edited by Bennett, J.H and foreword by Yates, F.

Fleiss, J. L. (1993) The statistical basis of meta-analysis. *Stat Methods Med Res*, **2**, 121–145.

Follmann, D. and Proschan, M. (1999) Valid inference in random effects meta-analysis. *Biometrics*, **55**, 723–7.

Fu, W., Dougherty, E., B, M. and RJ., C. (2005) How many samples are needed to build a classifier: a general sequential approach. *Bioinformatics*, **21**, 63–70.

Gentleman, R. and Carey, V. (2005) *Biobase: Base functions for Bioconductor*. Release 1.8.0.

Ginos, M. A., Page, G. P., Michalowicz, B. S., Patel, K. J., Volker, S. E., Pambuccian, S. E., Ondrey, F. G., Adams, G. L. and Gaffney, P. M. (2004) Identification of a gene expression signature associated with recurrent disease in squamous cell carcinoma of the head and neck. *Cancer Res*, **64**, 55–63.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D. and Lander, E. S. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Grützmann, R., Boriss, H., Ammerpohl, O., Lüttges, J., Kalthoff, H., Schackert, H. K., Klöppel, G., Saeger, H. D. and Pilarsky, C. (2005) Meta-analysis of microarray data on pancreatic cancer defines a set of commonly dysregulated genes. *Oncogene*, **24**, 5079–5088. URL `http://dx.doi.org/10.1038/sj.onc.1208696`.

Grützmann, R., Pilarsky, C., Ammerpohl, O., Lüttges, J., Böhme, A., Sipos, B., Foerder, M., Alldinger, I., Jahnke, B., Schackert, H. K., Kalthoff, H., Kremer, B., Klöppel, G. and Saeger, H. D. (2004) Gene expression profiling of microdissected pancreatic ductal carcinomas using high-density dna microarrays. *Neoplasia*, **6**, 611–622. URL `http://dx.doi.org/10.1593/neo.04295`.

Guenther, W. (1975) A sample size formula for a non-central t test. *The American Statistician*, **29**, 120–1.

Guenther, W. (1981) Sample size formulas for normal theory t tests. *The American Statistician*, **35**, 243–4.

Haslinger, C., Schweifer, N., Stilgenbauer, S., Döhner, H., Lichter, P., Kraut, N., Stratowa, C. and Abseher, R. (2004) Microarray gene expression profiling of b-cell chronic lymphocytic

leukemia subgroups defined by genomic aberrations and vh mutation status. *J Clin Oncol*, **22**, 3937–3949. URL http://dx.doi.org/10.1200/JCO.2004.12.133.

Hedges, L. V. (1981) Distribution theory for glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, **6**, 107–128. URL http://www.jstor.org/stable/1164588.

Hedges, L. V. and Olkin, I. (1985) *Statistical Methods for Meta-Analysis.* Academic Press. ISBN-13: 978-0123363800.

Hippo, Y., Taniguchi, H., Tsutsumi, S., Machida, N., Chong, J.-M., Fukayama, M., Kodama, T. and Aburatani, H. (2002) Global gene expression analysis of gastric cancer by oligonucleotide microarrays. *Cancer Res*, **62**, 233–240.

Hong, F., Breitling, R., McEntee, C. W., Wittner, B. S., Nemhauser, J. L. and Chory, J. (2006) Rankprod: a bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics*, **22**, 2825–2827. URL http://dx.doi.org/10.1093/bioinformatics/btl476.

Hu, J., Bianchi, F., Ferguson, M., Cesario, A., Margaritora, S., Granone, P., Goldstraw, P., Tetlow, M., Ratcliffe, C., Nicholson, A. G., Harris, A., Gatter, K. and Pezzella, F. (2005) Gene expression signature for angiogenic and nonangiogenic non-small-cell lung cancer. *Oncogene*, **24**, 1212–1219. URL http://dx.doi.org/10.1038/sj.onc.1208242.

Huang, D. W., Sherman, B. T. and Lempicki, R. A. (2009) Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nat Protoc*, **4**, 44–57. URL http://dx.doi.org/10.1038/nprot.2008.211.

Huang, E., Cheng, S., Dressman, H., Pittman, J., Tsou, M., Horng, C., Bild, A., Iversen, E., Liao, M., Chen, C., West, M., Nevins, J. and Huang, A. (2003) Gene expression predictors of breast cancer outcomes. *Lancet*, **361**, 1590–6.

Huang, Y., Prasad, M., Lemon, W. J., Hampel, H., Wright, F. A., Kornacker, K., LiVolsi, V., Frankel, W., Kloos, R. T., Eng, C., Pellegata, N. S. and de la Chapelle, A. (2001) Gene expression in papillary thyroid carcinoma reveals highly consistent profiles. *Proc Natl Acad Sci U S A*, **98**, 15044–15049. URL http://dx.doi.org/10.1073/pnas.251547398.

Huber, W., von Heydebreck, A., Sültmann, H., Poustka, A. and Vingron, M. (2002) Variance

stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, **18 Suppl 1**, S96–104.

Hwang, D., Schmitt, W., Stephanopoulos, G. and Stephanopoulos, G. (2002) Determination of minimum sample size and discriminatory expression patterns in microarray data. *Bioinformatics*, **18**, 1184–93.

Ikeo, K., Ishi-i, J., Tamura, T., Gojobori, T. and Tateno, Y. (2003) CIBEX: center for information biology gene expression database. *C R Biol.*, **326**, 1079–82.

Ioannidis, J. P. A., Polyzos, N. P. and Trikalinos, T. A. (2007) Selective discussion and transparency in microarray research findings for cancer outcomes. *Eur J Cancer*, **43**, 1999–2010. URL `http://dx.doi.org/10.1016/j.ejca.2007.05.019`.

Ioannidis, J. P. A., Rosenberg, P. S., Goedert, J. J., O'Brien, T. R. and International Meta-analysis of HIV Host Genetics (2002) Commentary: meta-analysis of individual participants' data in genetic epidemiology. *Am J Epidemiol*, **156**, 204–210.

Irizarry, R., Hobbs, B., Collin, F., Beazer-Barclay, Y., Antonellis, K., Scherf, U. and Speed, T. (2003a) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–64.

Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B. and Speed, T. P. (2003b) Summaries of affymetrix genechip probe level data. *Nucleic Acids Res*, **31**, e15.

Irizarry, R. A., Warren, D., Spencer, F., Kim, I. F., Biswal, S., Frank, B. C., Gabrielson, E., Garcia, J. G. N., Geoghegan, J., Germino, G., Griffin, C., Hilmer, S. C., Hoffman, E., Jedlicka, A. E., Kawasaki, E., Martínez-Murillo, F., Morsberger, L., Lee, H., Petersen, D., Quackenbush, J., Scott, A., Wilson, M., Yang, Y., Ye, S. Q. and Yu, W. (2005) Multiple-laboratory comparison of microarray platforms. *Nat Methods*, **2**, 345–350. URL `http://dx.doi.org/10.1038/nmeth756`.

Ishikawa, M., Yoshida, K., Yamashita, Y., Ota, J., Takada, S., Kisanuki, H., Koinuma, K., Choi, Y. L., Kaneda, R., Iwao, T., Tamada, K., Sugano, K. and Mano, H. (2005) Experimental trial for diagnosis of pancreatic ductal carcinoma based on gene expression profiles of pancreatic ductal cells. *Cancer Sci*, **96**, 387–393. URL `http://dx.doi.org/10.1111/j.1349-7006.2005.00064.x`.

Jafari, P. and Azuaje, F. (2006) An assessment of recently published gene expression data analyses: reporting experimental design and statistical factors. *BMC Med Inform Decis Mak*, **6**, 27. URL http://dx.doi.org/10.1186/1472-6947-6-27.

Jones, M. H., Virtanen, C., Honjoh, D., Miyoshi, T., Satoh, Y., Okumura, S., Nakagawa, K., Nomura, H. and Ishikawa, Y. (2004) Two prognostically significant subtypes of high-grade lung neuroendocrine tumours independent of small-cell and large-cell neuroendocrine carcinomas identified by gene expression profiles. *Lancet*, **363**, 775–781. URL http://dx.doi.org/10.1016/S0140-6736(04)15693-6.

Jung, S., Bang, H. and Young, S. (2005) Sample size calculation for multiple testing in microarray data analysis. *Biostatistics*, **6**, 157–169.

Kim, S.-Y., Kim, J.-H., Lee, H.-S., Noh, S.-M., Song, K.-S., Cho, J.-S., Jeong, H.-Y., Kim, W. H., Yeom, Y.-I., Kim, N.-S., Kim, S., Yoo, H.-S. and Kim, Y. S. (2007) Meta- and gene set analysis of stomach cancer gene expression data. *Mol Cells*, **24**, 200–209.

Klein, U., Tu, Y., Stolovitzky, G., Mattioli, M., Cattoretti, G., Husson, H., Freedman, A., Inghirami, G., Cro, L., Baldini, L., Neri, A., Califano, A. and Dalla-Favera, R. (2001) Gene expression profiling of B cell chronic lymphocytic leukemia reveals a homogeneous phenotype related to memory B cells. *J Exp Med.*, **194**, 1625–38.

Kruskal, J. B. and Wish, M. (1978) *Multidimensional Scaling*. SAGE Publications. ISBN:978-0803909403.

Kuriakose, M., Chen, W., He, Z., Sikora, A., Zhang, P., Zhang, Z., Qiu, W., Hsu, D., McMunn-Coffran, C., Brown, S., Elango, E., Delacure, M. and Chen, F. (2004) Selection and validation of differentially expressed genes in head and neck cancer. *Cell Mol Life Sci.*, **61**, 1372–83.

Lapointe, J., Li, C., Higgins, J. P., van de Rijn, M., Bair, E., Montgomery, K., Ferrari, M., Egevad, L., Rayford, W., Bergerheim, U., Ekman, P., DeMarzo, A. M., Tibshirani, R., Botstein, D., Brown, P. O., Brooks, J. D. and Pollack, J. R. (2004) Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proc Natl Acad Sci U S A*, **101**, 811–816. URL http://dx.doi.org/10.1073/pnas.0304146101.

Larkin, J. E., Frank, B. C., Gavras, H., Sultana, R. and Quackenbush, J. (2005) Indepen-

dence and reproducibility across microarray platforms. *Nat Methods*, **2**, 337–344. URL http://dx.doi.org/10.1038/nmeth757.

Larsson, O. and Sandberg, R. (2006) Lack of correct data format and comparability limits future integrative microarray research. *Nat Biotechnol*, **24**, 1322–1323. URL http://dx.doi.org/10.1038/nbt1106-1322.

Lee, H. K., Hsu, A. K., Sajdak, J., Qin, J. and Pavlidis, P. (2004) Coexpression analysis of human genes across many microarray data sets. *Genome Res*, **14**, 1085–1094. URL http://dx.doi.org/10.1101/gr.1910904.

Lee, M., Kuo, F., Whitmore, G. and Sklar, J. (2000) Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proc Natl Acad Sci U S A*, **97**, 9834–9.

Lee, M. and Whitmore, G. (2002) Power and sample size for DNA microarray studies. *Stat Med*, **21**, 3543–70.

Lenburg, M., Liou, L., Gerry, N., Frampton, G., Cohen, H. and Christman, M. (2003) Previously unidentified changes in renal cell carcinoma gene expression identified by parametric analysis of microarray data. *BMC Cancer*, **3**, 31.

Lennon, G., Auffray, C., Polymeropoulos, M. and Soares, M. (1996) The I.M.A.G.E. Consortium: an integrated molecular analysis of genomes and their expression. *Genomics*, **33**, 151–2.

Lewis, S. and Clarke, M. (2001) Forest plots: trying to see the wood and the trees. *BMJ*, **322**, 1479–80.

Li, C. and Wong, H. (2001) Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biology*, **2**, RESEARCH0032.

Lin, S. M., Du, P., Huber, W. and Kibbe, W. A. (2008) Model-based variance-stabilizing transformation for illumina microarray data. *Nucleic Acids Res*, **36**, e11. URL http://dx.doi.org/10.1093/nar/gkm1075.

Loh, M. and Mondry, A. (2008) Diagnostic robustness of DNA microarrays in the classification of acute leukemia. URL http://precedings.nature.com/documents/1056/version/1.

M. A. Q. C. Consortium, Shi, L., Reid, L. H., Jones, W. D., Shippy, R., Warrington, J. A., Baker, S. C., Collins, P. J., de Longueville, F., Kawasaki, E. S., Lee, K. Y., Luo, Y., Sun, Y. A., Willey, J. C., Setterquist, R. A., Fischer, G. M., Tong, W., Dragan, Y. P., Dix, D. J., Frueh, F. W., Goodsaid, F. M., Herman, D., Jensen, R. V., Johnson, C. D., Lobenhofer, E. K., Puri, R. K., Schrf, U., Thierry-Mieg, J., Wang, C., Wilson, M., Wolber, P. K., Zhang, L., Amur, S., Bao, W., Barbacioru, C. C., Lucas, A. B., Bertholet, V., Boysen, C., Bromley, B., Brown, D., Brunner, A., Canales, R., Cao, X. M., Cebula, T. A., Chen, J. J., Cheng, J., Chu, T.-M., Chudin, E., Corson, J., Corton, J. C., Croner, L. J., Davies, C., Davison, T. S., Delenstarr, G., Deng, X., Dorris, D., Eklund, A. C., hui Fan, X., Fang, H., Fulmer-Smentek, S., Fuscoe, J. C., Gallagher, K., Ge, W., Guo, L., Guo, X., Hager, J., Haje, P. K., Han, J., Han, T., Harbottle, H. C., Harris, S. C., Hatchwell, E., Hauser, C. A., Hester, S., Hong, H., Hurban, P., Jackson, S. A., Ji, H., Knight, C. R., Kuo, W. P., LeClerc, J. E., Levy, S., Li, Q.-Z., Liu, C., Liu, Y., Lombardi, M. J., Ma, Y., Magnuson, S. R., Maqsodi, B., McDaniel, T., Mei, N., Myklebost, O., Ning, B., Novoradovskaya, N., Orr, M. S., Osborn, T. W., Papallo, A., Patterson, T. A., Perkins, R. G., Peters, E. H., Peterson, R., Philips, K. L., Pine, P. S., Pusztai, L., Qian, F., Ren, H., Rosen, M., Rosenzweig, B. A., Samaha, R. R., Schena, M., Schroth, G. P., Shchegrova, S., Smith, D. D., Staedtler, F., Su, Z., Sun, H., Szallasi, Z., Tezak, Z., Thierry-Mieg, D., Thompson, K. L., Tikhonova, I., Turpaz, Y., Vallanat, B., Van, C., Walker, S. J., Wang, S. J., Wang, Y., Wolfinger, R., Wong, A., Wu, J., Xiao, C., Xie, Q., Xu, J., Yang, W., Zhang, L., Zhong, S., Zong, Y. and Slikker, W. (2006) The microarray quality control (maqc) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat Biotechnol*, **24**, 1151–1161. URL `http://dx.doi.org/10.1038/nbt1239`.

Ma, Y. and Hendershot, L. M. (2004) The role of the unfolded protein response in tumour development: friend or foe? *Nat Rev Cancer*, **4**, 966–977. URL `http://dx.doi.org/10.1038/nrc1505`.

Matsui, S. and Oura, T. (2009) Sample sizes for a robust ranking and selection of genes in microarray experiments. *Stat Med*. URL `http://dx.doi.org/10.1002/sim.3666`.

Matsui, S., Zeng, S., Yamanaka, T. and Shaughnessy, J. (2008) Sample size calculations based on ranking and selection in microarray experiments. *Biometrics*, **64**, 217–226. URL `http://dx.doi.org/10.1111/j.1541-0420.2007.00875.x`.

McShane, L., Shih, J. and Michalowska, A. (2003) Statistical issues in the design and analysis of gene expression microarray studies of animal models. *J Mammary Gland Biol Neoplasia*, **8**, 359–74.

Mehra, R., Varambally, S., Ding, L., Shen, R., Sabel, M. S., Ghosh, D., Chinnaiyan, A. M. and Kleer, C. G. (2005) Identification of gata3 as a breast cancer prognostic marker by global gene expression meta-analysis. *Cancer Res*, **65**, 11259–11264. URL http://dx.doi.org/10.1158/0008-5472.CAN-05-2495.

Michiels, S., Koscielny, S. and Hill, C. (2005) Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet*, **365**, 488–92.

Mondry, A., Loh, M. and Giuliani, A. (2008) DNA expression microarrays may be the wrong tool to identify biological pathways. URL http://precedings.nature.com/documents/1036/version/1.

Mondry, A., Loh, M., Liu, P., Zhu, A. and Nagel, M. (2005) Polymorphisms of the insertion / deletion ACE and M235T AGT genes and hypertension: surprising new findings and meta-analysis of data. *BMC Nephrology*, **6**, 1.

Morris, J., Yin, G., Baggerly, K., Wu, C. and Zhang, L. (2005) *Methods of microarray data analysis*, chap. Pooling information across different studies and oligonucleotide microarray chip types to identify prognostic genes for lung cancer., 51–66. Springer-Verlag, iv edn. Free reprint available at http://works.bepress.com/cgi/viewcontent.cgi?article=1005&context=jeffrey_s_morris.

Mosser, D. D. and Morimoto, R. I. (2004) Molecular chaperones and the stress of oncogenesis. *Oncogene*, **23**, 2907–2918. URL http://dx.doi.org/10.1038/sj.onc.1207529.

Mukherjee, S., Tamayo, P., Rogers, S., Rifkin, R., Engle, A., Campbell, C., Golub, T. and Mesirov, J. (2003) Estimating dataset size requirements for classifying DNA microarray data. *J. Comput Biol. 2003*, **10**, 119–42.

Muller, P., Parmigiani, G., Robert, C. and Rousseau, J. (2004) Optimal sample size for multiple testing: the case of gene expression microarrays. URL http://www.bepress.com/jhubiostat/paper31/. Department of Biostatistics Working Papers, John Hopkins University.

Nakagawa, S. (2004) A farewell to bonferroni: the problems of low statistical power and publication bias. *Behavioral Ecology*, **5**, 1044–1045.

Noth, S., Benecke, A. and Systems Epigenomics Group (2005) Avoiding inconsistencies over time and tracking difficulties in applied biosystems ab1700/panther probe-to-gene annotations. *BMC Bioinformatics*, **6**, 307.

Ntzani, E. and Ioannidis, J. (2003) Predictive ability of DNA microarrays for cancer outcomes and correlates: an empirical assessment. *Lancet*, **362**, 1439–44.

Oliveros, J. (2007) Venny: An interactive tool for comparing lists using venn diagrams. URL `http://bioinfogp.cnb.csic.es/tools/venny/index.html`.

Paik, S. (2007) Development and clinical utility of a 21-gene recurrence score prognostic assay in patients with early breast cancer treated with tamoxifen. *Oncologist*, **12**, 631–635. URL `http://dx.doi.org/10.1634/theoncologist.12-6-631`.

Pan, W., Lin, J. and Le, C. (2002) How many replicates of arrays are required to detect gene expression changes in microarray experiments? A mixture model approach. *Genome Biol*, **3**, RESEARCH0022.

Parmigiani, G., Garrett-Mayer, E., Anbazhagan, R. and Gabrielson, E. (2005) A cross-study comparison of gene expression studies for the molecular classification of lung cancer. *Clin. Cancer Res.*, **10**, 2922–7.

Pavlidis, P., Li, Q. and Noble, W. (2003) The effect of replication on gene expression microarray experiments. *Bioinformatics*, **19**, 1620–7.

Pawitan, Y., Michiels, S., Koscielny, S., Gusnanto, A. and Ploner, A. (2005) False discovery rate, sensitivity and sample size for microarray studies. *Bioinformatics*, **21**, 3017–3024.

Peduzzi, P., Concato, J., Kemper, E., Holford, T. and A.R., F. (1996) A simulation study of the number of events per variable in logistic regression analysis. *Journal Clinical Epidemiology*, **49**, 1373–9.

Pellagatti, A., Cazzola, M., Giagounidis, A. A. N., Malcovati, L., Porta, M. G. D., Killick, S., Campbell, L. J., Wang, L., Langford, C. F., Fidler, C., Oscier, D., Aul, C., Wainscoat, J. S. and Boultwood, J. (2006) Gene expression profiles of CD34+ cells in myelodysplastic syndromes: involvement of interferon-stimulated genes

and correlation to FAB subtype and karyotype. *Blood*, **108**, 337–345. URL http://dx.doi.org/10.1182/blood-2005-12-4769.

Perez-Iratxeta, C. and Andrade, M. A. (2005) Inconsistencies over time in 5of NetAffx probe-to-gene annotations. *BMC Bioinformatics*, **6**, 183. URL http://dx.doi.org/10.1186/1471-2105-6-183.

Perou, C. M., Jeffrey, S. S., van de Rijn, M., Rees, C. A., Eisen, M. B., Ross, D. T., Pergamenschikov, A., Williams, C. F., Zhu, S. X., Lee, J. C., Lashkari, D., Shalon, D., Brown, P. O. and Botstein, D. (1999) Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc Natl Acad Sci U S A*, **96**, 9212–9217.

Pilarsky, C., Wenzig, M., Specht, T., Saeger, H. D. and Grützmann, R. (2004) Identification and validation of commonly overexpressed genes in solid tumors by comparison of microarray data. *Neoplasia*, **6**, 744–750. URL http://dx.doi.org/10.1593/neo.04277.

Pinheiro, J. C. and Bates, D. M. (2000) *Mixed-Effects Models in S and S-PLUS*. Springer. ISBN 0-387-98957-0.

Piwowar, H. A., Day, R. S. and Fridsma, D. B. (2007) Sharing detailed research data is associated with increased citation rate. *PLoS ONE*, **2**, e308. URL http://dx.doi.org/10.1371/journal.pone.0000308.

Pounds, S. and Cheng, C. (2005) Sample size determination for the false discovery rate. *Bioinformatics*, **21**, 4263–4271. URL http://dx.doi.org/10.1093/bioinformatics/bti699.

Pounds, S. and Cheng, C. (2009) Erratum: sample size determination for the false discovery rate. *Bioinformatics*, **25**, 698–699.

Pruitt, K. D., Tatusova, T. and Maglott, D. R. (2007) Ncbi reference sequences (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*, **35**, D61–D65. URL http://dx.doi.org/10.1093/nar/gkl842.

R Development Core Team (2004) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org. ISBN 3-900051-00-3.

Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C., Angelo, M., Ladd, C.,

Reich, M., Latulippe, E., Mesirov, J., Poggio, T., Gerald, W., Loda, M., Lander, E. and Golub, T. (2001) Multiclass cancer diagnosis using tumor gene expression signatures. *Proc Natl Acad Sci U S A.*, **98**, 15149–54.

Rhodes, D., Barrette, T., Rubin, M., Ghosh, D. and Chinnaiyan, A. (2002) Meta-analysis of microarrays: inter-study validation of gene expression profiles reveals pathway dysregulation in prostate cancer. *Cancer Research.*, **62**, 4427–33.

Rhodes, D., Yu, J., Shanker, K., Deshpande, N., Varambally, R., Ghosh, D., Barrette, T., Pandey, A. and Chinnaiyan, A. (2004a) Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proc Natl Acad Sci U S A*, **101**, 9309–9314.

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

Satterthwaite, F. E. (1946) An approximate distribution of estimates of variance components. *International Biometric Society*, **2**, 110–114. URL `http://www.jstor.org/stable/view/3002019`.

Schena, M., Shalon, D., Davis, R. and Brown, P. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467–70.

Seo, J., Gordish-Dressman, H. and Hoffman, E. P. (2006) An interactive power analysis tool for microarray hypothesis testing and generation. *Bioinformatics*, **22**, 808–814. URL `http://dx.doi.org/10.1093/bioinformatics/btk052`.

Silva, G. L., Junta, C. M., Mello, S. S., Garcia, P. S., Rassi, D. M., Sakamoto-Hojo, E. T., Donadi, E. A. and Passos, G. A. S. (2007) Profiling meta-analysis reveals primarily gene coexpression concordance between systemic lupus erythematosus and rheumatoid arthritis. *Ann N Y Acad Sci*, **1110**, 33–46. URL `http://dx.doi.org/10.1196/annals.1423.005`.

Simon, R., Korn, E., McShane, L., Radmacher, M., Wright, G. and Zhao, Y. (2004) *Design and Analysis of DNA Microarray Investigations*. Springer, 1 edn.

Simon, R., Radmacher, M. and Dobbin, K. (2002) Design of studies using DNA microarrays. *Genet Epidemiol*, **23**, 21–36.

Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A. A., D'Amico, A. V., Richie, J. P., Lander, E. S., Loda, M., Kantoff, P. W., Golub, T. R. and Sellers, W. R. (2002) Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, **1**, 203–209.

Smid, M., Dorssers, L. C. J. and Jenster, G. (2003) Venn mapping: clustering of heterologous microarray data based on the number of co-occurring differentially expressed genes. *Bioinformatics*, **19**, 2065–2071.

Smyth, G. (2005) *LIMMA: linear models for microarray data.*, chap. 23, 297–420. Springer (New York). URL `http://www.statsci.org/smyth/pubs/limma-biocbook-reprint.pdf`.

Smyth, G. K. and Speed, T. (2003) Normalization of cDNA microarray data. *Methods*, **31**, 265–273.

Staunton, J. E., Slonim, D. K., Coller, H. A., Tamayo, P., Angelo, M. J., Park, J., Scherf, U., Lee, J. K., Reinhold, W. O., Weinstein, J. N., Mesirov, J. P., Lander, E. S. and Golub, T. R. (2001) Chemosensitivity prediction by transcriptional profiling. *Proc Natl Acad Sci U S A*, **98**, 10787–10792. URL `http://dx.doi.org/10.1073/pnas.191368598`.

Storey, J. (2002) A direct approach to false discovery rates. *Journal of the Royal Statistical Society Series B*, **64**, 479–498.

Stuart, J. M., Segal, E., Koller, D. and Kim, S. K. (2003) A gene-coexpression network for global discovery of conserved genetic modules. *Science*, **302**, 249–255. URL `http://dx.doi.org/10.1126/science.1087447`.

Student (1908) The probable error of a mean. *Biometrika*, **6**, 1–25. URL `http://biomet.oxfordjournals.org/cgi/reprint/6/1/1`.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S. and Mesirov, J. P. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, **102**, 15545–15550. URL `http://dx.doi.org/10.1073/pnas.0506580102`.

Suárez-Fariñas, M., Noggle, S., Heke, M., Hemmati-Brivanlou, A. and Magnasco, M. O.

(2005) Comparing independent microarray studies: the case of human embryonic stem cells. *BMC Genomics*, **6**, 99. URL http://dx.doi.org/10.1186/1471-2164-6-99.

Sutton, A., Abrams, K., Jones, D., Sheldon, T. and Song, F. (2000) *Methods for Meta-analysis in Medical Research*. John Wiley & Sons. (Based on the original NHS Health Technology Assessment funded Project (93/51/2) which is freely available online at http://www.hta.nhsweb.nhs.uk/fullmono/mon219.pdf).

Taylor, C. F., Field, D., Sansone, S.-A., Aerts, J., Apweiler, R., Ashburner, M., Ball, C. A., Binz, P.-A., Bogue, M., Booth, T., Brazma, A., Brinkman, R. R., Clark, A. M., Deutsch, E. W., Fiehn, O., Fostel, J., Ghazal, P., Gibson, F., Gray, T., Grimes, G., Hardy, N. W., Hermjakob, H., Julian, R. K., Jr., Kane, M., Kettner, C., Kinsinger, C., Kolker, E., andb, M. K., Novère, N. L., Leebens-Mack, J., Lewis, S. E., Lord, P., Mallon, A.-M., Marthandan, N., Masuya, H., McNally, R., Mehrle, A., Morrison, N., Quackenbush, J., Reecy, J. M., Robertson, D. G., Rocca-Serra, P., Rodriguez, H., Rosenfelder, H., Santoyo-Lopez, J., Scheuermann, R. H., Schober, D., Smith, B., Snape, J., Tipton, K., Sterk, P. and Wiemann, S. (2007) Promoting coherent minimum reporting requirements for biological and biomedical investigations: The mibbi project. URL http://mibbi.sourceforge.net/docs/papers/MIBBI.pdf. Submitted to Nature Biotechnology.

Tsai, C., Wang, S., Chen, D. and Chen, J. (2005) Sample size for gene expression microarray experiments. *Bioinformatics*, **21**, 1502–8.

Tsai, J., Sultana, R., Lee, Y., Pertea, G., Karamycheva, S., Antonescu, V., Cho, J., Parvizi, B., Cheung, F. and Quackenbush, J. (2001) Resourcerer: a database for annotating and linking microarray resources within and across species. *Genome Biol*, **2**, SOFTWARE0002.

Tusher, V. G., Tibshirani, R. and Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A*, **98**, 5116–5121. URL http://dx.doi.org/10.1073/pnas.091062498.

van de Vijver, M. J., He, Y. D., van't Veer, L. J., Dai, H., Hart, A. A. M., Voskuil, D. W., Schreiber, G. J., Peterse, J. L., Roberts, C., Marton, M. J., Parrish, M., Atsma, D., Witteveen, A., Glas, A., Delahaye, L., van der Velde, T., Bartelink, H., Rodenhuis, S., Rutgers, E. T., Friend, S. H. and Bernards, R. (2002) A gene-expression signature

as a predictor of survival in breast cancer. *N Engl J Med*, **347**, 1999–2009. URL http://dx.doi.org/10.1056/NEJMoa021967.

van 't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R. and Friend, S. H. (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–536. URL http://dx.doi.org/10.1038/415530a.

Venables, W. N. and Ripley, B. D. (2002) *Modern Applied Statistics with S*. Springer, 4th edn. ISBN:978-0387954578.

Venn, J. (1880) On the diagrammatic and mechanical representation of propositions and reasonings. *Dublin Philosophical Magazine and Journal of Science*, **9**, 1–18.

Ventura, B. (2005) Mandatory submission of microarray data to public repositories: how is it working? *Physiol Genomics*, **20**, 153–156. URL http://dx.doi.org/10.1152/physiolgenomics.00264.2004.

Wang, J., Coombes, K. R., Highsmith, W. E., Keating, M. J. and Abruzzo, L. V. (2004) Differences in gene expression between b-cell chronic lymphocytic leukemia and normal b cells: a meta-analysis of three microarray studies. *Bioinformatics*, **20**, 3166–3178. URL http://dx.doi.org/10.1093/bioinformatics/bth381.

Wang, S.-J. and Chen, J. J. (2004) Sample size for identifying differentially expressed genes in microarray experiments. *J Comput Biol*, **11**, 714–726. URL http://dx.doi.org/10.1089/1066527041887267.

Wang, Y., Joshi, T., Zhang, X.-S., Xu, D. and Chen, L. (2006) Inferring gene regulatory networks from multiple microarray datasets. *Bioinformatics*, **22**, 2413–2420. URL http://dx.doi.org/10.1093/bioinformatics/btl396.

Warnat, P., Eils, R. and Brors, B. (2005) Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes. *BMC Bioinformatics*, **6**, 265. URL http://dx.doi.org/10.1186/1471-2105-6-265.

Wei, C., Li, J. and Bumgarner, R. (2004) Sample size for detecting differentially expressed genes in microarray experiments. *BMC Genomics*, **5**, 87.

Welch, B. L. (1947) The generalization of "student's" problem when several different population variances are involved. *Biometrika*, **34**, 28–35. URL http://biomet.oxfordjournals.org/cgi/reprint/34/1-2/28.pdf.

Welsh, J., Sapinoso, L., Su, A., Kern, S., Wang-Rodriguez, J., Moskaluk, C., Frierson, H. and Hampton, G. (2001) Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer. *Cancer Res*, **61**, 5974–8.

Wenisch, L. (2002) Can replication save noisy microarray data. *Comparitive and Functional Genomics*, **3**, 372–4.

Wheeler, D., Church, D., Federhen, S., Lash, A., Madden, T., Pontius, J., Schuler, G., Schriml, L., Sequeira, E., Tatusova, T. and Wagner, L. (2003) Database resources of the National Center for Biotechnology. *Nucleic Acids Research*, **31**, 28–33.

Whitehead, A. (2002) *Meta-Analysis of Controlled Clinical Trials*. Wiley, 1 edn.

Wilson, C. L. and Miller, C. J. (2005) Simpleaffy: a bioconductor package for affymetrix quality control and data analysis. *Bioinformatics*, **21**, 3683–3685. URL http://dx.doi.org/10.1093/bioinformatics/bti605.

Winter, S. C., Buffa, F. M., Silva, P., Miller, C., Valentine, H. R., Turley, H., Shah, K. A., Cox, G. J., Corbridge, R. J., Homer, J. J., Musgrove, B., Slevin, N., Sloan, P., Price, P., West, C. M. L. and Harris, A. L. (2007) Relation of a hypoxia metagene derived from head and neck cancer to prognosis of multiple cancers. *Cancer Res*, **67**, 3441–3449. URL http://dx.doi.org/10.1158/0008-5472.CAN-06-3322.

Wit, E. and McClure, J. (2004) *Statistics for Microarrays: Design, Analysis and Inference.* John Wiley & Sons.

Wu, Z. and Irizarry, R. (2004) Preprocessing of oligonucleotide array data. *Nature Biotechnology*, **22**, 656–658.

Wu, Z., Irizarry, R., Gentleman, R., Murillo, F. and Spencer, F. (2004) A model based background adjustment for oligonucleotide expression arrays. Johns Hopkins University, Dept. of Biostatistics Working Paper.

Yang, M., Yang, J., McIndoe, R. and She, J. (2003) Microarray experimental design: power and sample size considerations. *Physiol Genomics*, **16**, 24–8.

Yang, X., Bentink, S. and Spang, R. (2005) Detecting common gene expression patterns in multiple cancer outcome entities. *Biomed Microdevices*, **7**, 247–251. URL http://dx.doi.org/10.1007/s10544-005-3032-7.

Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J. and Speed, T. P. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res*, **30**, e15.

Yeoh, E., Ross, M., Shurtleff, S., Williams, W., Patel, D., Mahfouz, R., Behm, F., Raimondi, S., Relling, M., Patel, A., Cheng, C., Campana, D., Wilkins, D., Zhou, X., Li, J., Liu, H., Pui, C., Evans, W., Naeve, C., Wong, L. and J.R., D. (2002) Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell*, **1**, 133–43.

Zeeberg, B. R., Riss, J., Kane, D. W., Bussey, K. J., Uchio, E., Linehan, W. M., Barrett, J. C. and Weinstein, J. N. (2004) Mistaken identifiers: gene name errors can be introduced inadvertently when using excel in bioinformatics. *BMC Bioinformatics*, **5**, 80. URL http://dx.doi.org/10.1186/1471-2105-5-80.

Zhang, S. and Gant, T. (2004) A statistical framework for the design of microarray experiments and effective detection of differential gene expression. *Bioinformatics*, **20**, 2821–8.

Zien, A., Fluck, J., Zimmer, R. and Lengauer, T. (2003) Microarrays: how many do you need? *J Comput Biol*, **10**, 653–67.

Zintzaras, E. and Ioannidis, J. P. A. (2008) Meta-analysis for ranked discovery datasets: Theoretical framework and empirical demonstration for microarrays. *Comput Biol Chem*, **32**, 38–46. URL http://dx.doi.org/10.1016/j.compbiolchem.2007.09.003.