Analysis of gene regulation using high throughput genomics

A thesis submitted

by

Paul Geeleher

to

The School of Mathematics, Statistics and Applied Mathematics,
National University *of* Ireland, Galway

In partial fulfilment of the requirements for the degree of
Ph.D. in Science

September 2012

Supervised by Professor Cathal Seoighe

Abstract

The recent development of high-throughput genomics techniques and their subsequent applications have completely transformed the study of biology. The analysis, interpretation and storage of the resulting large volumes of data have created a wide range of computational challenges and opportunities that have driven the majority of recent bioinformatics research. In this thesis we focus on four research questions grounded in functional genomics and epigenomics, yielding novel methodologies and biological insights.

The first research question relates to whether miRNA activity, as a general regulatory effect, is a heritable trait. To do this, we used Affymetrix Human Exon Microarray and RNA-seq data from the International HapMap project. We confirmed such an association in humans using the regulatory effect score (RE-score) of a miRNA, which has previously been defined as the difference in the gene expression rank of targets of the miRNA compared to non-targeted genes. We also identified a SNP in the miRNA processing gene *DROSHA*, which is associated with inter-individual difference in miRNA regulatory effect.

During this analysis we noted that correlations between gene expression measures from RNA-seq and gene expression microarray platforms were often relatively poor. This led us to develop a method to improve the estimation of gene expression from microarrays. Our method uses samples for which there is both microarray and RNA-seq data available and builds statistical models which learn the relationship between probe level gene expression, as measured by the microarrays, and gene level expression, as measured by RNA-seq. These models can then be used to estimate gene expression on separate sets of microarray samples. We have assessed the performance of our method in comparison to Affymetrix Power Tools (APT). To do this, we fitted models for all genes on a training set of the HapMap YRI samples and tested performance on the HapMap CEU (both microarray and RNA-seq data are available for all of these samples). Overall, our method improves within sample correlations with RNA-seq substantially, but does not achieve the same level of performance as APT in terms of across sample correlations.

The third research question aimed to determine whether or not it was possible to ascertain a consistent pattern of differential methylation in a limited number of ulcerative colitis (UC) biopsies, using data generated with the Agilent Human CpG Island microarray. Although there were no statistically significant differences between the sample groups at CpG island or probe level, we did uncover evidence of overall CpG island hypermethylation in UC. Subsequently, gene set

analysis (GSA) revealed highly significant results for several GO biological processes. It became apparent that these results were a consequence of a sampling effect, which stems from the large differences in numbers of probes (targeting CpG sites) associated with genes in different gene sets.

The fourth and final research question consisted of the development of a method to correct the bias in GSA analysis of these data. We applied our method to both the UC microarray dataset and a previously published genome-wide CpG island study of DNA methylation in lung cancer. We obtained novel biological insights into both of these conditions, consistent with their respective pathologies. Finally, we showed that this bias is also found with next generation sequencing based methylation assays, which we demonstrated using a HELP-seq dataset.

In conclusion, this thesis presents novel analytical strategies encompassing gene expression and genome-wide methylation, and it also introduces methodologies that link microarray and RNA-seq measures of expression. It documents for the first time a correction for an intrinsic bias in GSA associated with many CpG island methylation platforms, and yields results of biological consequence with regard to endogenous RNAi regulatory processes and the epigenetic characterization of several human diseases.

Contents

List of Figures

List of Tables

Acknowledgements

Firstly, I would like to thank Professor Cathal Seoighe. He has been an outstanding supervisor. His ability to solve problems quickly and effectively is completely unparalleled by anybody else that I have ever met. He has a passion for real science and a level of honesty and ethical integrity that is evident in the exceptional quality of all of his work. I believe that under his guidance, any student with a decent work ethic and a willingness to learn could be propelled to the top level of their field. I have been absolutely privileged and incredibly lucky to have had the opportunity to work with him over the past couple of years and I would be delighted to collaborate with him on future projects.

There have been a number of other important people who have been involved in various aspects of this project. Professor Larry Egan's group were closely involved in our work on ulcerative colitis and I would like to thank him and his colleagues Lori Hartnett and Raja Affendi Raja Ali (who generated the microarray and HELP-seq data). I would like to thank Dr. John Newell who provided some crucial statistical advice that made a big difference to my final chapter. I would like to thank Dr. Stephanie Huang who provided us with important data and some very useful advice on analysis. Last but certainly not least, I would like to thank Dr. Aaron Golden, who secured funding for the project and was originally my supervisor, before moving to the USA; he has remained involved in much of the work and has provided invaluable feedback and guidance throughout the entire course of this project.

I would also like to acknowledge my fellow bioinformatics Ph.D. students. They have always been happy to discuss my work and I have really valued their input and the exchange of ideas. There are obviously very bright futures for the current group of students and I look forward to collaborations with them in the years to come. I would also like to thank the other Ph.D. students in the Maths Department, who have always been easy to get along with and provided a relaxed and fun working environment.

Finally, I would like to thank my parents. Since my early childhood they have encouraged my interest in science and computing and have always supported (fi-

nancially and otherwise) my desire to pursue further education. I owe them an enormous debt of gratitude and promise that I will never forget the sacrifices that they have both made on my behalf.

Chapter 1

Introduction

1.1 Gene regulation

The regulation of gene expression is a complex process and is still only beginning to be fully understood. Gene expression can be modulated at any stage, from transcriptional control to post translational protein modification and degradation [13] [1] (Fig. 1.1). Expression levels can directly determine the phenotypic traits of a cell (e.g. cancer or normal) and thus, the molecular mechanisms controlling expression are key to all aspects of biology [14]. Much of the work in this thesis is based on high-throughput gene regulation and gene expression data; hence, this section will review some of the most widely studied aspects of gene regulation, with particular emphasis on microRNAs (miRNAs) and DNA methylation, which are discussed in chapters 2, 4 and 5.

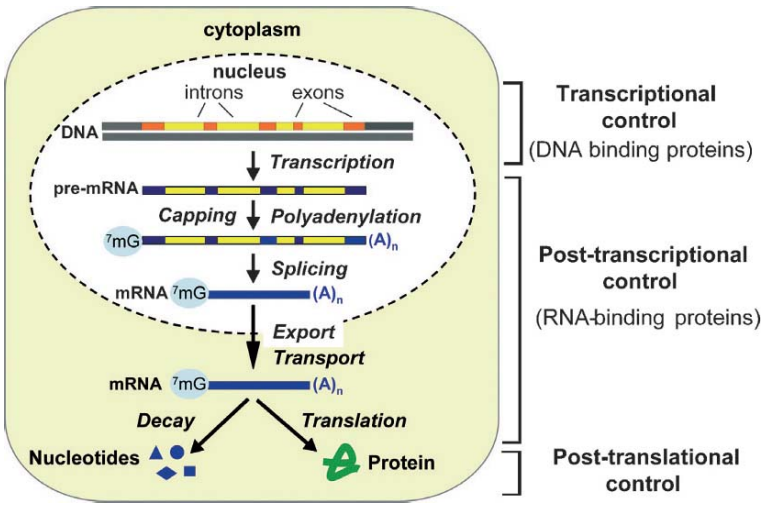

Figure 1.1: Types of gene regulation. Sourced from [1].

1.1.1 Transcriptional control by DNA binding proteins

Transcriptional control is the first possible stage of gene regulation and it is particularly important, as it is only at this point that a cell can ensure that it does not create superfluous transcripts [2]. In eukaryotes, protein coding genes are transcribed by the enzyme RNA polymerase II, which is aided by a set of proteins called general transcription factors. These are required for transcription of almost all genes [15]. General transcription factors are responsible for steps such as recognizing the transcription start site, positioning the RNA polymerase and unwinding the DNA [16]. In higher organisms, transcription is a highly complex process that varies subtely from gene to gene. In fact, the human genome is thought to encode as many as 2,000 proteins which perform some regulatory function [2]. Figure 1.2 (A) illustrates how a typical genomic locus may appear when a gene is primed for transcription. General transcription factors and RNA polymerase II have bound to the gene promoter region and regulatory proteins known as activators have bound to up- and downstream DNA sequences called enhancers. In some cases, these regulatory sequences may be as many as 50 kilobases [17] from the gene which they control and the expression of one gene is often influenced by many regulators [15].

DNA looping allows these regulatory proteins to interact with the proteins assembled at the transcription start site (TSS). In figure 1.2 (B) the DNA has looped to allow the regulatory proteins to initiate transcription, through the intermediary of a protein complex called Mediator [18]. Many regulatory proteins interact with the TSS through the Mediator complex, but some others directly influence RNA polymerase and/or general transcription factors [2]. There also exist a class of repressor proteins, which bind to similar regulatory sequences, but are associated with inhibition of transcription [19]. Differential expression of these enhancer and repressor proteins can have an effect on the expression of the genes that they regulate [20]. In eukaryotes, DNA is wrapped around histone proteins (described below) and this further complicates the transcription process, as transcription cannot occur until these regulatory regions can be bound by the necessary proteins. This is discussed in detail in the following subsections.

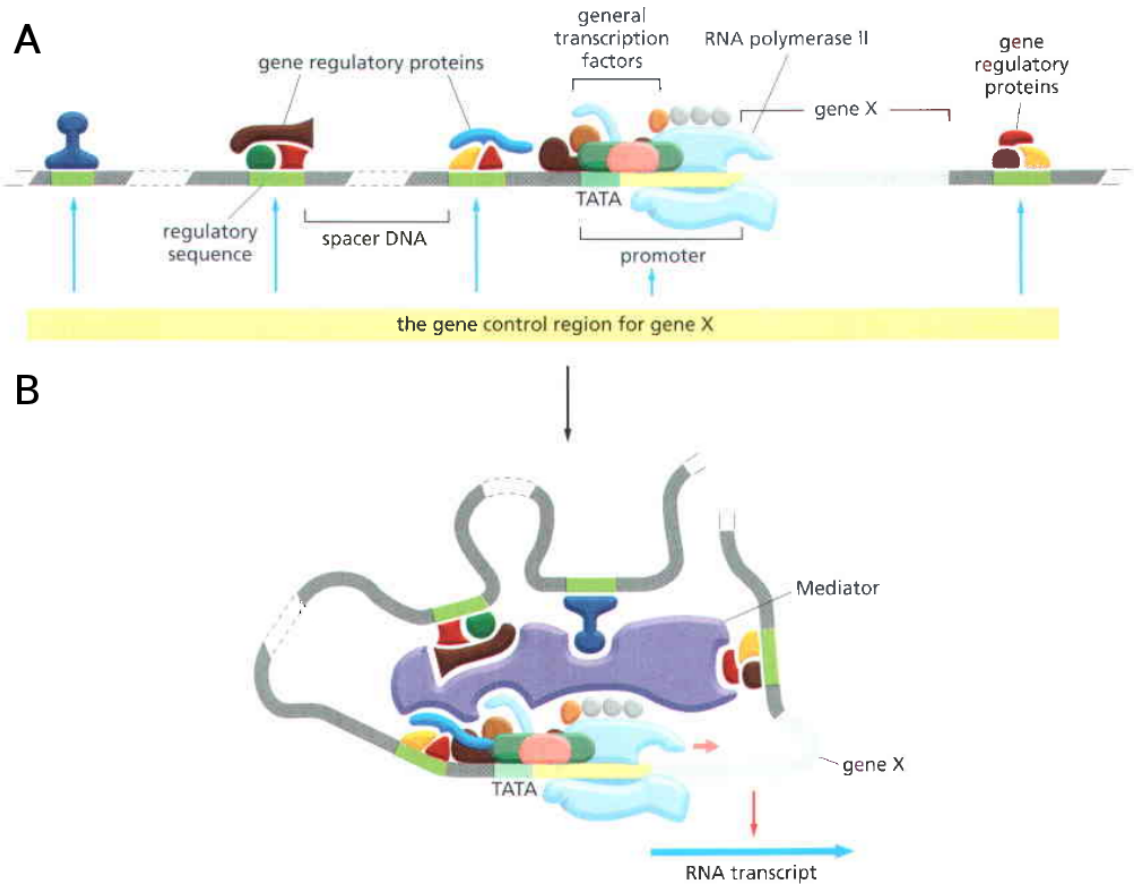

Figure 1.2: Transcriptional regulation by DNA binding proteins. Sourced from [2].

1.1.2 Epigenetics

The mechanisms that control transcription factor binding in gene promoters include histone modification and DNA methylation. These are a class of epigenetic modifications, which means that they may be passed to daughter cells and thus, these types of modifications in germ line cells can cause heritable changes in gene expression (and hence phenotype), without changing the underlying DNA sequence [21]. Epigenetic modifications have been linked to a number of human diseases, for example obesity [22] and cancer [23]. Because epigenetic changes are more easily reversed than genetic mutations, they are of particular interest as therapeutic targets, with a number of drugs already developed, for example, for the treatment of leukemia [21]. The following subsections discuss these regulatory mechanisms in detail.

Histone modification

Histones are a highly conserved class of protein that form the building blocks of nucleosomes; in eukaryotic cells, chromosomal DNA is wrapped around nucleosomes into a compact form called chromatin. This allows the DNA strands, which would normally total about 1.8 meters in length (in humans), to be stored in a tiny space in the cell nucleus [24]. Proteins H2A, H2B, H3 and H4 are known as the core histones and typically, 2 of each of these proteins assemble into a single nucleosome complex, along with a H1 protein, which binds to the assembled nucleosome to lock the formation in place [25].

Histone proteins are subject to post translational modifications, including methylation (not to be confused with DNA methylation), phosphorylation, acetylation and ubiquitation [26]. Modifications at the N-terminal histone tails are particularly important in controlling gene expression and have been extensively studied. These tails protrude from the body of the nucleosome and their modification can regulate the ability of the transcriptional machinery to access the underlying DNA and thus, regulate gene expression [27]. There are many examples of specific types of histone modifications and their affect on activation and repression of transcription; for example, methylation of H3K4 at promoters is associated with active transcription [28] and H3K27 tri-methylation has been linked to transcriptional repression [29].

DNA methylation

DNA methylation is another type of modification, that is also associated with changes in gene expression. It typically involves the addition of a methyl group to a cytosine nucleotide located in the context of a CpG site [30][31]. This results in two 5-methylcytosine (5mC) located diagonally across from each other on opposite DNA strands [32]. Methylation is primarily mediated by DNA methyltransferase (DNMT) enzymes. These include DNMT1, which is primarily responsible for maintenance of a methylated state, and DNMT3A and DNM3B which are responsible for *de novo* methylation, although there is though to be overlap between these roles [33] [34].

In some parts of the genome, particularly gene promoter regions, CpG sites tend to cluster in higher concentrations to form CpG islands [35] (normally defined as regions of length at least 200bp with GC content $> 50\%$ and observed/expected CpG ratio $> 60\%$ [36] [37]). The promoters of more than half of protein coding genes contain a region which satisfies this definition [38]. Methylation is most widely studied in this context and it has been observed that CpG sites within CpG islands, tend to be unmethylated approximately 90% of the time in healthy tissue [39][40]. Methylation of the promoter region of a gene is normally associated with silencing of expression; it is thought that this occurs either by directly

blocking the binding of transcription factors or by recruitment of methylated CpG binding proteins, that are associated with changes in chromatin structure, thus blocking transcription initiation [41] [36]. In healthy tissue, genes repressed due to methylated CpG islands are normally subject to long term silencing, for example, imprinted genes or genes that are only expressed during embryonic development [38]. Aberrant CpG island methylation has been observed in many diseases, including cancer [42][43], where it has been shown to be associated with silencing of tumor-suppressor genes [44][45][46].

The mechanisms that cause *de novo* DNMT enzymes to target particular CpG sites are not well understood [38], although recent research suggests that, at least in some circumstances, methylation is used as a type of "lock" to reinforce a silenced state that has already been induced by chromatin remodeling [47] (although this was observed in the context of X chromosome inactivation). Other observations of cellular dynamics during differentiation suggest that methylation itself is directly affecting expression, although further work is required to fully understand these processes [38].

In the remainder of the genome (outside of CpG islands), CpG sites are underrepresented, as a result of a deamination process, by which methylated cytosine can be spontaneously or enzymatically converted to thymine, thus altering the DNA sequence [48]. Despite this, methylation at these sites is still thought to play an important role. The study of methylation of CpG sites within gene bodies has recently revealed that these sites also affect gene regulation. However, gene body methylation is not associated with silencing of expression [49]; in fact, there have been reports of a positive association between gene body methylation and gene expression levels in human, plants and animals [50][51][52]. One confirmed function of gene body methylation is blocking transcription of parasitic DNA such as retrotransposons, whereby the methyl tags block transcription of these elements but allow for elongation of the host mRNA [53]. High throughput sequencing assays have shown that exons are more likely to be methylated than introns [54] and it has also been suggested that gene body methylation affects alternative splicing [55]. Recently, it was found that methylation affects CTCF (a transcriptional represser protein) binding on exons, which slows RNA polymerase II elongation, allowing the spliceosome time to recognize splice sites, leading to differential exon inclusion; thus for the first time elucidating one of the mechanisms by which gene body methylation influences alternative splicing [56].

CpG sites in intergenic regions also tend to be methylated in normal tissue. As with gene bodies, this is thought to prevent transcription from parasitic DNA. However, in cancer cells, the genome (outside of CpG islands) undergoes widespread hypomethylation and the absence of these methyl tags is associated with genomic instability [41]. This widespread genomic hypomethylation causes an increased tendency for mutation and is thus thought to be an important driving

force in cancer [57][12].

1.1.3 Post-transcriptional gene regulation

In complex organisms, there is an additional layer of gene regulation, which acts after RNA polymerase has begun RNA synthesis. This is known as post-transcriptional regulation and these mechanisms allow further fine tuning of gene expression levels, between transcription and translation [13]. Here, we will discuss some of the mechanisms which are currently of most interest to researchers.

miRNAs

microRNAs (miRNAs) are a class of small non-coding RNA molecule of approximately 21 nucleotides in length, that regulate gene expression. They typically bind to complementary loci (known as the seed region) in the $3'$ untranslated region (UTR) of target mRNA and prevent translation to mature protein. An individual miRNA can regulate the expression of hundreds of genes. Some genes, particularly those with longer $3'$ UTRs, are often the targets of multiple miRNAs and miRNA mediated regulation tends to result in the fine tuning of the expression of many proteins within a cell [58][59]. In mammals, miRNAs are thought to regulate the expression of as many as 50% of protein coding genes [60]. miRNA expression impacts on almost every cellular process and miRNA dysregulation has been implicated in many pathologies [61][58].

miRNAs regulate a range of biological pathways associated with cancer including apoptosis [62] and cell proliferation [63]; dysregulation of miRNAs has also been widely observed in cancer [64]. For example overexpression of miR-155 has been implicated in Hodgkin's and Burkitt's lymphoma [65], while miR-15 and miR-16, which target the anti-apoptotic gene BCL2, have been shown to be dysregulated in chronic lymphocytic leukemia [66]. miRNAs have been found in many of the genomic regions associated with chromosomal abnormalities in cancer, including regions of amplification, which may contain oncogenes, regions of loss of heterozygosity, which may harbor tumor suppressor genes and fragile sites which are preferential sites for translocation, deletion, amplification, sister chromatid exchange and insertion of tumor associated viruses like human papilloma virus [67].

While many specific maturation steps have been uncovered for different miRNAs, most known human miRNAs are thought to be processed in the same way by the miRNA biogenesis pathway. This process is as follows; miRNA precursors, known as primary miRNA (pri-miRNA) are transcribed by RNA polymerase II or III. These transcripts are subsequently cleaved by the microprocessor complex DROSHA-DGCR8 to form the pre-miRNA, which is transported from the

nucleus to the cytoplasm by XPO5-RAN-GTP. There, it is cleaved by DICER1-TRBP to form the two stranded miRNA duplex; the passenger strand is detached and normally degraded, although in some cases it acts as a separate functional miRNA. The remaining functional strand combines with E1F2C2 (AGO2) proteins and forms the RNA-induced silencing complex (RISC). The miRNA then guides RISC to prevent translation of target mRNAs. Translation is prevented by mRNA deadenylation, mRNA target cleavage or translational repression [3]. Of the mechanisms of post-transcriptional regulation by miRNAs, lowered mRNA levels (mRNA cleavage or deadenylation) accounts for most (>84%) of decreased protein production [68]. This implies that it is possible to assess levels of miRNA mediated gene silencing from the mRNA levels of a miRNA's target transcripts [69] and we have made use of this observation in chapter 2.

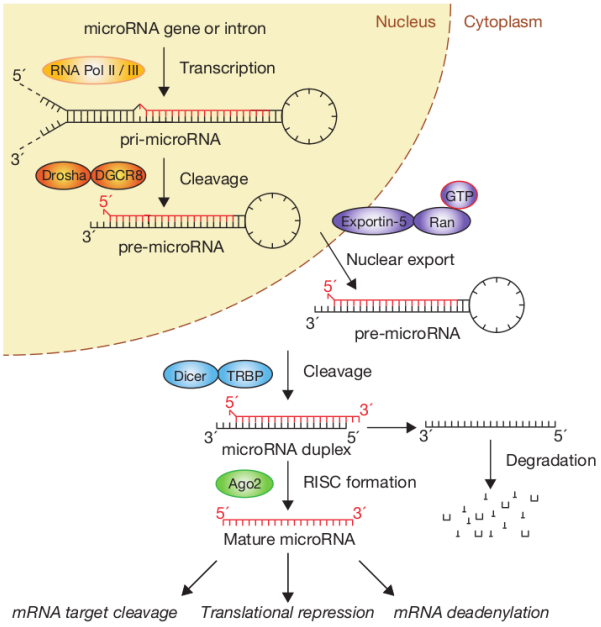

Figure 1.3: Canonical miRNA biogenesis pathway. Sourced from [3].

Several other classes of small RNAs have been discovered that regulate gene expression; of these, the most widely studied are siRNAs. These were originally reported in 1998, as a class of double stranded RNA (dsRNA), which inhibit transcription of target RNA through the RNAi pathway [70]. There is considerable overlap in the mechanisms which process siRNAs and miRNAs, with DICER and AGO proteins playing a vital role in both cases [71]. siRNAs were originally thought to be primarily exogenous in origin, for example, RNA transcribed by invading viruses [70][72]. This exogenous RNA is recognized by the RNAi pathway and used to directly silence matching transcripts, using mechanisms similar

to the miRNA biogenesis pathway [73] . It has been confirmed that RNAi plays a key role in genome defense, for example, silencing viral transcripts and transcription from parasitic DNA; however more recently, this picture has become more complex, with several reports of endogenous dsRNAs being incorporated into the RNAi pathway and these RNAs have also been shown (like miRNAs) to sometimes regulate endogenous genes [74][75]. RNAi is thought to possess therapeutic potential; siRNAs tend to bind with greater specificity than miRNAs [71], which means that, in theory, they can be engineered to target the expression of a single gene and RNAi based therapeutics are expected to be commercialized within the next 5-10 years [76].

miRNA Target Prediction

In chapter 2, we used a methodology which relies on the expression estimates of a miRNA's target genes to infer miRNA activity. These types of approaches rely on miRNA target prediction algorithms, which we will briefly discuss here. The miRNA seed region is small and in animals, miRNAs tend to bind imperfectly to their target mRNAs; these factors mean that prediction of miRNA targets using simple pattern matching is impossible, as the set of matches produced would include an enormous number of false positives [77]. Thus, many miRNA target prediction algorithms have been developed, that take account of additional information to identify true miRNA-mRNA interactions. Databases of validated miRNA targets also exist, but at present, the number of verified targets is small in comparison to the likely large number of true interactions [13]. Recently, *Baek et al.* [4] used a proteomics approach to compare the accuracy of some of the most widely used prediction algorithms. Their method used quantitative mass spectrometry to measure the response of proteins after introducing microRNAs into cultured cells. TargetScan [78] performed best, followed by PicTar [79]. Thus, our work in chapter 2 primarily uses TargetScan (which is also the most widely used algorithm [69]) and here, we briefly discuss how this algorithm works. Some other target prediction tools take quite a different approach, such as GenMiR++, which applies a Bayesian data analysis algorithm to samples for which there is both miRNA and mRNA expression data available to identify miRNA-mRNA interactions [80].

TargetScan takes account of several criteria when calculating the likelihood of an interaction between a miRNA and a mRNA. First, the seed match is considered, which describes how the miRNA binds to the mRNA target 3'UTR. The miRNA "seed" is a 7 base sequence, from base 2 to 8 in the 5' end of the miRNA, which forms a complementary bond to a seed match region in the 3'UTR of target mRNA. TargetScan recognizes four different possible types of interaction. These are (in order of strongest to weakest bond):

- "8mer" which is a match to positions 2 to 8 of the miRNA followed by an "A" in the mRNA sequence.
- "7mer-m8" which is an exact match to positions 2-8.
- "7mer-1A" which is a match in positions 2-7 followed by an "A".
- "3' compensatory" which is an imperfect match to the 5' seed region of the miRNA but when a pairing to the miRNA 3' region can compensate for the nucleotide mismatch in the seed.

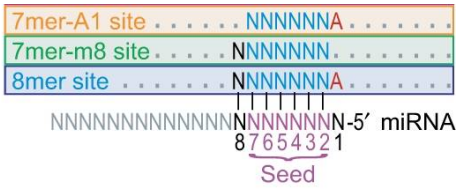

Figure 1.4: miRNA seed matches. Sourced from [4].

Because there are so many non-functional seed matches, TargetScan also calculates the "probability of conserved targeting" (PCT), which is an estimate of the level of evolutionary conservation (between human, mouse, rat, dog and chicken) of the target site, as it is known that true seed matches are more likely to be conserved across species. Finally, TargetScan calculates the "total context score", which is a measure of other characteristics, aside from the seed match, which are helpful in identifying true targets. The score is calculated as a function of five criteria, which are, AU-rich nucleotide composition near the seed site, proximity to other binding sites for co-expressed miRNAs, proximity to residues pairing to miRNA nucleotides 13-16, positioning within the 3'UTR and positioning away from the center of long UTRs [59].

Alternative splicing

Alternative splicing is the process by which a single gene can produce different RNA transcripts, and hence multiple proteins, by including/excluding different exons [81]. Splicing allows organisms to produce vastly more proteins than there are genes, thus may facilitate greater phenotypic diversity [82]. Splicing normally occurs co-transcriptionally, that is, while the gene is being transcribed by RNA polymerase [83]. The molecular machinery responsible for this process is a complex of RNA and proteins collectively known as the spliceosome [84]. Currently it is estimated that 92-94% of human genes are alternatively spliced [85] and that 15-60% of human genetic diseases involve splicing errors [86], for example increased

inclusion of exons 16 and 26 of ERBB4 in brain tissue has been shown to be associated with schizophrenia [87].

Other methods of post-transcriptional control of gene expression

Here, we will briefly describe some of the other post-transcriptional controls that occur in humans.

- Riboswitches are regions of mRNA which can bind small molecules such as metabolites and thus cause transcription to be aborted [88]. Most known riboswitches are in bacteria, but this mechanism has recently been discovered for the first time in humans [89]

- Post transcriptional cleavage of mature mRNA has been shown to sometimes produce functional byproducts, including small RNAs, coding RNAs and long non-coding RNAs. This means that cleaved RNA transcripts are not always degraded and recycled, but are also an important factor in transcriptome diversity [90].

- RNA editing is a process which can alter the sequence of RNA transcripts. Some RNA-editing is known to occur in humans [91], but how widespread this is has been a source of controversy in recent times. A 2011 study using high throughput sequencing technologies and mass spectrometry identified over 10,000 sites which were thought to be altered due to RNA editing [92]. However, this work has been widely criticized and it has been suggested by several groups that much of what was thought to be RNA editing can be explained by sequencing error [93][94].

- The decay rate of mRNA molecules is also subject to regulation, for example, by decapping enzymes, which can remove the $5'$ cap on an mRNA molecule and thus cause rapid degradation [95].

- Regulation of transcription is recognized as the primary means by which gene expression is adjusted, but there are also mechanism to regulate translation [2]. For example, translational represser proteins can bind to the $3'$ or $5'$ end of mRNA and prevent translation initiation. Sequences in the $3'$ ends of mRNA also determine to which cellular region a synthesized protein should be transported (known as localization) [96] and eukaryotic cells also possess mechanisms to modulate global levels of protein synthesis, for example in response to environmental stress [2].

1.2 Introduction to high throughput genomics techniques

The recent development of high throughput genomics techniques, such as DNA microarrays and next generation sequencing technologies (NGS) is transforming the study of biology, which is becoming an ever more quantitative science. These methods allow many thousands of simultaneous measurements, for example, the expression levels of all genes in a sample. Analysis of these data is generally highly computationally intensive, usually requiring specialized statistical techniques. As the cost of high throughput experiments drops and their applications are better understood, the volume of data produced continues to increase. This has driven the demand for novel bioinformatics approaches, which fully utilize the information captured. This thesis is focused on the development of improved methods for the analysis of these data and the application of these (and existing) methods to reveal novel biological insight. We focus mostly on problems relating to gene regulation, specifically miRNAs and DNA methylation, although much of the work is also applicable in a broader context. This Introduction chapter contains an overview of high throughput genomics platforms and a review of some of the tools, technologies and methodologies that have been used throughout the thesis.

1.3 DNA microarrays

DNA microarrays are a high throughput technology which are used to measure the quantity of many target DNA or RNA molecules in parallel. Their applications include measuring the expression levels of genes [97], genotyping [98], identification of protein binding sites [99] and quantification of small RNA expression [100]. A microarray contains many thousands of spots, each of which constitute millions of single stranded DNA oligonucleotides called probes. The sequence of these probes is complementary to the specific sequence which the spot targets; thus, the targeted sequence will tend to hybridize to a particular spot, forming covalent bonds. These target molecules will have first been labeled with a fluorescent dye. The level of fluorescence at each spot is thus indicative of the number of target molecules which have hybridized and hence, their abundance in the original sample. After sample preparation and hybridization, the florescence intensity at each spot is measured with a scanner, which outputs this information as a text file. These text files are generally the starting point for bioinformatics analysis [101] (although analysis of the raw image files is also an active area of research [102][103]).

Both single channel and two channel microarrays have been developed. Two channel arrays allow two different DNA samples to be hybridized to the same array,

where each sample is identified by labeling with a different colored fluorescent dye. These types of array are popular in applications like ChIP-chip or DNA-methylation analysis, where the level of DNA bound to a particular protein or methyl tag is usually compared to the level of all genomic DNA, known as "input DNA" [104]. Single channel arrays allow only one sample to be hybridized to each array and these are the more common choice in gene expression experiments. At the time of writing, the Affymetrix GeneChip, which is a single channel array for estimating gene expression, has been by far the most widely used platform, with expression estimates from over 100,000 human samples deposited in the online repository GEO.

1.3.1 The Affymetrix GeneChip and Exon microarrays

The GeneChip contains 1,000,000 distinct oligonucleotide features and each of these spots contains millions of copies of a particular 25 base DNA oligonucleotide. Each gene is typically (but not necessarily) targeted by 11 pairs of probes, collectively termed a probeset. This set of probes contains 11 perfect match (PM) probes, which are exactly complementary to a locus in the $3'$ region of the target mRNA. Each PM probe has a corresponding mismatch probe (MM), which contains the same 25 base sequence as the PM probe, except that the middle base, is substituted for its complement; for example, a G in the 13th base of a PM probe will be replaced with a C in the matching MM probe [105]. This allows estimation of the level of non-specific binding, which occurs when non-targeted mRNA binds to a PM probe [106].

Figure 1.5: Affymetrix GeneChip microarray shown with a matchstick for scale. Sourced from [5].

Affymetrix Exon arrays are a more recent technology, which allow measurement of expression from individual exonic regions. These arrays contain over 6

million probes and allow for the study of both gene expression and alternative splicing. Typically 4 probes (known as a probeset) target each exon with an average of approximately 40 probes for each gene (known as a metaprobeset). Non-specific binding is estimated using a set of background probes, which do not target a known exonic sequence [107]. In chapters 2 and 3 of this thesis, we have used of a set of 178 Affymetrix Human Exon 1.0 ST array samples, that were used to measure gene and exon expression in the lymphoblastoid cell lines of the International HapMap project.

1.3.2 miRNA expression arrays

Microarrays are also widely used to detect miRNA expression levels, with Affymetrix, Illumina, Agilent and Exiqon all having developed products. In chapter 2, we have used data generated from Exiqon miRCURY^{TM}LNA arrays to assess miRNA expression in the HapMap cell lines. These arrays use Locked Nucleic Acid (LNA) probes, which is a modified type of RNA that forms a more stable bond with miRNA than standard DNA probes, allowing for more accurate expression measures [108]. Other than this, the processes of labeling the sample with a fluorescent dye, hybridizing to the array and reading results with a scanner are similar to those of gene expression arrays. Analysis and interpretation of the results are also very similar.

1.3.3 Tiling arrays

Tiling arrays are similar in design to gene expression arrays, but instead of targeting sequences matching expressed mRNAs, their probes match sequences on the genome itself, such as promoter regions or even the whole genome [109]. Their utility depends on how the DNA hybridized to the array is isolated. For example, the ChIP-chip protocol uses an antibody to isolate DNA bound to a specific protein of interest. A tiling array can then be used to compare the level of this protein bound DNA to input DNA across all regions which are represented on the array [110], which gives an indication of where on the genome the protein is bound. Other examples of tiling array applications include analysis of structural variation in genes, such as copy number variation [109]. In chapter 4 of this thesis, we used a type of tiling array, which targets CpG sites located in CpG islands. This was used to assess differential methylation in sigmoid colon tissue of individuals suffering long standing ulcerative colitis. Methylated DNA was isolated using methylated DNA immunoprecipitation (MeDIP), which is similar to ChIP-chip, but uses an antibody which targets 5-methylcytosine (5mC). These arrays are also the main focus of chapter 5, where we describe a method to correct a severe bias that occurs when gene set analysis (GSA) is applied to these data.

1.3.4 Quality issues in microarray data

Several quality issues commonly arise in microarray experiments, which need to be addressed to ensure that results are reliable and reproducible [111]. Quality control may result in the exclusion of certain samples from further analysis, spots/genes being filtered out, or in extreme cases, the experiment being discarded altogether. The first issue is the phenomenon of ambient florescence surrounding each feature. This fluorescence is additive to the real signal at each spot, thus causing inaccurate measurement of the true intensity level. The causes of this include binding of labeled RNA to the glass surface of the array, contamination from the wash stage or even noise introduced by the scanner itself [112][113]. Some level of background noise affects all types of microarrays; in fact, even if labeled sterile water is hybridized to an array, some fluorescence is still measured [114]. The process of removing these signals is called "background correction" and there have been many of algorithms developed.

The most basic method of background correction is to subtract the local background estimates (returned by the scanner) from the corresponding intensity value for each spot; however, this method is not widely used and has even been shown to produce less reliable results than when no background correction is applied [115]. One popular method is implemented in the widely used Robust Multi-array Average (RMA) algorithm (which provides background correction, normalization and summarization in one function) [116]. This method works by assuming that the observed signal is the convolution of a normally distributed background and an exponential foreground signal [117]. A similar method has also been adopted for two channel arrays and tests comparing the performance of background correction algorithms (using spike-in control data), have shown this to be one of the most effective approaches [115].

The next common issue is differences in the overall brightness of microarrays in the same experiment. Reasons for this include slightly different quantities of starting DNA/RNA in samples or differences in detection efficiency between dyes used [118]. To allow biologically meaningful comparison across arrays, these differences must be corrected. As with background noise, this issue affects all microarray platforms and many different algorithms have been developed to address the problem. Severe scaling differences between one or more arrays could indicate a hybridization problem, that may not be correctable with normalization; hence it may be necessary to remove these arrays from the analysis [119]. The most convenient means to assess if severe scaling differences are evident, is to view boxplots of log-transformed probe level expression intensities, before and after normalization (Fig. 1.6). In the case of two channel microarrays, it may be necessary to visualize these data for both channels separately and for the log intensity ratios.

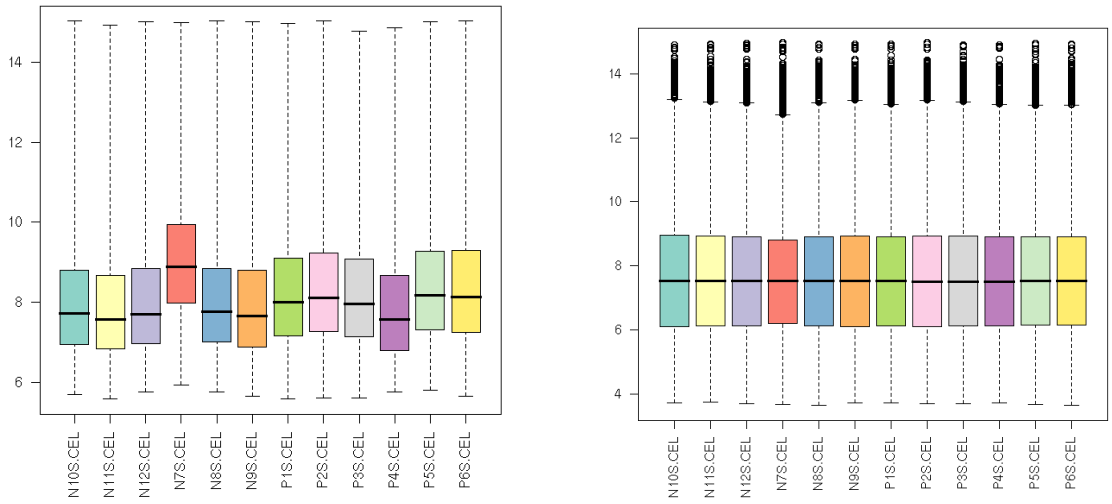

Figure 1.6: Boxplots of raw log_2 transformed probe intensity values (left) and boxplots of RMA preprocessed log_2 intensities (right) of the same 12 Affymetrix GeneChip arrays. Normalization has adjusted the small scaling differences between the arrays.

The most commonly used normalization algorithm is quantile normalization [120], which is also implemented as part of the RMA procedure [116]. This algorithm coerces the data in such way so that the gene at any particular rank, has the same expression level in all samples. This is achieved by calculating the mean expression level of the Nth ranked gene across all samples and then setting this value as the expression level for the Nth ranked gene in every sample. The algorithm can be applied to any type of expression matrix, be it probe (as is the case in RMA) or gene level intensities from a one channel array, or either channel or the log-ratios from a two channel array. Other popular normalization techniques include the VSN [121] and MAS5 [122] algorithms.

The next quality issue is spatial artifacts, whereby certain regions of an array have an obvious bias towards higher/lower expression levels, which can be caused by, for example, scratches, marks or poor handling of the chip [123]. These can be easily visualized by generating pseudo array images, which will highlight any obvious artifacts (Fig. 1.7). Severe spatial artifacts may lead to an array being discarded from further analysis, or a particular set of probes being filtered out.

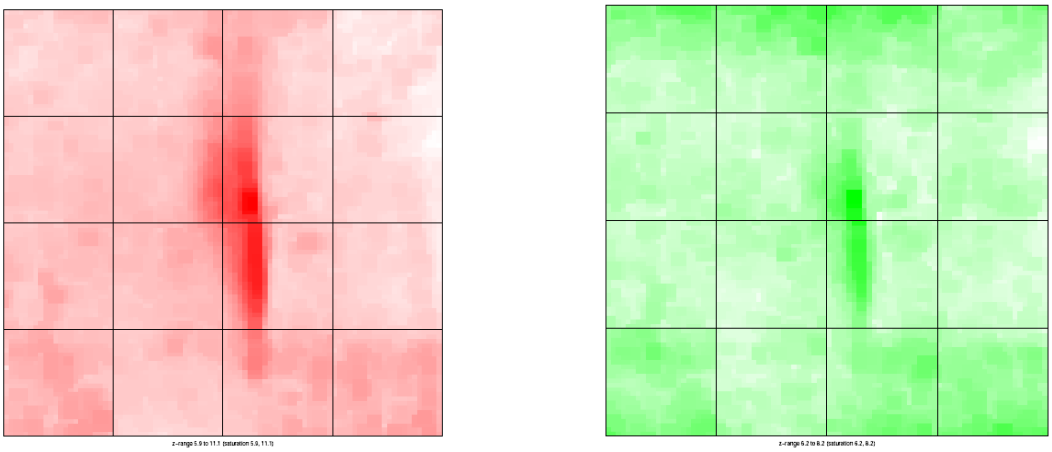

Figure 1.7: Pseudo array images of red and green channels of a two channel microarray. There is a clear spacial artifact in the center of the array, where a group of spots have higher expression than the surrounding regions.

The next issue is clustering of samples. It is normally expected that the quantity measured (e.g. gene expression) is more similar within a particular experimental condition. For example, if a gene expression microarray experiment consists of 12 samples, 6 of which are from lung cancer tissue and 6 of which are from healthy lung tissues, it would be expected that the gene expression profiles of the 6 lung cancer samples are more similar to each other, than they are to the normal lung tissue. Two of the more commonly used tools to create a visual representation of similarity between arrays are hierarchical clustering and principle component analysis (PCA) [124]. PCA is a technique that can reduce multidimensional datasets to lower dimensions for analysis and can determine the key features of high-dimensional datasets [125]. It works by identifying the directions of greatest variance in high dimensional datasets. In the context of gene expression data, it can be used to visualize the similarity of expression profiles between different samples in two or three dimensions. Hierarchical clustering attempts to build hierarchies of clusters; there are many methods available, which are broadly classed as either agglomerative, where each observation starts as its own cluster and these are subsequently joined together, or divisive, where all observations start as one cluster and these are split recursively [126]. Examples of these types of plots for an experiment containing 12 Affymetrix GeneChip samples are shown in figure 1.8. Samples labeled "Nxx.CEL" are from one experimental condition (cases) and "Pxx.CEL" are from the other (controls). Both plots convey similar information; in both cases, the sample "N7S.CEL" groups with samples of the wrong phenotype, indicating a possible problem with this array. Typically, if a small number of samples do not cluster as expected, this may indicate a problem

with these samples and further consideration should be taken before including these samples in downstream analysis. If samples of different phenotypes do not cluster at all, this may indicate that there are either technical problems with the arrays or that a confounding variable is affecting expression to a greater degree than is the experimental condition, in which case it may be difficult to achieve statistically significant results in, for example, a differential expression analysis [127].

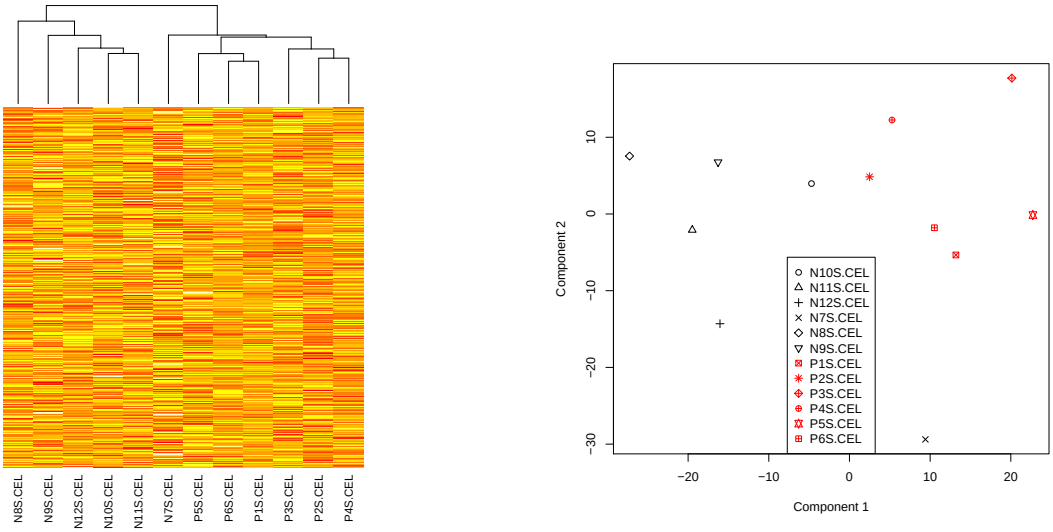

Figure 1.8: Heirarchical clustering (left) and PCA plot (right) of raw probe level data from 12 Affymetrix gene chip samples. Note non-grouping of sample NS7.CEL.

The next issue is genes whose expression cannot be reliably detected above background noise. As already outlined, labeled DNA will sometimes bind to probes for which it is not a target. This leads to a residual florescence from all probes, regardless of whether their target RNA is expressed. Methods for predicting whether a probe or set of probes are expressed above background are very much platform specific. In the case of the Affymetrix GeneChip, the MAS5 algorithm implements a method that considers probesets detected present if the expression of the PM probes is significantly higher than their matching MM probes [122]; eliminating these unreliable probesets from downstream differential expression analysis has been shown to significantly increase the ratio of true positive to false positive results [128]. Due to array design, the procedure for the Affymetrix Exon array is different. In this case, Affymetrix have included an explicit set of background probes on the array, with approximately 1000 of these probes per GC content count [129]. The Affymetrix "detected above background" (DABG) algorithm can be used to estimate the likelihood that a probeset's expression is

detectable above background, by comparing the probeset's signal intensities to those of the background probes (on the same array) with the same GC content [129].

Here, we have outlined many of the common quality issues, that are broadly applicable to most microarray platforms. However, at the time of writing, experiments using 765 different types of microarray platform from Affymetrix alone have been uploaded to GEO. The issues outlined here aside, a diligent researcher should always take account of the specific consideration of the platform being used and the experiment undertaken. For example, Affymetrix have released a whitepaper [130] which gives specific guidelines on plots and quality metrics for Exon and Gene arrays. With less established platforms, it is particularly important to give close attention to the manufacturers guidelines and if possible, to identify previously published studies of similar design in the scientific literature.

1.3.5 Summarization of gene expression microarray data

Gene expression microarrays target transcripts using several different probes, for example, the Affymetrix Exon array typically uses 4 probes to target each exon. These fluorescence intensity measures must be combined to yield a biologically meaningful measurement that is useful in downstream analysis. This is the final step in preprocessing gene expression microarray data and is known as "summarization". At the most basic level, summarization could be achieved by simply calculating the mean or median value of all probes which target a particular gene, but as with background correction and normalization, many more complex approaches have been developed. As an example, the popular RMA preprocessing function uses the median polish algorithm. This summarizes the expression of a gene, while taking account of probe and chip effects. The starting point is a matrix containing the gene's constituent probe expression levels (rows), in each sample (columns), then the following steps are applied [131].

1. Calculate the median for each row (the row effect). Subtract this value from each value in the row.

2. Calculate the median of the vector of row effects. Record this value as the "overall effect", then subtract this value from each of the row effects.

3. Calculate the median for each column (the column effect) then, for each column, subtract the column effect from each value in the column.

4. Calculate the median of the column effects and add this value to the overall effect, then subtract this value from each of the column medians.

5. Continue the above until the change in row/column medians drops below a predefined threshold, or a particular number of iterations has been completed.

6. The expression level for each gene, in any particular sample, is then calculated by the overall effect plus the column effect for that sample.

In chapter 2 of this thesis we propose a novel summarization method called *seqArray*. This method uses samples for which both RNA-seq and Affymetrix Exon array data are available, to build statistical models which learn the relationship between probe level gene expression, as measured by the microarrays, and gene level expression, as measured by RNA-seq. These models can then be used to estimate gene expression on separate sets of microarray samples.

1.3.6 Analysis of summarized microarray data

Typically, preprocessing of raw gene expression microarray data transforms a set of probe level fluorescence intensities into a matrix of normalized gene level expression estimates, which are the starting point for analysis. The most common type of assay is differential expression analysis, whereby gene expression levels in one set of samples are compared to those in another. For example, comparing gene expression between lung cancer samples and normal lung tissues could provide insight into which genes play a role in tumorigenesis and/or tumor function and hence potentially lead to novel interventions. Again, there are a large number of tools for these types of analysis, some of the most popular include the Bioconductor package *limma* [132] and the standalone package Significance Analysis of Microarrays (SAM) [133]. Some of the tools which we have used for summarization of gene expression and differential expression analysis are discussed later in this chapter. Gene expression microarrays have also been widely used for class prediction, whereby an algorithm can learn which gene's expression characterize, for example, a cancer subtype [134], or response to a treatment [135]. The most popular program for this type of analysis is "Prediction Analysis for Microarrays" (PAM) [136]. The high throughput nature of microarray expression data has also allowed the inference of gene regulatory networks [137].

There are many other interesting applications of gene expression arrays, particularly when these data are integrated with other microarray platforms. For example integration of gene expression microarray and SNP array data have lead to dramatically increased understanding of the genetic basis of gene expression [138]. Integration of gene expression and other types of tiling array platforms has lead to many more insights, for example the interaction between CpG island methylation, genetic polymorphisms and gene expression on a genome-wide scale [139].

1.4 High throughput DNA sequencing technologies

DNA sequencing is the process of determining the sequence of nucleotide bases in a DNA molecule. Until recently, Sanger sequencing (also known as chain-termination sequencing) had dominated this field for over 30 years and was a crucial technology in the completion of the Human Genome Project [140]. The method relies on radioactively or fluorescently labeled chain terminating nucleotides, which when applied to a PCR amplified DNA segment, results in a library of DNA fragments of varying length ending in one of the four nucleotides at each position; these fragments can be size separated by gel electrophoresis and the sequence read [141]. While a major advance over previous methods, by modern standards, Sanger sequencing is expensive and slow. A further notable advance was made in 1990, with the introduction of DNA sequencing by capillary electrophoresis, which drastically reduced the time required to separate DNA fragments (previously done by gel electrophoresis); the use of these capillary gels in electrophoresis and detection increased the speed of sequence analysis by over an order of magnitude [142]. A major advance occurred in 2005, when 454 Life Sciences released their new parallelized version of pyrosequencing, which reduced sequencing costs 6-fold compared to automated Sanger sequencing [143]. Other companies have since begun to market similar non-Sanger based high-throughput sequencing products; these platforms vary in characteristics like read length, read accuracy, production of paired-end reads and cost [144], but all share the ability to sequence a large number of DNA molecules in parallel. A typical run produces millions of sequence reads in only a few hours, dramatically increasing throughput and decreasing cost when compared to traditional approaches. Collectively these are frequently referred to as "High throughput sequencing" (HTS) technologies. The terms "Next generation sequencing" (NGS) or "deep sequencing" are also often used to describe these platforms. The rapid uptake of HTS poses many bioinformatics challenges, such as data storage, analysis and interpretation [145]; current HTS technologies produce reads which are shorter than those from previous Sanger sequencing experiments, which means that the algorithms used in their analysis are often considerably different, although newer HTS technologies are beginning to produce longer reads [144]. The continually growing rate of production of sequence data by centers all over the world means that this area will be a focus of research for the foreseeable future.

HTS is generally applied to either re-sequencing problems, where the read fragments are aligned to a reference genome and hence requires less coverage, or *de novo* sequencing, where a set of contigs or a whole genome is constructed without prior knowledge of the underlying sequence [143]. These technologies have

been applied to several types of problems, for example whole-genome sequencing [146], detection of DNA binding by proteins (ChIP-seq) [147], DNA methylation analysis [148] and sequencing of the transcriptome (RNA-seq) [149][85].

1.4.1 Illumina sequencing

At the time of writing, the Illumina Genome Analyzer has been by far the most widely used HTS platform (based on the number of samples uploaded to GEO). It was used to generate the RNA-seq data in chapters 2 and 3. Illumina provide a standard set of protocols for sequence library preparation [150], an overview of which is presented in figure 1.9, although researchers have found that there are adjustments that can be made to optimize these steps [6]. With RNA-seq, the first step is reverse transcribing extracted mRNA to its DNA complement. In the case of genomic sequencing, this step is unnecessary. DNA is then fragmented (usually by sonication), end-repaired, adenylated and adapter oligos are ligated to both ends of the molecules. These sequences are then size selected using gel electrophoresis [6] and the resulting size selected fragments are normally PCR amplified.

Prepared DNA is loaded onto a flowcell, which is a small glass slide (Fig. 1.10). A flowcell normally contains 8 separate lanes and a large number of oligonucleotides are bound to the surface of each lane. When the prepared DNA library is loaded, individual DNA molecules bind to the lawn of oligonucleotides on the flowcell (Fig. 1.11) [6]. Each of these molecules is extended and copied. Each copy is then bound to the flowcell surface, this process results in millions of unique clusters, with many copies of the same DNA fragment in the same cluster (Fig. 1.12). The reverse strand of each DNA molecule is removed and a sequencing primer is hybridized to the remaining strands. All clusters are now sequenced simultaneously. Each of the four nucleotides is fluorescently labeled with a different color and loaded onto the flowcell. During each round of sequencing the four nucleotides compete with each other to bind to the next position of each of the DNA templates, following which, the clusters are excited with a laser. The color of the florescence reveals the newly added base. The fluorescent label and blocking group are then removed and the process is repeated until the required number of bases have been sequenced [151].

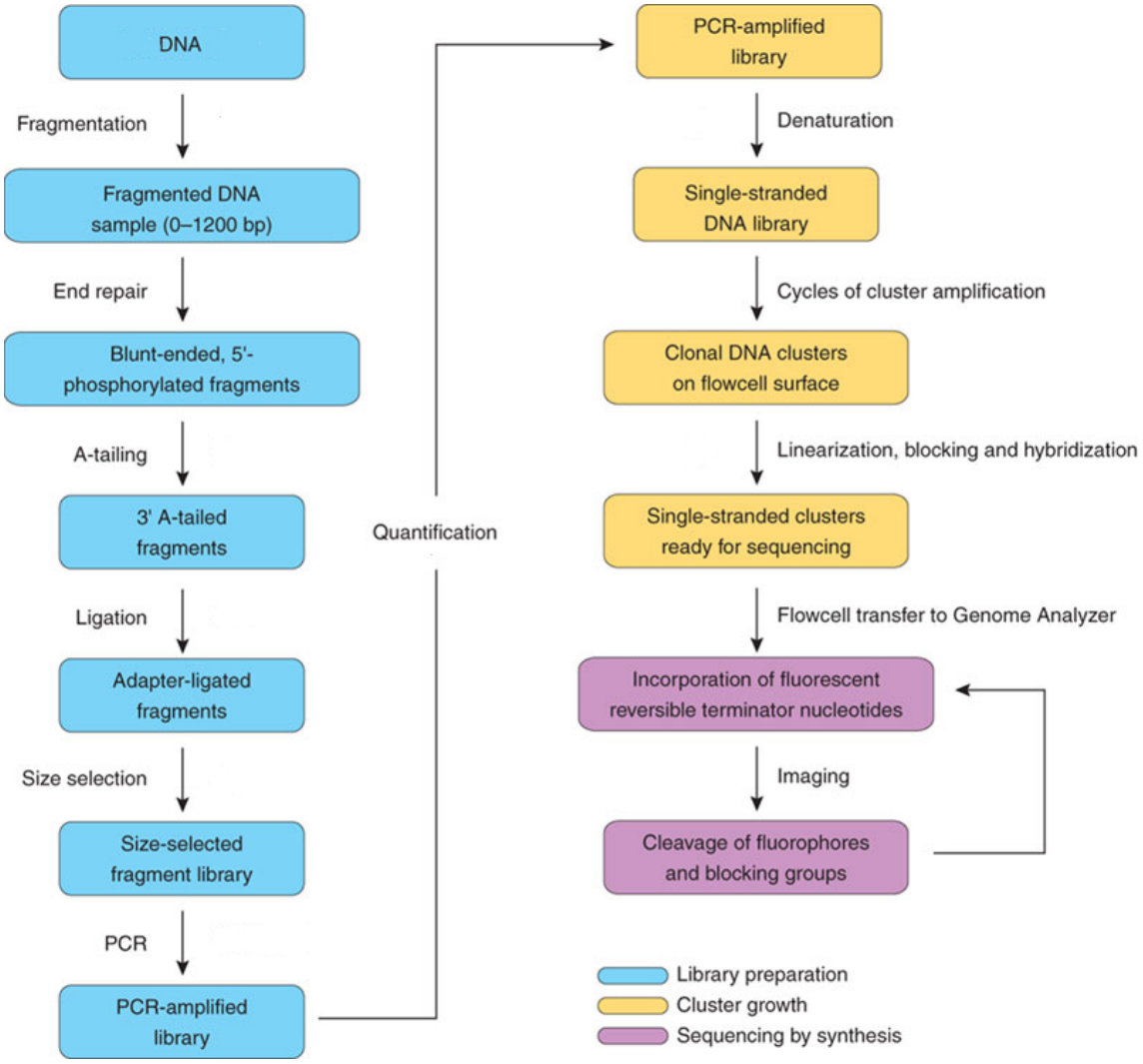

Figure 1.9: Typical Illumina Genome Analyzer sequencing workflow. Sourced from [6].

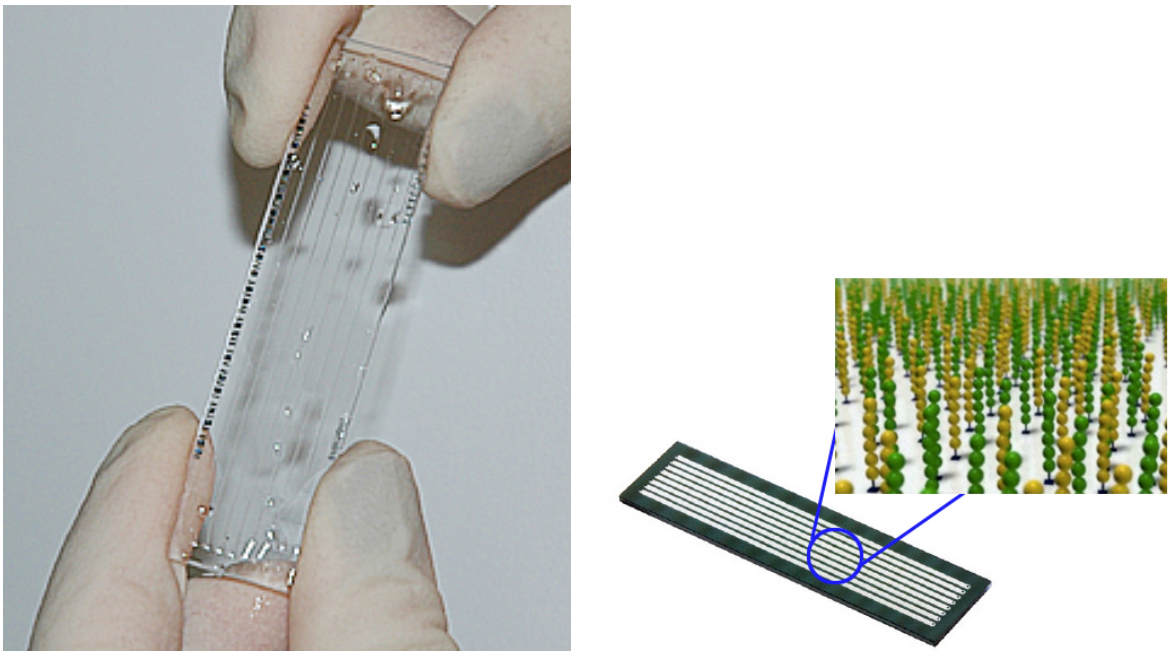

Figure 1.10: Photograph of an Illumina flowcell (sourced from [7]) and a diagrammatic representation of the millions of oligonucleotides present on the surface of the flowcell (sourced from [8]).

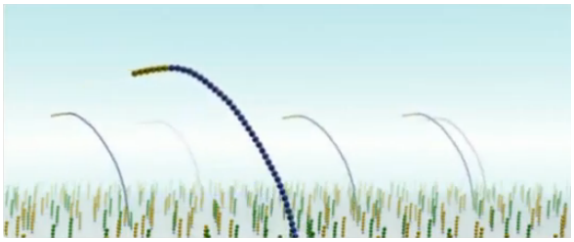

Figure 1.11: Single DNA fragments bound to the surface of the flowcell. Sourced from [8].

Figure 1.12: Clusters of DNA following bridge amplification of individual fragments. Sourced from [8].

1.4.2 Sequencing the transcriptome with RNA-seq

RNA-seq is a newly developed method of sequencing the transcriptome, to which in principle, any next-generation sequencing technology can be applied. The approach is normally used to quantify gene/transcript expression. RNA-seq overcomes many of the problems of gene expression microarrays, such as poor dynamic range, high levels of noise due to non-specific binding and the need for complicated normalizations when comparing expression levels across samples [152][153] and RNA-seq has been shown to provide a more accurate measure of absolute expression levels [154]. Unlike microarrays, which rely on previous annotations of the genome, RNA-seq can be used to discover novel transcripts in unannotated regions of the genome [153]. Reads mapping to splice junctions can be used to identify new transcripts, by determining novel connectivity patterns between exons. RNA-seq reads can also be used to identify sequence polymorphisms in transcribed regions [155].

Typically, an RNA-seq experiment begins by extracting the poly(A)+ fraction of RNA from a sample, as protein coding mRNAs normally have a poly(A)+ tail. The isolated RNA is reverse transcribed and the cDNA library is sequenced for the required number of bases, as described in the previous subsection. DNA fragments may be sequenced from one or both ends generating either single-end or paired-end reads [155]. The sequencer outputs millions of short sequence reads, which must be analysed by a computer in order to estimate gene expression levels.

For RNA-seq, the first step in data analysis is alignment of reads to the reference genome. This procedure attempts to match each short read to its most similar sequence in the genome and thus the most likely location from which the read originated. Once reads have been aligned, other tools are used to quantify gene and transcript expression, based on the locations of the mapped reads. Before and after alignment, steps should be taken to ensure the quality of the data. Read alignment, quality control and expression analysis are discussed in detail below, in subsections 1.4.3, 1.4.4 and 1.4.5. An overview of the tools used in these procedures is presented in section 1.5.

1.4.3 Aligning sequenced reads to the genome

Sequence mapping programs and algorithms have existed since Sanger sequencing, but HTS introduces new computational challenges because of the shorter read lengths and far greater number of reads [9]. Recently, there have been new tools developed for mapping these reads, which are aimed at meeting the challenges of the data. In the case of RNA-seq there is the additional challenge in mapping reads which span introns [9].

The major computational challenge faced by short-read mappers is handling

the volume of data in a practical manner. When billions of sequences are being mapped to a genome, memory and processing resources must be used highly efficiently if the data are to be processed in a reasonable time-frame using a typical desktop computer [156]. The first problem is handling reads which map equally well to multiple regions of the genome, which is inevitable given the short length of reads and the size of the the genome. The problem is exacerbated by the fact that sequencing errors sometimes occur and that the reference genome will not perfectly match the genome for the individual being mapped. Aligners use a number of different strategies to deal with this, including reporting multiple positions, picking one alignment at random or discarding multi-mapped reads[9]. Many programs will also attempt to calculate a quality score for each alignment, which gives an indication of the likelihood that a read is mapped to the correct position, which can be useful in downstream analysis.

RNA-seq reads that map to splice junctions contain information that facilitates the estimation of isoform expression and the identification of previously unknown gene transcripts. Specialized algorithms are required for this task. These spliced alignment algorithms fall into two categories, those that rely directly on known gene annotations and those that do not. Algorithms that rely on annotations are limited by those annotations, while alignment without known gene annotations are limited by the ability of the underlying algorithms generate an accurate representations of the data. Details of the algorithms used by TopHat and MAQ (TopHat is a spliced-aligner, MAQ is not) are discussed in section 1.5.

The main challenge facing sequence alignment in the near future will be the longer reads produced by newer platforms. Many of the current generation of tools are designed to handle reads of up to 100bp, but it is unclear how these algorithms will scale beyond this point. There is also currently poor support for insertions and deletions and the usefulness of mapping qualities in downstream analysis is yet to be fully explored. Ultimately however, the future of sequence mapping will be determined by the nature of the next batch of high throughput sequencing technologies produced by companies like Illumina and ABI Solid [9].

1.4.4 Quality issues in high throughput sequencing data

To ensure reliable results, it is important to quality assess sequence data. Given the divergent nature of microarray and HTS platforms, there is very little overlap in typical quality control procedures. When mapping to a reference genome, quality assessment steps should be taken before and after read mapping. Based on the results of quality control, a researcher will usually either remove a subset of bad reads from the sample, trim a number of bases from either end of reads, completely discard some or all samples from the experiment or take no action and proceed with the analysis. At the time of writing, HTS quality procedures

remain ad-hoc and many different approaches have been implemented for different platforms and types of experiments. Here, we outline some of the common known issues and methodologies.

The first issue is the number of reads in each sample, this should typically be comparable across samples, both before and after mapping (in the case of re-sequencing). If one or more samples have drastically different numbers of reads, or a much smaller proportion of reads align to the genome, this is likely indicative of a problem and further investigation should be carried out. The regions to which reads map is also of interest, for example, in an RNA-seq experiment, the majority of reads are expected to map to annotated exons. These can be easily tested by loading the data in R (or a similar programming environment) or by using a quality assessment tool like RNA-SeQC [157].

The next issue (which applies to both re-sequencing and *de novo* sequencing) is sequence degradation from $3'$ to $5'$ end. This is more commonly observed in RNA-seq and is manifested as poorer sequence quality scores and more N calls (where the sequencer was unable to make a base call with sufficient confidence) towards the $3'$ end of reads. Better quality samples will have higher and more consistent quality scores and less "N calls" across read bases. To assess the level of sequence degradation in a particular sample, one can create boxplots of sequence qualities and numbers of "N calls" across all reads in a sample (Fig. 1.13 and 1.14). There are no strict guidelines on how to deal with this issue, but for sequence quality degradation, the widely used program "FastQC" will flag a warning if the lower quartile quality score of any base is less than 10 or if the median is less than 25; a sample will fail this test if the lower quartile of any base is less than 5 or if the median is less than 20. For "N calls", FastQC will raise a warning if any position has more then 5% N content and failure for more than 20%. Commonly, if there is strong evidence that sequence quality degrades drastically, one may chose to trim the poor quality reads [158].

The next important issue is the sequence content of the reads and identifying over-representation of any of the four possible nucleotide bases. This is typically manifested as higher than expected GC content. The change in base content across reads in a sample can be plotted using FastQC or a similar program (Fig. 1.15 and 1.16). With RNA-seq, high GC content is commonly observed in the first few bases of reads, this is due to the nature of the primers used in many RNA-seq experiments [159]. When GC content is considered unacceptably high, one may again chose to trim the reads, or apply one of the normalization procedures which have been developed to correct for this [160].

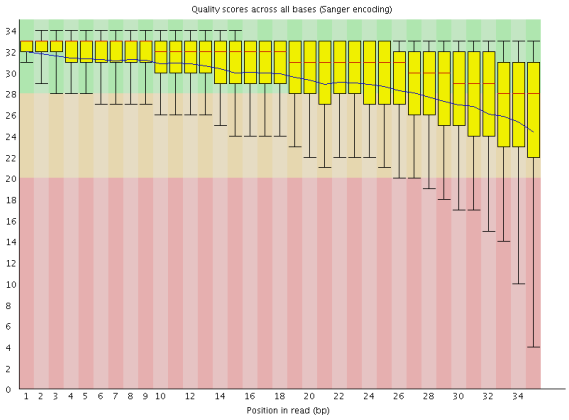

Figure 1.13: Typical per base sequence quality plot from an RNA-seq experiment. A trend towards loss of quality at the 3′ end of the reads is clear.

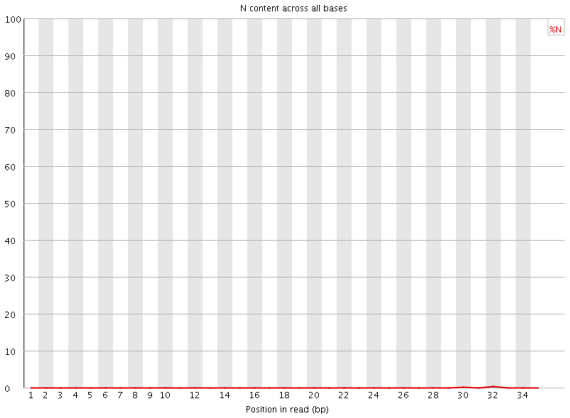

Figure 1.14: Per base N content plot from an RNA-seq experiment. On very close inspection, there is a slight wobble in the graph between positions 30 and 35, indicating an increase in N calls towards the 3′ end of reads.

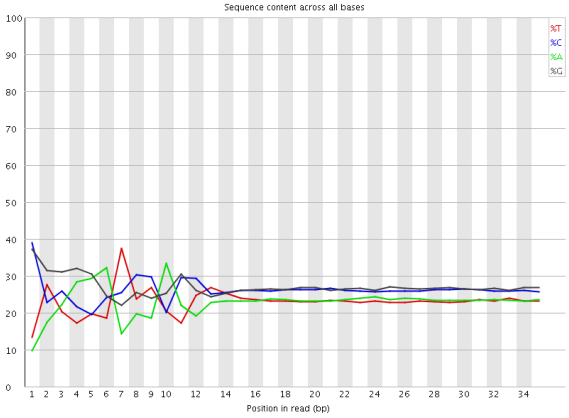

Figure 1.15: Per base sequence content from an RNA-seq experiment. This gives an indication of which bases are most likely to occur at which position of a read.

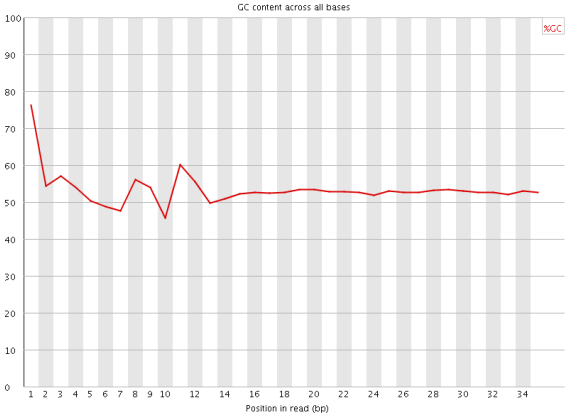

Figure 1.16: Per base GC content from an RNA-seq experiment. This shows the percentage GC content across the base positions in a sample. Higher GC content is evident in the first few bases.

The next issue is overall sequence quality in a each sample. It is expected that sequence quality scores and GC content should be similar across samples and above a certain threshold. Again, these thresholds will vary based on the type of experiment and the type of sequencing platforms used. Typically, a plot of sequence quality scores within a sample will yield a unimodal distribution (Fig. 1.17) and the mode of this distribution should be high (FastQC raises a warning if the mode of the distribution has quality score below 27 (equating to a .02% error rate) and raises failure if this is below 20, equating to a 1% error rate). A bimodal distribution indicates that there is a subset of poor quality sequences and it may be necessary to consider removing these sequences from further analysis. If all sequence quality scores in a sample are low, this may indicate a systematic problem and that sample may be omitted completely from further analysis. In each sample, sequence GC content is also expected to follow a normal distribution (Fig. 1.18); a shift to the left or right of GC count per read may indicate a systematic bias and an unusually shaped distribution may indicate the presence of some contaminant in the library.

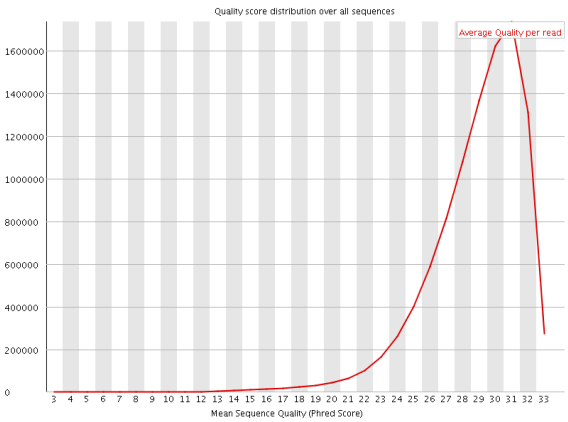

Figure 1.17: Per sequence quality plot for RNA-seq sample.

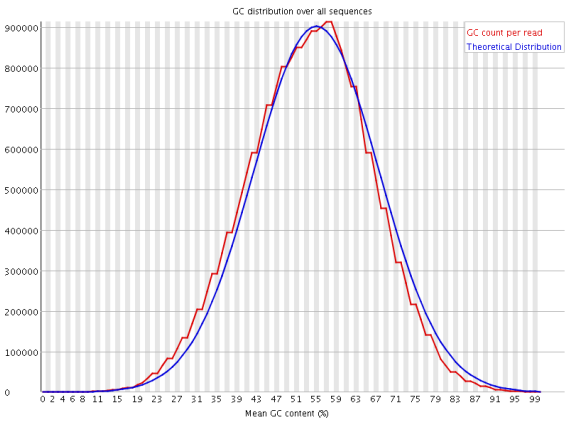

Figure 1.18: Per sequence GC content plot for RNA-seq sample.

The next issue is sequence duplication. The proportion of duplicate reads expected in a sample depends on the type of sequencing being undertaken. Normally, the majority of sequences in a library are expected to be unique. However, in most RNA-seq experiments, some level of duplication is inevitable; in order to detect expression of low copy number transcripts, highly expressed transcripts (for example housekeeping genes) must be over-sequenced, which means that, because of the large number of sequences originating from these transcripts, some duplication is expected. Very high levels of duplication may however indicate a PCR effect, whereby some sequences have been preferentially amplified. Badly

PCR duplicated libraries can produce extremely high levels of sequence duplication (> 90%) [161]. There are many tools which can be used to calculate the proportion of duplicate reads in a sample, for example SAMtools [162]; FastQC also allows the proportions of duplicates to be plotted (Fig 1.19). Samtools also implements a feature which can identify and remove potential PCR duplicates. Because of the problems introduced by PCR duplication, amplification-free sequence library preparation protocols have been developed and are beginning to gain in popularity [163]. It is also possible to further investigate exactly which sequences are overrepresented and identify sequences that may originate from highly expressed genes (in the case of RNA-seq) or possible containments; for example, FastQC has a feature which can identify the possible source of contamination by searching a database of known contaminants, such as PCR primers. Finally, a ligation bias can sometimes occur, whereby PCR primers may be more likely to bind to certain kinds of sequences, distorting the RNA profiles, these kinds of problems have been overcome by using different kinds of adapters [164][165].

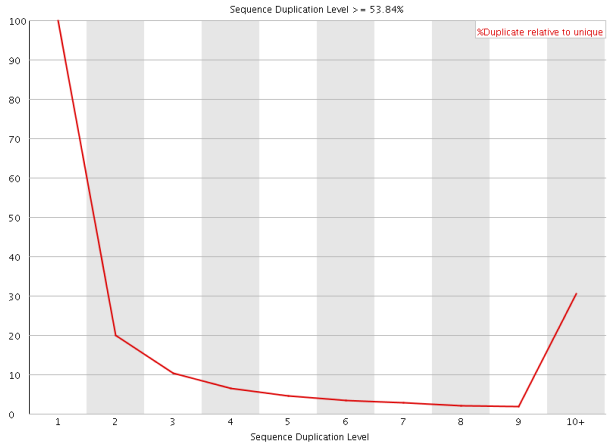

Figure 1.19: Duplication levels in an RNA-seq sample. In this case, the peak for sequences duplicated 10+ times is a result of a large proportion of reads originating from a small set of highly expressed genes.

1.4.5 Estimating gene and transcript expression from RNA-seq data

Once sequence reads are aligned to the genome, it is possible to quantify gene and/or transcript expression. In this thesis, we have used two approaches to quantifying expression in RNA-seq data. The first method is based on a custom pipeline in the programming language R; this pipeline counts the number

of mapped reads overlapping the exonic regions of each gene and normalizes using "reads per kilobase of exon model per million mapped reads" (RPKM) [152] procedure. We have also used Cufflinks [166], a standalone command line based program that can estimate both gene and transcript expression. Cufflinks and splice-junction read mapper TopHat (both tools developed by the same group), can be used together to identify novel transcripts and splice isoforms. The tools can also work from a known gene model. These are discussed in more detail in the next section.

1.5 Tools and technologies used in data analysis

1.5.1 R/Bioconductor

R is a free open-source software environment, for statistical computing and graphics, released under GNU General Public License (GPL) [167]. It runs on a wide variety of UNIX platforms, Windows and MacOS. The design of R has been heavily influenced by two existing languages, S and Scheme [167]. The main appeal is its functionality for statistical procedures, such as linear models, nonlinear regression models, time series analysis, classical parametric and nonparametric tests, clustering and smoothing. There is also a powerful graphical environment for creating publication standard plots [168]. The main disadvantage of R is that it is sometimes slow; as with any interpreted programming language, it carries much greater overhead than a compiled language like C. This means that for some tasks, for example looping operations, it can be 100s of times slower [169], although these types of bottlenecks can be avoided by coding them as functions in C, which can be called directly from R.

Bioconductor is an open source collection of R libraries that provide tools for the analysis of genomic data [170]. It is widely used in bioinformatics and support is provided for hundreds of different microarray platforms and many different types of sequence data. Examples of the types of analysis include data normalization and summarization, clustering, differential expression analysis, gene annotation, data visualization and gene set analysis. Bioconductor is constantly updated with new materials, such as novel analysis procedures and pipelines; all of these improvements are released free of charge as part of the biannual releases of R [171].

R and Bioconductor have been the primary environment used in all data analysis in this thesis. We have, on occasion, built external functions in C; Python has also been used for some tasks. For computationally intensive tasks, it has also been possible to split R jobs across multiple processing nodes of a *Beowulf* cluster, using the *Rmpi* [172] library. Other Bioconductor packages used extensively include *affy* [114] and *limma* [173] for microarray analysis, *GOstats* [174] for gene

set analysis and *GenomicRanges* [175] and *GenomicFeatures* [176] for sequence analysis.

1.5.2 Tools for Microarray analysis

The limma package

Limma [132] is an R library, which is part of the Bioconductor project and is used primarily for differential expression analysis of gene expression microarray data. It is based on linear models and Bayesian statistics. The procedure first fits a linear model to the expression level of each gene, dependent on phenotype. Next limma uses a method called Empirical Bayes shrinkage, which computes p-values for each gene, by shrinkage of the standard errors towards a common value, which is estimated borrowing information from the expression levels of all genes. This method has the advantage of providing a stable result, even when the number of arrays in an experiment is small [132]. The package includes an extensive collection of other functions for analysis of microarray data, including background correction, normalization and visualization.

Affymetrix Powertools

Affymetrix Powertools (APT) is a standalone command line based application, developed by Affymetrix. It is used for preprocessing of data from their various gene expression microarray platforms. As Bioconductor support for the Affymetrix Exon microarray is relatively poor in comparison to other platforms, in chapters 2 and 3, we have used APT for preprocessing of these arrays. APT provides the RMA preprocessing and DABG algorithms (discussed earlier in this chapter) and, for Exon arrays, is capable of summarizing expression estimates at gene or exon level. These results are output to flat text files, which can be subsequently imported and analysed using R or other tools.

1.5.3 Tools for aligning RNA-seq data

MAQ

MAQ [177] is a popular aligner that can be used with single or paired-end short sequences. It aligns reads to a reference genome for subsequent analysis (for example quantification of gene expression) and can also infer variants like SNPs and indels. It was the first short read aligner to incorporate the concept of a quality score; instead of discarding poorly aligned data, the probability that a read is mapped to the correct position is calculated and retained for downstream analysis, allowing for better use of these reads than simply rejecting them.

To align a read, MAQ first searches for the ungapped alignments with the lowest number of mismatches using an algorithm called spaced seed indexing, this algorithm is described in detail in figure 1.20. Next, the probability that a read is correctly mapped is calculated using a Bayesian statistical model [177]. The probability that a read is correctly mapped to a given position is calculated from the error probabilities of the mismatch bases; however, the final quality score for any alignment is also a function of the quality scores at all other alignment positions in the reference genome. As this is impractical to calculate, the authors have derived a formula which approximates this based on the quality score at the second best hit and the number of other hits having the same number of mismatches as the second best hit. A quality score of 0 is assigned if a read maps equally well to multiple positions. In many analysis, these reads will be later discarded, but in some circumstances may provide useful information and some tools have begun to make use of these reads [177].

Tophat

TopHat is a splice junction mapper. It features the ability to discover novel splice junctions by aligning reads without prior knowledge of splice sites in the reference genome. While not the first *de novo* splice junction mapper, TopHat shows impressive performance gains over its competitors, which allows many typical RNA-seq experiments to be analysed in less than a day on a standard desktop computer. TopHat is built on the short-read aligner Bowtie [178], which is developed and maintained by the same group.

The TopHat pipeline first uses Bowtie to map non-junction reads to the genome. Based on the locations of these aligned sequences, TopHat invokes a function (which is part of MAQ) to identify contiguous "islands" of sequences which it infers to be exons. The remaining unmapped reads are then mapped to putative exon-exon boundaries in order to identify gene splicing patterns. TopHat's alignment algorithm is based on a technique known as the Burrows-Wheeler transform, this is described in more detail in figure 1.21.

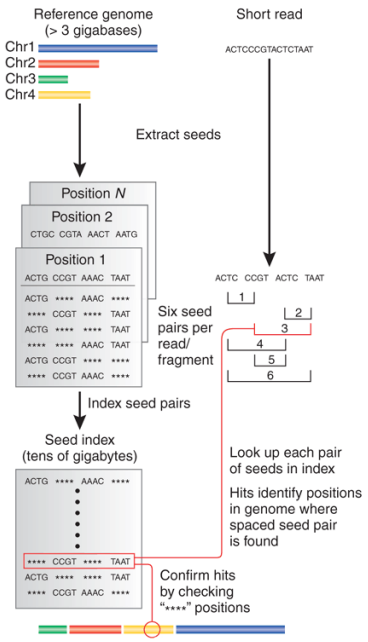

Figure 1.20: Spaced Seeds Indexing (used by MAQ). This algorithm first divides the reference genome into paired seeds and stores them in a lookup table, which allows fast searching. Then, each read is divided into four equally sized seeds and each seed pair is aligned to the reference. For each read, there are 6 possible combinations of pairs of seeds, each of which is aligned using the lookup table. As MAQ allows at most two mismatches in the read sequence, 4 of the 6 possible seed combinations must align perfectly to a particular locus, if the read is to have a chance of mapping there. After this initial pass, the resulting set of candidate regions is small enough that other seed regions can be checked individually and the read mapped to the best matching region. Sourced from [9].

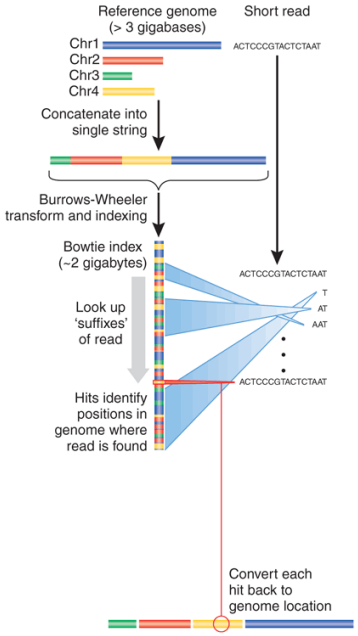

Figure 1.21: Burrows-Wheeler Transform (used by TopHat). This algorithm first creates a memory efficient representation of the reference genome using the Burrows-Wheeler transform, a technique originally developed to improve data compression. Using this method the entire human reference genome can be stored in under 2 gigabytes of memory, small enough for analysis on a typical desktop computer. The search algorithm works by aligning reads one character at a time, with each successive character narrowing down the likely positions that the read may map. While more complicated than the spaced seeds algorithm used by MAQ, this approach is also more than 30 times faster, owing to the speed of the Burrows-Wheeler search algorithm. Sourced from [9]

1.5.4 Tools for estimating gene expression in RNA-seq data

Custom R pipeline

Bioconductor provides a framework for analysis of short-read sequence data. When estimating expression in RNA-seq data, there are advantages to using this approach in parallel with the TopHat/Cufflinks pipeline. R/Bioconductor affords an experienced user much greater power and flexibility for closely examining and manipulating data; for example, counting reads which map to individual exons of a gene, investigating reads mapped to regions like introns and $3'$ and $5'$ untranslated regions, or manually computing metrics which may be useful for interpreting the data. These outputs can then be seamlessly integrated with subsequent analysis pipelines, taking full advantage of the plethora of statistical analysis tools available in R.

Bioconductor is beginning to implement some functionality for sequence alignment (for example in the *Biostrings* library). However, these tools are still in their infancy, when compared to some of the more sophisticated aligners available, so for this pipeline, reads were aligned using MAQ. The *ShortRead* [179] library (which is capable of handling aligned reads in a number of different formats) was used to import aligned data and perform various quality assessment and filtering steps. *GenomicFeatures* [176] was used to download reference genome annotations and *GenomicRanges* [175] was used to quantify reads mapped to genomic locations of interest, such as exons. As this pipeline was computationally expensive, the *Rmpi* [172] library was used to divide the workload across nodes of a high-performance *Beowulf* cluster. Differential expression analysis of RNA-seq data is also possible in R, for example, using the *DEseq* [180] or *edgeR* [181] libraries.

Cufflinks

Cufflinks is a standalone command line based program available on Linux and Mac OS X platforms [166]. It can be used to estimate gene and/or transcript expression in RNA-seq experiments and can also perform differential expression analysis. It works from single or paired-end RNA-seq spliced alignments (for example from TopHat) and works with or without a known gene model. To estimate transcript abundance, Cufflinks uses an algorithm based on graph theory, to construct a set of parsimonious transcripts which best explain the observed alignments. Expression levels are normalized using the "Fragments Per Kilobase of exon per Million fragments mapped" (FPKM) procedure, which is equivalent to the aforementioned RPKM (but "fragment" may refer to a single read or to two paired reads). Cufflinks' statistical models estimate the likelihood that the abundance assigned to any particular transcript is accurate, which means that all expression estimates

are accompanied by a confidence interval. Sequence bias is also corrected (this occurs when certain sequences are preferentially amplified, usually during PCR or reverse transcription steps); the algorithm learns which sequences introduce bias and incorporates this information into expression estimates. Cufflinks also attempts to deal with the aforementioned problem that some reads map equally well to multiple positions. These multi-mapped reads are initially down-weighted based on the number of positions to which they map, for example a read which maps equally well to 4 positions will be given 25% the weight of a normal read at each of those four positions; then, after the initial expression estimate of each of the transcripts to which the read maps, Cufflinks will attempt to recalculate the likelihood that the sequence maps to any one of the putative locations, based on the expression of each transcript, the inferred fragment length (in the case of paired-end reads) and sequence bias estimations [166].

1.6 Applications of high throughput genomics techniques

1.6.1 Gene set analysis

Gene-set analysis (GSA) is frequently used to discover meaningful biological patterns from lists of genes generated from microarray or high-throughput sequencing experiments. The objective is typically to identify similarities between the genes in the list, with respect to annotations available from sources such as the Gene Ontology (GO) [182] or Kyoto Encyclopedia of Genes and Genomes (KEGG) [183]. GSA approaches fall into two categories, "overrepresentation analysis" (ORA) and "functional class scoring" (FSC) [184]. ORA (the more popular approach) begins with a user supplied list of "foreground" genes, for example, genes that are differentially expressed in a set of samples at some arbitrary p-value and/or fold change threshold. These genes are then tested for over- or under-representation in biologically meaningful gene sets (e.g. genes annotated with specific GO terms), compared to a 'background' set of genes (which could be all genes in a genome or all genes represented on a microarray). Popular tools that make use of this approach include *GOstats* [174] and *DAVID* [185][186]. The FSC approach does not divide genes into foreground and background sets, but rather scores each gene (e.g. by statistical significance, in a differential gene expression setting) and from that attributes a score to each functional category, based on the scores of the individual genes in the category. FSC methods can be further divided into "competitive", where significance of differential expression in the gene set is compared to that of all other genes and "self-contained", where only the genes within the gene set are examined [187]. The most popular FSC method is Gene-set Enrichment Analysis

(GSEA) [188]. A significant result for a gene set is typically interpreted as evidence that the corresponding biological function or process is perturbed in the experimental condition. Chapter 5 of this thesis discusses a severe bias when GSA is applied to high throughput genome wide methylation data. We also propose a method to correct this bias.

1.6.2 The International HapMap project and genome wide associated studies (GWAS)

The International HapMap Project was initiated to develop a haplotype map of the human genome, which catalogs the pattern of common human genetic variation. The project aimed to identify a set of tag SNPs, which would allow the determination of an individuals haplotypes at less expense than previously [189][190]. The HapMap project used DNA isolated from lymphoblasts and the common sites of genomic variation were identified using SNP microarrays; during the project, over 1,000 individuals in 11 populations have been genotyped. The HapMap project also genotyped some related individuals, which allows researchers to investigate patterns of heritability. Other groups have performed gene and miRNA expression analysis on some of the HapMap cell lines and all of this data is freely available.

GWAS

The HapMap project has facilitated genome-wide association studies (GWAS), because large cohorts can now be genotyped (capturing most common variants using only tag SNPs) at less expense and this data can be used to assess the genetic basis for various traits and diseases. GWAS examines the association of common variation in genome sequence, with phenotypic traits (e.g. height, BMI, disease etc.). The result of over 1,000 GWAS studies have been deposited in the online database "GWAS catalog" and this continues to grow [191]. One example is the 2007 Welcome Trust Case Control Consortium [10]. This study genotyped a subset of 15,000 individuals from the British population, to assess the genetic basis of 7 diseases. The study was successful in uncovering many new susceptibility loci. A summary of these results is plotted in Fig.1.22.

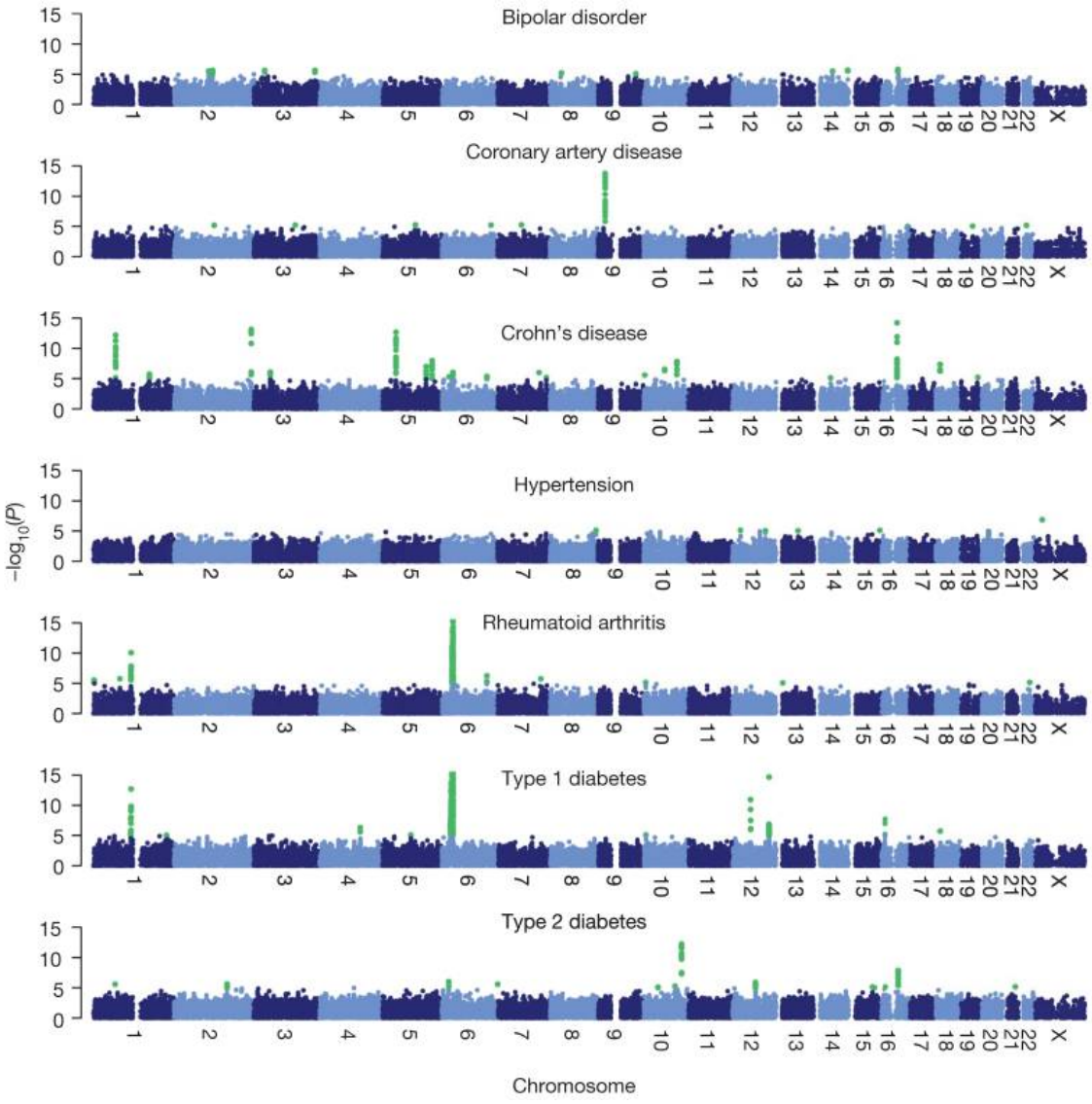

Figure 1.22: Manhattan plots showing p-values for SNP associations with each of seven diseases, as identified by Welcome Trust Case Control Consortium. P-values of $< 10^{-5}$ are highlighted in green [10].

While GWAS has been successful in identifying a large number of new disease susceptibility loci, some authors have deemed the approach a failure [192], as in almost all cases, the SNPs identified still only explain a small fraction of the total genetic component (as calculated from pedigree studies) of diseases. For example, twin studies have revealed that between 60% and 80% of susceptibility to Crohn's disease can be explained by genetics [193], but the SNPs identified by GWAS have

only captured approximately 10% of this [194]. The remaining genetic variance has been referred to as the "missing heritability" [195]. One explanation is that a large number of SNPs have a very small effect, below the threshold of what can be detected by GWAS. This has been shown to be the case for many diseases and in the case of Crohn's disease, models considering all SNPs simultaneously have explained up to 40% of the genetic variance [196]. The remaining heritability is thought to be explained by rare variants, which are not genotyped on current SNP arrays. As the cost of high-throughput sequencing continues to drop, it is hoped that it will soon be possible to identify many of these rare variants [192]. In this thesis, we have used the genotype and gene expression data from the HapMap project to assess the genetic basis for inter-individual differences in miRNA regulatory effect. In the following subsection we will briefly discuss some of the key issues when performing GWAS analysis.

GWAS data quality control and filtering

Prior to GWAS, it is important to filter and quality assess SNP data. In most cases there are no universally accepted thresholds or guidelines for this [197], but some of the important criteria to consider are as follows. Firstly SNPs with low minor allele frequency are normally filtered out, as these SNPs have little power to detect an association, but contribute to the multiple testing problem. Next, both SNPs and individuals with low call rate (where the SNP could not be determined with high confidence) are normally removed, as these cannot provide reliable results. Also, each SNP is normally tested for deviation from Hardy-Weinberg equilibrium. This describes the relationship between genotype distribution seen in the data and allelic frequency in an idealized population. Deviation from the expected proportions can theoretically occur because of selective pressure, mixture of genetically heterogeneous populations, cryptic relatedness (kinship among the individuals that is not known), or genotyping errors due to limitations in microarray or other experimental technologies [198].

Population structure is also an important issue in GWAS. Normally, it must be ensured that individuals originate from a genetically homogeneous population, as population stratification will cause spurious results. This happens when a trait varies between sub-populations and these populations happen to have different allele frequencies at a genotyped SNP [197]. Because of this, different populations are normally analysed separately.

Chapter 2

The regulatory effect of miRNAs is a heritable genetic trait in humans

The content of this chapter was published as:

Geeleher, P., Huang RS., Gamazon ER., Golden, A. and Seoighe, C. (2012).
The regulatory effect of miRNAs is a heritable genetic trait in humans.
BMC Genomics, 13:383.

2.1 Abstract

2.1.1 Background

microRNAs (miRNAs) have been shown to regulate the expression of a large number of genes and play key roles in many biological processes. Several previous studies have quantified the inhibitory effect of a miRNA indirectly by considering the expression levels of genes that are predicted to be targeted by the miRNA. This approach has been shown to be robust to the choice of prediction algorithm. Given a gene expression dataset, Cheng *et al.* defined the regulatory effect score (RE-score) of a miRNA as the difference in the gene expression rank of targets of the miRNA compared to non-targeted genes.

2.1.2 Results

Using microarray data from parent-offspring trios from the International HapMap project, we show that the RE-score of most miRNAs is correlated between parents and offspring and, thus, inter-individual variation in RE-score has a genetic

component in humans. Indeed, the mean RE-score across miRNAs is correlated between parents and offspring, suggesting genetic differences in the overall efficiency of the miRNA biogenesis pathway between individuals. To explore the genetics of this quantitative trait further, we carried out a genome-wide association study of the mean RE-score separately in two HapMap populations (CEU and YRI). No genome-wide significant associations were discovered; however, a SNP rs17409624, in an intron of *DROSHA*, was significantly associated with mean RE-score in the CEU population following permutation-based control for multiple testing based on all SNPs mapped to the canonical miRNA biogenesis pathway; of 244 individual miRNA RE-scores assessed in the CEU, 214 were associated ($p < 0.05$) with rs17409624. The SNP was also nominally significantly associated ($p = 0.04$) with mean RE-score in the YRI population. Interestingly, the same SNP was associated with 17 (8.5% of all expressed) miRNA expression levels in the CEU. We also show here that the expression of the targets of most miRNAs is more highly correlated with global changes in miRNA regulatory effect than with the expression of the miRNA itself.

2.1.3 Conclusions

We present evidence that miRNA regulatory effect is a heritable trait in humans and that a polymorphism of the *DROSHA* gene contributes to the observed inter-individual differences.

2.2 Background

Of the mechanisms of post-transcriptional regulation by miRNAs, lowered mRNA levels (mRNA cleavage or deadenylation) accounts for most (>84%) of decreased protein production [68]. This implies that it is possible to assess levels of miRNA mediated gene silencing from the mRNA levels of a miRNA's target transcripts. Cheng *et al.* quantified miRNA activity in this way by defining the regulatory effect score (RE-score). For each miRNA in each sample, this is calculated by the average expression rank of genes that are not predicted to be targeted by the miRNA, minus the average expression rank of the predicted targets of the miRNA [69]. Thus, the RE-score is intended to measure the extent to which targets of the miRNA are downregulated in a sample relative to other genes. It is not informative to compare the RE-scores of different miRNAs, but comparison of the RE-score of a given miRNA between samples can provide an indication of a difference in the repressive effect of the miRNA between the samples. For example, if the targets of a given miRNA relative to non-targets are ranked higher in a set of cancer samples than in comparable normal tissues, this suggests that

the miRNA exerts less control over gene expression in the cancer samples. There have been numerous other studies published that have also investigated miRNA regulation by assessing changes in expression of mRNA targets [199] [200] [201] [202] [203] [204].

Messenger RNA targets of each microRNA are deduced using a prediction algorithm. If a specific microRNA's inhibitory effect in a particular sample is large one would expect that the RE-score for that miRNA would also be large, as its targets would tend to be ranked lower. Conversely, if a microRNA's inhibitory effect is small, its targets will tend to be ranked higher and hence the RE-score will be small.

We sought to investigate whether there is evidence of natural variation in this phenotype between human individuals, using RE-scores calculated from microarray and RNA-seq data generated from the CEU (Utah residents with ancestry from northern and western Europe) and YRI (Yoruba in Ibadan, Nigeria) lymphoblastoid cell lines of the HapMap project [189][190][205][206][207]. Microarray data were available for 56 trios of related individuals in these populations (consisting of two parents and an offspring). We used these data to investigate the genetic component of the variation in RE-scores. Positive correlation between the value of a phenotype in an offspring and the mean value in parents provides evidence of a heritable component in the variation of the phenotype and the slope of the linear regression line can be used as an estimate of the narrow-sense heritability [208][209][210] .

2.3 Results and Discussion

2.3.1 Heritability of the regulatory effect of miRNAs

Microarray data [207] were obtained for 56 trios (both parents and an offspring) from the CEU and YRI populations of the HapMap project [189][190]. Using miRNA targets predicted by TargetScan (for conserved miRNAs) [211][59] we compared RE-scores between parents and offspring. For 51% of miRNAs the mean RE-score of parents and the RE-score of the offspring were significantly ($p < 0.05$) positively correlated (Table 2.1). Population of origin was included in these regressions to model biological and technical differences between the CEU and YRI cell lines. Histograms of regression p-values for heritability of individual miRNA RE-scores from TargetScan and a second miRNA prediction algorithm (PicTar [79]) are shown in figure 2.1.

Number of miRNAs	244
Average number of target genes per miRNA	437
RE-score positively correlated between mean of parent and offspring	235
Positively correlated ($p < 0.05$)	124
Average Heritability (S.D)	0.30 (0.15)

Table 2.1: Summary of results for individual miRNA RE-scores calculated for conserved miRNAs using TargetScan.

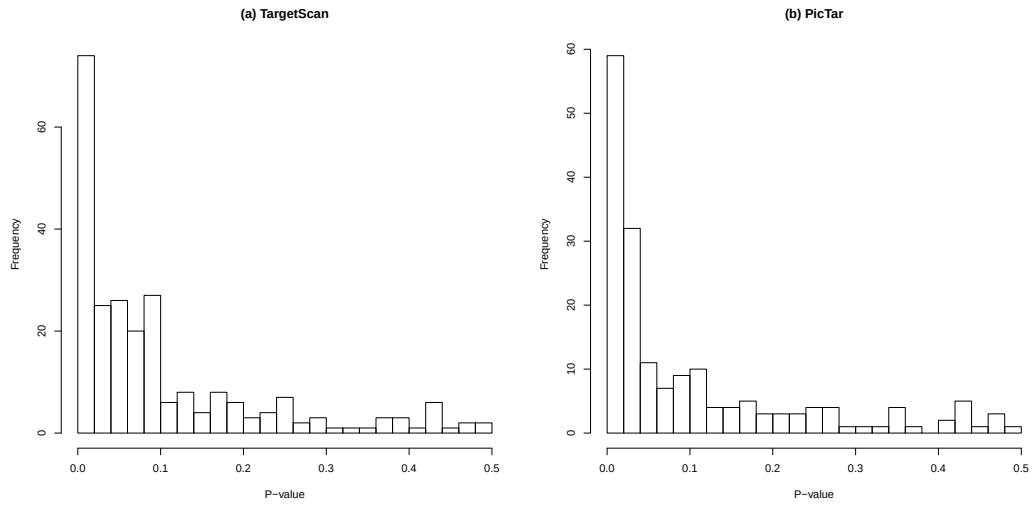

Figure 2.1: Heritability for individual RE-scores. Histograms of p-values for tests of heritability of individual RE-scores for (a) TargetScan and (b) PicTar algorithms.

We calculated the mean of the RE-score over all miRNAs. Unsurprisingly, the mean RE-score is also strongly correlated between parents and offspring in HapMap trios (Fig. 2.2). This correlation is statistically significant using mean RE-scores calculated from targets predicted by TargetScan ($slope = 0.68 \pm 0.34$; $p = 2 \times 10^{-4}$). The slopes of these regression lines provide estimates of the narrow-sense heritability of the mean RE-score. We also assessed mean RE-score heritability based on targets predicted by three other algorithms (although TargetScan has previously been found to be most acurate in predicting change in protein levels during miRNA transfection [4]). PicTar ($slope = 0.62 \pm 0.36$; $p = 1.3 \times 10^{-3}$), miRanda [212]($slope = 0.40 \pm 0.37$; $p = 3.6 \times 10^{-2}$) and mirTarget2 [213] ($slope = 0.35 \pm 0.32$; $p = 2.8 \times 10^{-2}$) all showed significant evidence of heritability. The Pearson's correlation (against TargetScan) of mean RE-score calculated using mirTarget2, miRanda and PicTar, are 0.73, 0.89 and 0.94 respectively. This indicates that the mean RE-score is relatively robust to choice of

prediction algorithm.

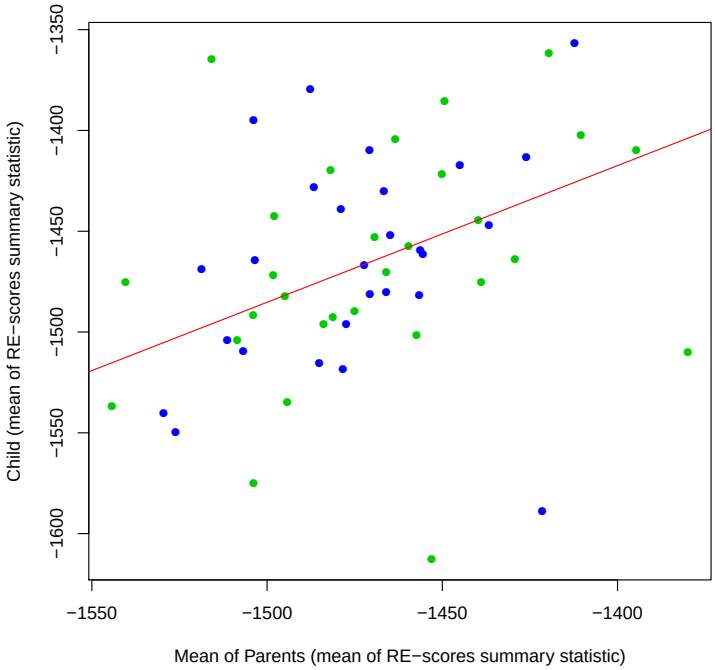

Figure 2.2: Heritability of mean RE-score using TargetScan. The scatter plot shows child values of mean RE-score against mean value of both parents. Points from the CEU are colored blue and YRI are green. The estimated regression line is shown in red.

It is possible that the apparent genetic contribution to the regulatory effect of miRNAs is a consequence of the heritability of gene expression, rather than a novel molecular phenotype. Since the expression levels of a large proportion of human genes have a strong genetic component [214][215][216], the correlation in RE-score between parents and offspring could simply reflect the correlation in the expression levels of a proportion of the genes targeted by the miRNA. We devised a permutation test to evaluate this possibility. For each set of mRNAs predicted to be targeted by a given miRNA we replaced predicted target genes by genes chosen at random (details in Methods). If the apparent heritability of RE-scores is merely a consequence of heritability of individual gene expression levels, the RE-scores obtained from sets of random genes should exhibit similar levels of heritability to the RE-scores based on the true predicted target sets. Greater evidence of heritability from true predicted targets compared to sets of randomly selected genes suggests that the RE-score heritability cannot be explained by

the heritability of individual gene expression levels. Of 1,000 randomizations, just eight ($p = 0.008$) reached a regression p-value as extreme as the target sets predicted by TargetScan. This approach could be criticized on the grounds that the TargetScan predictions used conservation to identify putative miRNA binding sites, as it has previously been shown that conserved genes are more likely to be highly expressed [217] and also that more highly expressed genes are more likely to be heritable [218], these factors introduce an obvious bias. To address this issue, we repeated the analysis, but separated genes into 5 bins based on their median expression levels across all samples; this time, during the permutations, genes were only replaced with genes of similar expression level. This approach did not affect the results, indicating that this potential bias is not an issue.

2.3.2 Genome-wide association of mean RE-score

In order to explore the genetic contribution to RE-score variation further, we carried out a genome-wide association (GWA) test, treating mean RE-score, calculated using miRNA targets predicted by TargetScan, as a quantitative trait, and using genotype data from the HapMap project [189][190]. To avoid artifacts resulting from population structure, we carried out these tests separately on the CEU and YRI samples and excluded related individuals (offspring of the HapMap trios). RE-scores were recalculated using expression data derived from RNA-seq [205][206], which was available for parents but not for offspring of HapMap trios. Histograms and Manhattan plots of p-values are shown in figure 2.3. The p-value distributions show a peak towards low p-values, suggesting the presence of some true positive associations. However, none of these associations remained significant following a permutation-based correction for multiple testing. This is not surprising given the relatively small number of samples compared to typical GWA studies.

2.3.3 Association of mean RE-score with SNPs in the miRNA biogenesis pathway

In their original study, Cheng *et al.* [69] used the RE-score metric to compare miRNA repression in Estrogen Receptor Positive (ER+) and Estrogen Receptor Negative (ER-) breast cancers. They found that miRNAs tended to have higher RE-scores in ER- and hypothesized that differences between the two cancer subtypes may be attributable to dysregulation of key genes in the microRNA biogenesis pathway [69]. Thus, we used linear regression to investigate the relationships between the expression levels of seven key genes in the miRNA biogenesis pathway, (*DICER1*, *EIF2C2*, *DROSHA*, *DGCR8*, *XPO5*, *RAN* and *TRBP*) and mean RE-score. We first used all samples from both populations pooled (including pop-

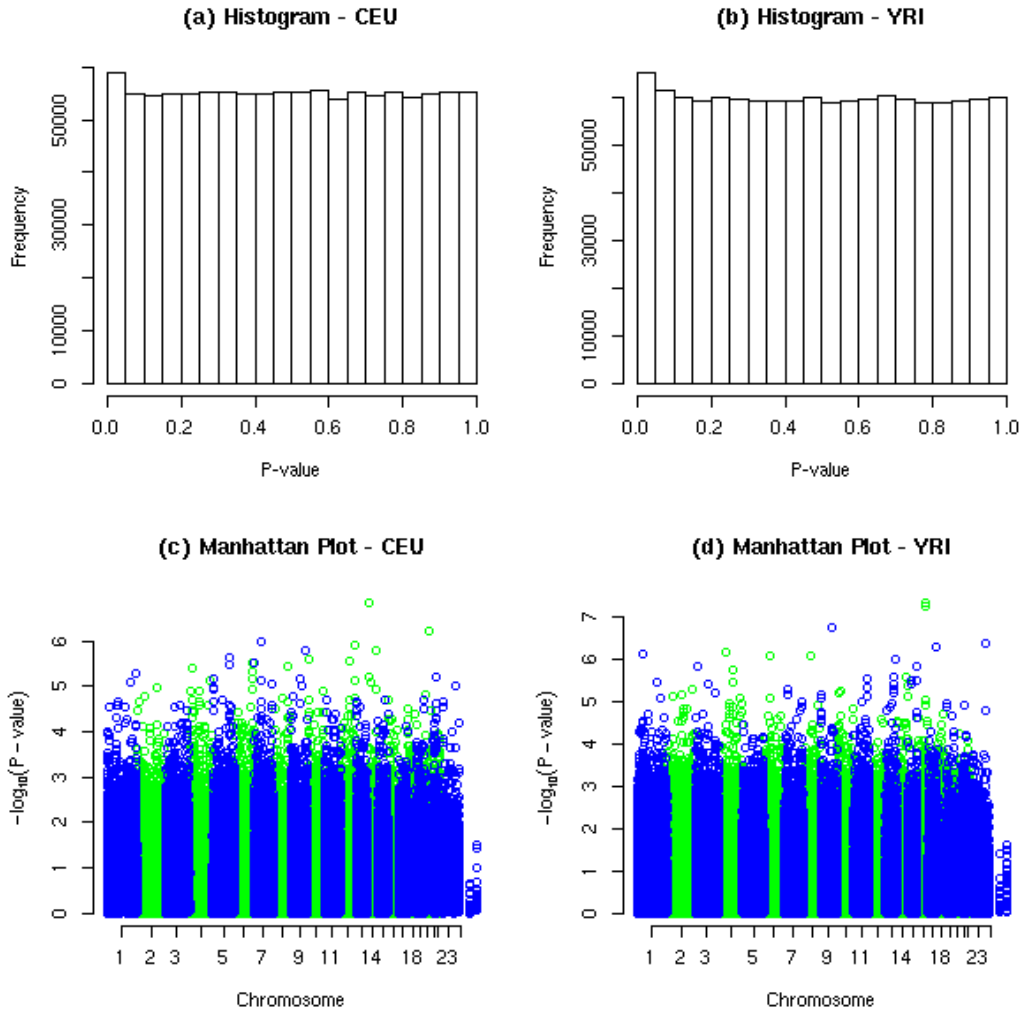

Figure 2.3: Histograms (a & b) of p-values for tests of association between all SNP markers and mean RE-score and Manhattan plots (c & d) of p-values across the genome in the CEU and YRI respectively.

ulation of origin as a factor in the model) and also in each of the populations separately. Expression levels of five of these seven genes were significantly correlated with mean RE-score (Table 2.2), consistent with a contribution of differential regulation of the miRNA biogenesis pathway to differences in mean RE-score. In fact, a large proportion (37.8%) of all genes were significantly associated ($p < 0.05$) with mean RE-score; however, this proportion was somewhat higher (five out of seven, or 71.4%) for genes in the miRNA biogenesis pathway. Given this relationship between RE-score and the activities of genes in the miRNA biogenesis

pathway, these genes are worthy of closer examination for genetic association with mean RE-score.

	CEU		YRI		Pooled	
	Bonferroni P	**Slope**	**Bonferroni P**	**Slope**	**Bonferroni P**	**Slope**
DROSHA	9.42×10^{-03}	-10.23	1.37×10^{-05}	-22.12	5.19×10^{-06}	-15.64
DGCR8	0.036	11.57	0.95	-0.46	0.37	6.23
XPO5	0.47	-3.03	1.38×10^{-04}	-17.85	2.17×10^{-03}	-10.74
RAN	0.27	0.49	0.14	-0.94	0.75	-0.12
DICER1	8.51×10^{-03}	-13.77	1.97×10^{-09}	-26.18	5.57×10^{-10}	-21.72
TRBP	2.95×10^{-05}	12.26	0.085	8.12	2.68×10^{-04}	10.60
EIF2C2	0.022	-6.25	1.39×10^{-07}	-9.07	1.88×10^{-08}	-8.41

Table 2.2: P-values and slopes from the linear regression of expression level of genes in the miRNA biogenesis pathway against mean RE-score, in the CEU, YRI and for both populations pooled.

We carried out a second test of association, restricting to 336 SNPs that map to the genomic regions (according to dbSNP) of these seven key genes involved in the miRNA biogenesis pathway. A SNP is mapped a gene by dbSNP if it lies between 2kb upstream and 500bp downstream of the gene. Again there appear to be more low p-values than would be expected under the uniform distribution, pointing to a proportion of true positive associations in both populations (Fig. 2.4). The ten SNPs most strongly associated with mean RE-score in CEU and YRI are shown in Tables 2.3 and 2.4, respectively. One SNP, rs17409624, in an intron of *DROSHA* remained statistically significantly ($p_{adjusted} < 0.05$) associated with mean RE-score in the CEU following Bonferroni and permutation-based control for multiple testing. This SNP was also nominally significantly associated with mean RE-score in the YRI ($p = 0.04$); however, the minor allele frequency was much lower in YRI, limiting the power to detect an association with a significance that could survive multiple test correction. The magnitude and direction of the RE-score differences between genotypes are consistent across the two populations (Fig. 2.5). Taken individually, the vast majority (214 of 244) of RE-scores are associated ($p < 0.05$) with this SNP in the CEU. This number drops to 36 of 244 in the YRI, however the lower minor allele frequency in the YRI again limits the power to detect the association.

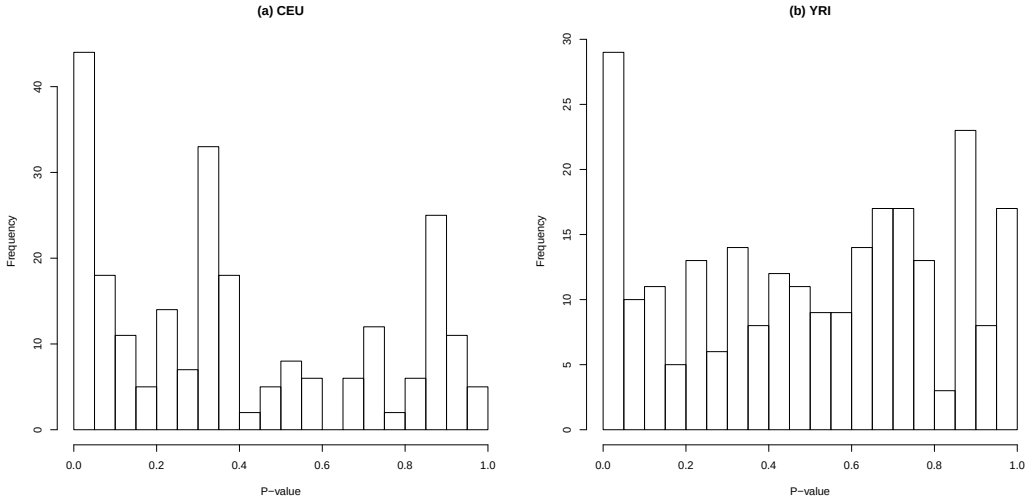

Figure 2.4: Histograms of p-values for the tests of association between SNP markers mapped to the miRNA biogenesis pathway and mean RE-score in the (a) CEU and (b) YRI populations.

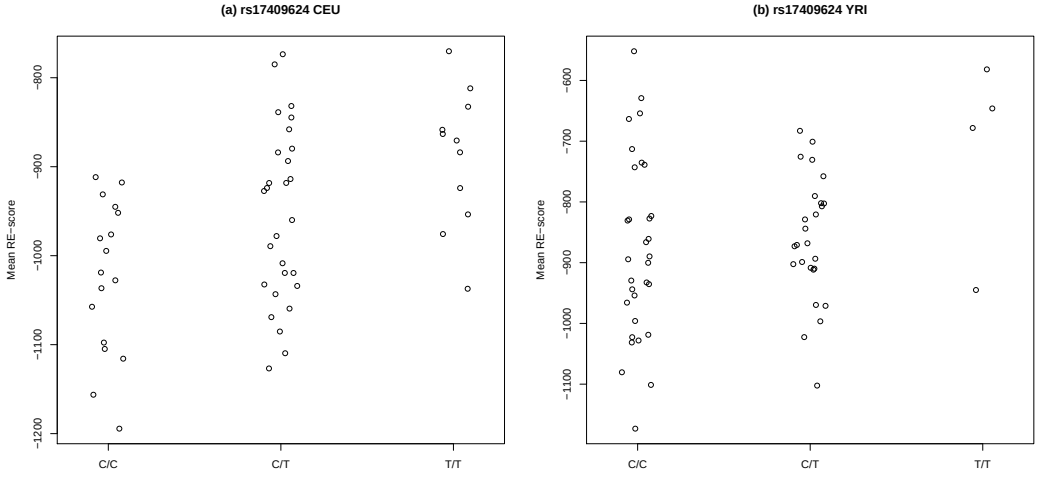

Figure 2.5: Stripcharts of mean RE-score against genotype at rs17409624 in the (a) CEU and (b) YRI populations.

	Location	Associated Gene	P-value	Bonferroni P	Q-value	Permutation p-value
rs17409624	chr5:31,528,733	*DROSHA*	1.81×10^{-04}	0.043	0.018	0.03
rs10078886	chr5:31,427,441	*DROSHA*	3.32×10^{-04}	0.079	0.018	0.051
rs16901121	chr5:31,418,215	*DROSHA*	3.32×10^{-04}	0.079	0.018	0.051
rs2279797	chr5:31,428,028	*DROSHA*	3.32×10^{-04}	0.079	0.018	0.051
rs13183642	chr5:31,511,106	*DROSHA*	1.25×10^{-03}	0.3	0.054	0.16
rs3805516	chr5:31,420,670	*DROSHA*	1.56×10^{-03}	0.37	0.056	0.2
rs4867349	chr5:31,536,327	*DROSHA*	1.82×10^{-03}	0.43	0.056	0.23
rs2287584	chr5:31,423,007	*DROSHA*	3.27×10^{-03}	0.78	0.073	0.33
rs615344	chr5:31,425,788	*DROSHA*	3.39×10^{-03}	0.8	0.073	0.34
rs682902	chr5:31,423,694	*DROSHA*	3.39×10^{-03}	0.8	0.073	0.34

Table 2.3: Top 10 associations for miRNA biogenesis pathway related SNPs (CEU).

	Location	Associated Gene	P-value	Bonferroni P	Q-value	Permutation p-value
rs6994531	chr8:141,544,476	*EIF2C2*	4.57×10^{-03}	1	0.38	0.55
rs1633445	chr22:20,100,596	*DGCR8*	0.011	1	0.38	0.77
rs17409275	chr5:31,514,127	*DROSHA*	0.012	1	0.38	0.8
rs1209904	chr14:95,563,712	*DICER1*	0.015	1	0.38	0.86
rs1187650	chr14:95,551,554	*DICER1*	0.018	1	0.38	0.9
rs1187655	chr14:95,565,556	*DICER1*	0.018	1	0.38	0.9
rs6575499	chr14:95,622,200	*DICER1*	0.018	1	0.38	0.9
rs12881840	chr14:95,568,003	*DICER1*	0.020	1	0.38	0.9
rs12889800	chr14:95,618,898	*DICER1*	0.020	1	0.38	0.9
rs2292780	chr8:141,561,357	*EIF2C2*	0.022	1	0.38	0.93

Table 2.4: Top 10 associations for miRNA biogenesis pathway related SNPs (YRI).

As a further test of the association between rs17409624 and mean RE-score, we investigated the RE-scores of a particular class of intronic miRNAs (mirtrons), which are not processed by *DROSHA* [219]. If the association between the SNP and mean RE-score is real and is mediated by an effect on miRNA processing by *DROSHA*, the SNP should not be associated with the RE-scores of mirtrons. Consistent with this prediction, we found that a much lower proportion of mirtron RE-scores (based on TargetScan predictions from CEU RNA-seq data) are associated (at $\alpha = 0.05$) with the *DROSHA* SNP (5 out of 13 mirtrons, compared to 214 out of 244 conventional miRNAs; $p = 0.0004$, from a two-sided Fisher's exact test). We have found evidence that the subset of mirtrons that do show an association with the SNP do so because of an overlap between their target gene sets and the target gene sets of conventional miRNAs. The mirtrons which are most significantly associated with rs17409624 tend to target genes that are also targeted by many other miRNAs; and mirtrons that target genes that are targeted by few conventional miRNAs are less significantly associated with rs17409624 (Fig. 2.6).

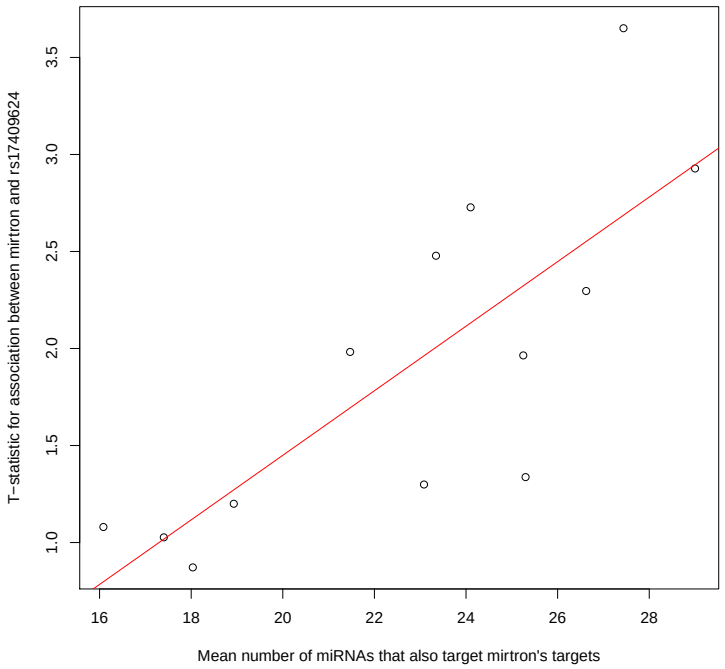

Figure 2.6: Relationship between the strength of association with rs17409624 for mirtrons and the average number of conventional miRNAs that also target the mirtron's target genes. This figure is based on TargetScan predictions for conserved miRNA families on HapMap CEU data ($p = 1.2 \times 10^{-3}$).

2.3.4 Searching for causal SNPs

We investigated the function of SNP rs17409624 using the "SNP Function Prediction" tool, which is part of the SNPinfo suite [220]; however, no significant results were identified. We also searched the "GWAS Catalog" but did not find any previous studies which had identified this SNP[221]. To search for other SNPs that may be causally responsible for this association we used confidence intervals [222] as implemented in HaploView to calculate haplotype blocks for the CEU HapMap data. rs17409624 is located within a haplotype block that includes the *DROSHA* promoter region (Fig. 2.7). We verified that this is the active promoter of *DROSHA* using data recently released by the ENCODE project *et al.* [223]. Chromatin states for this locus are shown in figure 2.8. The expression level of *DROSHA* is significantly associated with mean RE-score (Table 2.2); however, the genotype of this locus was not significantly correlated with *DROSHA* expression level ($p = 0.39$); nor is it correlated with the relative expression level of any *DROSHA* transcript isoforms (identified using Cufflinks [224]) or the inclusion of any of the individual *DROSHA* exons. A further possibility is that rs17409624 is in linkage disequilibrium (LD) with an exonic SNP that was not genotyped on the HapMap microarrays. Using SNP calls from genome sequence data released by the 1,000 Genomes Project [225], we found no coding SNPs with a stronger association to mean RE-score than rs17409624, the regions assayed included the $3'$ and $5'$ UTRs. We caution however, that there was much less statistical power to detect an association using the 1,000 Genomes data, as there was an overlap of only 45 samples between the 1,000 Genomes Project dataset (versus 59 for the HapMap microarray data) and the RNA-seq samples from the CEU used to calculate RE-scores. This means that it is difficult to rule out the possibility of linkage of rs17409624 with a causative SNP in the coding region. Thus, the causal mechanism linking genetic variation at the *DROSHA* locus to variation in the RE-score remains unclear.

2.3.5 Integrative analysis of miRNA expression and RE-score data

miRNA expression data has recently been generated for some of the HapMap CEU and YRI cell lines [226]. In the majority of cases, miRNA expression levels and their corresponding RE-scores were not significantly correlated. Average Spearman correlation between miRNA expression and corresponding TargetScan based RE-score from the RNA-seq data is only 0.009 in the CEU and -0.0003 in the YRI. Although surprising, this observation is consistent with the findings of Cheng *et al.* [69], who, for the original RE-score study, performed Spearman correlations of the t-scores of comparisons of miRNA expression and RE-scores

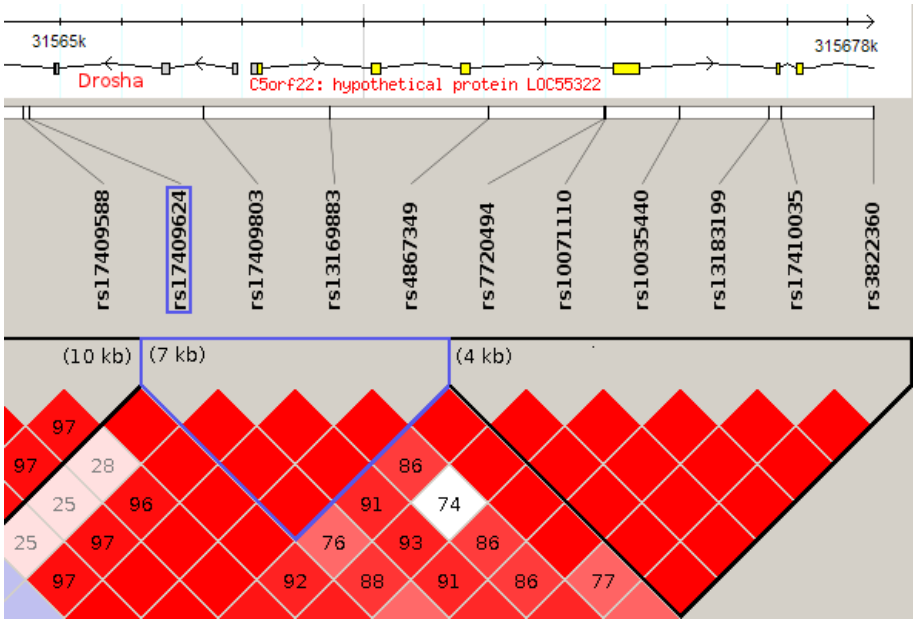

Figure 2.7: Haplotype blocks around rs17409624 as calculated by Haploview, using the HapMap CEU data. The block which includes rs17409624 is highlighted in blue; this block also includes the *DROSHA* promoter region.

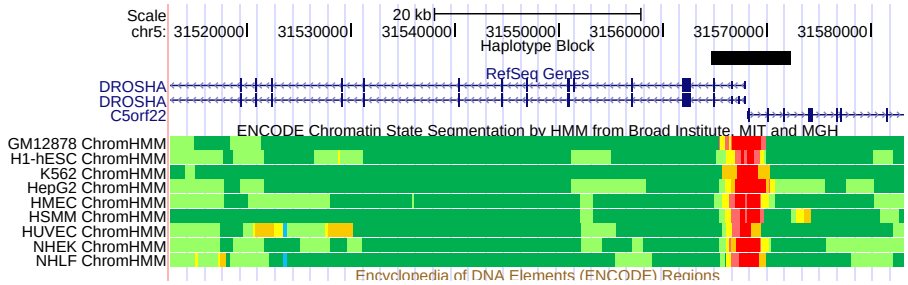

Figure 2.8: *DROSHA* promoter region. Chromatin state of *DROSHA* region for nine cell lines from the ENCODE project. Active promoter is shown in bright red. The haplotype block for rs17409624 is shown in black and clearly overlaps the promoter region.

between ER- and ER+ breast cancers, finding only very weak positive correlation. Similar results have also been observed on two separate datasets by Liang *et al.* [227]. However, we find that in the CEU, the expression of 17 of 201 miRNAs (Table 6.3 in Appendix A) that were consistently expressed across the cell lines is associated ($p < 0.05$) with rs17409624 and that 13 of these associations are in the same direction as mean RE-score. One miRNA is associated with the SNP in the YRI, but once again, the lower minor allele frequency of rs17409624 in

the YRI limits the power to identify associations. Thus, this SNP represents a trans-eQTL cluster for miRNA gene expression. We hypothesize that this trans-eQTL reflects inter-individual differences in the efficiency of miRNA processing by *DROSHA*. Given that miRNA expression measurements are relative (in this case miRNA expression was measured using a pooled reference microarray design), it is possible that this polymorphism may affect the absolute copy numbers of a large fraction of miRNAs, even though an association between miRNA expression and the SNP is detectable for a relatively small fraction of miRNAs. This hypothesis could be tested using transcriptome sequencing strategies designed to measure the abundance of miRNAs relative to other RNA species. Indeed, given a global and consistent change in expression of all miRNAs in a sample, one may not expect the expression of any miRNAs to be associated with rs17409624, as the proportion of the transcript pool occupied by any given miRNA, would remain unchanged. However, the miRNA regulatory effect polymorphism need not affect the expression of all miRNAs to exactly the same degree, potentially leading to both positive and negative associations of miRNA expression with the SNP.

As discussed above, RE-scores of the majority of miRNAs were not correlated with miRNA expression. This remained the case when we restricted to miRNAs whose expression varied most across samples. However, the RE-scores of individual miRNAs were correlated with the mean RE-score calculated across all miRNAs. We restricted this analysis to the 20 most variable miRNAs. Of the top 20 in either population, 14 in the CEU and 13 in the YRI had TargetScan prediction data and therefore RE-scores could be calculated. We only considered these highly variable miRNAs because quantities that are relatively constant across samples are not expected to be correlated, given the noise inherent in microarray data. The correlation between mean and individual miRNA RE-scores is not simply a consequence of overlaps in genes targeted by different miRNAs, since it holds true even when the mean RE-score is recalculated, for each miRNA correlation test, after all of the individual miRNAs targets have been subtracted from the target sets of the remaining miRNAs. 13 of the 14 highly varying miRNAs in the CEU and all 13 of 13 in the YRI show a stronger association between the individual RE-score and (subtracted) mean RE-score, than between the individual RE-score and the expression of the miRNA itself. In most cases this difference is large (Tables 6.1 and 6.2 in Appendix A), hence, the mean RE-score in a sample may be a much better predictor of the expression level of the targets of any particular miRNA, than is the expression profile of the miRNA itself. It is, perhaps, not surprising that the expression level of an individual miRNA is not indicative of the expression of its target genes, given that genes are often targets of a large number of miRNAs. Of 11,759 genes which are predicted to be targeted by at least one miRNA (by the full TargetScan set), the average number of miRNAs targeting each gene is 17.48. In this context, the fact that the mean RE-score has power to

predict the expression levels of a miRNA targets, even when the mean RE-score is calculated without considering the targets of that miRNA, is interesting and points to differences in the effect of the miRNA pathway on target genes across the samples.

2.4 Conclusions

We have found evidence of heritability of the regulatory effect of miRNAs in human. We have also identified an association between the regulatory effect of miRNAs and a SNP in the miRNA processing gene *DROSHA*. This association was identified in lymphoblastoid cell lines and it remains to be seen whether and in which primary cells the regulatory effect of miRNAs is associated with the *DROSHA* locus. As noted in the Background, Cheng *et al.* had observed that there is a change in miRNA RE-scores between ER- and ER+ breast cancer subtypes. Thomsom *et al.* showed that mature miRNA levels are generally lower in several human primary cancers, despite unchanged pri-miRNA levels and this has been attributed to defective processing by *DROSHA*[228], while *DROSHA* and *DICER* have also been shown to be downregulated in endometrial cancer and specific subgroups of breast cancer[229][230]. Thus, it will be important to investigate further the phenotypic consequences of inter-individual differences in miRNA regulatory efficiency and the influence on gene expression, possible tumorigenesis and the impact of such inter-individual differences in the context of the use of miRNAs as biomarkers.

2.5 Methods

2.5.1 Data

Raw gene expression microarray data of related individuals from the CEU and YRI populations of the HapMap project were downloaded from GEO under accession number *GSE7792*, these data were generated by Huang *et al.* [207] using Affymetrix Human Exon 1.0 ST microarrays. Prior to calculating gene expression level estimates, the data were RMA normalized [116] and genes whose expression level were below the detection threshold, as estimated by the DABG algorithm ($p < 0.05$), were set to zero; these steps were performed using Affymetrix Power Tools and R as described in [129]. RNA-seq data for unrelated individuals of the HapMap YRI population were generated by Pickrell *et. al* [205] and we obtained these aligned data from GEO under accession number *GSE19480*. Similarly, Montgomery *et al.* [206] used RNA-seq to assess gene expression of unrelated CEU samples and these data were obtained from ArrayExpress under

accession number *E-MTAB-197*. All data were aligned to *hg18* using MAQ [177]. We performed gene expression analysis using R/Bioconductor. Data were loaded in R [167] using the *ShortRead* [179] library. Following Montgomery *et al.*, only reads that had a mapping quality score of greater than or equal to 10 were included. The *GenomicRanges* [175] library was used to compute the number of reads mapping to exons of each gene and expression values were normalized the using the RPKM [152] procedure. miRNA prediction data were obtained using the R library *RmiR.Hs.miRNA* [231] which provides a database of miRNA targets for several widely used algorithms. The HapMap release 28 (merged data for phases I, II and III) [189][190] SNP data were downloaded from the HapMap website, converted to GenABEL format and trimmed to include only samples in the CEU and YRI populations for which there was matching RNA-seq data.

2.5.2 Estimating Heritability of mean RE-score

Narrow sense heritability of individual miRNA RE-scores and mean RE-score was estimated using a robust linear regression model [209][208]. The *rlm()* function from the R library *MASS* was used to fit regression model for child value dependent on mean of parents. Population of origin was included as a factor in the models. The slope of the regression line provides an estimate of heritability.

2.5.3 Permutation testing of heritability of mean RE-score

To calculate a corrected p-value for heritability of mean RE-score of a miRNA prediction algorithm, we performed 1,000 permutations of the prediction algorithm's miRNA gene target sets and recalculated heritability of mean RE-score following each permutation; the permutation p-value was the proportion of permuted sets that return p-values which are equal to, or lower than, the original raw p-value for that algorithm. To perform a permutation, we replace each gene target of each miRNA's target set with a randomly chosen gene, but only genes for which expression data is available are replaced or used for replacement, as only these can affect RE-scores. If a gene is a target of multiple microRNAs, it is replaced by the same randomly chosen gene in every target set, so as to maintain the structure of the data.

2.5.4 Genome-wide association test

The R package *GenABEL* [232][233] was used for filtering and tests of association. Prior to testing for association, genotype data were filtered as follows. Obvious close relatives are removed by discarding the child samples and to avoid the effects

of population stratification CEU and YRI samples are assayed separately. Markers with a low minor allele frequency were filtered by excluding SNPs for which there were less than 5 copies of the minor allele across all samples. We used only SNPs genotyped as part of HapMap phase III. Individuals or SNPs were excluded for a call rate of < 0.95. Tests for Hardy-Weinberg equilibrium were conducted using Pearson's χ^2, comparing observed genotype frequencies in the data to the calculated expected frequencies; a cut-off FDR level of 0.2 was applied. To assess if any remaining relatedness exists among the samples, the pairwise proportion of alleles identical-by-state (IBS) was calculated between all individuals based on 2,000 randomly chosen autosomal markers, ensuring IBS < 0.95 for all samples. For multiple testing correction of association p-values, permutations were calculated by permuting phenotype labels and performing tests of association as normal; for each raw p-value, we computed the number of permutations for which a p-value equal to, or lower than, the original raw p-value was reached and divide this by the number of permutations, the result of which is the adjusted p-value. False discovery rates were also assessed using the R package *qvalue* [234].

2.5.5 Calculating association between individual miRNA RE-score, mean RE-score and miRNA expression

For each of 14 highly varying miRNAs in the CEU samples and 13 in the YRI, we fit a multiple linear regression model of individual miRNA RE-score dependent on the expression of the miRNA and the mean RE-score. For each fit of the model, mean RE-score was re-calculated with the genes that are targets of the particular individual miRNA removed from the gene expression matrix, so as to avoid a bias in the association between the two variables.

Chapter 3

Improving gene expression estimates from DNA microarrays using machine learning

3.1 Abstract

3.1.1 Background

We have devised a method called *SeqArray*, which aims to improve gene expression estimates from microarrays, by learning the relationships between probe-level expression intensities and gene expression estimates obtained using RNA-seq, from samples for which both microarray and RNA-seq data have been generated. We have used a flexible regression technique, Multivariate Adaptive Regression Splines (MARS), to learn these relationships for each gene. In the training phase, the models learn how microarray probe intensities respond to varying expression levels of their corresponding genes. The trained models can then be used to estimate gene expression in samples for which only microarray data are available. One of the goals of this approach is to reproduce the association of mean RE-score with rs17409624 (identified using RNA-seq data) using gene expression microarray data.

3.1.2 Results

We identified 52 and 40 samples from the HapMap YRI and CEU populations, respectively, for which both microarray and RNA-seq gene expression data were available. Using the YRI data only we built MARS models for each gene. We then used these models to estimate gene expression levels in the CEU samples. We compared the performance of *SeqArray* and Affymetrix Power Tools (APT)

by determining the within sample and between sample correlations between gene expression estimated using these methods and using RNA-seq. The across sample correlation (between expression estimates of a given gene across all samples) was slightly higher for APT; however, gene expression estimates calculated using *SeqArray* had much higher within sample correlation with RNA-seq, indicating that it has performed much better in estimating the absolute expression level of each gene. We also developed a related method, that improves the performance of APT, by omitting probes that are not strongly correlated with the expression of their target gene (as measured by RNA-seq) in a training set of samples.

3.1.3 Conclusions

We have developed a method which dramatically improves the ability of microarrays to measure absolute gene expression levels. We also developed a method which improves the performance of APT, a well established tool for estimating gene expression from raw microarray probe intensities.

3.2 Background

In this Chapter, we propose *SeqArray* as a novel method to estimate gene expression levels, from probe-level fluorescent intensity data generated by microarrays. In Chapter 2, we had difficulty in reproducing the RNA-seq based GWAS result using the equivalent microarray data. *SeqArray* attempts to address this, improving gene expression estimates in microarrays by learning the relationship between probe level intensity on a microarray and gene expression level, as measured by an optimal technique. Here we treat RNA-seq as the optimal technique for gene expression estimation; however, gene expression estimates obtained using any technique could play this role.

Currently, the most common method of summarizing individual probe expression measures into gene level estimates is the median-polish algorithm, often as part of the RMA method. This algorithm takes account of probe and sample level effects in summarizing gene expression (see Introduction chapter). While there have been some efforts in recent years to introduce improved techniques [235][236][237], median-polish remains dominant. One of the few novel approaches in the literature is SCOREM [238], which is for Affymetrix GeneChip microarrays. It uses a statistical test, Kendall's W coefficient of concordance, to consolidate expression estimates from redundant probesets to provide a more reliable measure of gene expression. Frozen RMA (fRMA) [239] is another novel approach, which can be used to improve normalization of experiments with small sample size and allows comparisons between batches. It works by using information in large publicly

available databases to estimate variances and probe specific effects for different microarray platforms (ordinarily these parameters are calculated separately for each batch by the median-polish algorithm). Outside these developments however, microarray summarization methods have remained largely stable in the last number of years.

Several studies have compared gene expression estimates from RNA-seq and gene expression microarrays on the same samples. The results have been consistent in different experiments, with Spearman correlation coefficients between expression estimates from the two platforms in a given sample, typically of approximately 0.75. However, correlation between expression estimates for a given gene across samples is typically lower, at approximately 0.15 [160][240][154]. This indicates that there is often considerable disagreement concerning the change of expression levels across samples, as measured by the two platforms. As detailed in the Introduction, RNA-seq has been shown to produce more accurate estimates of gene expression than microarrays [241][154][242]. Our proposed method makes use of overlapping gene expression datasets, obtained using both RNA-seq and microarrays applied to the same set of samples. Using these overlapping datasets, we build statistical models to learn the relationships between individual microarray probe intensities and the RNA-seq gene expression estimates. The method could be applied to any microarray platform and it is not limited to training on RNA-seq data; in theory, any more accurate technology developed in the future could be used as the basis to train the statistical models. It is also unnecessary to restrict the method to any particular modeling technique and it is possible that an as yet undeveloped or untested algorithm will outperform what we have used in this study. Moreover the application of this approach should allow expression estimates to be compared more easily between different laboratories and even different microarray platforms.

Here, we have chosen to use Multivariate Adaptive Regression Splines (MARS) [243] to learn the relationships between the probe expression intensities and the gene expression levels. MARS is a flexible extension of linear models that allows the modeling of non-linearities using splines. A spline is a piecewise function, meaning that its definition changes based on the value of the independent variable [244]. For most applications of splines, the piecewise function is constructed from a set of polynomials (which describe a set of curves); however MARS uses a set of straight lines. MARS can also be applied to high dimensional data, which makes it suitable for probe level microarray data. MARS builds models of the following form:

$$\hat{f}(x) = \sum_{i=1}^{k} c_i B_i(x)$$

where B_i is a set of basis functions (the constituent elements of the piecewise spline function), with each function modeling the relationship between input and response variables over a particular range of values. A juxtaposition of a linear model and a MARS model for simulated two-dimensional data is shown below (Fig. 3.1 and Fig. 3.2); these figures give a clear visual indication of the ability of MARS to model non-linearities. An illustration of MARS in three dimensions is provided in figure 3.3. It becomes impossible to visualise these models beyond three dimensions.

The MARS model in figure 3.2 is represented by the following equation:

$$\hat{y} = 20.40 + 5.06 \max(0, x - 12) - 2.55 \max(0, 12 - x)$$

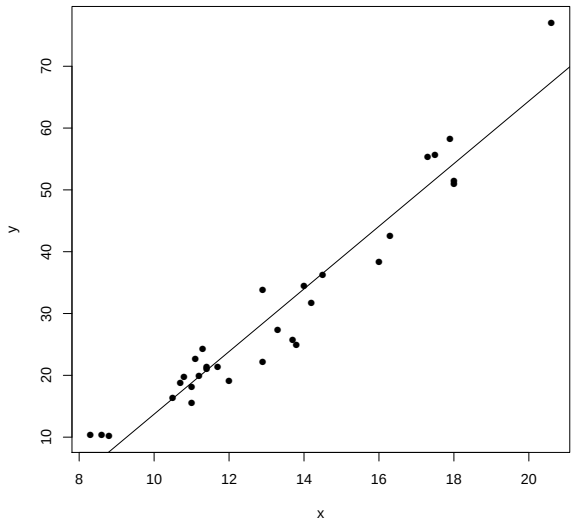

Figure 3.1: Linear Regression on simulated two dimensional data.

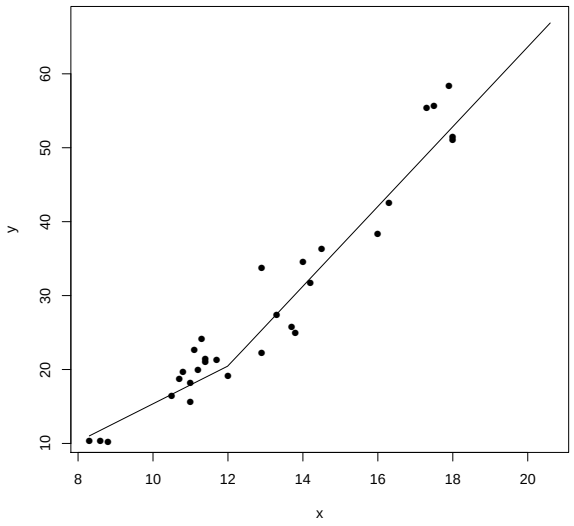

Figure 3.2: MARS model on the same set of two dimensional data.

The positions in the model where MARS infers non-linearities are known as knots, the basis functions which produce the knots in the graph are known as hinge functions. In the example given, MARS has identified a knot at $x = 12$ and the hinge functions indicate that the slope of the line has changed from 2.55 to 5.06 at this point. MARS calculates the values for these hinge functions in two steps, the forward pass and the backward pass.

During the forward pass, a MARS model begins as a straight horizontal line through the mean value of the response variables. MARS then repeatedly adds

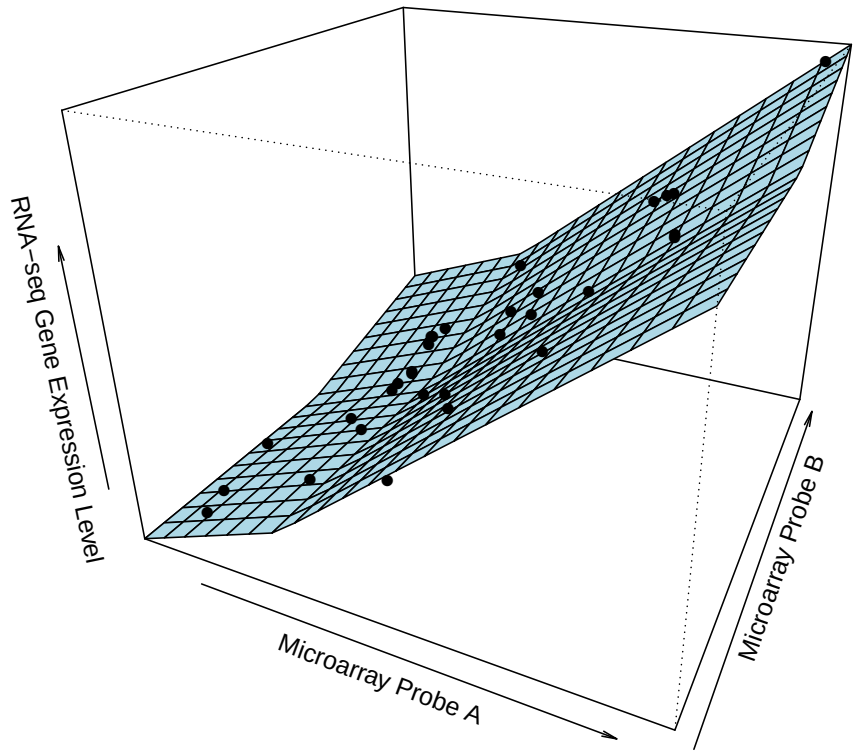

Figure 3.3: A MARS model fitted on simulated data for the expression level of two hypothetical microarray probes against the RNA-seq measured expression level of their corresponding gene.

pairs of mirrored hinge functions, identifying which pairs of functions give the biggest reduction in sum-of-squares residual error. This process is repeated until the maximum number of iterations is reached or the residual error reaches a cutoff threshold. To avoid overfitting MARS implements a backwards pass step, in which the least effective terms are iteratively removed. These models are then compared using generalized cross validation (GCV), which estimates model performance by rewarding for goodness-of-fit while penalizing for model complexity (i.e. the number of knots). Using these calculations, the best overall model is identified.

GCV is calculated as follows:

$$GCV = \frac{SS_{err}}{N \times (\frac{1-E}{N})^2}$$

$$E = NTerms + Penalty \times \frac{NTerms-1}{2}$$

Where "SS_{err}" is the residual sum of squares, "N" is the number of observations (i.e. the number of microarray probes) and "NTerms" is the number of hinge function knots in the model, so that GCV penalizes for the number of knots in the model. "Penalty" can be set to adjust how severely model complexity is penalized.

Previously, MARS has been successfully applied to many types of problems across multiple disciplines, for example predicting climate change [245], forecasting energy prices [246] and breast cancer diagnosis [247]. MARS is useful for predicting gene expression from microarray probe level data because probe intensities do not necessarily scale linearly with gene expression (Fig. 3.4). Expression profiles may also be significantly different for probes annotated to the same gene because of different probe sequences, which leads to differences in levels of cross hybridization and polymorphisms in the target gene sequence [248]. MARS is capable of capturing this information and learning how each individual probe responds to changes in overall gene expression.

To test the method, we used RNA-seq and Affymetrix Human Exon 1.0 ST data generated from the HapMap project [240][160]. This is the same data that was used in Chapter 2. These datasets are useful for testing *SeqArray*, as they consist of RNA-seq as well as microarray data from a substantial number of samples. We used the YRI samples as a training set to build the statistical models and then used these models to predict gene expression in the CEU. In the training phase, for each gene, a MARS model was fitted for gene expression level (as estimated by RNA-seq) as a function of microarray probe level expression intensities. In the prediction phase, the trained models were used to estimate gene expression from probe level expression, in the CEU microarrays. We then compared the performance of *SeqArray* and Affymetrix Power Tools (APT), which is a standard package for estimating gene expression levels in Affymetrix Human Exon microarrays. APT was run using the RMA background correction/normalization/summarization method. Our objective was to develop a microarray preprocessing method that mirrors the RNA-seq expression estimates as closely as possible. Therefore, as an evaluation of performance, we compared the extent of correlation between the RNA-seq and the microarrays, preprocessed using our method or APT. As a further evaluation we considered expression quantitative trait loci (eQTLs) inferred in the CEU sam-

ples. Method performance is evaluated based on similarity in the eQTLs identified using *SeqArray* and APT to the eQTLs identified from the RNA-seq data. Note, however, that the method does not assume that RNA-seq provides accurate gene expression estimates and results obtained using any alternative gene expression measurement could be substituted for RNA-seq at the training stage.

3.3 Results and Discussion

3.3.1 *SeqArray* improves within-sample correlation with RNA-seq

In our tests *SeqArray* outperforms APT for within-sample correlation with RNA-seq. Scatterplots of gene expression levels, for CEU sample "NA06985", of RNA-seq against APT and *SeqArray* respectively, are shown below (Fig. 3.4). There is a much more linear and tightly clustered relationship between RNA-seq and microarray expression estimates for *SeqArray* than for APT. This is further supported by the increase in Spearman and Pearson correlations with RNA-seq, from 0.85 to 0.90 and 0.19 to 0.92, respectively, when microarray data are processed using *SeqArray*, rather than APT. This means that if we assume the RNA-seq expression estimates are accurate, *SeqArray* has outperformed APT in correctly estimating the expression levels of genes within samples. Because of differences in probe sequences and varying levels of cross-hybridization, microarrays are not expected to accurately assess the relative expression level of genes within a sample, meaning that using existing methods, the expression levels of two different genes, can not be reliably compared within the same sample. Differences in probe sequence and array design also mean that expression levels from different microarray platforms cannot be reliably compared. This is not the case with RNA-seq, where normalized expression measures (for example FPKM) for different genes within a sample are indicative of relative transcript abundance. Our results show that, following processing with *SeqArray*, the distribution of microarray expression estimates now closely matches that of RNA-seq. This indicates that *SeqArray* can potentially address these problems and, given the availability of suitable training data, can allow comparison of gene expression between different microarray experiments and potentially even between different microarray platforms.

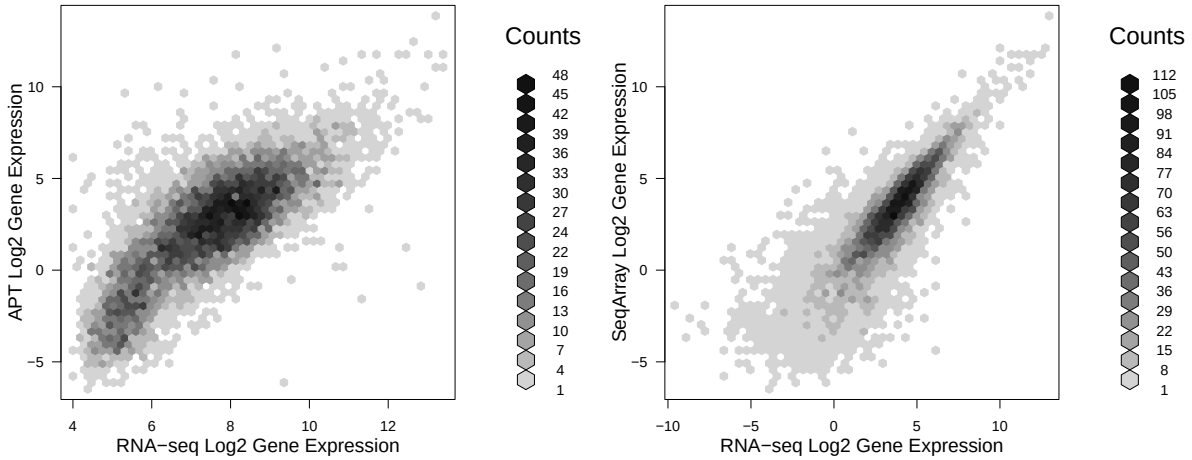

Figure 3.4: Scatterplots of log transformed gene expression levels, for CEU sample "NA06985", of RNA-seq against APT (left) and *SeqArray* (right). Note the much more linear and tightly clustered relationship for gene expression estimates calculated using *SeqArray*.

3.3.2 Across-sample correlation is not improved

The vastly improved within-sample correlation is, however, not matched with an overall improvement in across-sample correlation. This is often the major requirement of gene expression experiments, for example, differential expression studies, as these kinds of analysis compare expression levels of genes across different samples. Fig. 3.5 and 3.6 illustrate histograms of Spearman correlations with RNA-seq, of each gene, across all samples of the CEU for APT and *SeqArray* respectively. The average across sample Spearman correlation for APT is 0.15, clearly outperforming *SeqArray*, whose mean across sample correlation is only 0.06.

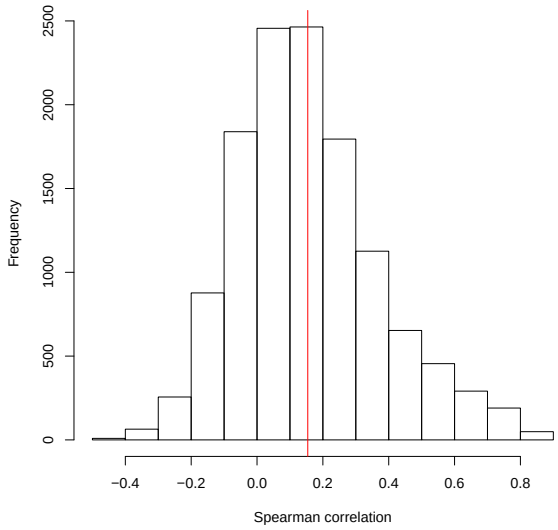

Figure 3.5: Spearman correlations between APT and RNA-seq gene expression levels, across all samples of HapMap CEU population.

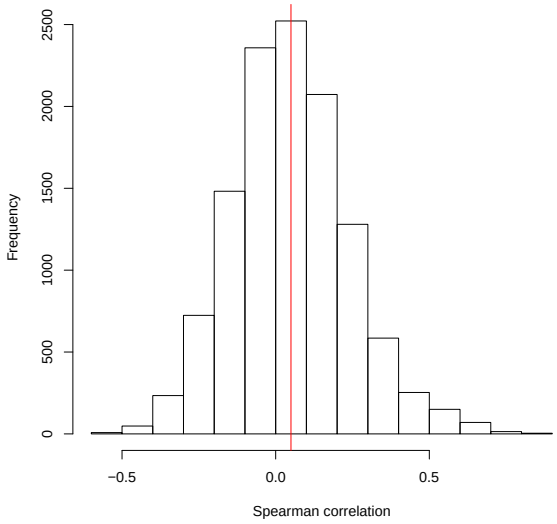

Figure 3.6: Spearman correlations between *SeqArray* and RNA-seq gene expression levels, across all samples of HapMap CEU population.

3.3.3 Adjustments which improve across sample correlation with RNA-seq

We have identified several steps that improve the performance of *SeqArray*. These are, (1) silencing probes which are not detected above background noise, (2) setting all genes whose expression is predicted to be negative to zero and (3) fitting the models using only probes whose expression is strongly correlated with that of their target gene in the training set.

Affymetrix also annotate unambiguous exonic probesets as belonging to one of three evidence levels, "core", "extended" or "full". The core group are supported by RefSeq annotations and these are typically used in APT analysis; the extended probesets are supported by EST evidence and the full set includes computationally predicted probesets. The average number of probes per gene, for core, extended and full probesets is 59.58, 73.19 and 75.96 respectively. However, the evidence suggests that "core" probes perform best, suggesting that including these additional probes in the models does not improve performance.

To silence probes not expressed above background noise, we used the "Detected above background" (DABG) algorithm, which was developed by Affymetrix and can estimate the probability that probesets expression is detectable above back-

ground noise. It works by comparing the expression level of each probe, to the expression level of a set of background probes, which have similar GC content, but whose sequence does not target a real exon. We set the expression level of all probes that belonged to probesets not detected at a $p > 0.05$ threshold to zero. This is done in both the training and prediction phases.

The expression levels of some genes will also have been predicted to be less than zero by the MARS models. As a negative value for gene expression cannot exist in reality, the expression level of these genes is set to zero.

Finally, we implemented an approach whereby only probes that are likely to contain useful information were used in fitting the models; these are identified by calculating a Spearman correlation between each probe and its target gene (in the training set) and only using probes which showed a positive correlation. We examined the effect of using only probes with a positive correlation and probes with a high positive correlation (selecting only probes in the top 50% of positive correlations). Adding this step gives the biggest improvement in performance.

A summary of the improvements in correlation with RNA-seq gained by these adjustments is presented in Tables 6.4 to 6.7 in Appendix B. In all cases applying DABG and setting non-expressed genes to zero has improved correlations. Within sample Spearman correlations reach 0.93, but across sample correlations remains behind APT, with a maximum across sample Spearman correlation of 0.12. Core probesets yield the best correlations with RNA-seq, suggesting that including the additional (extended and full) probesets does not improve performance.

When we compare the best *SeqArray* based approach to APT we see that across sample correlations are still well behind. The number of significant ($p < 0.05$) across sample Spearman correlations for *SeqArray* is also far less, at 1,925, versus 3,025 for APT ($p < 2.2 \times 10^{-16}$ from Fisher's exact test)

As noted, the performance of *SeqArray* improves dramatically when the models are fitted using only probes which are highly correlated with their targeted gene in the training set. In the analysis above, we arbitrarily selected the top 50% of probes with positive correlations. We have assessed whether a different cutoff might lead to improved performance. To do this we have selected between 10% and 90% (in steps of 10%) of positively correlated probes and compared the performance of those with APT. However, while using ever smaller numbers of highly correlated probes does improve performance, APT also performs better for these subsets of genes (Fig. 3.7). As the number of probes uses to fit the models decreases, so does the number of genes for which we have enough probes to fit a model, e.g. when using only 10% of positively correlated probes, it is only possible to fit a model for 2500 genes.

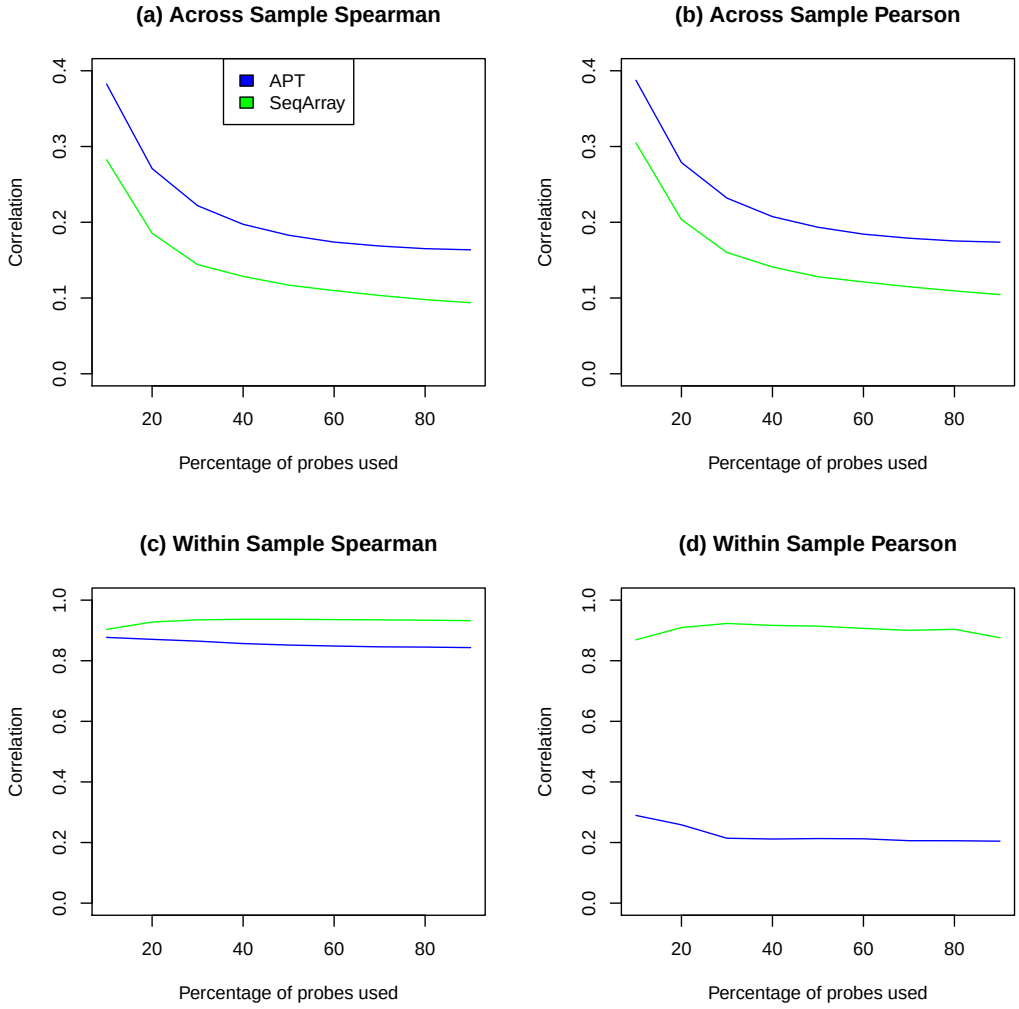

Figure 3.7: Correlations with RNA-seq when fitting MARS models using an ever smaller number of more highly correlated probes (in increments of 10%), on the HapMap CEU samples. Correlations for APT are are shown as a blue line and *SeqArray* as a green line.

3.3.4 Comparing the performance of *SeqArray* and APT on eQTL finding

Next, we compared the performance of *SeqArray* and APT for a real-world application, specifically, the ability of the two approaches to identify *cis*-eQTLs in the HapMap CEU data. For this, we used the best performing *SeqArray* expression matrix (in terms of correlation with RNA-seq); this was the data for which only

the "core" probes are used, where probes belonging to probesets not detected above background were set to zero, negatively expressed genes were set to zero and only probes with high (top 50%) positive Spearman correlation with their target gene in the YRI training set were used to fit the models. We then compared eQTL finding for the set of genes for which we had expression data from *SeqArray*, APT and RNA-seq platforms.

Using an additive linear model (see Methods for details), the CEU RNA-seq data identified a total of 941 significant *cis*-eQTLs ($p_{adj} < 0.05$). APT finds 733 significant eQTLs with 35.3% of these interactions also identified by RNA-seq. Unsurprisingly, given the lower across sample correlations, *SeqArray* does not outperform APT, identifying 533 significant eQTLs, with an overlap of 23.8% with the RNA-seq. This suggests that, at least as it is currently implemented, the poorer across sample correlation will limit *SeqArray* in many real world applications.

3.3.5 Identifying genes for which MARS fits better models

As outlined in the background section, MARS uses Generalized Cross Validation (GCV) to estimate model performance. This method rewards for goodness-of-fit while penalizing for model complexity (i.e. the number of knots). Lower values of GCV are indicative of a better model. We have investigated whether a subset of genes whose associated MARS models have low GCV, can outperform APT. To do this, we devised a simple analysis, whereby genes were segmented into 100 bins, based on the GCV value of their associated model. We then calculated the mean across sample correlation of the genes in each bin with RNA-seq. Figure 3.8 shows the change in across sample correlation across these bins. Genes with low GCV are in lower bins (i.e. genes with the lowest GCV values are in bin 1 and genes with the highest GCV are in bin 100). It is clear that genes with low GCV do show higher across sample correlation with RNA-seq, indicating that these models are performing better. However, APT also performs better for these subsets of genes, suggesting that the factors which allow these genes to estimate expression more accurately using our method, also apply to APT.

3.3.6 Comparing the performance of MARS models and linear models

As outlined in the background section, MARS models could be easily substituted for a different machine learning technique. As such, we have assessed the change in performance when MARS is substituted with linear models. Whether linear models outperform MARS is contingent on how linear the relationship between probe level expression and gene level expression (as measured by RNA-seq) is for

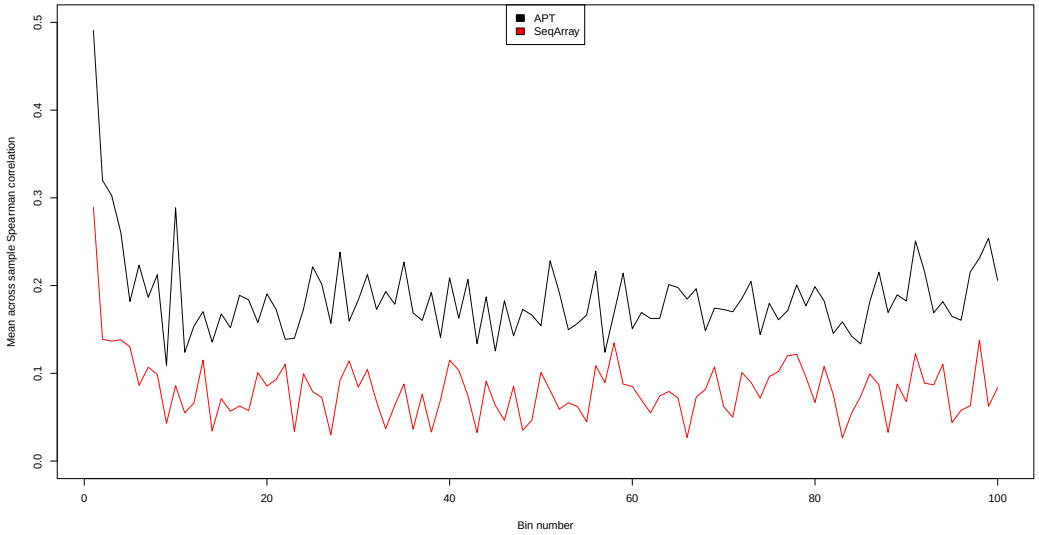

Figure 3.8: GCV against across sample Spearman correlation for *SeqArray* and APT. Performance for core probes with DABG and zeroing applied are shown. Models with low GCV clearly perform better, but the performance of APT also increases on this subset of genes.

most probes. If the relationship is, for the most part, highly linear, MARS may attempt to over-fit the data, incorporating knots where there is no need, which may mean that linear models will perform better.

Linear models pose a problem, because the number of covariates in the model must be less than the number of data points. This means that it is not possible to fit a linear model for a gene which has more probes than the number of samples available in the training data (52 samples in our YRI training set). Hence, for the core probesets, when all probes are included, it is possible to fit models in the YRI training set for only 5,347 genes. Training and testing are conducted as before; core probesets are used, DABG applied, negatively expressed genes are set to zero and we assess performance by using the models to predict expression on the CEU samples and comparing across sample correlations with RNA-seq. For this subset of genes linear models do not outperform MARS (Table 3.1)

	Linear	MARS
Across sample Spearman	0.048	0.056
Across sample Pearson	0.054	0.060
Within sample Spearman	0.854	0.908
Within sample Pearson	0.909	0.927

Table 3.1: Comparison of linear and MARS models.

We have also fit the linear models for only probes which show high positive correlation with the expression of their target genes (as measured by RNA-seq) in the training set. We compare these to our MARS models, which were fit with only highly positively correlated probes (the best performing MARS models). The linear models are fit for probes with high Pearson correlation, as opposed to Spearman correlation for the MARS models, as Pearson correlation also provides a measure of linearity. Again, the top 50% of probes with highest positive correlation are selected. Using only a subset of probes now means that it is possible to fit linear models for 8943 genes. Again, the performance (as measured by correlation with RNA-seq on this subset of genes) of MARS remains superior to that of linear models (Table 3.2). The ability to detect significant ($p < 0.05$) correlations with RNA-seq is only slightly improved using MARS compared to linear models, with 1345 as opposed to 1272 significant correlations detected ($p = 0.12$ from Fisher's exact test).

	Linear	MARS
Across sample Spearman	0.069	0.096
Across sample Pearson	0.080	0.101
Within sample Spearman	0.870	0.941
Within sample Pearson	0.844	0.949

Table 3.2: Comparison of linear and MARS models for models fit using only highly positively correlated probes.

3.3.7 Improving the performance of APT

As outlined above, fitting the MARS models using only highly positively correlated probes in the training set, markedly improves performance of the models when tested against RNA-seq. We have also found that a similar approach can be used to improve the performance of APT. More recent versions of APT (>= 1.8.0) allow the user to mask out individual probes when estimating gene expression in a set of samples. This functionality has previously been used to mask probes which overlap SNPs [249]. We have found that masking probes which do not

show high (top 50%) Spearman correlation with RNA-seq in the YRI training set, improves across sample correlations of gene expression estimates with RNA-seq in the CEU. The mean across sample Spearman correlation increases from 0.164 to 0.176, on a set of approximately 12,000 Entrez genes, for which there are enough probes remaining to estimate expression. The number of genes that were significantly ($p < 0.05$) positively correlated (Spearman correlation) between the microarray and RNA-seq has increased from 3,318 to 3,727 ($p = 4.034 \times 10^{-9}$ from Fisher's exact test). This suggests that, using this smaller and potentially more informative set of probes to estimate gene expression, can more accurately measure change in gene expression between different samples/phenotypes.

3.3.8 Using *SeqArray* to the investigate the genetics of miRNA regulatory effect

In the previous chapter, we were unable to achieve a significant association between the SNP rs17409624 in *DROSHA* and mean RE-score using the microarray data. No significant SNPs were identified using the array data which suggests that, unsurprisingly, RNA-seq may have more power to detect such associations. Here, we try again to reproduce the result, using expression estimates from *SeqArray*.

We trained the models used to calculate expression in the CEU data have been trained on the YRI data and the models used to calculate expression in the YRI have been trained on the CEU. The core probesets were used, the DABG algorithm was applied and negatively expressed genes were set to zero. RE-scores were calculated using the TargetScan prediction algorithm. Analysis of association of mean RE-score with SNPs in the miRNA biogenesis pathway was performed as previously described. This first approach did not successfully reproduce the significant result obtained for rs17409624 using RNA-seq ($p = 0.89$ in the CEU and $p = 0.88$ in the YRI; Fig. 3.9)

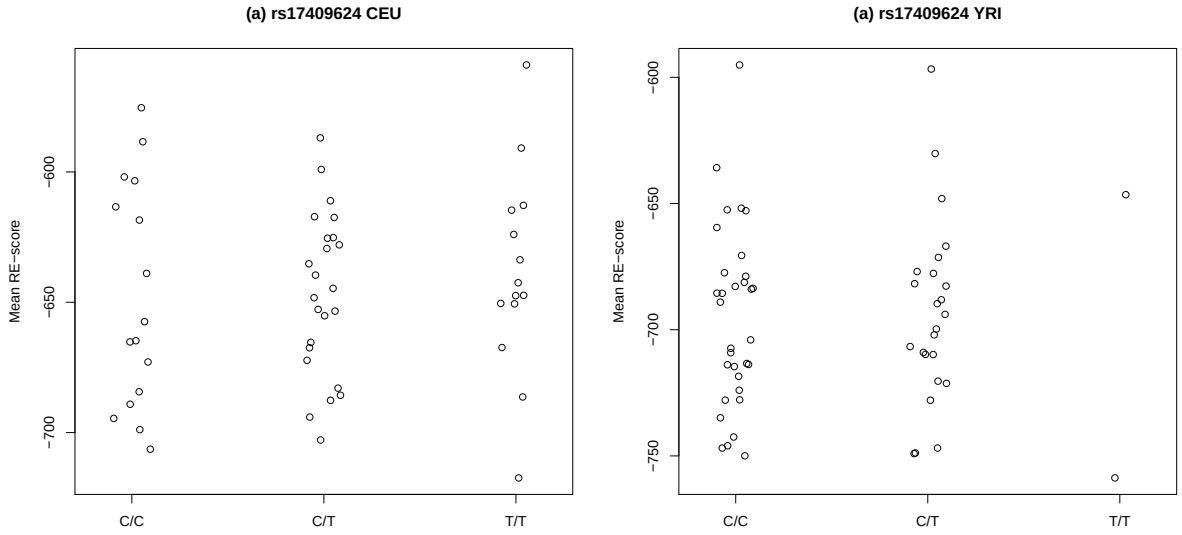

Figure 3.9: Stripcharts of mean RE-score against rs17409624 genotypes for the CEU and YRI, calculated using *seqArray* gene expression estimates.

Finally, we attempted to reproduce the result by using as much training data as possible, when predicting expression for each sample. This time, instead of training the models used to estimate expression in the YRI samples on the CEU and vise versa, we trained on all available samples (both CEU and YRI), leaving out the one sample for which we wish to predict expression. Using this method, gene expression on each sample is estimated using a different set of models, which were trained on all other available samples. An approach like this could never have any useful real world application, as clearly one could simply use the RNA-seq data, but it is implemented here, to give an indication of whether it may be possible to improve the performance of *SeqArray*, by adding more training data and to establish if it is possible to reproduce the results of the RNA-seq association analysis using the microarray data. We used the core probesets, applied DABG, set negatively expressed genes to zero and fit the models using only probes which showed a high positive correlation with the RNA-seq estimated gene expression level across all available samples. However, this approach does not improve the results obtained for rs17409624 ($p = 0.89$ in the CEU and $p = 0.88$ in the YRI; Fig. 3.10). As RE-scores are only meaningful when compared across samples, the poor across sample correlation achieved between RNA-seq and *SeqArray* is the most plausible explanation for this result.

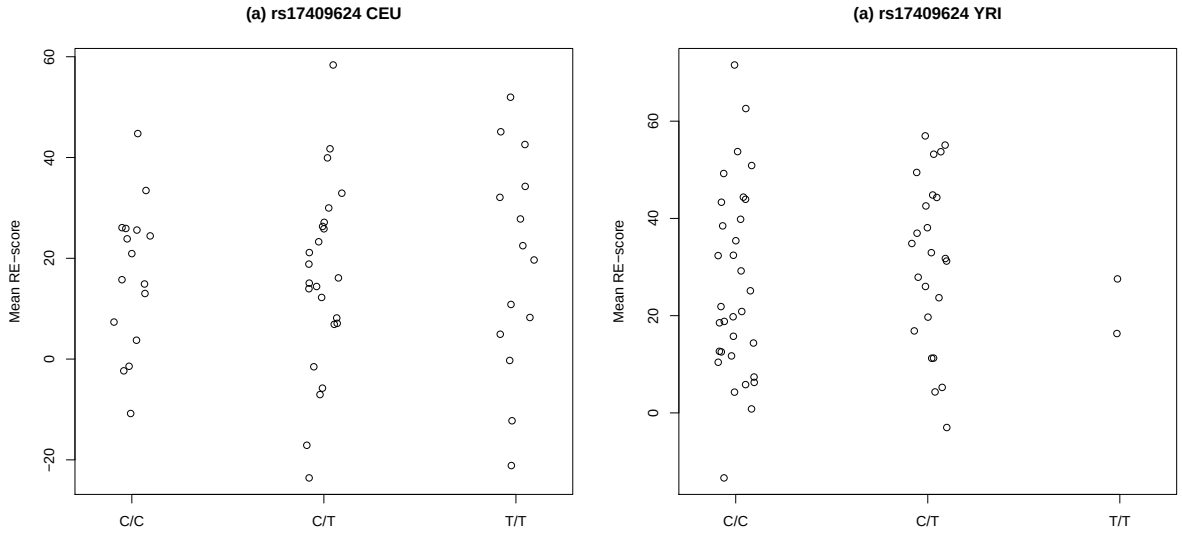

Figure 3.10: Stripcharts of mean RE-score against rs17409624 genotypes for the CEU and YRI, calculated using *seqArray* gene expression estimates and using as much training data as possible.

3.4 Conclusions and future work

In conclusion, we have developed a novel method to estimate gene expression from microarray probe-level fluorescence intensities. We have applied this method to Affymetrix Human Exon ST 1.0 arrays, using matching RNA-seq data as a training set; we used MARS models to relate microarray probe level intensities to RNA-seq gene level expression measures. To test the method we used these models to estimate gene expression on a separate set of samples for which exon array and RNA-seq data were also available. Our method has not outperformed APT, in terms of across sample correlation with RNA-seq. This means that our method is not as sensitive to change in gene expression between different samples. Most real world applications (for example differential expression analysis) are primarily concerned with comparing expression levels across samples, which means that we have, as yet, been unable to develop *SeqArray* to a level where it is more useful in these kinds of scenarios.

However *SeqArray* has far outperformed APT in both Pearson and Spearman within sample correlation with RNA-seq. This means that it is more accurately estimating the expression and rank of genes within samples. Because of differences in hybridization affinity between different probes, estimating absolute expression level of genes is not an application for which microarrays are often used. This

is illustrated by the poor within sample Pearson's correlation, averaging at only 0.19 observed in the HapMap CEU, between expression levels estimated by APT and RNA-seq. This is improved dramatically by our method, which shows an average within sample Pearson's correlation > 0.9. Although less common, there are undoubtedly some applications where accurately estimating the absolute level of expression of a gene may be useful. One example is a previous study which assessed the relationship between intron length and gene expression level (which were measured using microarrays); highly expressed genes were found to tend to have shorter introns [250]. In this case, our method would have been far more accurate in estimating which genes were in fact highly expressed.

The training data that we have used here is also very limited. All of the gene expression estimates are from lymphoblastoid cell lines, which means that, for the majority of genes, a great deal of variation in gene expression between the different samples cannot be expected. Without at least some variation in expression in the training set it is impossible to fit useful, broadly applicable models. This is also supported by our observation that across sample correlations improve for genes that have more variable expression in the training set (as measured by RNA-seq). There are some other datasets that are becoming available which have assessed gene expression using both RNA-seq and Affymetrix Human Exon ST 1.0 arrays on the same samples; future work may focus on incorporating more of these datasets in the training phase, assuming that they are of sufficient quality.

A method similar to what we have outlined in this chapter may also be capable of estimating transcript level expression in exon microarrays, as opposed to just gene level expression. There are well established methods of estimating transcript abundance in RNA-seq data, but at the time of writing we are not aware of a tool which can accurately estimate the expression of different isoforms from exon array expression data. The main challenge in this will be taking account of the expression of overlapping transcripts when fitting the models. This could perhaps be achieved using models that allow for a multi-variate response (i.e. multiple transcripts dependent on the expression of multiple probes).

Finally, we have devised a separate method, by which to improve across sample correlations between APT gene expression estimates and RNA-seq. This works by estimating gene expression using only probes which show a positive correlation with their target gene in a training set. As with *SeqArray*, it is possible that this method could also be further improved using a more diverse set of training data.

3.5 Methods

3.5.1 Data Analysis

All data analysis were performed in R. We used the same gene expression data as in Chapter 2. RNA-seq data were processed as outlined in the Methods section of that chapter. Microarray data were processed by Affymetrix Powertools as previously described. Training and prediction microarray data were quantile normalized together at probe level, using the Bioconductor library *XPS* [251]. This library allows very large datasets to be processed using only a small amount of memory. It is built on top of the ROOT data analysis framework [252], an application developed by CERN, which works by indexing the raw data on the hard disk instead of loading it into memory. This is necessary, as the data are too large to be loaded on any machine to which we had access, the largest of which had 32gb of RAM. *XPS* also implements the DABG algorithm. MARS models were fit using the *mda* [253] library in R.

3.5.2 eQTL Finding

We used an additive linear model [254] to identify eQTLs. P-values were corrected using the Benjamini and Hochberg method and eQTLs were called significant at an FDR < 0.05. This is a highly computationally intensive task; hence, we used the ".C" interface in R, to call a function, written in the C programming language, which performed the calculations. This function was based on the GNU Scientific Library (GLS) [255]. We only considered *cis*-eQTLs and regarded a SNP as associated with a particular gene if it is between 2kb upstream and 500bp downstream of the gene.

Chapter 4

Ulcerative Colitis is associated with CpG island hypermethylation in sigmoid colon tissue

Some of the contents of this chapter have been submitted for publication as part of:

Affendi, RRA., Hartnett, L., Newell J., Golden, A. Seoighe C, Geeleher P., Egan, LJ. Analysis of the interleukin-6/STAT3 signalling pathway and DNA methylation patterns in patients with short or longstanding inflammatory bowel disease. *Inflammatory Bowel Diseases*

4.1 Abstract

4.1.1 Background

Individuals suffering long standing ulcerative colitis (UC) have an increased risk of developing colorectal cancer [256]. It has been previously shown that aberrant DNA methylation is associated with colorectal cancer. Previous work has also shown that there is likely a causative link between DNA methylation in UC and the increased risk of colorectal cancer. Hence, we have investigated whether there is evidence of differential CpG island methylation in sufferers of UC, by comparing the genome-wide methylation pattern in sigmoid colon tissue of 5 UC patients and 5 healthy controls, using data generated from the Agilent Human CpG Island microarray.

4.1.2 Results

We have found evidence of overall CpG Island hypermethylation in UC. This result is consistent with previous findings in colorectal and other cancers. However, differential methylation at the level of individual probes or CpG islands was not significant following correction for multiple testing.

4.1.3 Conclusions

CpG Island hypermethylation is associated with UC, which may be a factor in the increased risk of carcinogenesis. In this analysis, we have been unable to identify which individual genes are more likely to be targeted by this methylation. This is likely due to the small sample size, but it may also be the case that overall CpG island hypermethylation in UC is not targeted at any particular genes.

4.2 Background

Inflammation is a normal biological response to damage caused by external factors such as physical injury, burns, toxins and infection by pathogens. In humans, acute inflammation is a vital defense mechanism which is crucial to survival. It causes infiltration of white blood cells to the affected area and leads to redness, swelling, heat, pain and disturbance of function [257] [258]. In some cases, a persistent or prolonged inflammatory response may occur; this is known as chronic inflammation and is the cause of harmful disorders such as celiac disease, asthma and inflammatory bowel disease (IBD). Chronic inflammation is also associated with approximately 20% of cancers in humans [259] [260]. Ulcerative colitis (UC) is a common chromic inflammatory disease. It is a class of IBD and affects up to 200 in 100,000 individuals in some populations [261]; it is characterized by chronic inflammation of the gastrointestinal tract [262]. Common symptoms include abdominal pain, blood and pus in stools, diarrhea and weight loss. Medical interventions include drugs, dietary change [263], or in severe cases, surgical removal of part of the bowel [264].

The causes of UC are not well understood, but early studies on twins revealed a likely genetic component [265]. More recently, genome-wide association studies (GWAS) have identified approximately 100 loci which are associated with UC, but the individual contribution of these alleles is small, suggesting that susceptibility is influenced by many genomic variants [266] [267] [268]. Sufferers of UC are 2- to 3- times more likely than average to develop colorectal cancer (CRC), over the course of their lifetime. Colitis associated cancer (CAC) is similar to sporadic CRC, although the frequency and timing of malignancy is thought to be influenced by the inflammation [256] [269] [270]. The most important factors determining

risk are the duration, extent and severity of the UC [271] [272]. The use of anti-inflammatory drugs has also been shown to significantly decrease the risk of cancer in individuals suffering chronic inflammation, which establishes strong evidence for a causal link between chronic inflammation and cancer [273].

Epigenetic changes, particularly DNA methylation and histone modification, play a key role in cancer [274] (see Introduction chapter for detailed discussion). CRC is among the cancers affected [275]. Methylation in cancer involves genomic hypomethylation and hypermethylation of gene promoter regions. These lead to genomic instability and downregulation of tumor suppressor genes; factors which, followed by a selection process, contribute to uncontrolled cell proliferation [44][45][46]. Chronic inflammation is associated with aberrant DNA methylation in several conditions, such as Barrett's esophagus [276], chronic biliary tract inflammation [277] and IBD [271]. Several previous studies have identified an overlap between genes hypermethylated in UC and CRC (Table 1). This suggests that inflammation associated DNA methylation in UC, may be a factor in the increased risk of neoplastic transformation.

Gene Symbol	**Methylation Status**	**Disease**	**Reference**	**Array**
CDKN2A	Promoter hypermethylation	UC and CAC	[278][279][280]	Y
ESR1	Promoter hypermethylation	UC and CAC	[279][281]	Y
APC1A	Promoter hypermethylation	IBD and CAC	[282]	N
APC2	Promoter hypermethylation	IBD and CAC	[282]	Y
SFRP1	Promoter hypermethylation	IBD and CAC	[282]	Y
SFRP2	Promoter hypermethylation	IBD and CAC	[282]	Y
F2RL1	Promoter hypermethylation	UC	[283]	Y
SOCS3	Hypermethylation	CAC	[284]	Y
TUSC3	Promoter hypermethylation	UC	[285]	Y
CDH1	Hypermethylation	UC	[286]	Y
GDNF	Hypermethylation	UC	[286]	Y
ABCB1	Hypermethylation	UC	[287]	N

Table 4.1: Genes previously identified as differentially methylated in IBD, UC or CAC [12]. The "Array" column refers to whether a gene is represented on the Agilent Human CpG Island microarray.

In this chapter, we present the results of a study to characterize the genome-wide differential CpG island methylation in UC, using the Agilent Human CpG Island microarray. We compared promoter methylation in sigmoid colon tissue, between five individuals suffering UC and five healthy controls. Methylated DNA was isolated using the methylated DNA immunoprecipitation (MeDIP) protocol and input DNA and DNA isolated using a 5-methylcytosine (5mC) antibody are

competitively hybridized to each array [288][289]. These data were generated in the National Centre for Biomedical Engineering Science, NUI Galway (see Methods for details). Each microarray probe is targeted to a particular CpG site, within a CpG island; the log ratio intensity for each probe indicates the level of methylation of the corresponding targeted region. Comparing these levels of methylation between UC and control samples gives an indication of the level of differential methylation between the two phenotypes.

The majority of previous MeDIP-chip based experiments have set out to characterize a methylation pattern in a particular tissue or cell type [290][291]. Tools such as Agilent Genomic Workbench [292], Ringo [293][294] and Batman [295][296] have been used to analyse these experiments. Most commonly, methylation microarrays are used in an experimental design in which isolated methylated DNA from two different phenotypes are hybridized to the same microarray [297][298]; in that case, genes of interest are usually identified using an arbitrary probe level fold-change cutoff. The dataset discussed in this chapter is different, because samples of each phenotype have been hybridized to different arrays. At the time of writing, we are not aware of an equivalent published study, that has used the MeDIP-chip protocol and this experimental design, with this platform. Thus, we have analysed these data by developing methods based on existing strategies for analysis of ChIP-chip and gene expression microarrays. We have compared the performance of the various approaches by assessing their ability to find genes which have previously been identified as differentially methylated in UC (Table 4.1).

4.3 Results and Discussion

4.3.1 Data quality assessment

Quality assessment steps for CpG Island methylation microarrays are slightly different to those typically used for gene expression arrays. We used some of the steps suggested by Palmke *et al.* [290] and also applied some methods tailored specifically to this dataset. Boxplots of raw log intensity ratios did not indicate any sizable scaling differences between arrays (Fig. 4.1). Quantile normalization corrected the small differences that were evident (Fig. 4.2). We have also assessed pseudo array images, which did not reveal any spacial hybridization artifacts (Fig. 6.2 in Appendix C).

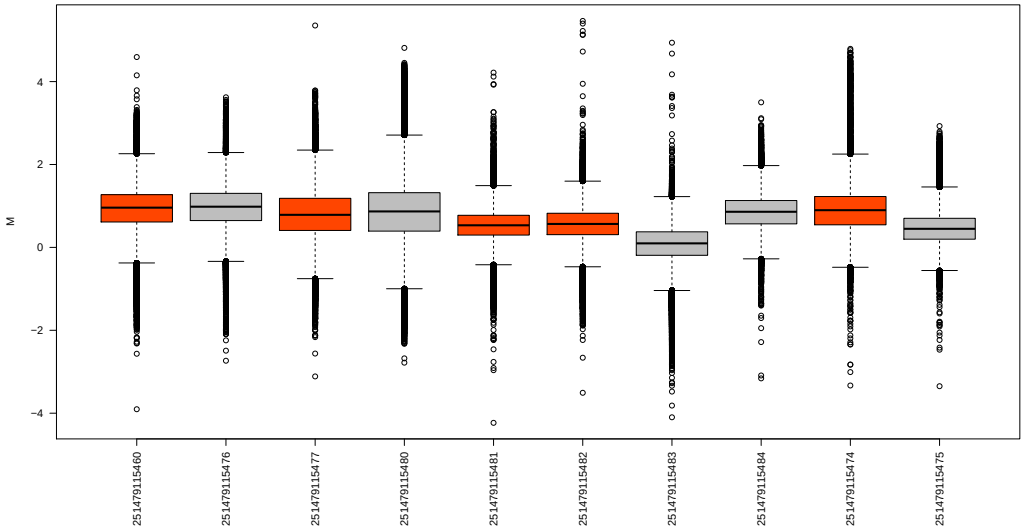

Figure 4.1: Boxplot of raw log intesity ratios. UC samples are highlighted in red.

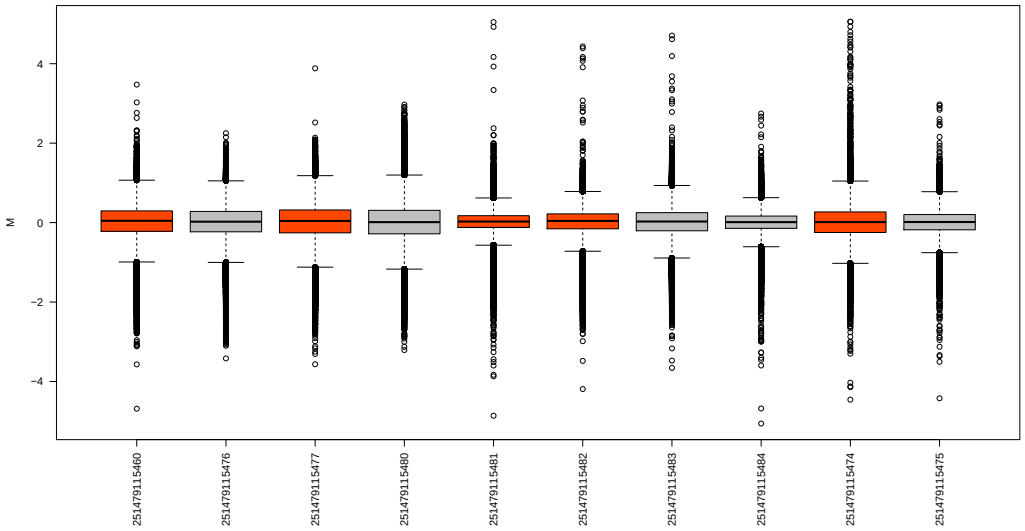

Figure 4.2: Boxplot of quantile normalized log intensity ratios. UC samples are highlighted in red.

We used principle components analysis (PCA) to visualize the similarity between log intensity ratio patterns in UC and normal samples. If the experimental condition was the main source of variation within the data, one would expect that

UC and normal samples cluster on opposite ends of principle component 1 (PC1). This is not the case for the raw or normalized data (Fig. 4.3 and 4.4). This suggests that some unknown factor(s) has a greater influence over the variability in measured methylation levels in the 10 samples, than their status as UC or normal. Any number of factors which influence methylation may be responsible, known examples include smoking [299] and diet [300]. Only age and gender are available as additional phenotypic information and we have included this on the PCA plots, but it is clear that neither of these is correlated with PC1. This indicates that there may be difficulty in achieving statistically significant results. However, UC and normal samples cluster on PC7, which means that PCA analysis has identified that there is potentially some subset of probes which vary consistently between the two phenotypes. These could be identified by statistical analysis. PCs 1-10 are shown in figure 6.1 in Appendix C. The proportion of variance captured by any of PCs 2-10 is much lower than that of PC1 (Fig. 4.5)

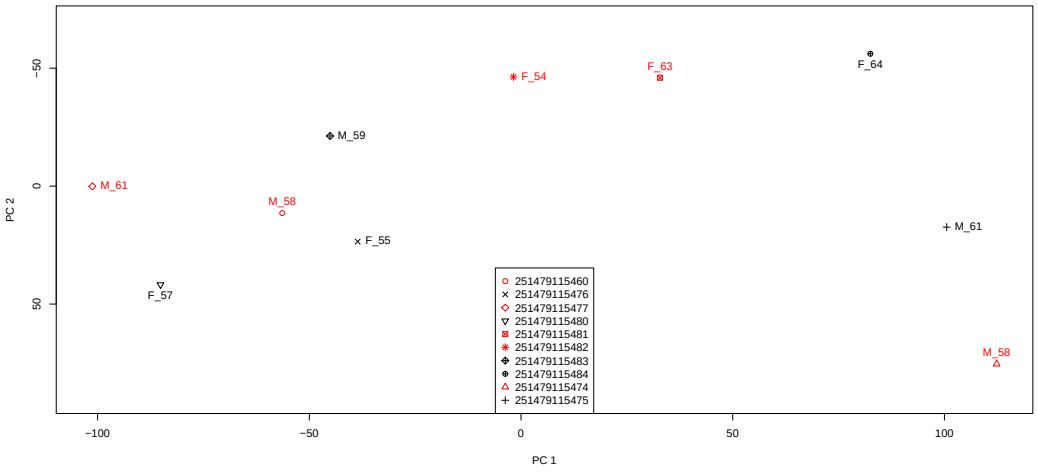

Figure 4.3: PCA plot for raw log intensity ratio data, showing PC1 (x-axis) and PC2 (y-axis). UC samples are highlighted in red.

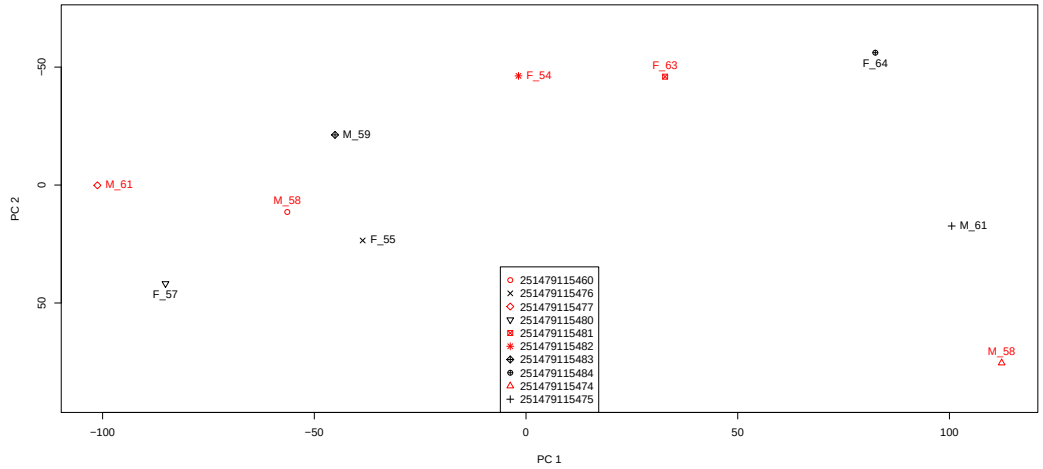

Figure 4.4: PCA plot quantile normalized log intensity ratio data, showing PC1 (x-axis) and PC2 (y-axis). UC samples are highlighted in red.

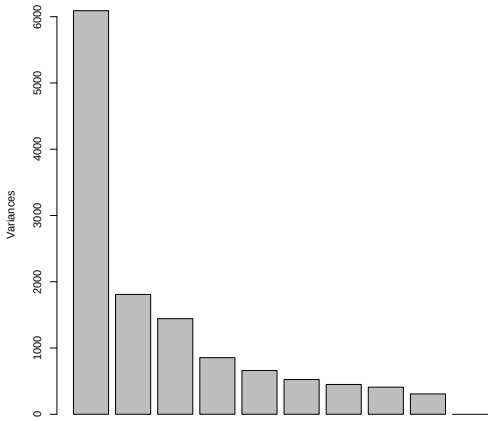

Figure 4.5: PCA scree plot, showing the proportion of variance captured by each PC on the normalized data. This result is similar for the raw data.

4.3.2 Adapting a ChIP-chip approach to identify methylated loci

We have first used a ChIP-chip type approach to quantify the number of methylated loci in each sample. This will give an indication of whether overall hypermethylation is evident in UC. To do this, we used Ringo, which is a Bioconductor

package designed for the analysis of two color ChIP-chip microarrays and genomic tiling arrays, but can also be applied to some types of methylation experiments [293]. Typical ChIP-chip experimental design is similar to MeDIP-chip; the difference is that, in the case of ChIP-chip, an antibody is used to isolate DNA bound to a protein of interest, but with MeDIP-chip, an antibody is used to isolate methylated DNA. Thus, Ringo can be adapted to characterize methylation profiles in MeDIP-chip samples [294][290].

The log intensity ratio distribution within a sample, for a typical ChIP-chip experiment is expected to follow a bimodal distribution [293], with the enriched probes contained in the smaller of two Gaussian distributions. Unsurprisingly, this also applies to MeDIP-based two channel microarray experiments [11](Fig. 4.6), but in this case the smaller distribution contains the probes enriched for methylation. The log intensity ratio distributions of our 10 samples are shown in figure 4.7. Some of these distributions appear somewhat bimodal, while others do not. This is not uncommon and the Ringo library implements functionality that takes account of this when identifying enriched probes. The Agilent documentation provides an explanation for why a clear bimodal distribution is not always evident, calling attention to the fact that different probes have different binding affinities (referred to as "melting temperature"). Binning probes based on melting temperature leads to cleaner enriched distributions and it is recommended that direct probe comparisons are only made within these bins [11]. Consequently, on each array, we divided probes into 17 bins of 0.1°C (Fig. 4.8). Using this approach, we determined the number of enriched probes in each sample (see Methods for details). A one sided t-test shows a greater number of enriched probes in the UC ($p = 0.036$; Fig. 4.9), which is consistent with our expectations. To establish if there is evidence of differential methylation between different loci, we have implemented statistical tests, which are discussed below.

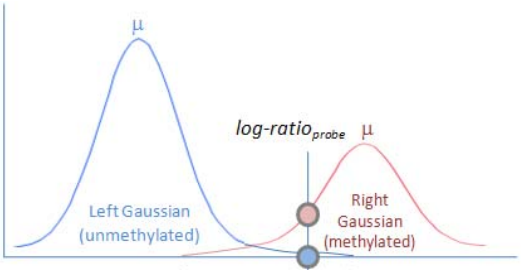

Figure 4.6: Typical bimodal distribution of log-ratios expected from MeDIP-chip experiment [11].

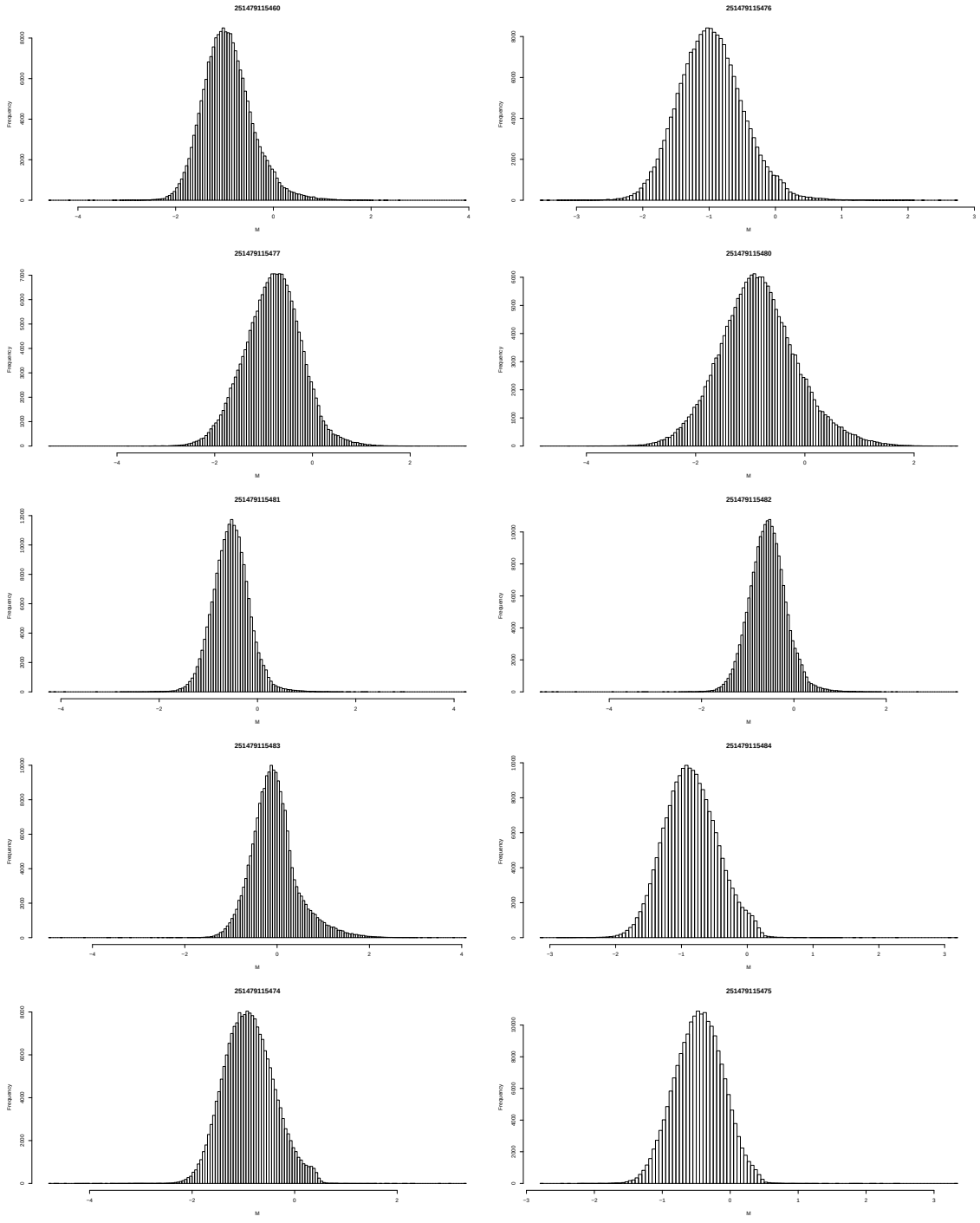

Figure 4.7: Log-ratio distribution for the 10 UC samples; in theory, enriched probes are in a smaller right Gaussian distribution.

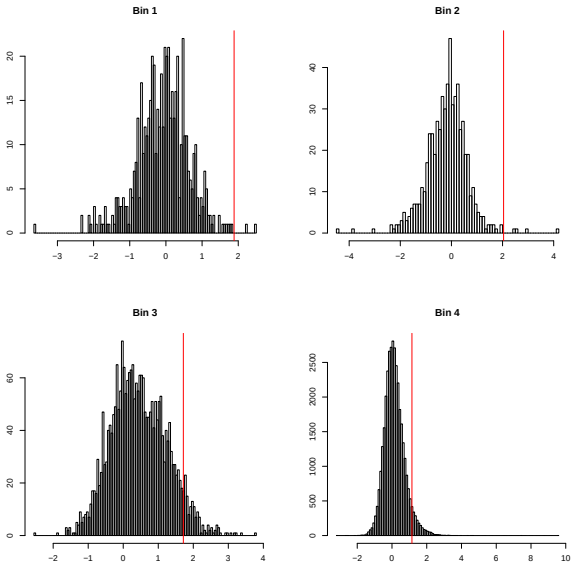

Figure 4.8: Histograms of bins 1, 2, 3 and 4 from sample 251479115460. The cutoff identified by the "*upperBoundNull()*" function from the Ringo package is included as a red vertical line.

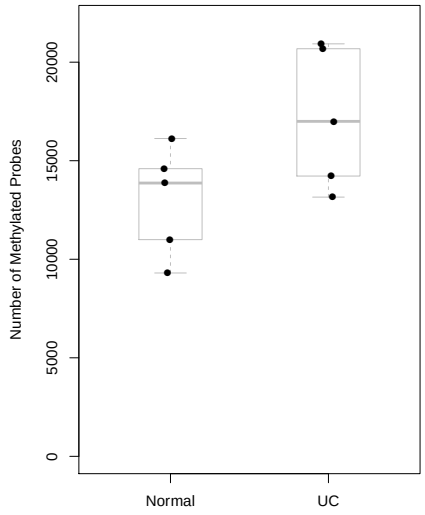

Figure 4.9: Number of methylated probes identified by Ringo in UC and normal samples ($p = 0.036$).

4.3.3 Statistical inference of differentially methylated probes

We have used several different methods to identify candidate lists of differentially methylated genes. None of these methods have been verified experimentally, nor has any method that we are aware of for this type of experimental design. Because of this we have used a list of genes previously found to be differentially methylated in UC, IBD or CAC (Table 4.1) as a benchmark by which to compare the various methods; 10 of the 12 genes in this table are represented on the Agilent Human CpG Island microarray.

Identifying differential probe methylation using *limma*

We used the Bioconductor library *limma* (normally used for analysis of gene expression microarrays) to assess differential probe methylation between UC and normal phenotypes. Unlike Ringo, this approach ignores the shape of the log intensity ratio distribution and instead attempts to directly identify statistically significant differences between UC and normal, in the quantile normalized log intensity ratios of probes or groups of probes.

Unsurprisingly, given the results of PCA, none of the individual probes reach the $p < 0.05$ threshold, following correction for multiple testing (Table 4.2). A histogram of the raw p-values reveals that there are many more high p-values (and hence much fewer low p-values) than would be expected by chance (Fig. 4.10). Nonetheless, we know that there is some consistent variation between the two phenotypes (identified by PC7), hence, we identified a list of nominally significant probes by using the $p < 0.05$ threshold. This identified 615 probes. Of these 380 showed hypermethylation in the UC, compared to 235 in controls and this difference is statistically significant ($p = 5 \times 10^{-9}$ from a binomial exact test). again suggesting CpG island hypermethylation in UC. Of the 380 probes, 2 map to genes previously identified as hypermethylated in UC (*GDNF* and *APC2*). The probability of 2 or more of these 10 genes being identified by chance is 0.09 (this is calculated using the method developed in the next chapter). This test assumes that all 10 previously reported genes are truly hypermethylated in all 5 of our UC samples; in reality, this is unlikely and as such 0.09 can be considered a conservative p-value estimate.

Next, we conducted a similar analysis, but this time included PC1 as a covariate in the linear model (which is fitted for each probe using *limma*); this approach will improve power to detect differential methylation of probes that are strongly influenced by PC1, while sacrificing power to detect those which are not. These types of methods, which use PCA to identify hidden confounding factors, that are subsequently included in statistical models, have previously been applied to, for example, eQTL analysis [301]. Again, no significant probes remained following multiple testing correction. This time, 1,633 hypermethylated probes reached the

$p < 0.05$ threshold. However, this did not appear to improve the power to detect true associations, with only 3 of these probes mapping to the same set of 10 genes which yields a less significant p-value than previously ($p = 0.32$). This suggests that factoring in PC1 has not improved the analysis.

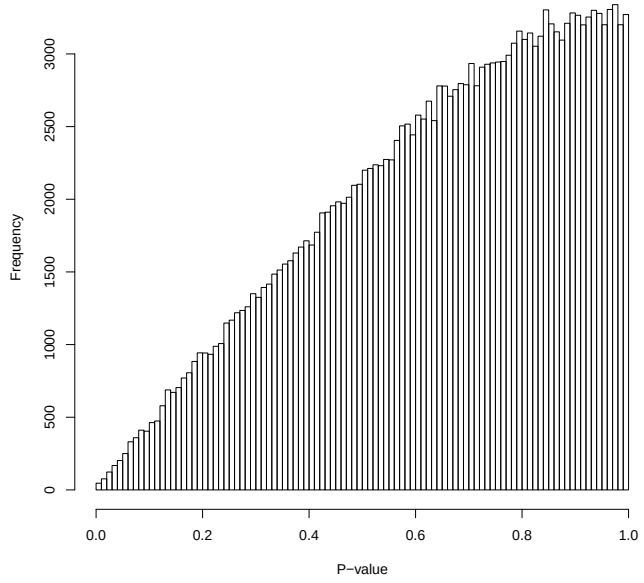

Figure 4.10: P-value distribution for all probes, from the *limma* differential methylation analysis.

Probe ID	**Gene Symbol**	**Log FC**	**P-Value**
A_17_P17272535	None	−0.41	1.09×10^{-03}
A_17_P15810386	None	−0.56	1.15×10^{-03}
A_17_P15865576	SLC35D3	−0.56	1.35×10^{-03}
A_17_P16393414	LHPP	−0.37	1.87×10^{-03}
A_17_P16975164	PIGW	−0.65	2.18×10^{-03}
A_17_P16879789	FUS	−0.39	2.70×10^{-03}
A_17_P15219918	FAM84A	−0.49	2.86×10^{-03}
A_17_P16897414	IRX3	−0.49	3.32×10^{-03}
A_17_P16464772	KIAA1394	−0.82	3.50×10^{-03}
A_17_P16645611	PCDH9	−0.42	3.53×10^{-03}

Table 4.2: P-values for top 10 probes identified by *limma*. Negative fold-change implies hypermethylation in UC. False discovery rates are 1 for all of these probes.

Identifying differential probe methylation using fold changes and a non-stringent p-value threshold

The MicroArray Quality Control (MAQC) project has suggested, that in the case of gene expression microarrays, identifying differentially expressed genes using a fold change and non-stringent p-value cutoff leads to improved reproducibility, although strict p-value based methods are still much more commonly used [302]. Consequently, we also tried a fold change based approach. We used probes with $p < 0.25$ (as identified by *limma*) and selected 615 probes with the highest absolute value of log fold change. This number of probes was selected to facilitate comparison with the previous analysis, which identified 615 probes with $p < 0.05$. Again, many more of the probes identified are hypermethylated in UC (457 vs 158). Two of the set of ten genes of interest are identified, *APC2* (also identified by the p-value based approach) and *SFRP2*. This indicates that, for these data, a fold change based cutoff has approximately the same power to detect differential methylation as a strict p-value cutoff. Importantly, this alternate approach is again suggestive of CpG island hypermethylation in UC.

Identifying differential probe methylation using variance filter and t-tests

We have also attempted to identify differentially methylated probes using t-tests. To reduce the multiple testing problem, we first filter out 50% of probes, which show the lowest variance across all samples, as these probes are unlikely to be differentially methylated [303]. 214 probes reached the $p < 0.05$ threshold and of these 165 showed hypermethylation in the UC (again consistent with the previous approaches). Unsurprisingly, the probes identified are very similar to those obtained using *limma*, with 211 of the 214 probes also having $p < 0.05$ in the *limma* analysis. However, the high false discovery rates remain and none of the probes survive Benjamini and Hochberg correction for multiple testing (Table 4.3). T-tests do not improve on *limma* as regards finding genes which were previously identified as hypermethylated in UC, with only one (*APC2*) of the two genes identified by *limma* reaching $p < 0.05$.

4.3.4 Assessing differential methylation at CpG island level

The Agilent documentation states that "It is generally observed that CpG islands, measured on the Agilent catalogue CH3 design array, are either fully methylated or fully unmethylated". This statement is not referenced, nor have we found any supporting evidence in the literature. In fact, there are studies which show differential methylation occurring preferentially at CpG island shores [304]. This has also been shown to be the case in colon cancer [305]. Nonetheless, we have applied

Probe ID	Gene Symbol	Log FC	P-Value
A_17_P16393414	LHPP	−0.37	5.27×10^{-04}
A_17_P17272535	None	−0.41	5.96×10^{-04}
A_17_P15000049	PPP2R3B	0.34	1.36×10^{-03}
A_17_P16879789	FUS	−0.39	2.21×10^{-03}
A_17_P15366352	ASNSD1	−0.38	3.60×10^{-03}
A_17_P15810386	None	−0.56	3.64×10^{-03}
A_17_P15865576	SLC35D3	−0.56	4.37×10^{-03}
A_17_P16645611	PCDH9	−0.42	5.20×10^{-03}
A_17_P16672113	None	−0.41	5.33×10^{-03}
A_17_P15427385	STAC	−0.41	5.58×10^{-03}

Table 4.3: P-values and false discovery rates for top 10 probes from t-tests. Negative fold-change implies hypermethylation in UC. False discovery rates are 1 for all of these probes.

two methods to assess whether there is evidence of CpG island level differential methylation and tested performance as previously. CpG island level also has the advantage of drastically reducing the multiple testing problem, as there are far fewer CpG islands than individual microarray probes.

Comparing the median expression level of CpG islands

The first method has summarized each CpG island by their median probe expression level and used *limma* to assess differential methylation between UC and normal phenotypes. We included only CpG islands which had at least two associated microarray probes and where all probes on the island were annotated to the same gene. This yielded a set of 17,322 CpG islands which were associated with 11,494 genes. Differential methylation analysis revealed a total of 78 CpG island regions with $p < 0.05$, but as with the probe level analysis, none of these was significant following multiple testing correction (Table 4.4). Also, none of the 10 genes previously identified as hypermethylated in UC were identified as nominally significant by this method, which gives no evidence that this method provides meaningful insight.

CpG Island Location	Gene Symbol	Log FC	P-Value
chr11:56950916-56951226	MSX1	-0.26	4.61×10^{-03}
chr7:72364760-72365376	CENTA1	0.18	6.62×10^{-03}
chr12:50938285-50939010	TRIM50	-0.19	8.02×10^{-03}
chr10:126268288-126268493	FPGS	-0.20	8.34×10^{-03}
chr18:2903059-2903307	LHPP	0.20	9.48×10^{-03}
chr9:129604684-129605387	SLC43A3	0.17	9.96×10^{-03}
chr15:89298998-89299619	KRT7	0.20	1.22×10^{-02}
chrX:48974153-48974464	RCCD1	0.19	1.24×10^{-02}
chr7:909233-909466	EMILIN2	0.18	1.24×10^{-02}
chr4:4917339-4917714	CACNA1F	0.22	1.39×10^{-02}

Table 4.4: Top 10 results for differential methylation of CpG islands assessed by *limma* from the median log ratio intensity of each CpG island. False discovery rates are 1 for all of these probes.

Assessing all probe fold changes on CpG islands

We have developed a second method for assessing differential methylation at CpG island level. This attempts to identify CpG islands where many of the probes are concordantly differentially hyper- or hypomethylated. To do so, we calculated a fold change, between UC and normal, for each individual probe; then, for each CpG island, used a one sample t-test to assess the degree to which these fold changes deviate from zero. Again, this approach did not identify any significant differential methylation, following correction for multiple testing; nor were any of the genes previously identified as methylated in UC found among the nominally significant genes. This indicates that it is unlikely that this method is useful for identifying differential methylation in this case.

4.3.5 Gene set analysis

In order to determine the functional significance of the genes identified by our analysis, we subjected the results to gene set analysis (GSA). We have found that a severe bias affects this kind of assay. This stems from the fact that different genes often have very different numbers of associated probes. Several previous studies have applied GSA to high throughput methylation data and many of these results are inaccurate. The next chapter contains a detailed discussion of this bias and we have also proposed a novel method for an unbiased GSA in genome wide methylation datasets. This approach reveals results which were previously unidentifiable in this dataset.

4.4 Conclusions

We have found evidence of hypermethylation of CpG island regions in UC. This conclusion is supported by two different approaches; firstly a ChIP-chip assay, analyzed using the R package *Ringo* and secondly, a statistical analysis, using *limma*. We were unable to identify statistically significant evidence of differential methylation at the level of individual probes or CpG islands, after correction for multiple testing. This may be due to the small numbers of samples. However, of the approaches that we implemented, differential methylation analysis at probe level, using *limma*, performed best in identifying genes that were previously found to be hypermethylated in UC. We selected the genes identified by *limma* for GSA and identified a severe bias which affects the analysis. This is the subject of the next chapter.

4.5 Methods

4.5.1 Identifying methylation using Ringo

Raw microarray data were loaded in R and log intensity ratios were calculated. As suggested by Agilent, in order to allow comparison, probes were binned by melting temperature, into 17 bins of 0.1°C. These distributions were centered and enriched probes were identified using the *upperBoundNull()* function from the *Ringo* package. The numbers of enriched probes identified in samples of each phenotype were then compared using a t-test.

4.5.2 UC Microarray Data

MeDIP was performed to capture methylated DNA sequence as previously described by Weber et al with slight modifications. Briefly, 10μg of 5-methylcytosine antibody was incubated with 50μl of Dynabeads M-28 Sheep anti-mouse IgG for 5 hours in immunoprecipitate (IP) buffer at 4°C. Genomic DNA was sonicated using the Branson digital sonifier and 4μg of genomic DNA was incubated with the antibody-beads complex overnight at 4°C. Then, the DNA-antibody-dynabeads complex was washed three times with IP buffer and incubated with 5μl of proteinase K for 2 hours at 55°C. In our experiment, we labeled the IP DNA with fluorescent dye, cyanine 3 and reference (R) DNA with cyanine 5 and co-hybridized to the Agilent microarrays. The MeDIP followed by CpG island microarray analysis enables us to identify the methylated and unmethylated CpG islands between long standing UC patients and age-matched control patients. Purification of labeled products, array hybridization and scanning were performed at the functional genomics and high throughput screening facility at the National

Centre for Biomedical Engineering Science, NUI Galway. These data have been uploaded to GEO and are available under accession number *GSE39188*.

Chapter 5

Severe bias in gene set analysis applied to high-throughput methylation data

The content of this chapter has been submitted for publication as:

Geeleher, P., Hartnett, L., Affendi, RRA., Egan, LJ., Golden, A. and Seoighe, C. Severe Bias in Gene-Set Analysis Applied to High-throughput Methylation Data. *Oncogene*

5.1 Abstract

5.1.1 Background

Changes in DNA methylation play a major role in the development of cancer, for example by silencing tumor suppressor genes. Because of its biological and clinical significance, several previous studies have compared genome-wide patterns of methylation between phenotypes. The application of gene set analysis to identify biological processes that are enriched for differentially methylated genes is often a major component of these analysis. This can be used, for example, to identify biological processes that are perturbed by methylation in cancer development. This chapter discusses gene set analysis applied to high throughput methylation data in the context of the results from the previous chapter, as well as other publicly available datasets.

5.1.2 Results

Gene set analysis, as it is typically applied to genome-wide methylation assays is severely biased as a result of differences in the numbers and sizes of CpG islands associated with different classes of genes. We demonstrate this bias using published data from a study of differential CpG Island methylation in lung cancer and a data set we generated to study methylation changes in patients with long-standing ulcerative colitis (both experiments used the Agilent Human CpG Island microarray) and show that several of the gene sets that appear enriched would also be identified with randomized data. We also report a method to correct the bias. Application of the corrected method to the lung cancer and ulcerative colitis data sets provides novel biological insights into the role of methylation in cancer development and chronic inflammation.

5.1.3 Conclusions

Gene set analysis is a very widely used tool in genomics research, but is unreliable when genes belonging to different gene sets have, *a priori*, different probabilities of appearing in the foreground list. We show that this is a particularly important effect in genome-wide methylation analysis and provide a method to correct the bias. Our results have significant implications for several prior genome-wide methylation studies that have inadvertently drawn conclusions on the basis of strongly biased gene set analysis.

5.2 Background

The application of gene set analysis (GSA) is not restricted to the results of high-throughput gene expression measurements; the same approach is used for many other high-throughput experiments. We focus on the application of GSA to the results of high-throughput DNA methylation experiments. Microarrays have frequently been used to assess the methylation status of CpG sites and CpG islands genome-wide. Platforms for this purpose have been developed by Agilent, Illumina and NimbleGen. Here, we analyze data generated using the Agilent Human CpG Island microarray, which measures methylation levels at 244,000 CpG sites, located within CpG islands. Applications of this platform have included the study of CpG island methylation in leukemia [306] and in lung [307], prostate [297] and breast [308] cancers. Several previous studies have applied GSA to lists of differentially methylated genes obtained using the Agilent Human CpG Island array, including Helman *et al.* [298], who applied GSA to hypermethylated genes in lung cancer, using the Bioconductor package *GOstats* and Dunwell *et al.* [309], who assessed differentially methylated gene sets in childhood acute lymphoblastic

leukemia (ALL) using the bioinformatics tool DAVID. In both cases, several functional categories of genes were identified as highly enriched among differentially methylated genes.

We show that when GSA is applied to differentially methylated genes, detected using microarrays, the results are strongly biased towards certain gene sets. The bias stems from large differences between genes in the number of probes on the array that map to CpG islands in the promoter region of the gene. This is because different gene promoters have very different CpG content [310], thus GSA applied to data from any microarray designed to assay methylation of these CpG dinucleotides may be affected by this bias. On the Agilent array, the number of probes associated with each gene ranges from 1 to 285 (Fig. 5.1). Similar platforms by NimbleGen (Human DNA Methylation 385K Promoter Plus CpG Island Array) and Illumina (Infinium HumanMethylation450 BeadChip) contain from 1-80 and 1-1288 probes per gene, respectively (Fig. 6.3 in Appendix D). Given the criteria typically used to designate genes as differentially methylated (e.g. [298][309]) genes with higher numbers of associated probes are more likely to meet the criteria by chance (higher false positive rate). Indeed, there may also be more power to detect a real differential methylation signal for these genes. These factors combined mean that genes with more associated CpG probes are more likely to be called as differentially methylated. Critically, there are often large differences between the mean numbers of probes that map to genes in different gene sets; hence, gene sets that contain genes with large numbers of associated probes are more likely to be identified as significant during GSA. We demonstrate this problem for two kinds of experimental design using the Agilent platform. First, for the lung cancer study described above and next for a study that we have recently performed on methylation changes associated with ulcerative colitis (UC). We also demonstrate that the same bias affects DNA methylation data generated using high-throughput sequencing. Finally, we propose a method of unbiased gene set analysis that uncovers previously unidentified, plausible and biologically relevant patterns of differential methylation in these datasets.

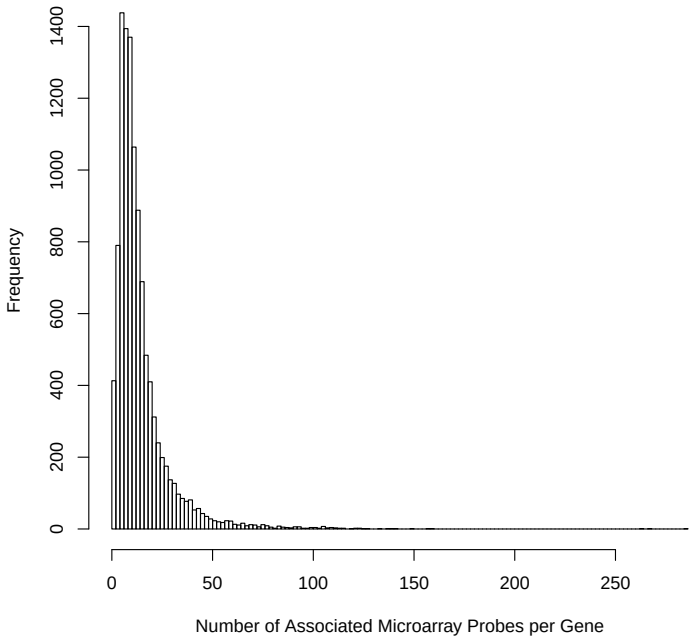

Figure 5.1: A histogram illustrating the distribution of the numbers of microarray probes associated with each gene on the Agilent Human CpG Island Array.

5.3 Results and Discussion

Rauch *et al.* [307] used the Agilent Human CpG Island microarray to assess methylation in five lung cancer samples compared to normal lung tissue and Hel-man *et al.* [298] applied GSA to identify hypermethylated gene sets in this dataset. CpG islands were called as hypermethylated in a sample "when at least two adjacent probes, allowing a one-probe gap, within the CpG island scored a fold-difference factor of > 2 when comparing tumor and normal tissue DNA" [307]. Genes were considered hypermethylated in lung cancer if any associated CpG island met this criterion in four out of the five samples. A total of 102 hypermethylated genes were identified in this way. The R package *GOstats* was then used to assess aberrant methylation of GO biological processes (BP) containing between 100 and 1,000 genes. Differentiation/developmental and transcription factor activity related gene sets were identified as highly significantly enriched among the hypermethylated genes.

5.3.1 Genes identified as hypermethylated have more associated probes

We obtained the dataset of Rauch *et al.* [307] from GEO (accession number *GSE9622*) and, following their methodology (summarized above), we identified 73 genes hypermethylated in lung cancer, 71 of which overlap the 102 identified in the original study. Our results are slightly different from the results reported by Rauch *et al.* [307], as we mapped microarray probes to CpG island regions of the *hg18* human genome build (as opposed to *hg17* in the original study). The 73 hypermethylated genes identified had, on average, a far higher number of associated probes than non-hypermethylated genes (39.6, compared to 9.7; $p < 2.2 \times 10^{-16}$ from Wilcoxon rank sum test; Fig. 5.2(a)), suggesting that genes with more associated reporters are more likely than other genes to appear in the foreground list.

5.3.2 Strong bias in the results of GSA

We used *GOstats* to identify enriched gene sets for GO BPs in the size class range used above and obtained results similar to Helman *et al.* [298]. The top ten most significantly enriched GO categories are shown in Table 5.1. The mean number of probes per gene in these ten gene sets was significantly higher, compared to the mean for all other genes (15.5, compared to 8.8; $p < 2.2 \times 10^{-16}$ from Wilcoxon rank sum test), suggesting that larger numbers of associated probes may be at least partially responsible for the enrichment of these gene sets among the hypermethylated genes. We performed a permutation test to investigate this hypothesis further. Log-intensity ratios associated with each probe were permuted randomly 100 times within each sample; for each permutation we repeated the inference of differential methylation followed by GSA. To ensure that the results were comparable with the original data, we modified the fold change cutoff for differential methylation so that the average number of hypermethylated genes was the same as in the original data. Given these random permutations, one would expect approximately 5% of gene sets tested to be significantly enriched at $p < 0.05$. However, we found that, in 100 permutations, 34% of GO BP terms tested had a median p-value less than 0.05. In particular, the gene sets in Table 5.1 all showed evidence of enrichment among the genes called as differentially methylated in most or all permutations (Fig. 5.2(b)). These results demonstrate that, using the existing methodology, many of the gene sets tested achieve significance, even when the input probe log intensity ratios are essentially randomly generated noise.

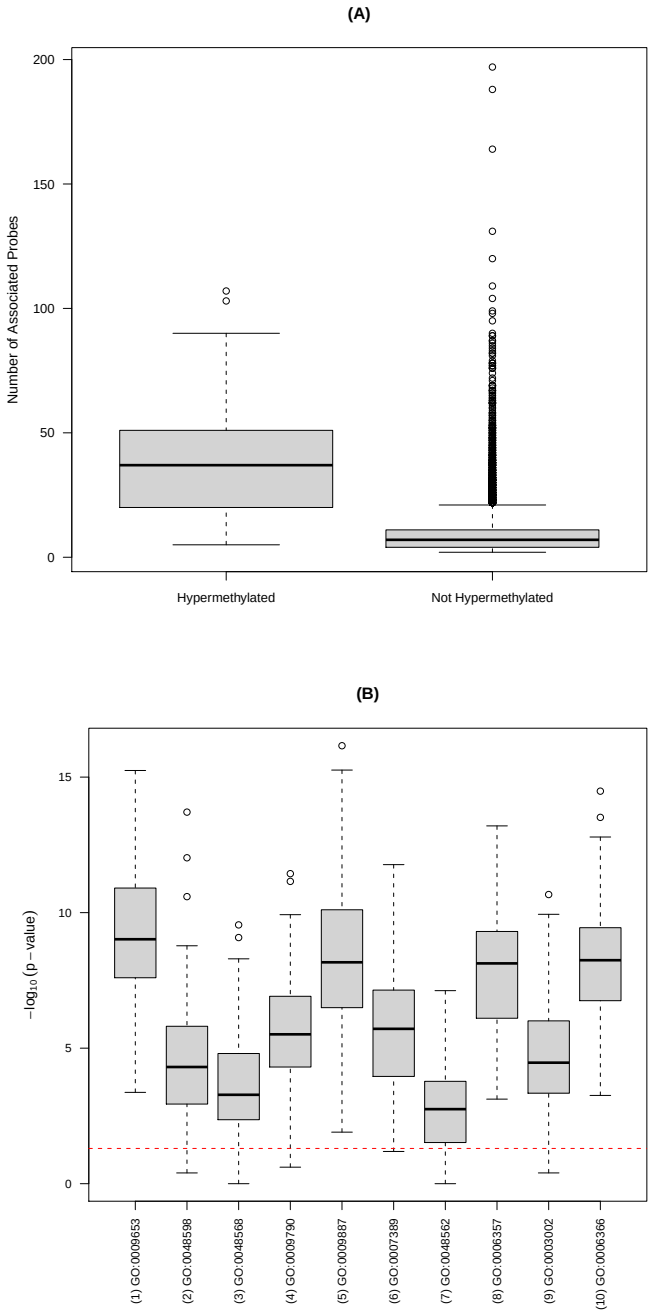

Figure 5.2: (A) Number of probes associated with genes called as hypermethylated and not hypermethylated in the lung cancer dataset. (B) Boxplots of $-\log_{10}$ p-values for the top 10 GO BPs (from Table 5.1) obtained from 100 random permutations of probe values. The dashed red line shows the $p = 0.05$ threshold.

5.3.3 Bias correction

We used logistic regression to model the association between the odds of a gene appearing in the foreground list (i.e. being detected as hypermethylated) and the log-transformed number of microarray probes associated with the gene (Fig. 5.3). The model can be used to predict the probability of a gene appearing in the foreground list as a function of the number of associated probes. Given these probabilities, we calculated the expected number of foreground genes in each gene set, under the null hypothesis of no association between gene set membership and differential methylation, by summing the probabilities corresponding to each gene in the gene set. Expected and observed numbers of foreground genes in each gene set were then compared, either using a chi-squared goodness of fit test or by simulation (see Methods). However, since the chi-squared goodness of fit test is unreliable when expected counts are small [311] only the results of the simulation-based method are discussed in the main text.

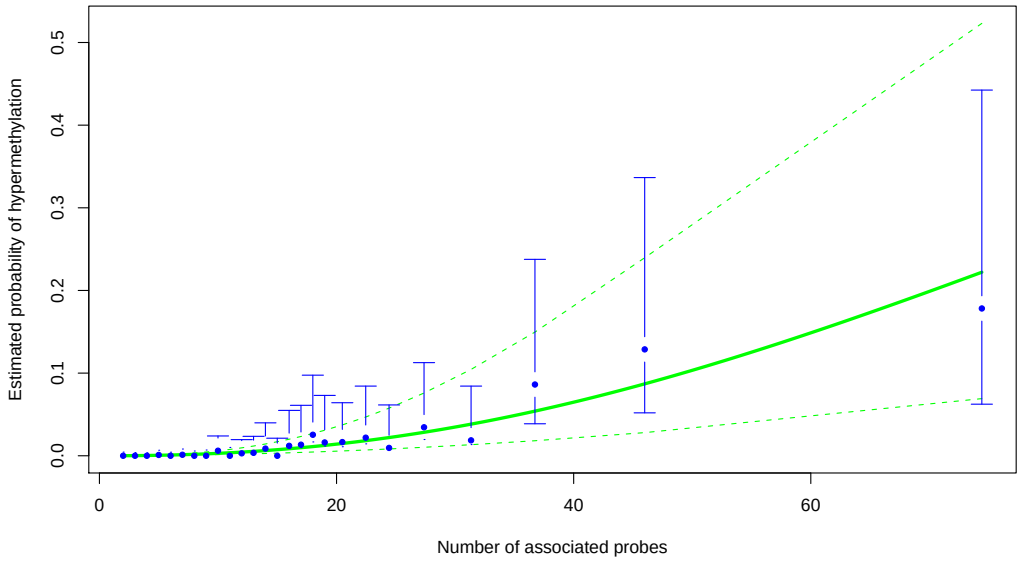

Figure 5.3: Fit of the logistic regression to the lung cancer data. The logistic regression is shown as the solid green line, with 95% confidence intervals shown as dashed green lines. The blue points show the proportion of hypermethylated genes, in bins of minimum size 100 genes. 95% confidence intervals for the bins are shown as blue lines.

GOBPID	Count	Expected Count	P-value	FDR	Term
GO:0009653	35	7.40	7.48×10^{-17}	1.18×10^{-14}	Anatomical Structure Morphogenesis
GO:0048598	21	1.94	8.07×10^{-17}	1.18×10^{-14}	Embryonic Morphogenesis
GO:0048568	17	1.10	2.98×10^{-16}	2.59×10^{-14}	Embryonic Organ Development
GO:0009790	25	3.36	3.74×10^{-16}	2.59×10^{-14}	Embryonic Development
GO:0009887	25	3.38	4.42×10^{-16}	2.59×10^{-14}	Organ Morphogenesis
GO:0007389	19	1.63	7.88×10^{-16}	3.85×10^{-14}	Pattern Specification Process
GO:0048562	15	0.85	2.56×10^{-15}	1.07×10^{-13}	Embryonic Organ Morphogenesis
GO:0006357	24	4.15	4.84×10^{-13}	1.77×10^{-11}	Regulation of Transcription from RNA Polymerase II Promoter
GO:0003002	15	1.29	1.34×10^{-12}	4.36×10^{-11}	Regionalization
GO:0006366	25	5.15	6.79×10^{-12}	1.99×10^{-10}	Transcription from RNA Polymerase II Promoter

Table 5.1: Top 10 GO BP categories for uncorrected gene set analysis (lung cancer).

GOBPID	Count	Expected Count	P-value	FDR	Term
GO:0048598	21	7.34	1.00×10^{-06}	9.77×10^{-05}	Embryonic Morphogenesis
GO:0048568	17	4.68	1.00×10^{-06}	9.77×10^{-05}	Embryonic Organ Development
GO:0048562	15	3.56	1.00×10^{-06}	9.77×10^{-05}	Embryonic Organ Morphogenesis
GO:0009790	25	10.87	4.00×10^{-06}	2.93×10^{-04}	Embryonic Development
GO:0007389	19	7.75	1.80×10^{-05}	1.05×10^{-03}	Pattern Specification Process
GO:0001501	15	5.45	7.30×10^{-05}	3.56×10^{-03}	Skeletal System Development
GO:0009653	35	20.91	1.15×10^{-04}	4.81×10^{-03}	Anatomical Structure Morphogenesis
GO:0009887	25	13.06	1.75×10^{-04}	5.70×10^{-03}	Organ Morphogenesis
GO:0003002	15	6.08	1.69×10^{-04}	5.70×10^{-03}	Regionalization
GO:0007423	13	5.21	3.61×10^{-04}	1.06×10^{-02}	Sensory Organ Development
GO:0009952	11	4.34	7.87×10^{-04}	2.10×10^{-02}	Anterior/Posterior Pattern Formation

Table 5.2: GO BP categories with $FDR < 0.05$ for logistic regression corrected gene set analysis (lung cancer).

The number of GO biological process categories that were significantly enriched ($FDR < 0.05$) following correction was much smaller than the number that were significant if the correction for the number of probes associated with each gene was not carried out (11, compared to 72). In Table 5.2 "Embryonic Morphogenesis" is now the most significant category ($p = 9.8 \times 10^{-5}$, compared to $p = 8.1 \times 10^{-17}$, prior to correction). The expected number of hypermethylated probes for "Embryonic Morphogenesis" rose from 1.94 to 7.34, but this is still considerably less than the observed number of hypermethylated genes for this category, which is 21. This suggests that the reported hypermethylation of developmental associated genes in lung cancer is not an artifact of the higher numbers of associated probes. However, several of the gene sets identified in the original analysis are no longer significant. These include gene sets related to transcription factor activity and, perhaps importantly, the gene sets related directly to differentiation. The p-value for "Regulation of Cell Differentiation" increased from $p = 3 \times 10^{-4}$ to $p = 0.17$ and "Cell Morphogenesis Involved in Differentiation" from $p = 8.8 \times 10^{-3}$ to $p = 0.34$. This may bring into question the validity of the original conclusions of Helman *et al*, that hypermethylation silences genes required for maintenance of the differentiated state.

5.3.4 Application of corrected GSA to differential methylation in ulcerative colitis

We also applied the corrected GSA approach to a dataset that we generated to assess differential methylation in patients with long standing (more than 25 years) ulcerative colitis (UC). These patients are at risk of developing colorectal cancer (CRC) [271]. We used Agilent Human CpG Island microarrays to compare methylation patterns in sigmoid colon tissue between five individuals suffering from ulcerative colitis and five healthy age-matched controls. This is a different experimental design to the dataset discussed above. Cases and control samples were hybridized to different microarrays. The Cy3 channels were hybridized with immunoprecipitated methylated DNA (isolated using the MeDIP [312] approach) from sigmoid colon tissue and the Cy5 channels were hybridized with input DNA (both methylated and unmethylated) from sigmoid colon tissue of the same individual (see Methods for details). Thus, the log intensity ratio of a probe is, in this case, indicative of the extent of methylation of a probe in a given sample. This is in contrast to the lung cancer dataset, discussed above, which, like the majority of datasets in the literature, was generated by hybridizing methylated DNA from cancerous and normal tissue of the same individual to separate channels of the same two-channel microarray. In this case the log intensity ratio on an array is indicative of the level of differential methylation between lung and normal tissue. The experimental design of the UC dataset allows results from gene set analysis

to be corrected using sample label permutations, an established approach in both gene expression and gene set analysis, used by tools such as SAM [133] and GSEA [188]. We can thus assess the performance of our model-based correction for differences in numbers of mapped probes, by comparing to the results obtained from sample label permutation.

We used the Bioconductor package *limma* [173] to identify probes that were hypermethylated in the UC samples ($p < 0.05$), and all genes associated with at least one hypermethylated probe were considered hypermethylated (see Methods for details). This approach identified a foreground list of 380 genes for gene set analysis, which was carried out using *GOstats* (Table 5.3). It is clear that many of the functional categories identified in the lung cancer study are again among the most significant. In fact, the Pearson correlation between the GO BP $-\log_{10}$ p-values from the lung cancer dataset and the UC dataset is 0.69 ($p < 2.2 \times 10^{-16}$; Fig. 5.4(a)). This indicates that both experiments are finding similar p-values for the same gene sets. This similarity may be biological, so that similar genes are differentially methylated in the lung cancer and UC datasets. This would be of interest since it might shed light on the involvement of chronic UC in the development of colon cancer. However, because of the strong bias, discussed above, this similarity could be completely artifactual – a result simply of the tendency for genes associated with a large number of probes on the microarray to be called differentially methylated.

To address these issues we performed a corrected GSA on the UC data set. The logistic regression model again indicates a strong association between the number of associated probes and probability of differential methylation ($p < 2.2 \times 10^{-16}$; Fig. 5.5). Corrected p-values and corrected expected values for all 22 gene sets with $p < 0.05$ are provided in Table 5.5. The Pearson correlation between the corrected p-values in the UC and lung cancer datasets is considerably lower than for the uncorrected results (r = 0.17, compared to r = 0.69; Fig. 5.4(b)). Furthermore, after correction, the number of gene sets with $p < 0.05$ in the UC dataset drops from 262 to only 22, suggesting that a large proportion of the results from the original analysis were indeed artefacts.

GOBPID	Count	Expected Count	P-value	Term
GO:0003002	14	4.23	8.52×10^{-05}	Regionalization
GO:0048562	11	2.79	1.04×10^{-04}	Embryonic Organ Morphogenesis
GO:0009790	25	11.30	1.57×10^{-04}	Embryonic Development
GO:0009952	11	3.06	2.37×10^{-04}	Anterior/Posterior Pattern Formation
GO:0048568	11	3.59	9.44×10^{-04}	Embryonic Organ Development
GO:0007389	14	5.43	1.10×10^{-03}	Pattern Specification Process
GO:0048598	15	6.38	1.86×10^{-03}	Embryonic Morphogenesis
GO:0009887	21	11.03	3.45×10^{-03}	Organ Morphogenesis
GO:0016055	10	3.67	3.73×10^{-03}	Wnt Receptor Signaling Pathway
GO:0031589	8	2.54	3.90×10^{-03}	Cell-Substrate Adhesion

Table 5.3: Top 10 GO BP categories for uncorrected gene set analysis (UC).

GOBPID	Count	P-value	Term
GO:0031589	8	1.59×10^{-02}	Cell-Substrate Adhesion
GO:0008610	12	2.38×10^{-02}	Lipid Biosynthetic Process
GO:0006644	7	2.38×10^{-02}	Phospholipid Metabolic Process
GO:0019637	7	2.38×10^{-02}	Organophosphate Metabolic Process
GO:0006629	20	3.17×10^{-02}	Lipid Metabolic Process
GO:0044255	14	3.17×10^{-02}	Cellular Lipid Metabolic Process
GO:0046486	5	3.17×10^{-02}	Glycerolipid Metabolic Process
GO:0043085	17	3.97×10^{-02}	Positive Regulation of Catalytic Activity
GO:0006913	8	3.97×10^{-02}	Nucleocytoplasmic Transport
GO:0051169	8	3.97×10^{-02}	Nuclear Transport
GO:0009607	9	3.97×10^{-02}	Response to Biotic Stimulus

Table 5.4: GO BPs with $p < 0.05$ for label permutation corrected gene set analysis (UC).

GOBPID	Count	Expected Count	P-value	Term
GO:0008610	12	5.37	5.01×10^{-03}	Lipid Biosynthetic Process
GO:0031589	8	3.15	7.24×10^{-03}	Cell-Substrate Adhesion
GO:0006644	7	2.78	1.27×10^{-02}	Phospholipid Metabolic Process
GO:0043085	17	9.91	1.54×10^{-02}	Positive Regulation of Catalytic Activity
GO:0019637	7	2.92	1.65×10^{-02}	Organophosphate Metabolic Process
GO:0048562	11	5.93	2.15×10^{-02}	Embryonic Organ Morphogenesis
GO:0044093	18	11.37	2.70×10^{-02}	Positive Regulation of Molecular Function
GO:0044087	7	3.26	2.72×10^{-02}	Regulation of Cellular Component Biogenesis
GO:0016053	6	2.57	2.80×10^{-02}	Organic Acid Biosynthetic Process
GO:0046394	6	2.57	2.80×10^{-02}	Carboxylic Acid Biosynthetic Process
GO:0044255	14	8.30	2.89×10^{-02}	Cellular Lipid Metabolic Process
GO:0006629	20	13.04	2.94×10^{-02}	Lipid Metabolic Process
GO:0006913	8	3.99	3.07×10^{-02}	Nucleocytoplasmic Transport
GO:0051169	8	4.01	3.13×10^{-02}	Nuclear Transport
GO:0009952	11	6.33	3.32×10^{-02}	Anterior/Posterior Pattern Formation
GO:0030003	8	4.13	3.76×10^{-02}	Cellular Cation Homeostasis
GO:0016055	10	5.65	3.79×10^{-02}	Wnt Receptor Signaling Pathway
GO:0009123	7	3.46	3.79×10^{-02}	Nucleoside Monophosphate Metabolic Process
GO:0046486	5	2.17	4.01×10^{-02}	Glycerolipid Metabolic Process
GO:0009607	9	4.91	4.02×10^{-02}	Response To Biotic Stimulus
GO:0003002	14	9.02	4.72×10^{-02}	Regionalization
GO:0043065	14	8.98	4.97×10^{-02}	Positive Regulation of Apoptosis

Table 5.5: GO BPs with with $p < 0.05$ for logistic regression corrected gene set analysis (UC).

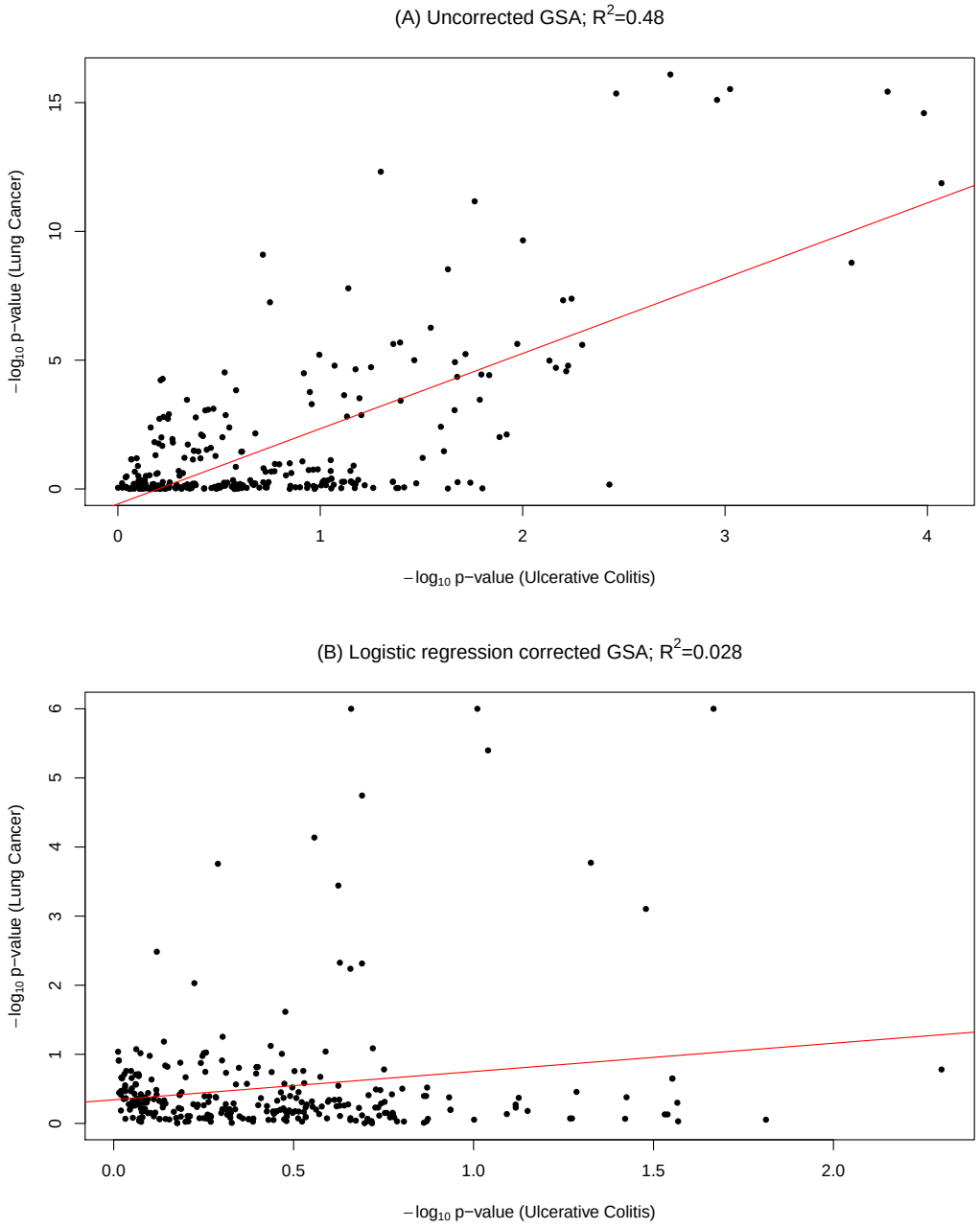

Figure 5.4: Scatterplot of $-\log_{10}$ p-values for each GO BP category tested in the lung cancer and ulcerative colitis datasets for (A) the uncorrected GSA and (B) the corrected GSA. In each plot, a linear regression line is shown in red.

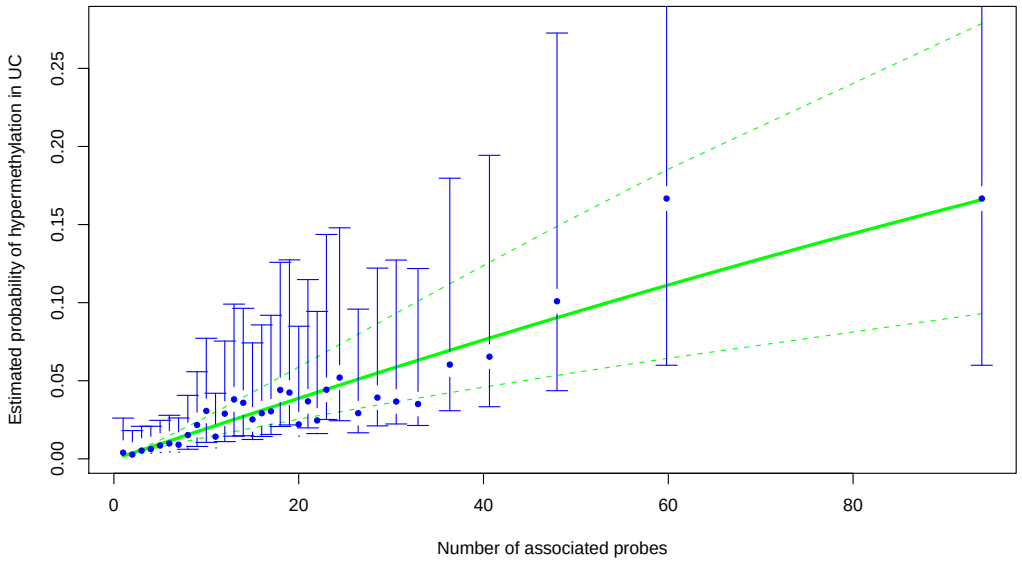

Figure 5.5: Fit of the logistic regression to the UC data. The logistic regression is shown as the solid green line, with 95% confidence intervals shown as dashed green lines. The blue points show the probability of differential methylation, as calculated by grouping the data by number of associated probes, in bins of minimum size 100 genes. 95% confidence intervals for the bins are shown as blue lines.

We also corrected the UC GSA results using sample label permutation (Table 5.4). All 11 of the gene sets identified by label permutation are also identified by our logistic regression-based correction for differences in numbers of mapped probes and the top three most significantly enriched gene sets identified by the two methods are the same. The similarity to the results of an independent and robust permutation-based approach provides good evidence that our corrected GSA has performed well. By comparison, only 7 of the 11 gene sets identified using label permutation were detected using the uncorrected GSA, despite an order of magnitude more processes reported significant in the latter analysis. The results identified by the corrected method are also highly plausible. Although, the false-discovery rates are high (Table 5.4), a number of biological processes related to lipids appear to be hypermethylated in UC. It has recently been shown that colonic mucus from UC patients has decreased phosphatidylcholine (a class of phospholipid) content and that the addition of phosphatidylcholine reduced inflammation [313]. Our results suggest, for the first time, that DNA hypermethylation of genes involved in lipid biosynthesis may be one of the causes of decreased

phosphatidylcholine. The gene *CHPT1* (cholinephosphotransferase 1) appears in our foreground list (ranked 42nd at $p = 9.1 \times 10^{-3}$) of candidate genes hypermethylated in UC. This gene is known to catalyze phosphatidylcholine biosynthesis [314] and is included in the foreground list on account of microarray probe (*A_17_P08575667*), which targets part of a CpG island in the promoter region of this gene. This suggests that hypermethylation may be the causative mechanism in the disruption of this process; however, the p-value associated with this probe is not significant following correction for multiple testing and the relationship requires further investigation. "Cell-substrate adhesion" was also enriched for hypermethylated genes. Increased intestinal permeability is known to be associated with UC [315] [316] and hypermethylation of these genes may play a role in this association.

Where label permutation is possible (e.g. in the UC experimental design but not in the case of the lung cancer data set) it can be used to perform GSA in a way that is robust to the differences in the number of probes per gene. However, it is likely that the corrected method that we describe provides better power to detect gene sets that are enriched for differentially methylated genes. The statistical significance that can be achieved by a permutation method can be very limited due to the relatively small sample sizes that are often encountered in genome-wide methylation experiments. For example, the minimum p-value that can be obtained by label permutations in our case with two groups of five samples is 7.9×10^{-3} (see Methods). Approaches that make use of the extent of over-representation of a gene class (beyond comparing the over-representation between observed and permuted-label data) can achieve much higher power, but at the expense of lower robustness, for example to confounding factors such as the differences in detection power between different gene classes (discussed here). By correcting for the bias caused by differences in the number of associated probes, we detected a larger number of gene sets that were nominally significant than were detected using permutation (22 gene sets reached the nominally significant $p < 0.05$ threshold, compared to 11 using label permutation). There is evidence to suggest that some of the additional gene sets are biologically relevant. Using our method we identified the "Wnt Receptor Signaling Pathway" ($p = 3.7 \times 10^{-2}$), but this reached only $p = 0.36$ using the permutation based approach. Hypermethylation of genes in this pathway has been previously identified in IBD and these genes have been shown to become progressively more methylated during IBD associated neoplastic transformation [282]. This suggests that correcting for the bias in the number of probes per gene is still potentially insightful, even when the option of label permutation is available.

5.3.5 GSA bias in methylation analysis using high-throughput sequencing

We have recently performed a HELP-seq assay to study the effect of pro-inflammatory factor IL-6 on DNA methylation in human epithelial colorectal adenocarcinoma cells (*in preparation*). The experiment compared methylation levels in three IL-6 treated samples and three controls. The methylation of each CCGG site was estimated from the angle subtended by the two-dimensional vector comprising counts of short reads derived from samples digested with MSP1 and HPAII as described by Suzuki *et al.* [317]. A gene was considered differentially methylated if any one of its associated CCGG sites consistently achieved an angle value of > 60 in one set of samples and < 30 in the other set. Similarly to the case of the microarray probes, there are large differences in the numbers of CCGG sites associated with each gene (Fig. 5.6). We assessed hyper- and hypomethylation of both gene promoters and gene bodies. Unsurprisingly, the same bias that affected the microarrays is evident in HELP-seq, in that genes with more associated CCGG sites are more likely to be called as "differentially methylated" and hence more likely to appear in the foreground list for GSA. The corrected analysis was performed exactly as described for microarrays, except that the number of associated microarray probes per gene is replaced with the number of associated CCGG sites. Plots of the logistic regression models are provided in figures 6.4 to 6.7 in Appendix D. Tables of the corrected results for GO BPs (of the same size class as before) are provided in tables 6.8 to 6.11 in Appendix D. Again, there is clear evidence of a bias and hence, researchers applying GSA to the results of high-throughput sequencing for genome-wide methylation analysis should account for this bias either using a permutation strategy, which may have low power, particularly for modest numbers of samples, or by correcting the bias using a method such as the one we propose.

In general, when different genes and gene sets are associated with different *a priori* probabilities of appearing in the foreground list as a consequence of factors other than those that are of biological interest there is the potential for bias. This arises in many, if not most GSA applications, including applications of GSA to the results of genome-wide association studies as well as genome-wide analysis. It is common in GSA to associate multiple and different numbers of features with each gene; typically, multiple features are collapsed onto single gene identifiers. For instance, the popular web-based GSA tool, *DAVID* [185][186] offers the option to use microarray probe IDs (e.g. from methylation or gene expression arrays) as foreground and background lists. These are converted to unique gene IDs prior to statistical analysis. When there is a difference in the numbers of probes associated with each gene this can give rise to the bias that we have outlined in this paper. The approach we used here can account for this bias by modeling the

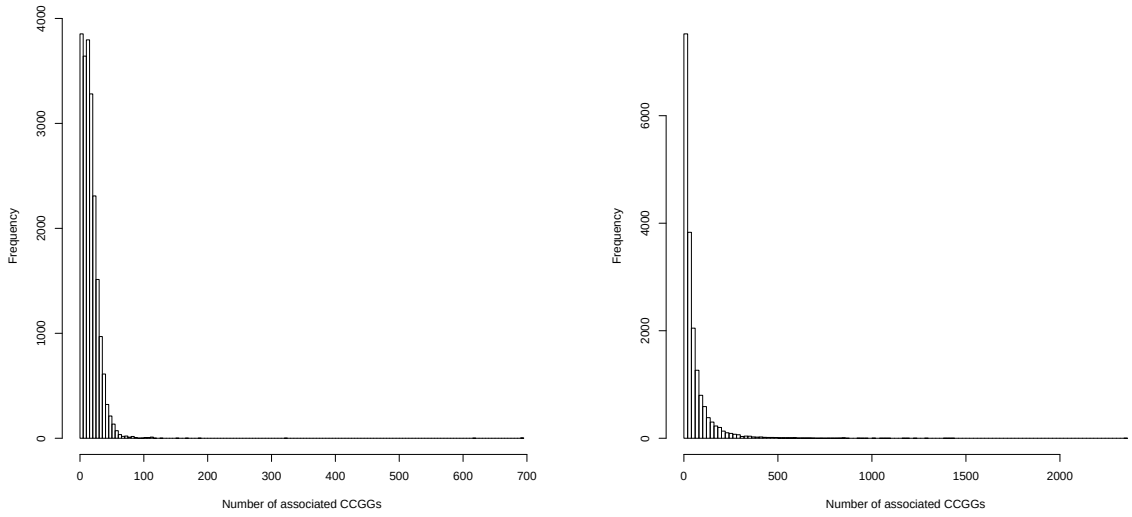

Figure 5.6: Histograms illustrating the distribution of the numbers of CCGG sites associated with each (left) promotor regions and (right) gene body region.

relationship between the number of features (e.g. probes) associated with a gene and its probability of appearing in the foreground list.

5.4 Conclusions

We have identified a severe bias that causes spurious results in gene set analysis of results from microarray and high throughput sequencing based DNA methylation assays. This bias is caused by differences in the numbers of CpG sites associated with each gene. We have developed a method to correct for this and applied it to reanalyze the results of a published study of differential methylation in lung cancer and to a dataset that we generated from ulcerative colitis samples. Based on our analysis, there is no evidence that the hypermethylation in lung cancer is targeted towards genes involved in maintaining a differentiated state or towards transcription factors, as was originally reported, although the evidence for hypermethylation of genes associated with embryonic development remains. The application of our corrected GSA method to the UC dataset revealed evidence of hypermethylation of some highly relevant biological processes in ulcerative colitis, including lipid biosynthesis and cell-substrate adhesion.

Several other published studies have applied gene set analysis to DNA methylation microarray data, using similar experimental designs and analysis methodologies to the lung cancer study discussed here. We have demonstrated the bias

in the case of one DNA methylation microarray, but the same considerations apply to other similar microarray platforms (from Illumina and NimbleGen) and to methylation data generated using high-throughput sequencing strategies. Results obtained by applying uncorrected GSA to these data are likely to be affected by the bias we describe and should be reanalyzed, taking into account the numbers of probes or CpG sites per gene. Given the increasing popularity of these types of experiments, researchers should be aware of and correct for this severe bias. The method that we have developed can be applied to both microarray and HTS based assays and has the potential to uncover biologically relevant results that would otherwise be overlooked.

5.5 Methods

5.5.1 Logistic regression model

We used the *glm()* function in R [167] to fit a logistic regression model to gene hypermethylation status as a function of the log-transformed number of probes on the array that map to the gene. Log-transformation of the explanatory variable, which had a heavy tailed distribution (Fig. S1), was found to improve the fit of the model substantially, as judged by the Akaike Information Criterion [318]. Given the regression model we estimated the probability of hypermethylation of each gene as a function of the number of probes that were mapped to the gene. In order to visualize the fit of the model to the data (e.g. Fig. 2), we grouped genes according to the number of associated probes, in bins of minimum size 100 genes. The probability of methylation was calculated for each bin from the proportion of hypermethylated genes.

5.5.2 Corrected gene set analysis

Each gene was associated with a probability of hypermethylation from the logistic regression model, based on the number of probes that map to the gene, independently of gene set membership. Thus, the expected number of hypermethylated genes in the gene set under the null hypothesis (that hypermethylation is independent of gene set membership, given the number of probes that map to the gene) is the sum of these probabilities. This expected number of hypermethylated genes can be compared to the number observed using the chi-squared goodness of fit test. The chi-squared statistic is

$\chi^2 = \frac{O_{set}-E_{set}}{E_{set}} + \frac{O_{other}-E_{other}}{E_{other}}$

where

O_{set} and E_{set} are the observed and expected numbers of hypermethylated

genes in a gene set and O_{other} and E_{other} are the observed and expected numbers of hypermethylated genes not in the gene set.

Because the chi-squared goodness of fit test is unreliable when expected counts are small [311], we also developed an approach based on randomization. One million foreground/background groups were generated randomly such that the probability of a gene appearing in the foreground list was equal to its probability of hypermethylation, given the number of associated probes. GSA was applied to each randomly generated foreground/background group using *GOstats* version 2.14.0, with annotations from version 2.4.1 of the *org.Hs.eg.db* library. A corrected GSA p-value for each gene set was calculated as the proportion of random data sets with an enrichment odds ratio at least as extreme as the observed enrichment odds ratio.

5.5.3 Gene set analysis using label permutation

Sample labels were rearranged in all possible combinations. As there were a total of 10 samples, split equally between UC and control phenotypes, this yielded a total of 126 distinct arrangements of the samples. For each of these arrangements the enrichment odds ratio test statistic was recalculated for each gene set. P-values were calculated as the proportion of the test statistics that were as extreme, or more extreme, than the test statistic corresponding to the observed data.

Chapter 6

Conclusions and scope for future work

In this thesis we have uncovered novel insights into the genetics of miRNA regulatory effect (chapter 2) and CpG island hypermethylation in ulcerative colitis (chapter 4). In chapter 3 we have developed a method to improve the absolute level of gene expression estimates in microarrays and in chapter 5 we have developed a new method to correct gene set analysis on high throughput genome wide methylation platforms. Much of this work leaves questions which could be pursued as part of future research projects.

In chapter 2, we discovered that the regulatory effect of miRNAs is a heritable trait in humans and uncovered an association with a SNP (rs17409624) in and intronic region of the gene *DROSHA*. This association was evident in both populations assayed. However, we have thus far been unable to establish the mechanism by which this SNP affects miRNAs. Interestingly, since the publication of this work, it has come to our attention that this same SNP is also associated with BMI in British populations ($p = 0.01$) [319]. There have also been studies which have reported that miRNAs play a key role in the differentiation of adipose tissue and a number of miRNA have been shown to be differentially expressed in these tissues in overweight individuals [320] [321]. These observations suggest a possible link between obesity and the regulatory effect of miRNAs, although further research will be required to establish a clear link.

Chapter 3 developed a novel method called *seqArray* for estimating gene expression in microarray experiments. This method dramatically improved the estimation of absolute gene expression level within samples, which is not an application for which these arrays have traditionally been useful. We have also developed a method that improves sensitivity to change in gene expression level across samples, by discarding probes which are likely not to be informative. Both of the approaches would likely benefit from more training data. Presently, there are

very few datasets publicly available for which both microarray and RNA-seq data are available and are of suitable quality. A more diverse set of training data would allow the *seqArray* models to predict gene expression in a more diverse range of conditions, which should allow for more broadly applicable models. Similarly for the alternate method, more training data would allow greater power in detecting which probes are more useful in estimating expression. It should also be possible to adapt *seqArray* to estimate transcript expression. This would require transcript expression to be estimated on the RNA-seq training set using an application like Cufflinks and a modeling technique which allows for a multivariate response would be needed for training. Currently, there is no way of reliably estimating transcript expression from exon microarray data.

Chapters 4 and 5 analysed CpG island methylation in ulcerative colitis. We first established that there is evidence of CpG island hypermethylation genome wide, but we were unable to identify which regions are targeted at gene level. We also established that a severe bias affects GSA when applied to this type of data and we developed a method to correct this bias. Then, using the results of our modified GSA approach, we identified some individual genes which may be targets of methylation. Among the interesting results are the genes relating to lipid proteins, with hypermethylation of *CHPT1* perhaps particularly relevant, because of its involvement in catalyzing phosphatidylcholine biosynthesis. However, these results have as yet not been validated experimentally. There are many other studies in the literature which have applied GSA to methylation data in a similar way to the lung cancer study which we have discussed, these studies should be re-analysed and it is likely that in many cases biologically relevant findings will be uncovered when an unbiased approach is applied.

The method that we developed could in theory be applied to any GSA where there is an obvious confounding variable. One example of this is RNA-seq, where it is known that there is more power to detect differential expression from genes which have more mapped reads [322]. Our method could easily model the probability of a gene appearing in the GSA foreground list, based on its number of mapped reads; it would then be possible to correct for this confounder in the same way as we corrected for the number of associated probes in the methylation data. With RNA-seq, our method could be particularly useful when sample sizes are small and sample label permutations have little power. Also, at present our GSA method has only been implemented as a basic R script, but this could be implemented as an R package as part of future work.

Overall, this thesis has demonstrated the utility of high throughput genomics techniques in deriving biological insights, much of which would have been impossible only a few years ago. Our work into the genetics of miRNA regulatory effect and genome wide CpG island methylation in ulcerative colitis are examples of this. The major drawback of these kinds of data is that it has never been eas-

ier to generate seemingly meaningful, statistically significant results, using flawed analysis approaches. Gene set analysis applied to high throughput methylation data was one example of this. These types of issues are inevitable, but provide us with the opportunity to develop novel analytical approaches, that properly account for biases in the data. We have done this in correcting the bias in high throughput methylation data, which has revealed previously undetectable insight into the pathology of lung cancer and ulcerative colitis.

Bibliography

[1] R. E. Halbeisen, A. Galgano, T. Scherrer, and A. P. Gerber. Post-transcriptional gene regulation: from genome-wide studies to principles. *Cell. Mol. Life Sci.*, 65(5):798–813, Mar 2008.

[2] Julian Lewis Martin Raff Keith Roberts Bruce Alberts, Alexander Johnson and Peter Walter. *Molecular Biology of the Cell*. Garland Science, 2007.

[3] Julia Winter, Stephanie Jung, Sarina Keller, Richard I Gregory, and Sven Diederichs. Many roads to maturity: microrna biogenesis pathways and their regulation. *Nat Cell Biol*, 11(3):228–234, Mar 2009.

[4] Daehyun Baek, Judit Villén, Chanseok Shin, Fernando D Camargo, Steven P Gygi, and David P Bartel. The impact of microrna on protein output. *Nature*, 455(7209):64–71, Sep 2008.

[5] Schutz. Two affymetrix chips. *Wikipedia*, 2006.

[6] Michael A Quail, Iwanka Kozarewa, Frances Smith, Aylwyn Scally, Philip J Stephens, Richard Durbin, Harold Swerdlow, and Daniel J Turner. A large genome center's improvements to the illumina sequencing system. *Nat Methods*, 5(12):1005–1010, Dec 2008.

[7] Emily Singer. Chinese, african genomes sequenced.

[8] giga.ulg.ac.be. Illumina genome analyser sequencing technology: how it works.

[9] Cole Trapnell and Steven L Salzberg. How to map billions of short reads onto genomes. *Nat Biotechnol*, 27(5):455–457, May 2009.

[10] Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, Jun 2007.

[11] Agilent Technologies. *Methylation Analysis, Agilent Genomic Workbench, User Guide*. Agilent, 1 edition, 2009.

[12] Lori Hartnett and Laurence J Egan. Inflammation, dna methylation and colitis-associated cancer. *Carcinogenesis*, 33(4):723–731, Apr 2012.

[13] Kevin Chen and Nikolaus Rajewsky. The evolution of gene regulation by transcription factors and microrna. *Nat Rev Genet*, 8(2):93–103, Feb 2007.

[14] T. M. Keane, L. Goodstadt, P. Danecek, M. A. White, K. Wong, B. Yalcin, A. Heger, A. Agam, G. Slater, M. Goodson, N. A. Furlotte, E. Eskin, C. Nellaker, H. Whitley, J. Cleak, D. Janowitz, P. Hernandez-Pliego, A. Edwards, T. G. Belgard, P. L. Oliver, R. E. McIntyre, A. Bhomra, J. Nicod, X. Gan, W. Yuan, L. van der Weyden, C. A. Steward, S. Bala, J. Stalker, R. Mott, R. Durbin, I. J. Jackson, A. Czechanski, J. A. Guerra-Assuncao, L. R. Donahue, L. G. Reinholdt, B. A. Payseur, C. P. Ponting, E. Birney, J. Flint, and D. J. Adams. Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature*, 477(7364):289–294, Sep 2011.

[15] M. Sawadogo and A. Sentenac. Rna polymerase b (ii) and general transcription factors. *Annu Rev Biochem*, 59:711–754, 1990.

[16] P. Komarnitsky, E.J. Cho, and S. Buratowski. Different phosphorylated forms of rna polymerase ii and associated mrna processing factors during transcription. *Genes & development*, 14(19):2452–2460, 2000.

[17] Suzuki DT Griffiths AJF, Miller JH. *An Introduction to Genetic Analysis*. W. H. Freeman, 2000.

[18] José M G Vilar and Stanislas Leibler. Dna looping and physical constraints on transcription regulation. *J Mol Biol*, 331(5):981–989, Aug 2003.

[19] D. H. Ohlendorf, W. F. Anderson, R. G. Fisher, Y. Takeda, and B. W. Matthews. The molecular basis of dna-protein recognition inferred from the structure of cro repressor. *Nature*, 298(5876):718–723, Aug 1982.

[20] Laura A Lettice, Simon J H Heaney, Lorna A Purdie, Li Li, Philippe de Beer, Ben A Oostra, Debbie Goode, Greg Elgar, Robert E Hill, and Esther de Graaff. A long-range shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum Mol Genet*, 12(14):1725–1735, Jul 2003.

[21] G. Garcia-Manero, S. Assouline, J. Cortes, Z. Estrov, H. Kantarjian, H. Yang, W. M. Newsome, W. H. Miller, C. Rousseau, A. Kalita, C. Bonfils,

M. Dubay, T. A. Patterson, Z. Li, J. M. Besterman, G. Reid, E. Laille, R. E. Martell, and M. Minden. Phase 1 study of the oral isotype specific histone deacetylase inhibitor MGCD0103 in leukemia. *Blood*, 112(4):981–989, Aug 2008.

[22] X. Wang, H. Zhu, H. Snieder, S. Su, D. Munn, G. Harshfield, B. L. Maria, Y. Dong, F. Treiber, B. Gutin, and H. Shi. Obesity related methylation changes in DNA of peripheral blood leukocytes. *BMC Med*, 8:87, 2010.

[23] M. Rodriguez-Paredes and M. Esteller. Cancer epigenetics reaches mainstream oncology. *Nat. Med.*, 17(3):330–339, Mar 2011.

[24] Christophe Redon, Duane Pilch, Emmy Rogakou, Olga Sedelnikova, Kenneth Newrock, and William Bonner. Histone h2a variants h2ax and h2az. *Curr Opin Genet Dev*, 12(2):162–169, Apr 2002.

[25] Manoj Bhasin, Ellis L Reinherz, and Pedro A Reche. Recognition and classification of histones using support vector machine. *J Comput Biol*, 13(1):102–112, 2006.

[26] Bradley E Bernstein, Michael Kamal, Kerstin Lindblad-Toh, Stefan Bekiranov, Dione K Bailey, Dana J Huebert, Scott McMahon, Elinor K Karlsson, Edward J Kulbokas, Thomas R Gingeras, Stuart L Schreiber, and Eric S Lander. Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell*, 120(2):169–181, Jan 2005.

[27] T. Jenuwein and C. D. Allis. Translating the histone code. *Science*, 293(5532):1074–1080, Aug 2001.

[28] Elizaveta V Benevolenskaya. Histone h3k4 demethylases are essential in development and differentiation. *Biochem Cell Biol*, 85(4):435–443, Aug 2007.

[29] Artem Barski, Suresh Cuddapah, Kairong Cui, Tae-Young Roh, Dustin E Schones, Zhibin Wang, Gang Wei, Iouri Chepelev, and Keji Zhao. High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4):823–837, May 2007.

[30] A. Razin. Cpg methylation, chromatin structure and gene silencing—a three-way connection. *The EMBO journal*, 17(17):4905–4908, 1998.

[31] P. A. Jones and S. M. Taylor. Cellular differentiation, cytidine analogs and dna methylation. *Cell*, 20(1):85–93, May 1980.

[32] Theresa Phillips. The role of methylation in gene expression. *Nature Education*, 1:1, 2008.

[33] Mehrnaz Fatemi, Andrea Hermann, Humaira Gowher, and Albert Jeltsch. Dnmt3a and dnmt1 functionally cooperate during de novo methylation of dna. *Eur J Biochem*, 269(20):4981–4984, Oct 2002.

[34] Albert Jeltsch. On the enzymatic properties of dnmt1: specificity, processivity, mechanism of linear diffusion and allosteric regulation of the enzyme. *Epigenetics*, 1(2):63–66, 2006.

[35] A. P. Bird. Cpg-rich islands and the function of dna methylation. *Nature*, 321(6067):209–213, 1986.

[36] P. A. Jones and D. Takai. The role of dna methylation in mammalian epigenetics. *Science*, 293(5532):1068–1070, Aug 2001.

[37] M. Gardiner-Garden and M. Frommer. Cpg islands in vertebrate genomes. *J Mol Biol*, 196(2):261–282, Jul 1987.

[38] Peter A Jones. Functions of dna methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet*, 13(7):484–492, Jul 2012.

[39] Adrian Bird. Dna methylation patterns and epigenetic memory. *Genes Dev*, 16(1):6–21, Jan 2002.

[40] J. F. Costello, M. C. Frühwald, D. J. Smiraglia, L. J. Rush, G. P. Robertson, X. Gao, F. A. Wright, J. D. Feramisco, P. Peltomaki, J. C. Lang, D. E. Schuller, L. Yu, C. D. Bloomfield, M. A. Caligiuri, A. Yates, R. Nishikawa, H. Su Huang, N. J. Petrelli, X. Zhang, M. S. O'Dorisio, W. A. Held, W. K. Cavenee, and C. Plass. Aberrant cpg-island methylation has non-random and tumour-type-specific patterns. *Nat Genet*, 24(2):132–138, Feb 2000.

[41] P. A. Jones and P. W. Laird. Cancer epigenetics comes of age. *Nat Genet*, 21(2):163–167, Feb 1999.

[42] Peter A Jones and Stephen B Baylin. The epigenomics of cancer. *Cell*, 128(4):683–692, Feb 2007.

[43] Manel Esteller. Aberrant dna methylation as a cancer-inducing mechanism. *Annu Rev Pharmacol Toxicol*, 45:629–656, 2005.

[44] Andrew P Feinberg and Benjamin Tycko. The history of cancer epigenetics. *Nat Rev Cancer*, 4(2):143–153, Feb 2004.

[45] Manel Esteller. Cpg island hypermethylation and tumor suppressor genes: a booming present, a brighter future. *Oncogene*, 21(35):5427–5440, Aug 2002.

[46] A. Merlo, J. G. Herman, L. Mao, D. J. Lee, E. Gabrielson, P. C. Burger, S. B. Baylin, and D. Sidransky. 5' cpg island methylation is associated with transcriptional silencing of the tumour suppressor p16/cdkn2/mts1 in human cancers. *Nat Med*, 1(7):686–692, Jul 1995.

[47] L. F. Lock, N. Takagi, and G. R. Martin. Methylation of the hprt gene on the inactive x occurs after chromosome inactivation. *Cell*, 48(1):39–46, Jan 1987.

[48] Sébastien A Smallwood, Shin-Ichi Tomizawa, Felix Krueger, Nico Ruf, Natasha Carli, Anne Segonds-Pichon, Shun Sato, Kenichiro Hata, Simon R Andrews, and Gavin Kelsey. Dynamic cpg island methylation landscape in oocytes and preimplantation embryos. *Nat Genet*, 43(8):811–814, Aug 2011.

[49] Y. Zhou. "permutation tests", in berger, v. (ed.), design and analysis of randomized clinical trials: Design, analysis & theory, the biomedical & life sciences collection, henry stewart talks ltd, london (online at http://hstalks.com/bio).

[50] Asaf Hellman and Andrew Chess. Gene body-specific methylation on the active x chromosome. *Science*, 315(5815):1141–1143, Feb 2007.

[51] Shawn J Cokus, Suhua Feng, Xiaoyu Zhang, Zugen Chen, Barry Merriman, Christian D Haudenschild, Sriharsa Pradhan, Stanley F Nelson, Matteo Pellegrini, and Steven E Jacobsen. Shotgun bisulphite sequencing of the arabidopsis genome reveals dna methylation patterning. *Nature*, 452(7184):215–219, Mar 2008.

[52] Suhua Feng, Shawn J Cokus, Xiaoyu Zhang, Pao-Yang Chen, Magnolia Bostick, Mary G Goll, Jonathan Hetzel, Jayati Jain, Steven H Strauss, Marnie E Halpern, Chinweike Ukomadu, Kirsten C Sadler, Sriharsa Pradhan, Matteo Pellegrini, and Steven E Jacobsen. Conservation and divergence of methylation patterning in plants and animals. *Proc Natl Acad Sci U S A*, 107(19):8689–8694, May 2010.

[53] J. A. Yoder, C. P. Walsh, and T. H. Bestor. Cytosine methylation and the ecology of intragenomic parasites. *Trends Genet*, 13(8):335–340, Aug 1997.

[54] Schraga Schwartz, Eran Meshorer, and Gil Ast. Chromatin organization marks exon-intron structure. *Nat Struct Mol Biol*, 16(9):990–995, Sep 2009.

[55] Louise Laurent, Eleanor Wong, Guoliang Li, Tien Huynh, Aristotelis Tsirigos, Chin Thing Ong, Hwee Meng Low, Ken Wing Kin Sung, Isidore Rigoutsos, Jeanne Loring, and Chia-Lin Wei. Dynamic changes in the human methylome during differentiation. *Genome Res*, 20(3):320–331, Mar 2010.

[56] Sanjeev Shukla, Ersen Kavak, Melissa Gregory, Masahiko Imashimizu, Bojan Shutinoski, Mikhail Kashlev, Philipp Oberdoerffer, Rickard Sandberg, and Shalini Oberdoerffer. Ctcf-promoted rna polymerase ii pausing links dna methylation to splicing. *Nature*, 479(7371):74–79, Nov 2011.

[57] Zhiyuan Shen. Genomic instability and cancer: an introduction. *J Mol Cell Biol*, 3(1):1–3, Feb 2011.

[58] Jacek Krol, Inga Loedige, and Witold Filipowicz. The widespread regulation of microrna biogenesis, function and decay. *Nat Rev Genet*, 11(9):597–610, Sep 2010.

[59] Andrew Grimson, Kyle Kai-How Farh, Wendy K Johnston, Philip Garrett-Engele, Lee P Lim, and David P Bartel. Microrna targeting specificity in mammals: determinants beyond seed pairing. *Mol Cell*, 27(1):91–105, Jul 2007.

[60] Marina Chekulaeva and Witold Filipowicz. Mechanisms of mirna-mediated post-transcriptional regulation in animal cells. *Curr Opin Cell Biol*, 21(3):452–460, Jun 2009.

[61] David P Bartel. Microrna: target recognition and regulatory functions. *Cell*, 136(2):215–233, Jan 2009.

[62] Peizhang Xu, Ming Guo, and Bruce A Hay. Microrna and the regulation of cell death. *Trends Genet*, 20(12):617–624, Dec 2004.

[63] Angie M Cheng, Mike W Byrom, Jeffrey Shelton, and Lance P Ford. Antisense inhibition of human mirnas and indications for an involvement of mirna in cell growth and apoptosis. *Nucleic Acids Res*, 33(4):1290–1297, 2005.

[64] Rosa Visone and Carlo M Croce. Mirnas and cancer. *Am J Pathol*, 174(4):1131–1138, Apr 2009.

[65] Peggy S Eis, Wayne Tam, Liping Sun, Amy Chadburn, Zongdong Li, Mario F Gomez, Elsebet Lund, and James E Dahlberg. Accumulation of mir-155 and bic rna in human b cell lymphomas. *Proc Natl Acad Sci U S A*, 102(10):3627–3632, Mar 2005.

[66] Amelia Cimmino, George Adrian Calin, Muller Fabbri, Marilena V Iorio, Manuela Ferracin, Masayoshi Shimizu, Sylwia E Wojcik, Rami I Aqeilan, Simona Zupo, Mariella Dono, Laura Rassenti, Hansjuerg Alder, Stefano Volinia, Chang-Gong Liu, Thomas J Kipps, Massimo Negrini, and Carlo M Croce. mir-15 and mir-16 induce apoptosis by targeting bcl2. *Proc Natl Acad Sci U S A*, 102(39):13944–13949, Sep 2005.

[67] G. A. Calin and C. M. Croce. Micrornas and chromosomal abnormalities in cancer cells. *Oncogene*, 25(46):6202–6210, Oct 2006.

[68] H. Guo, N. T. Ingolia, J. S. Weissman, and D. P. Bartel. Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature*, 466:835–840, Aug 2010.

[69] Chao Cheng, Xuping Fu, Pedro Alves, and Mark Gerstein. mrna expression profiles show differential regulatory effects of micrornas between estrogen receptor-positive and estrogen receptor-negative breast cancer. *Genome Biol*, 10(9):R90, 2009.

[70] A. Fire, S. Xu, M. K. Montgomery, S. A. Kostas, S. E. Driver, and C. C. Mello. Potent and specific genetic interference by double-stranded rna in caenorhabditis elegans. *Nature*, 391(6669):806–811, Feb 1998.

[71] Richard W Carthew and Erik J Sontheimer. Origins and mechanisms of mirnas and sirnas. *Cell*, 136(4):642–655, Feb 2009.

[72] Craig C Mello and Darryl Conte. Revealing the world of rna interference. *Nature*, 431(7006):338–342, Sep 2004.

[73] Gunter Meister and Thomas Tuschl. Mechanisms of gene silencing by double-stranded rna. *Nature*, 431(7006):343–349, Sep 2004.

[74] Edwards Allen, Zhixin Xie, Adam M Gustafson, and James C Carrington. microrna-directed phasing during trans-acting sirna biogenesis in plants. *Cell*, 121(2):207–221, Apr 2005.

[75] Daniel E Golden, Vincent R Gerbasi, and Erik J Sontheimer. An inside job for sirnas. *Mol Cell*, 31(3):309–312, Aug 2008.

[76] Dirk Haussecker. The business of rnai therapeutics in 2012. *Molecular Therapy*, 1:1, 2012.

[77] Marc Rehmsmeier, Peter Steffen, Matthias Hochsmann, and Robert Giegerich. Fast and effective prediction of microrna/target duplexes. *RNA*, 10(10):1507–1517, Oct 2004.

[78] Benjamin P Lewis, I hung Shih, Matthew W Jones-Rhoades, David P Bartel, and Christopher B Burge. Prediction of mammalian microrna targets. *Cell*, 115(7):787–798, Dec 2003.

[79] Azra Krek, Dominic Grün, Matthew N Poy, Rachel Wolf, Lauren Rosenberg, Eric J Epstein, Philip MacMenamin, Isabelle da Piedade, Kristin C Gunsalus, Markus Stoffel, and Nikolaus Rajewsky. Combinatorial microrna target predictions. *Nat Genet*, 37(5):495–500, May 2005.

[80] Jim C Huang, Tomas Babak, Timothy W Corson, Gordon Chua, Sofia Khan, Brenda L Gallie, Timothy R Hughes, Benjamin J Blencowe, Brendan J Frey, and Quaid D Morris. Using expression profiling data to identify human microrna targets. *Nat Methods*, 4(12):1045–1049, Dec 2007.

[81] Douglas L Black. Mechanisms of alternative pre-messenger rna splicing. *Annu Rev Biochem*, 72:291–336, 2003.

[82] Benjamin J Blencowe. Alternative splicing: new insights from global analyses. *Cell*, 126(1):37–47, Jul 2006.

[83] A. C. Goldstrohm, A. L. Greenleaf, and M. A. Garcia-Blanco. Co-transcriptional splicing of pre-messenger rnas: considerations for the mechanism of alternative splicing. *Gene*, 277(1-2):31–47, Oct 2001.

[84] Harald König, Nathalie Matter, Rüdiger Bader, Wilko Thiele, and Ferenc Müller. Splicing segregation: the minor spliceosome acts outside the nucleus and controls cell proliferation. *Cell*, 131(4):718–729, Nov 2007.

[85] Eric T Wang, Rickard Sandberg, Shujun Luo, Irina Khrebtukova, Lu Zhang, Christine Mayr, Stephen F Kingsmore, Gary P Schroth, and Christopher B Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–476, Nov 2008.

[86] Gretchen Edwalds-Gilbert. Regulation of mrna splicing by signal transduction. *Nature Education*, 3:43, 2010.

[87] Amanda J Law, Joel E Kleinman, Daniel R Weinberger, and Cynthia Shannon Weickert. Disease-associated intronic variants in the erbb4 gene are related to altered erbb4 splice-variant expression in the brain in schizophrenia. *Hum Mol Genet*, 16(2):129–141, Jan 2007.

[88] Evgeny Nudler and Alexander S Mironov. The riboswitch control of bacterial metabolism. *Trends Biochem Sci*, 29(1):11–17, Jan 2004.

[89] Partho Sarothi Ray, Jie Jia, Peng Yao, Mithu Majumder, Maria Hatzoglou, and Paul L Fox. A stress-responsive rna switch regulates vegfa expression. *Nature*, 457(7231):915–919, Feb 2009.

[90] Tim R Mercer, Marcel E Dinger, Cameron P Bracken, Gabriel Kolle, Jan M Szubert, Darren J Korbie, Marjan E Askarian-Amiri, Brooke B Gardiner, Gregory J Goodall, Sean M Grimmond, and John S Mattick. Regulated post-transcriptional rna cleavage diversifies the eukaryotic transcriptome. *Genome Res*, 20(12):1639–1650, Dec 2010.

[91] Jin Billy Li, Erez Y Levanon, Jung-Ki Yoon, John Aach, Bin Xie, Emily Leproust, Kun Zhang, Yuan Gao, and George M Church. Genome-wide identification of human rna editing sites by parallel dna capturing and sequencing. *Science*, 324(5931):1210–1213, May 2009.

[92] Mingyao Li, Isabel X Wang, Yun Li, Alan Bruzel, Allison L Richards, Jonathan M Toung, and Vivian G Cheung. Widespread rna and dna sequence differences in the human transcriptome. *Science*, 333(6038):53–58, Jul 2011.

[93] Jonathan K. Pritchard Joseph K. Pickrell, Yoav Gilad. Comment on "widespread rna and dna sequence differences in the human transcriptome". *Science*, 335:1302, 2012.

[94] Jacek Majewski Claudia L. Kleinman. Comment on "widespread rna and dna sequence differences in the human transcriptome". *Science*, 335:1302, 2012.

[95] Roy Parker and Haiwei Song. The enzymes and control of eukaryotic mrna turnover. *Nat Struct Mol Biol*, 11(2):121–127, Feb 2004.

[96] Kellie A Dean, Aneel K Aggarwal, and Robin P Wharton. Translational repressors in drosophila. *Trends Genet*, 18(11):572–577, Nov 2002.

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

[98] D. G. Wang, J. B. Fan, C. J. Siao, A. Berno, P. Young, R. Sapolsky, G. Ghandour, N. Perkins, E. Winchester, J. Spencer, L. Kruglyak, L. Stein, L. Hsie, T. Topaloglou, E. Hubbell, E. Robinson, M. Mittmann, M. S. Morris, N. Shen, D. Kilburn, J. Rioux, C. Nusbaum, S. Rozen, T. J. Hudson, R. Lipshutz, M. Chee, and E. S. Lander. Large-scale identification, mapping,

and genotyping of single-nucleotide polymorphisms in the human genome. *Science*, 280(5366):1077–1082, May 1998.

[99] Peter J Park. Chip-seq: advantages and challenges of a maturing technology. *Nat Rev Genet*, 10(10):669–680, Oct 2009.

[100] Jun Lu, Gad Getz, Eric A Miska, Ezequiel Alvarez-Saavedra, Justin Lamb, David Peck, Alejandro Sweet-Cordero, Benjamin L Ebert, Raymond H Mak, Adolfo A Ferrando, James R Downing, Tyler Jacks, H. Robert Horvitz, and Todd R Golub. Microrna expression profiles classify human cancers. *Nature*, 435(7043):834–838, Jun 2005.

[101] Michael J Heller. Dna microarray technology: devices, systems, and applications. *Annu Rev Biomed Eng*, 4:129–153, 2002.

[102] Edward L Korn, Jens K Habermann, Madhvi B Upender, Thomas Ried, and Lisa M McShane. Objective method of comparing dna microarray image analysis systems. *Biotechniques*, 36(6):960–967, Jun 2004.

[103] Eugene Novikov and Emmanuel Barillot. Software package for automatic microarray image analysis (maia). *Bioinformatics*, 23(5):639–640, Mar 2007.

[104] Edison T Liu, Sebastian Pott, and Mikael Huss. Q&a: Chip-seq technologies and the study of gene regulation. *BMC Biol*, 8:56, 2010.

[105] Affymetrix. Data sheet: Genechip human genome u133 arrays. *Affymetrix*, 2003.

[106] R. J. Lipshutz, S. P. Fodor, T. R. Gingeras, and D. J. Lockhart. High density synthetic oligonucleotide arrays. *Nat Genet*, 21(1 Suppl):20–24, Jan 1999.

[107] Affymetrix. Identifying and validating alternative splicing events. *Technical Note*.

[108] Yong You, Bernardo G Moreira, Mark A Behlke, and Richard Owczarzy. Design of lna probes that improve mismatch discrimination. *Nucleic Acids Res*, 34(8):e60, 2006.

[109] Todd C Mockler, Simon Chan, Ambika Sundaresan, Huaming Chen, Steven E Jacobsen, and Joseph R Ecker. Applications of dna tiling arrays for whole-genome analysis. *Genomics*, 85(1):1–15, Jan 2005.

[110] Oscar Aparicio, Joseph V Geisberg, and Kevin Struhl. Chromatin immunoprecipitation for determining the association of proteins with specific genomic sequences in vivo. *Curr Protoc Cell Biol*, Chapter 17:Unit 17.7, Sep 2004.

[111] L. Shi, G. Campbell, W. D. Jones, F. Campagne, Z. Wen, S. J. Walker, Z. Su, T. M. Chu, F. M. Goodsaid, L. Pusztai, J. D. Shaughnessy, A. Oberthuer, R. S. Thomas, R. S. Paules, M. Fielden, B. Barlogie, W. Chen, P. Du, M. Fischer, C. Furlanello, B. D. Gallas, X. Ge, D. B. Megherbi, W. F. Symmans, M. D. Wang, J. Zhang, H. Bitter, B. Brors, P. R. Bushel, M. Bylesjo, M. Chen, J. Cheng, J. Cheng, J. Chou, T. S. Davison, M. Delorenzi, Y. Deng, V. Devanarayan, D. J. Dix, J. Dopazo, K. C. Dorff, F. Elloumi, J. Fan, S. Fan, X. Fan, H. Fang, N. Gonzaludo, K. R. Hess, H. Hong, J. Huan, R. A. Irizarry, R. Judson, D. Juraeva, S. Lababidi, C. G. Lambert, L. Li, Y. Li, Z. Li, S. M. Lin, G. Liu, E. K. Lobenhofer, J. Luo, W. Luo, M. N. McCall, Y. Nikolsky, G. A. Pennello, R. G. Perkins, R. Philip, V. Popovici, N. D. Price, F. Qian, A. Scherer, T. Shi, W. Shi, J. Sung, D. Thierry-Mieg, J. Thierry-Mieg, V. Thodima, J. Trygg, L. Vishnuvajjala, S. J. Wang, J. Wu, Y. Wu, Q. Xie, W. A. Yousef, L. Zhang, X. Zhang, S. Zhong, Y. Zhou, S. Zhu, D. Arasappan, W. Bao, A. B. Lucas, F. Berthold, R. J. Brennan, A. Buness, J. G. Catalano, C. Chang, R. Chen, Y. Cheng, J. Cui, W. Czika, F. Demichelis, X. Deng, D. Dosymbekov, R. Eils, Y. Feng, J. Fostel, S. Fulmer-Smentek, J. C. Fuscoe, L. Gatto, W. Ge, D. R. Goldstein, L. Guo, D. N. Halbert, J. Han, S. C. Harris, C. Hatzis, D. Herman, J. Huang, R. V. Jensen, R. Jiang, C. D. Johnson, G. Jurman, Y. Kahlert, S. A. Khuder, M. Kohl, J. Li, L. Li, M. Li, Q. Z. Li, S. Li, Z. Li, J. Liu, Y. Liu, Z. Liu, L. Meng, M. Madera, F. Martinez-Murillo, I. Medina, J. Meehan, K. Miclaus, R. A. Moffitt, D. Montaner, P. Mukherjee, G. J. Mulligan, P. Neville, T. Nikolskaya, B. Ning, G. P. Page, J. Parker, R. M. Parry, X. Peng, R. L. Peterson, J. H. Phan, B. Quanz, Y. Ren, S. Riccadonna, A. H. Roter, F. W. Samuelson, M. M. Schumacher, J. D. Shambaugh, Q. Shi, R. Shippy, S. Si, A. Smalter, C. Sotiriou, M. Soukup, F. Staedtler, G. Steiner, T. H. Stokes, Q. Sun, P. Y. Tan, R. Tang, Z. Tezak, B. Thorn, M. Tsyganova, Y. Turpaz, S. C. Vega, R. Visintainer, J. von Frese, C. Wang, E. Wang, J. Wang, W. Wang, F. Westermann, J. C. Willey, M. Woods, S. Wu, N. Xiao, J. Xu, L. Xu, L. Yang, X. Zeng, J. Zhang, L. Zhang, M. Zhang, C. Zhao, R. K. Puri, U. Scherf, W. Tong, and R. D. Wolfinger. The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nat. Biotechnol.*, 28(8):827–838, Aug 2010.

[112] H. Bengtsson, G. Jonsson, and J. Vallon-Christersson. Calibration and as-

sessment of channel-specific biases in microarray data with extended dynamical range. *BMC Bioinformatics*, 5:177, Nov 2004.

[113] A. Bengtsson and H. Bengtsson. Microarray image analysis: background estimation using quantile and morphological filters. *BMC Bioinformatics*, 7:96, 2006.

[114] Laurent Gautier, Leslie Cope, Benjamin M Bolstad, and Rafael A Irizarry. affy–analysis of affymetrix genechip data at the probe level. *Bioinformatics*, 20(3):307–315, Feb 2004.

[115] Matthew E Ritchie, Jeremy Silver, Alicia Oshlack, Melissa Holmes, Dileepa Diyagama, Andrew Holloway, and Gordon K Smyth. A comparison of background correction methods for two-colour microarrays. *Bioinformatics*, 23(20):2700–2707, Oct 2007.

[116] RA Irizarry, Laurent Gautier, and L. Cope. *The Analysis of Gene Expression Data*. Springer, 2003.

[117] Benjamin Milo Bolstad. *Low-level Analysis of High-density Oligonucleotide Array Data: Background, Normalization and Summarization*. PhD thesis, University of California, Berkely, 2004.

[118] J. Quackenbush. Microarray data normalization and transformation. *Nat. Genet.*, 32 Suppl:496–501, Dec 2002.

[119] G. K. Smyth and T. Speed. Normalization of cDNA microarray data. *Methods*, 31(4):265–273, Dec 2003.

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

[121] Wolfgang Huber, Anja von Heydebreck, Holger Sültmann, Annemarie Poustka, and Martin Vingron. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18 Suppl 1:S96–104, 2002.

[122] Affymetrix. Statistical algorithms description document. Affymetrix White Paper, 2002.

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

[124] Alexander Sturn, John Quackenbush, and Zlatko Trajanoski. Genesis: cluster analysis of microarray data. *Bioinformatics*, 18(1):207–208, Jan 2002.

[125] Markus Ringnér. What is principal component analysis? *Nat Biotechnol*, 26(3):303–304, Mar 2008.

[126] F. Corpet. Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res*, 16(22):10881–10890, Nov 1988.

[127] Jeremy Gollub and Gavin Sherlock. Clustering microarray data. *Methods Enzymol*, 411:194–213, 2006.

[128] Jeanette N McClintick and Howard J Edenberg. Effects of filtering by present call on analysis of microarray experiments. *BMC Bioinformatics*, 7:49, 2006.

[129] Helen E Lockstone. Exon array data analysis using affymetrix power tools and r statistical software. *Brief Bioinform*, 12(6):634–644, Nov 2011.

[130] Affymetrix. Quality assessment of exon and gene 1.0 st arrays. *Affymetrix White Paper*, 2010.

[131] Daniel Holder, Richard F. Raubertas, V. Bill Pikounis, Vladimir Svetnik, and Keith Soper. Statistical analysis of high density oligonucleotide arrays: a safer approach. 2001.

[132] Gordon K Smyth. Linear models and empirical bayes methods for assessing differential expression of microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1):Article 3, 2004.

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

[134] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, Oct 1999.

[135] T. Sørlie, C. M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, T. Thorsen, H. Quist, J. C. Matese, P. O. Brown, D. Botstein, P. E. Lønning, and A. L. Børresen-Dale. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A*, 98(19):10869–10874, Sep 2001.

[136] Robert Tibshirani, Trevor Hastie, Balasubramanian Narasimhan, and Gilbert Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci U S A*, 99(10):6567–6572, May 2002.

[137] David Oviatt, Mark Clement, Quinn Snell, Kenneth Sundberg, Chun Wan J Lai, Jared Allen, and Randall Roper. Inferring gene regulatory networks from asynchronous microarray data with airnet. *BMC Genomics*, 11 Suppl 2:S6, 2010.

[138] Jun Li and Margit Burmeister. Genetical genomics: combining genetics with gene expression analysis. *Hum Mol Genet*, 14 Spec No. 2:R163–R169, Oct 2005.

[139] Jordana T Bell, Athma A Pai, Joseph K Pickrell, Daniel J Gaffney, Roger Pique-Regi, Jacob F Degner, Yoav Gilad, and Jonathan K Pritchard. Dna methylation patterns associate with genetic and gene expression variation in hapmap cell lines. *Genome Biol*, 12(1):R10, 2011.

[140] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. Di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R. R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Wang, A. Wang, X. Wang, J. Wang, M. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May,

S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y. H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guigó, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y. H. Chiang, M. Coyne, C. Dahlke, A. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh, and X. Zhu. The sequence of the human genome. *Science*, 291(5507):1304–1351, Feb 2001.

[141] F. Sanger, S. Nicklen, and A. R. Coulson. Dna sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A*, 74(12):5463–5467, Dec 1977.

[142] J.A. Luckey, H. Drossman, A.J. Kostichka, D.A. Mead, J. D'Cunha, T.B. Norris, and L.M. Smith. High speed dna sequencing by capillary electrophoresis. *Nucleic acids research*, 18(15):4417–4421, 1990.

[143] Stephan C Schuster. Next-generation sequencing transforms today's biology. *Nat Methods*, 5(1):16–18, Jan 2008.

[144] Daniel MacLean, Jonathan D G Jones, and David J Studholme. Application of 'next-generation' sequencing technologies to microbial genetics. *Nat Rev Microbiol*, 7(4):287–296, Apr 2009.

[145] Fatih Ozsolak and Patrice M Milos. Rna sequencing: advances, challenges and opportunities. *Nat Rev Genet*, 12(2):87–98, Feb 2011.

[146] Timothy J Ley, Elaine R Mardis, Li Ding, Bob Fulton, Michael D McLellan, Ken Chen, David Dooling, Brian H Dunford-Shore, Sean McGrath, Matthew Hickenbotham, Lisa Cook, Rachel Abbott, David E Larson, Dan C

Koboldt, Craig Pohl, Scott Smith, Amy Hawkins, Scott Abbott, Devin Locke, Ladeana W Hillier, Tracie Miner, Lucinda Fulton, Vincent Magrini, Todd Wylie, Jarret Glasscock, Joshua Conyers, Nathan Sander, Xiaoqi Shi, John R Osborne, Patrick Minx, David Gordon, Asif Chinwalla, Yu Zhao, Rhonda E Ries, Jacqueline E Payton, Peter Westervelt, Michael H Tomasson, Mark Watson, Jack Baty, Jennifer Ivanovich, Sharon Heath, William D Shannon, Rakesh Nagarajan, Matthew J Walter, Daniel C Link, Timothy A Graubert, John F DiPersio, and Richard K Wilson. Dna sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature*, 456(7218):66–72, Nov 2008.

[147] David S Johnson, Ali Mortazavi, Richard M Myers, and Barbara Wold. Genome-wide mapping of in vivo protein-dna interactions. *Science*, 316(5830):1497–1502, Jun 2007.

[148] Y. Ruike, Y. Imanaka, F. Sato, K. Shimizu, and G. Tsujimoto. Genome-wide analysis of aberrant methylation in human breast cancer cells using methyl-DNA immunoprecipitation combined with high-throughput sequencing. *BMC Genomics*, 11:137, 2010.

[149] Ugrappa Nagalakshmi, Zhong Wang, Karl Waern, Chong Shou, Debasish Raha, Mark Gerstein, and Michael Snyder. The transcriptional landscape of the yeast genome defined by rna sequencing. *Science*, 320(5881):1344–1349, Jun 2008.

[150] D. R. Bentley. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–59, Nov 2008.

[151] Martin Kircher, Udo Stenzel, and Janet Kelso. Improved base calling for the illumina genome analyzer using machine learning strategies. *Genome Biol*, 10(8):R83, 2009.

[152] Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by rna-seq. *Nat Methods*, 5(7):621–628, Jul 2008.

[153] J. H. Malone and B. Oliver. Microarrays, deep sequencing and the true measure of the transcriptome. *BMC Biol.*, 9:34, 2011.

[154] Xing Fu, Ning Fu, Song Guo, Zheng Yan, Ying Xu, Hao Hu, Corinna Menzel, Wei Chen, Yixue Li, Rong Zeng, and Philipp Khaitovich. Estimating accuracy of rna-seq and microarrays with proteomics. *BMC Genomics*, 10:161, 2009.

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

[156] Heng Li and Nils Homer. A survey of sequence alignment algorithms for next-generation sequencing. *Brief Bioinform*, 11(5):473–483, Sep 2010.

[157] David S DeLuca, Joshua Z Levin, Andrey Sivachenko, Timothy Fennell, Marc-Danie Nazaire, Chris Williams, Michael Reich, Wendy Winckler, and Gad Getz. Rna-seqc: Rna-seq metrics for quality control and process optimization. *Bioinformatics*, 28(11):1530–1532, Jun 2012.

[158] Robert Schmieder and Robert Edwards. Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, 27(6):863–864, Mar 2011.

[159] Yuval Benjamini and Terence P Speed. Summarizing and correcting the gc content bias in high-throughput sequencing. *Nucleic Acids Res*, 40(10):e72, May 2012.

[160] J. K. Pickrell, J. C. Marioni, A. A. Pai, J. F. Degner, B. E. Engelhardt, E. Nkadori, J. B. Veyrieras, M. Stephens, Y. Gilad, and J. K. Pritchard. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, 464:768–772, Apr 2010.

[161] M. S. Burriesci, E. M. Lehnert, and J. R. Pringle. Fulcrum: condensing redundant reads from high-throughput sequencing studies. *Bioinformatics*, 28(10):1324–1327, May 2012.

[162] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, Aug 2009.

[163] I. Kozarewa, Z. Ning, M. A. Quail, M. J. Sanders, M. Berriman, and D. J. Turner. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nat. Methods*, 6(4):291–295, Apr 2009.

[164] A.D. Jayaprakash, O. Jabado, B.D. Brown, and R. Sachidanandam. Identification and remediation of biases in the activity of rna ligases in small-rna deep sequencing. *Nucleic acids research*, 39(21):e141–e141, 2011.

[165] D. Aird, M.G. Ross, W.S. Chen, M. Danielsson, T. Fennell, C. Russ, D.B. Jaffe, C. Nusbaum, and A. Gnirke. Analyzing and minimizing pcr amplification bias in illumina sequencing libraries. *Genome Biol*, 12(2):R18, 2011.

[166] Adam Roberts, Harold Pimentel, Cole Trapnell, and Lior Pachter. Identification of novel transcripts in annotated genomes using rna-seq. *Bioinformatics*, 27(17):2325–2329, Sep 2011.

[167] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2010. ISBN 3-900051-07-0.

[168] Kurt Hornik. *The R FAQ*. http://CRAN.R-project.org/doc/FAQ/R-FAQ.html, 2008. ISBN 3-900051-08-9.

[169] Uwe Ligges and John Fox. R Help Desk: How can I avoid this loop or make it faster? *R News*, 8(1):46–50, May 2008.

[170] Robert C Gentleman, Vincent J. Carey, Douglas M. Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, Kurt Hornik, Torsten Hothorn, Wolfgang Huber, Stefano Iacus, Rafael Irizarry, Friedrich Leisch, Cheng Li, Martin Maechler, Anthony J. Rossini, Gunther Sawitzki, Colin Smith, Gordon Smyth, Luke Tierney, Jean Y. H. Yang, and Jianhua Zhang. Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, 5:R80, 2004.

[171] Gentleman, Rossini, Dudoit, and Hornik. *The Bioconductor FAQ*. http://www.bioconductor.org/docs/faq/, 2008.

[172] Hao Yu. Rmpi: Interface (wrapper) to mpi (message-passing interface). 2001. R package version 0.5-8.

[173] Gordon K. Smyth. Limma: linear models for microarray data. pages 397–420, 2005.

[174] S. Falcon and R. Gentleman. Using gostats to test gene lists for go term association. *Bioinformatics*, 23(2):257–258, Jan 2007.

[175] P. Aboyoun, H. Pages, and M. Lawrence. *GenomicRanges: Representation and manipulation of genomic intervals*. R package version 1.0.1.

[176] M. Carlson, H. Pages, P. Aboyoun, S. Falcon, M. Morgan, D. Sarkar, and M. Lawrence. *GenomicFeatures: Tools for making and manipulating transcript centric annotations*. R package version 1.0.10.

[177] Heng Li, Jue Ruan, and Richard Durbin. Mapping short dna sequencing reads and calling variants using mapping quality scores. *Genome Res*, 18(11):1851–1858, Nov 2008.

[178] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L Salzberg. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biol*, 10(3):R25, 2009.

[179] Martin Morgan, Michael Lawrence, and Simon Anders. *ShortRead: Base classes and methods for high-throughput short-read sequencing data.* R package version 1.6.2.

[180] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biol*, 11(10):R106, 2010.

[181] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, Jan 2010.

[182] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, 25(1):25–29, May 2000.

[183] M. Kanehisa and S. Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 28(1):27–30, Jan 2000.

[184] Paul Pavlidis, Jie Qin, Victoria Arango, John J Mann, and Etienne Sibille. Using the gene ontology for microarray data mining: a comparison of methods and application to age effects in human prefrontal cortex. *Neurochem Res*, 29(6):1213–1222, Jun 2004.

[185] Da Wei Huang, Brad T Sherman, and Richard A Lempicki. Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nat Protoc*, 4(1):44–57, 2009.

[186] Da Wei Huang, Brad T Sherman, and Richard A Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res*, 37(1):1–13, Jan 2009.

[187] Jelle J Goeman and Peter Bühlmann. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, 23(8):980–987, Apr 2007.

[188] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy,

Todd R Golub, Eric S Lander, and Jill P Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, 102(43):15545–15550, Oct 2005.

[189] International HapMap Consortium. The international hapmap project. *Nature*, 426(6968):789–796, Dec 2003.

[190] International HapMap Consortium, Kelly A Frazer, Dennis G Ballinger, David R Cox, David A Hinds, Laura L Stuve, Richard A Gibbs, John W Belmont, Andrew Boudreau, Paul Hardenbol, Suzanne M Leal, Shiran Pasternak, David A Wheeler, Thomas D Willis, Fuli Yu, Huanming Yang, Changqing Zeng, Yang Gao, Haoran Hu, Weitao Hu, Chaohua Li, Wei Lin, Siqi Liu, Hao Pan, Xiaoli Tang, Jian Wang, Wei Wang, Jun Yu, Bo Zhang, Qingrun Zhang, Hongbin Zhao, Hui Zhao, Jun Zhou, Stacey B Gabriel, Rachel Barry, Brendan Blumenstiel, Amy Camargo, Matthew Defelice, Maura Faggart, Mary Goyette, Supriya Gupta, Jamie Moore, Huy Nguyen, Robert C Onofrio, Melissa Parkin, Jessica Roy, Erich Stahl, Ellen Winchester, Liuda Ziaugra, David Altshuler, Yan Shen, Zhijian Yao, Wei Huang, Xun Chu, Yungang He, Li Jin, Yangfan Liu, Yayun Shen, Weiwei Sun, Haifeng Wang, Yi Wang, Ying Wang, Xiaoyan Xiong, Liang Xu, Mary M Y Waye, Stephen K W Tsui, Hong Xue, J. Tze-Fei Wong, Luana M Galver, Jian-Bing Fan, Kevin Gunderson, Sarah S Murray, Arnold R Oliphant, Mark S Chee, Alexandre Montpetit, Fanny Chagnon, Vincent Ferretti, Martin Leboeuf, Jean-François Olivier, Michael S Phillips, Stéphanie Roumy, Clémentine Sallée, Andrei Verner, Thomas J Hudson, Pui-Yan Kwok, Dongmei Cai, Daniel C Koboldt, Raymond D Miller, Ludmila Pawlikowska, Patricia Taillon-Miller, Ming Xiao, Lap-Chee Tsui, William Mak, You Qiang Song, Paul K H Tam, Yusuke Nakamura, Takahisa Kawaguchi, Takuya Kitamoto, Takashi Morizono, Atsushi Nagashima, Yozo Ohnishi, Akihiro Sekine, Toshihiro Tanaka, Tatsuhiko Tsunoda, Panos Deloukas, Christine P Bird, Marcos Delgado, Emmanouil T Dermitzakis, Rhian Gwilliam, Sarah Hunt, Jonathan Morrison, Don Powell, Barbara E Stranger, Pamela Whittaker, David R Bentley, Mark J Daly, Paul I W de Bakker, Jeff Barrett, Yves R Chretien, Julian Maller, Steve McCarroll, Nick Patterson, Itsik Pe'er, Alkes Price, Shaun Purcell, Daniel J Richter, Pardis Sabeti, Richa Saxena, Stephen F Schaffner, Pak C Sham, Patrick Varilly, David Altshuler, Lincoln D Stein, Lalitha Krishnan, Albert Vernon Smith, Marcela K Tello-Ruiz, Gudmundur A Thorisson, Aravinda Chakravarti, Peter E Chen, David J Cutler, Carl S Kashuk, Shin Lin, Gonçalo R Abecasis, Weihua Guan, Yun Li, Heather M Munro, Zhaohui Steve Qin, Daryl J Thomas, Gilean McVean, Adam Auton, Leonardo Bottolo, Niall Cardin, Susana Eyheramendy, Colin Freeman, Jonathan Marchini, Simon Myers, Chris

Spencer, Matthew Stephens, Peter Donnelly, Lon R Cardon, Geraldine Clarke, David M Evans, Andrew P Morris, Bruce S Weir, Tatsuhiko Tsunoda, James C Mullikin, Stephen T Sherry, Michael Feolo, Andrew Skol, Houcan Zhang, Changqing Zeng, Hui Zhao, Ichiro Matsuda, Yoshimitsu Fukushima, Darryl R Macer, Eiko Suda, Charles N Rotimi, Clement A Adebamowo, Ike Ajayi, Toyin Aniagwu, Patricia A Marshall, Chibuzor Nkwodimmah, Charmaine D M Royal, Mark F Leppert, Missy Dixon, Andy Peiffer, Renzong Qiu, Alastair Kent, Kazuto Kato, Norio Niikawa, Isaac F Adewole, Bartha M Knoppers, Morris W Foster, Ellen Wright Clayton, Jessica Watkin, Richard A Gibbs, John W Belmont, Donna Muzny, Lynne Nazareth, Erica Sodergren, George M Weinstock, David A Wheeler, Imtaz Yakub, Stacey B Gabriel, Robert C Onofrio, Daniel J Richter, Liuda Ziaugra, Bruce W Birren, Mark J Daly, David Altshuler, Richard K Wilson, Lucinda L Fulton, Jane Rogers, John Burton, Nigel P Carter, Christopher M Clee, Mark Griffiths, Matthew C Jones, Kirsten McLay, Robert W Plumb, Mark T Ross, Sarah K Sims, David L Willey, Zhu Chen, Hua Han, Le Kang, Martin Godbout, John C Wallenburg, Paul L'Archevêque, Guy Bellemare, Koji Saeki, Hongguang Wang, Daochang An, Hongbo Fu, Qing Li, Zhen Wang, Renwu Wang, Arthur L Holden, Lisa D Brooks, Jean E McEwen, Mark S Guyer, Vivian Ota Wang, Jane L Peterson, Michael Shi, Jack Spiegel, Lawrence M Sung, Lynn F Zacharia, Francis S Collins, Karen Kennedy, Ruth Jamieson, and John Stewart. A second generation human haplotype map of over 3.1 million snps. *Nature*, 449(7164):851–861, Oct 2007.

[191] Andrew D Johnson and Christopher J O'Donnell. An open access database of genome-wide association results. *BMC Med Genet*, 10:6, 2009.

[192] Peter M Visscher, Matthew A Brown, Mark I McCarthy, and Jian Yang. Five years of gwas discovery. *Am J Hum Genet*, 90(1):7–24, Jan 2012.

[193] G. Järnerot B. Flodérus-Myrhed C. Tysk, E. Lindberg. Ulcerative colitis and crohn's disease in an unselected population of monozygotic and dizygotic twins. a study of heritability and the influence of smoking. *Gut*, 29:990–996, 1988.

[194] Andre Franke, Dermot P B McGovern, Jeffrey C Barrett, Kai Wang, Graham L Radford-Smith, Tariq Ahmad, Charlie W Lees, Tobias Balschun, James Lee, Rebecca Roberts, Carl A Anderson, Joshua C Bis, Suzanne Bumpstead, David Ellinghaus, Eleonora M Festen, Michel Georges, Todd Green, Talin Haritunians, Luke Jostins, Anna Latiano, Christopher G Mathew, Grant W Montgomery, Natalie J Prescott, Soumya Raychaudhuri,

Jerome I Rotter, Philip Schumm, Yashoda Sharma, Lisa A Simms, Kent D Taylor, David Whiteman, Cisca Wijmenga, Robert N Baldassano, Murray Barclay, Theodore M Bayless, Stephan Brand, Carsten Büning, Albert Cohen, Jean-Frederick Colombel, Mario Cottone, Laura Stronati, Ted Denson, Martine De Vos, Renata D'Inca, Marla Dubinsky, Cathryn Edwards, Tim Florin, Denis Franchimont, Richard Gearry, Jürgen Glas, Andre Van Gossum, Stephen L Guthery, Jonas Halfvarson, Hein W Verspaget, Jean-Pierre Hugot, Amir Karban, Debby Laukens, Ian Lawrance, Marc Lemann, Arie Levine, Cecile Libioulle, Edouard Louis, Craig Mowat, William Newman, Julián Panés, Anne Phillips, Deborah D Proctor, Miguel Regueiro, Richard Russell, Paul Rutgeerts, Jeremy Sanderson, Miquel Sans, Frank Seibold, A. Hillary Steinhart, Pieter C F Stokkers, Leif Torkvist, Gerd Kullak-Ublick, David Wilson, Thomas Walters, Stephan R Targan, Steven R Brant, John D Rioux, Mauro D'Amato, Rinse K Weersma, Subra Kugathasan, Anne M Griffiths, John C Mansfield, Severine Vermeire, Richard H Duerr, Mark S Silverberg, Jack Satsangi, Stefan Schreiber, Judy H Cho, Vito Annese, Hakon Hakonarson, Mark J Daly, and Miles Parkes. Genome-wide meta-analysis increases to 71 the number of confirmed crohn's disease susceptibility loci. *Nat Genet*, 42(12):1118–1125, Dec 2010.

[195] Teri A Manolio, Francis S Collins, Nancy J Cox, David B Goldstein, Lucia A Hindorff, David J Hunter, Mark I McCarthy, Erin M Ramos, Lon R Cardon, Aravinda Chakravarti, Judy H Cho, Alan E Guttmacher, Augustine Kong, Leonid Kruglyak, Elaine Mardis, Charles N Rotimi, Montgomery Slatkin, David Valle, Alice S Whittemore, Michael Boehnke, Andrew G Clark, Evan E Eichler, Greg Gibson, Jonathan L Haines, Trudy F C Mackay, Steven A McCarroll, and Peter M Visscher. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, Oct 2009.

[196] Carl A Anderson, Gabrielle Boucher, Charlie W Lees, Andre Franke, Mauro D'Amato, Kent D Taylor, James C Lee, Philippe Goyette, Marcin Imielinski, Anna Latiano, Caroline Lagacé, Regan Scott, Leila Amininejad, Suzannah Bumpstead, Leonard Baidoo, Robert N Baldassano, Murray Barclay, Theodore M Bayless, Stephan Brand, Carsten Büning, Jean-Frédéric Colombel, Lee A Denson, Martine De Vos, Marla Dubinsky, Cathryn Edwards, David Ellinghaus, Rudolf S N Fehrmann, James A B Floyd, Timothy Florin, Denis Franchimont, Lude Franke, Michel Georges, Jürgen Glas, Nicole L Glazer, Stephen L Guthery, Talin Haritunians, Nicholas K Hayward, Jean-Pierre Hugot, Gilles Jobin, Debby Laukens, Ian Lawrance, Marc Lémann, Arie Levine, Cecile Libioulle, Edouard Louis, Dermot P McGovern, Monica Milla, Grant W Montgomery, Katherine I Morley, Craig Mowat, Aylwin Ng, William Newman, Roel A Ophoff, Laura Papi, Orazio Palmieri,

Laurent Peyrin-Biroulet, Julián Panés, Anne Phillips, Natalie J Prescott, Deborah D Proctor, Rebecca Roberts, Richard Russell, Paul Rutgeerts, Jeremy Sanderson, Miquel Sans, Philip Schumm, Frank Seibold, Yashoda Sharma, Lisa A Simms, Mark Seielstad, A. Hillary Steinhart, Stephan R Targan, Leonard H van den Berg, Morten Vatn, Hein Verspaget, Thomas Walters, Cisca Wijmenga, David C Wilson, Harm-Jan Westra, Ramnik J Xavier, Zhen Z Zhao, Cyriel Y Ponsioen, Vibeke Andersen, Leif Torkvist, Maria Gazouli, Nicholas P Anagnou, Tom H Karlsen, Limas Kupcinskas, Jurgita Sventoraityte, John C Mansfield, Subra Kugathasan, Mark S Silverberg, Jonas Halfvarson, Jerome I Rotter, Christopher G Mathew, Anne M Griffiths, Richard Gearry, Tariq Ahmad, Steven R Brant, Mathias Chamaillard, Jack Satsangi, Judy H Cho, Stefan Schreiber, Mark J Daly, Jeffrey C Barrett, Miles Parkes, Vito Annese, Hakon Hakonarson, Graham Radford-Smith, Richard H Duerr, Séverine Vermeire, Rinse K Weersma, and John D Rioux. Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. *Nat Genet*, 43(3):246–252, Mar 2011.

[197] G. Pare. Genome-wide association studies–data generation, storage, interpretation, and bioinformatics. *J Cardiovasc Transl Res*, 3:183–188, Jun 2010.

[198] D. G. Cox and P. Kraft. Quantification of the power of Hardy-Weinberg equilibrium testing to detect genotyping error. *Hum. Hered.*, 61:10–14, 2006.

[199] Lee P Lim, Nelson C Lau, Philip Garrett-Engele, Andrew Grimson, Janell M Schelter, John Castle, David P Bartel, Peter S Linsley, and Jason M Johnson. Microarray analysis shows that some micrornas downregulate large numbers of target mrnas. *Nature*, 433(7027):769–773, Feb 2005.

[200] Zhenbao Yu, Zhaofeng Jian, Shi-Hsiang Shen, Enrico Purisima, and Edwin Wang. Global analysis of microrna target gene expression reveals that mirna targets are lower expressed in mature mouse and drosophila tissues than in the embryos. *Nucleic Acids Res*, 35(1):152–164, 2007.

[201] Chao Cheng and Lei M Li. Inferring microrna activities by combining gene expression with microrna target prediction. *PLoS One*, 3(4):e1989, 2008.

[202] Amit Arora and David Ac Simpson. Individual mrna expression profiles reveal the effects of specific microrna. *Genome Biol*, 9(5):R82, 2008.

[203] Xiaowei Wang and Xiaohui Wang. Systematic identification of microrna functions by combining target prediction and expression profiling. *Nucleic Acids Res*, 34(5):1646–1652, 2006.

[204] Francesca M Buffa, Carme Camps, Laura Winchester, Cameron E Snell, Harriet E Gee, Helen Sheldon, Marian Taylor, Adrian L Harris, and Jiannis Ragoussis. microrna associated progression pathways and potential therapeutic targets identified by integrated mrna and microrna expression profiling in breast cancer. *Cancer Res*, Jul 2011.

[205] Joseph K Pickrell, John C Marioni, Athma A Pai, Jacob F Degner, Barbara E Engelhardt, Everlyne Nkadori, Jean-Baptiste Veyrieras, Matthew Stephens, Yoav Gilad, and Jonathan K Pritchard. Understanding mechanisms underlying human gene expression variation with rna sequencing. *Nature*, 464(7289):768–772, Apr 2010.

[206] Stephen B Montgomery, Micha Sammeth, Maria Gutierrez-Arcelus, Radoslaw P Lach, Catherine Ingle, James Nisbett, Roderic Guigo, and Emmanouil T Dermitzakis. Transcriptome genetics using second generation sequencing in a caucasian population. *Nature*, 464(7289):773–777, Apr 2010.

[207] R. Stephanie Huang, Shiwei Duan, Wasim K Bleibel, Emily O Kistner, Wei Zhang, Tyson A Clark, Tina X Chen, Anthony C Schweitzer, John E Blume, Nancy J Cox, and M. Eileen Dolan. A genome-wide approach to identify genetic variants that contribute to etoposide-induced cytotoxicity. *Proc Natl Acad Sci U S A*, 104(23):9758–9763, Jun 2007.

[208] Peter M Visscher, William G Hill, and Naomi R Wray. Heritability in the genomics era–concepts and misconceptions. *Nat Rev Genet*, 9(4):255–266, Apr 2008.

[209] P. Wray, N. & Visscher. Estimating trait heritability. *Nature Education*, 2008.

[210] Matthew B. Hamilton. *Population Genetics*. John Wiley & Sons, April 2009.

[211] Benjamin P Lewis, Christopher B Burge, and David P Bartel. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microrna targets. *Cell*, 120(1):15–20, Jan 2005.

[212] Bino John, Anton J Enright, Alexei Aravin, Thomas Tuschl, Chris Sander, and Debora S Marks. Human microrna targets. *PLoS Biol*, 2(11):e363, Nov 2004.

[213] Xiaowei Wang and Issam M El Naqa. Prediction of both conserved and nonconserved microrna targets in animals. *Bioinformatics*, 24(3):325–332, Feb 2008.

[214] Vivian G Cheung, Laura K Conlin, Teresa M Weber, Melissa Arcaro, Kuang-Yu Jen, Michael Morley, and Richard S Spielman. Natural variation in human gene expression assessed in lymphoblastoid cells. *Nat Genet*, 33(3):422–425, Mar 2003.

[215] Eric E Schadt, Stephanie A Monks, Thomas A Drake, Aldons J Lusis, Nam Che, Veronica Colinayo, Thomas G Ruff, Stephen B Milligan, John R Lamb, Guy Cavet, Peter S Linsley, Mao Mao, Roland B Stoughton, and Stephen H Friend. Genetics of gene expression surveyed in maize, mouse and man. *Nature*, 422(6929):297–302, Mar 2003.

[216] Rachel B Brem, Gaël Yvert, Rebecca Clinton, and Leonid Kruglyak. Genetic dissection of transcriptional regulation in budding yeast. *Science*, 296(5568):752–755, Apr 2002.

[217] J. Mata and J. Bähler. Correlations between gene expression and gene conservation in fission yeast. *Genome research*, 13(12):2686–2690, 2003.

[218] SA Monks, A. Leonardson, H. Zhu, P. Cundiff, P. Pietrusiak, S. Edwards, JW Phillips, A. Sachs, and EE Schadt. Genetic inheritance of gene expression in human cell lines. *The American Journal of Human Genetics*, 75(6):1094–1105, 2004.

[219] Eugene Berezikov, Wei-Jen Chung, Jason Willis, Edwin Cuppen, and Eric C Lai. Mammalian mirtron genes. *Mol Cell*, 28(2):328–336, Oct 2007.

[220] Zongli Xu and Jack A Taylor. Snpinfo: integrating gwas and candidate gene information into functional snp selection for genetic association studies. *Nucleic Acids Res*, 37(Web Server issue):W600–W605, Jul 2009.

[221] Lucia A Hindorff, Praveen Sethupathy, Heather A Junkins, Erin M Ramos, Jayashri P Mehta, Francis S Collins, and Teri A Manolio. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A*, 106(23):9362–9367, Jun 2009.

[222] Stacey B Gabriel, Stephen F Schaffner, Huy Nguyen, Jamie M Moore, Jessica Roy, Brendan Blumenstiel, John Higgins, Matthew DeFelice, Amy Lochner, Maura Faggart, Shau Neen Liu-Cordero, Charles Rotimi, Adebowale Adeyemo, Richard Cooper, Ryk Ward, Eric S Lander, Mark J Daly, and David Altshuler. The structure of haplotype blocks in the human genome. *Science*, 296(5576):2225–2229, Jun 2002.

[223] Jason Ernst, Pouya Kheradpour, Tarjei S Mikkelsen, Noam Shoresh, Lucas D Ward, Charles B Epstein, Xiaolan Zhang, Li Wang, Robbyn Issner, Michael Coyne, Manching Ku, Timothy Durham, Manolis Kellis, and

Bradley E Bernstein. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, 473(7345):43–49, May 2011.

[224] Cole Trapnell, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*, 28(5):511–515, May 2010.

[225] 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, Oct 2010.

[226] R. Stephanie Huang, Eric R Gamazon, Dana Ziliak, Yujia Wen, Hae Kyung Im, Wei Zhang, Claudia Wing, Shiwei Duan, Wasim K Bleibel, Nancy J Cox, and M. Eileen Dolan. Population differences in microrna expression and biological implications. *RNA Biol*, 8(4):692–701, Jul 2011.

[227] Zhi Liang, Hong Zhou, Haoran Zheng, and Jiarui Wu. Expression levels of micrornas are not associated with their regulatory activities. *Biol Direct*, 6:43, 2011.

[228] J. Michael Thomson, Martin Newman, Joel S Parker, Elizabeth M Morin-Kensicki, Tricia Wright, and Scott M Hammond. Extensive post-transcriptional regulation of micrornas and its implications for cancer. *Genes Dev*, 20(16):2202–2207, Aug 2006.

[229] Anna Torres, Kamil Torres, Tomasz Paszkowski, Barbara Jodlowska-Jedrych, Tomasz Radomanski, Andrzej Ksiazek, and Ryszard Maciejewski. Major regulators of micrornas biogenesis dicer and drosha are down-regulated in endometrial cancer. *Tumour Biol*, 32(4):769–776, Aug 2011.

[230] Konstantin J Dedes, Rachael Natrajan, Maryou B Lambros, Felipe C Geyer, Maria Angeles Lopez-Garcia, Kay Savage, Robin L Jones, and Jorge S Reis-Filho. Down-regulation of the mirna master regulators drosha and dicer is associated with specific subgroups of breast cancer. *Eur J Cancer*, 47(1):138–150, Jan 2011.

[231] Francesco Favero. *RmiR.Hs.miRNA: Various databases of microRNA Targets*. R package version 1.0.6.

[232] Y. S. Aulchenko, S. Ripke, A. Isaacs, and C. M. van Duijn. GenABEL: an R library for genome-wide association analysis. *Bioinformatics*, 23:1294–1296, May 2007.

[233] Y. S. Aulchenko, M. V. Struchalin, and C. M. van Duijn. ProbABEL package for genome-wide association analysis of imputed data. *BMC Bioinformatics*, 11:134, 2010.

[234] John D Storey and Robert Tibshirani. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A*, 100(16):9440–9445, Aug 2003.

[235] Zhongxue Chen, Monnie McGee, Qingzhong Liu, and Richard H Scheuermann. A distribution free summarization method for affymetrix genechip arrays. *Bioinformatics*, 23(3):321–327, Feb 2007.

[236] Koji Kadota, Yuji Nakai, and Kentaro Shimizu. A weighted average difference method for detecting differentially expressed genes from microarray data. *Algorithms Mol Biol*, 3:8, 2008.

[237] C. Li and W. H. Wong. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc Natl Acad Sci U S A*, 98(1):31–36, Jan 2001.

[238] Stephanie Schneider, Temple Smith, and Ulla Hansen. Scorem: statistical consolidation of redundant expression measures. *Nucleic Acids Res*, Dec 2011.

[239] Matthew N McCall, Benjamin M Bolstad, and Rafael A Irizarry. Frozen robust multiarray analysis (frma). *Biostatistics*, 11(2):242–253, Apr 2010.

[240] S. B. Montgomery, M. Sammeth, M. Gutierrez-Arcelus, R. P. Lach, C. Ingle, J. Nisbett, R. Guigo, and E. T. Dermitzakis. Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature*, 464:773–777, Apr 2010.

[241] John C Marioni, Christopher E Mason, Shrikant M Mane, Matthew Stephens, and Yoav Gilad. Rna-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res*, 18(9):1509–1517, Sep 2008.

[242] John H Malone and Brian Oliver. Microarrays, deep sequencing and the true measure of the transcriptome. *BMC Biol*, 9:34, 2011.

[243] Jerome H. Friedman. Multivariate adaptive regression splines. *The Annals of Statistics*, 19:1–141, 1991.

[244] C.H. Reinsch. Smoothing by spline functions. *Numerische Mathematik*, 10(3):177–183, 1967.

[245] Paul Duffy-Mike Flannigan John Walsh Jerry Melill Michael S. Balshi, A. David Mcguire3. Assessing the response of area burned to changing climate in western boreal north america using a multivariate adaptive regression splines (mars) approach. *Global Change Biology*, 15:578–600, 2008.

[246] Bhattacharya K. Canizares C.A. Zareipour, H. Forecasting the hourly ontario energy price by multivariate adaptive regression splines. *Power Engineering Society General Meeting*, 2006.

[247] Yuehjen E. Shaoc Fei Chenb Shieu-Ming Choua, Tian-Shyug Leeb. Mining the breast cancer pattern using artificial neural networks and multivariateadaptiveregressionsplines. *Expert Systems with Applications*, 27:133–142, 2004.

[248] G. A. Held, G. Grinstein, and Y. Tu. Relationship between gene expression and observed intensities in dna microarrays–a modeling study. *Nucleic Acids Res*, 34(9):e70, 2006.

[249] S. Duan, W. Zhang, W. K. Bleibel, N. J. Cox, and M. E. Dolan. Snpinprobe: a database for filtering out probes in the affymetrix genechip human exon 1.0 st array potentially affected by snps.

[250] C. I. Castillo-Davis, S. L. Mekhedov, D. L. Hartl, E. V. Koonin, and F. A. Kondrashov. Selection for short introns in highly expressed genes. *Nat. Genet.*, 31(4):415–418, Aug 2002.

[251] Christian Stratowa, Vienna, and Austria. *xps: Processing and Analysis of Affymetrix Oligonucleotide Arrays including Exon Arrays, Whole Genome Arrays and Plate Arrays*. R package version 1.14.0.

[252] Rene Brun & Fons Rademakers. Root: An object oriented data analysis framework. *Linux Journal*, 389:81–86, 1997.

[253] S original by Trevor Hastie & Robert Tibshirani. Original R port by Friedrich Leisch, Kurt Hornik, and Brian D. Ripley. *mda: Mixture and flexible discriminant analysis*, 2011. R package version 0.4-2.

[254] Kevin Bullaughey, Claudia I Chavarria, Graham Coop, and Yoav Gilad. Expression quantitative trait loci detected in cell lines are often present in primary tissues. *Hum Mol Genet*, 18(22):4296–4303, Nov 2009.

[255] M. Galassi. *GNU Scientific Library Reference Manual*. ISBN 0954612078, 3 edition.

[256] Matthew Rutter, Brian Saunders, Kay Wilkinson, Steve Rumbles, Gillian Schofield, Michael Kamm, Christopher Williams, Ashley Price, Ian Talbot, and Alastair Forbes. Severity of inflammation is a risk factor for colorectal neoplasia in ulcerative colitis. *Gastroenterology*, 126(2):451–459, Feb 2004.

[257] L. J. Rather. Disturbance of function (functio laesa): the legendary fifth cardinal sign of inflammation, added by galen to the four cardinal signs of celsus. *Bull N Y Acad Med*, 47(3):303–322, Mar 1971.

[258] P S Andersen L Ferrero-Miliani, O H Nielsen and S E Girardin. Chronic inflammation: importance of nod2 and nalp3 in interleukin-1b generation. *Clinical and Experimental Immunology*, 147:227–235, 2007.

[259] Michael Karin and Florian R Greten. Nf-kappab: linking inflammation and immunity to cancer development and progression. *Nat Rev Immunol*, 5(10):749–759, Oct 2005.

[260] Alberto Mantovani, Paola Allavena, Antonio Sica, and Frances Balkwill. Cancer-related inflammation. *Nature*, 454(7203):436–444, Jul 2008.

[261] Conor G Loftus, Edward V Loftus, W. Scott Harmsen, Alan R Zinsmeister, William J Tremaine, L. Joseph Melton, and William J Sandborn. Update on the incidence and prevalence of crohn's disease and ulcerative colitis in olmsted county, minnesota, 1940-2000. *Inflamm Bowel Dis*, 13(3):254–261, Mar 2007.

[262] Eduardo Garcia Vilela, Henrique Osvaldo da Gama Torres, Fabiana Paiva Martins, Maria de Lourdes de Abreu Ferrari, Marcella Menezes Andrade, and Aloísio Sales da Cunha. Evaluation of inflammatory activity in crohn's disease and ulcerative colitis. *World J Gastroenterol*, 18(9):872–881, Mar 2012.

[263] R Wright. A controlled therapeutic trial of various diets in ulcerative colitis. *British medical journal*, 2:138–141, 1965.

[264] Daniel Burger and Simon Travis. Conventional medical management of inflammatory bowel disease. *Gastroenterology*, 140(6):1827–1837.e2, May 2011.

[265] M. Orholm, V. Binder, T. I. Sørensen, L. P. Rasmussen, and K. O. Kyvik. Concordance of inflammatory bowel disease among danish twins. results of a nationwide study. *Scand J Gastroenterol*, 35(10):1075–1081, Oct 2000.

[266] Alexandra I Thompson and Charlie W Lees. Genetics of ulcerative colitis. *Inflamm Bowel Dis*, 17(3):831–848, Mar 2011.

[267] Judy H Cho and Steven R Brant. Recent insights into the genetics of inflammatory bowel disease. *Gastroenterology*, 140(6):1704–1712, May 2011.

[268] Stephen B Hanauer. Inflammatory bowel disease: epidemiology, pathogenesis, and therapeutic opportunities. *Inflamm Bowel Dis*, 12 Suppl 1:S3–S9, Jan 2006.

[269] A. S. Fleisher, M. Esteller, N. Harpaz, A. Leytin, A. Rashid, Y. Xu, J. Liang, O. C. Stine, J. Yin, T. T. Zou, J. M. Abraham, D. Kong, K. T. Wilson, S. P. James, J. G. Herman, and S. J. Meltzer. Microsatellite instability in inflammatory bowel disease-associated neoplastic lesions is associated with hypermethylation and diminished expression of the dna mismatch repair gene, hmlh1. *Cancer Res*, 60(17):4864–4868, Sep 2000.

[270] Steven H Itzkowitz and Xianyang Yio. Inflammation and cancer iv. colorectal cancer in inflammatory bowel disease: the role of inflammation. *Am J Physiol Gastrointest Liver Physiol*, 287(1):G7–17, Jul 2004.

[271] C. N. Bernstein, J. F. Blanchard, E. Kliewer, and A. Wajda. Cancer risk in patients with inflammatory bowel disease: a population-based study. *Cancer*, 91(4):854–862, Feb 2001.

[272] Natalie A Molodecky, Ing Shian Soon, Doreen M Rabi, William A Ghali, Mollie Ferris, Greg Chernoff, Eric I Benchimol, Remo Panaccione, Subrata Ghosh, Herman W Barkema, and Gilaad G Kaplan. Increasing incidence and prevalence of the inflammatory bowel diseases with time, based on systematic review. *Gastroenterology*, 142(1):46–54.e42; quiz e30, Jan 2012.

[273] Elizabeth H Ruder, Adeyinka O Laiyemo, Barry I Graubard, Albert R Hollenbeck, Arthur Schatzkin, and Amanda J Cross. Non-steroidal anti-inflammatory drugs and colorectal cancer risk in a large, prospective cohort. *Am J Gastroenterol*, 106(7):1340–1350, Jul 2011.

[274] Yutaka Kondo. Epigenetic cross-talk between dna methylation and histone modifications in human cancers. *Yonsei Med J*, 50(4):455–463, Aug 2009.

[275] M. van Rijnsoever, F. Grieu, H. Elsaleh, D. Joseph, and B. Iacopetta. Characterisation of colorectal cancers showing hypermethylation at multiple cpg islands. *Gut*, 51(6):797–802, Dec 2002.

[276] Hui Ying Zhang, Stuart Jon Spechler, and Rhonda F Souza. Esophageal adenocarcinoma arising in barrett esophagus. *Cancer Lett*, 275(2):170–177, Mar 2009.

[277] Hector Alvarez, Joanna Opalinska, Li Zhou, Davendra Sohal, Melissa J Fazzari, Yiting Yu, Christina Montagna, Elizabeth A Montgomery, Marcia Canto, Kerry B Dunbar, Jean Wang, Juan Carlos Roa, Yongkai Mo, Tushar Bhagat, K. H. Ramesh, Linda Cannizzaro, J. Mollenhauer, Reid F Thompson, Masako Suzuki, Stephen J Meltzer, Stephen Meltzer, Ari Melnick, John M Greally, Anirban Maitra, and Amit Verma. Widespread hypomethylation occurs early and synergizes with gene amplification during esophageal carcinogenesis. *PLoS Genet*, 7(3):e1001356, Mar 2011.

[278] C. J. Hsieh, B. Klump, K. Holzmann, F. Borchard, M. Gregor, and R. Porschen. Hypermethylation of the p16ink4a promoter in colectomy specimens of patients with long-standing and extensive ulcerative colitis. *Cancer Res*, 58(17):3942–3945, Sep 1998.

[279] Fang-Yu Wang, Tomiyasu Arisawa, Tomomitsu Tahara, Kazuya Takahama, Makoto Watanabe, Ichiro Hirata, and Hiroshi Nakano. Aberrant dna methylation in ulcerative colitis without neoplasia. *Hepatogastroenterology*, 55(81):62–65, 2008.

[280] Tomohiko Moriyama, Takayuki Matsumoto, Shotaro Nakamura, Yukihiko Jo, Ryuichi Mibu, Takashi Yao, and Mitsuo Iida. Hypermethylation of p14 (arf) may be predictive of colitic cancer in patients with ulcerative colitis. *Dis Colon Rectum*, 50(9):1384–1392, Sep 2007.

[281] Keiichi Tominaga, Shigehiko Fujii, Kenichiroh Mukawa, Mikio Fujita, Kazuhito Ichikawa, Shigeki Tomita, Yasuo Imai, Kazunari Kanke, Yuko Ono, Akira Terano, Hideyuki Hiraishi, and Takahiro Fujimori. Prediction of colorectal neoplasia by quantitative methylation analysis of estrogen receptor gene in nonneoplastic epithelium from patients with ulcerative colitis. *Clin Cancer Res*, 11(24 Pt 1):8880–8885, Dec 2005.

[282] Mashaal Dhir, Elizabeth A Montgomery, Sabine C Glöckner, Kornel E Schuebel, Craig M Hooker, James G Herman, Stephen B Baylin, Susan L Gearhart, and Nita Ahuja. Epigenetic regulation of wnt signaling pathway genes in inflammatory bowel disease (ibd) associated neoplasia. *J Gastrointest Surg*, 12(10):1745–1753, Oct 2008.

[283] Tomomitsu Tahara, Tomoyuki Shibata, Masakatsu Nakamura, Hiromi Yamashita, Daisuke Yoshioka, Masaaki Okubo, Naoko Maruyama, Toshiaki Kamano, Yoshio Kamiya, Hiroshi Fujita, Yoshihito Nakagawa, Mitsuo Nagasaka, Masami Iwata, Kazuya Takahama, Makoto Watanabe, Hiroshi Nakano, Ichiro Hirata, and Tomiyasu Arisawa. Promoter methylation of

protease-activated receptor (par2) is associated with severe clinical phenotypes of ulcerative colitis (uc). *Clin Exp Med*, 9(2):125–130, Jun 2009.

[284] Yi Li, Colin de Haar, Min Chen, Jasper Deuring, Monique M Gerrits, Ron Smits, Bing Xia, Ernst J Kuipers, and C. Janneke van der Woude. Disease-related expression of the il6/stat3/socs3 signalling pathway in ulcerative colitis and ulcerative colitis-related carcinogenesis. *Gut*, 59(2):227–235, Feb 2010.

[285] Ramesh P Arasaradnam, Kevin Khoo, Mike Bradburn, John C Mathers, and Seamus B Kelly. Dna methylation of esr-1 and n-33 in colorectal mucosa of patients with ulcerative colitis (uc). *Epigenetics*, 5(5):422–426, Jul 2010.

[286] Shunsuke Saito, Jun Kato, Sakiko Hiraoka, Joichiro Horii, Hideyuki Suzuki, Reiji Higashi, Eisuke Kaji, Yoshitaka Kondo, and Kazuhide Yamamoto. Dna methylation of colon mucosa in ulcerative colitis patients: correlation with inflammatory status. *Inflamm Bowel Dis*, 17(9):1955–1965, Sep 2011.

[287] Tomomitsu Tahara, Tomoyuki Shibata, Masakatsu Nakamura, Hiromi Yamashita, Daisuke Yoshioka, Masaaki Okubo, Naoko Maruyama, Toshiaki Kamano, Yoshio Kamiya, Yoshihito Nakagawa, Hiroshi Fujita, Mitsuo Nagasaka, Masami Iwata, Kazuya Takahama, Makoto Watanabe, Ichiro Hirata, and Tomiyasu Arisawa. Effect of mdr1 gene promoter methylation in patients with ulcerative colitis. *Int J Mol Med*, 23(4):521–527, Apr 2009.

[288] Ian M Wilson, Jonathan J Davies, Michael Weber, Carolyn J Brown, Carlos E Alvarez, Calum MacAulay, Dirk Schübeler, and Wan L Lam. Epigenomics: mapping the methylome. *Cell Cycle*, 5(2):155–158, Jan 2006.

[289] Michael Weber, Jonathan J Davies, David Wittig, Edward J Oakeley, Michael Haase, Wan L Lam, and Dirk Schübeler. Chromosome-wide and promoter-specific analyses identify sites of differential dna methylation in normal and transformed human cells. *Nat Genet*, 37(8):853–862, Aug 2005.

[290] Nina Palmke, Diana Santacruz, and Jorn Walter. Comprehensive analysis of dna-methylation in mammalian tissues using medip-chip. *Methods*, 53(2):175–184, Feb 2011.

[291] Yi Shu, Bing Wang, Ji Wang, Jian-Ming Wang, and Sheng-Quan Zou. Identification of methylation profile of hox genes in extrahepatic cholangiocarcinoma. *World J Gastroenterol*, 17(29):3407–3419, Aug 2011.

[292] Audrey Vincent, Noriyuki Omura, Seung-Mo Hong, Andrew Jaffe, James Eshleman, and Michael Goggins. Genome-wide analysis of promoter methylation associated with gene expression profile in pancreatic adenocarcinoma. *Clin Cancer Res*, 17(13):4341–4354, Jul 2011.

[293] Joern Toedling, Oleg Skylar, Oleg Sklyar, Tammo Krueger, Jenny J Fischer, Silke Sperling, and Wolfgang Huber. Ringo–an r/bioconductor package for analyzing chip-chip readouts. *BMC Bioinformatics*, 8:221, 2007.

[294] Julie Borgel, Sylvain Guibert, Yufeng Li, Hatsune Chiba, Dirk Schübeler, Hiroyuki Sasaki, Thierry Forné, and Michael Weber. Targets and dynamics of promoter dna methylation during early mouse development. *Nat Genet*, 42(12):1093–1100, Dec 2010.

[295] Thomas A Down, Vardhman K Rakyan, Daniel J Turner, Paul Flicek, Heng Li, Eugene Kulesha, Stefan Graf, Nathan Johnson, Javier Herrero, Eleni M Tomazou, Natalie P Thorne, Liselotte Backdahl, Marlis Herberth, Kevin L Howe, David K Jackson, Marcos M Miretti, John C Marioni, Ewan Birney, Tim J P Hubbard, Richard Durbin, Simon Tavare, and Stephan Beck. A bayesian deconvolution strategy for immunoprecipitation-based dna methylome analysis. *Nat Biotechnol*, 26(7):779–785, Jul 2008.

[296] Mehregan Movassagh, Mun-Kit Choy, David A Knowles, Lina Cordeddu, Syed Haider, Thomas Down, Lee Siggens, Ana Vujic, Ilenia Simeoni, Chris Penkett, Martin Goddard, Pietro Lio, Martin R Bennett, and Roger S-Y Foo. Distinct epigenomic features in end-stage failing human hearts. *Circulation*, 124(22):2411–2422, Nov 2011.

[297] Ken Kron, Vaijayanti Pethe, Laurent Briollais, Bekim Sadikovic, Hilmi Ozcelik, Alia Sunderji, Vasundara Venkateswaran, Jehonathan Pinthus, Neil Fleshner, Theodorus van der Kwast, and Bharati Bapat. Discovery of novel hypermethylated genes in prostate cancer using genomic cpg island microarrays. *PLoS One*, 4(3):e4830, 2009.

[298] E. Helman, K. Naxerova, and I. S. Kohane. Dna hypermethylation in lung cancer is targeted at differentiation-associated genes. *Oncogene*, 31(9):1181–1188, Mar 2012.

[299] Bogdan C Paun, Debra Kukuruga, Zhe Jin, Yuriko Mori, Yulan Cheng, Mark Duncan, Sanford A Stass, Elizabeth Montgomery, David Hutcheon, and Stephen J Meltzer. Relation between normal rectal methylation, smoking status, and the presence or absence of colorectal adenomas. *Cancer*, 116(19):4495–4501, Oct 2010.

[300] F. I. Milagro, J. Campion, D. F. Garcia-Diaz, E. Goyenechea, L. Paternain, and J. A. Martinez. High fat diet-induced obesity modifies the methylation pattern of leptin promoter in rats. *J Physiol Biochem*, 65(1):1–9, Mar 2009.

[301] Oliver Stegle, Leopold Parts, Richard Durbin, and John Winn. A bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eqtl studies. *PLoS Comput Biol*, 6(5):e1000770, May 2010.

[302] L. Shi, L. H. Reid, W. D. Jones, R. Shippy, J. A. Warrington, S. C. Baker, P. J. Collins, F. de Longueville, E. S. Kawasaki, K. Y. Lee, Y. Luo, Y. A. Sun, J. C. Willey, R. A. Setterquist, G. M. Fischer, W. Tong, Y. P. Dragan, D. J. Dix, F. W. Frueh, F. M. Goodsaid, D. Herman, R. V. Jensen, C. D. Johnson, E. K. Lobenhofer, R. K. Puri, U. Schrf, J. Thierry-Mieg, C. Wang, M. Wilson, P. K. Wolber, L. Zhang, S. Amur, W. Bao, C. C. Barbacioru, A. B. Lucas, V. Bertholet, C. Boysen, B. Bromley, D. Brown, A. Brunner, R. Canales, X. M. Cao, T. A. Cebula, J. J. Chen, J. Cheng, T. M. Chu, E. Chudin, J. Corson, J. C. Corton, L. J. Croner, C. Davies, T. S. Davison, G. Delenstarr, X. Deng, D. Dorris, A. C. Eklund, X. H. Fan, H. Fang, S. Fulmer-Smentek, J. C. Fuscoe, K. Gallagher, W. Ge, L. Guo, X. Guo, J. Hager, P. K. Haje, J. Han, T. Han, H. C. Harbottle, S. C. Harris, E. Hatchwell, C. A. Hauser, S. Hester, H. Hong, P. Hurban, S. A. Jackson, H. Ji, C. R. Knight, W. P. Kuo, J. E. LeClerc, S. Levy, Q. Z. Li, C. Liu, Y. Liu, M. J. Lombardi, Y. Ma, S. R. Magnuson, B. Maqsodi, T. McDaniel, N. Mei, O. Myklebost, B. Ning, N. Novoradovskaya, M. S. Orr, T. W. Osborn, A. Papallo, T. A. Patterson, R. G. Perkins, E. H. Peters, R. Peterson, K. L. Philips, P. S. Pine, L. Pusztai, F. Qian, H. Ren, M. Rosen, B. A. Rosenzweig, R. R. Samaha, M. Schena, G. P. Schroth, S. Shchegrova, D. D. Smith, F. Staedtler, Z. Su, H. Sun, Z. Szallasi, Z. Tezak, D. Thierry-Mieg, K. L. Thompson, I. Tikhonova, Y. Turpaz, B. Vallanat, C. Van, S. J. Walker, S. J. Wang, Y. Wang, R. Wolfinger, A. Wong, J. Wu, C. Xiao, Q. Xie, J. Xu, W. Yang, L. Zhang, S. Zhong, Y. Zong, and W. Slikker. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat. Biotechnol.*, 24(9):1151–1161, Sep 2006.

[303] Richard Bourgon, Robert Gentleman, and Wolfgang Huber. Independent filtering increases detection power for high-throughput experiments. *Proc Natl Acad Sci U S A*, 107(21):9546–9551, May 2010.

[304] Akiko Doi, In-Hyun Park, Bo Wen, Peter Murakami, Martin J Aryee, Rafael Irizarry, Brian Herb, Christine Ladd-Acosta, Junsung Rho, Sabine Loewer,

Justine Miller, Thorsten Schlaeger, George Q Daley, and Andrew P Feinberg. Differential methylation of tissue- and cancer-specific cpg island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts. *Nat Genet*, 41(12):1350–1353, Dec 2009.

[305] R. A. Irizarry, C. Ladd-Acosta, B. Wen, Z. Wu, C. Montano, P. Onyango, H. Cui, K. Gabo, M. Rongione, M. Webster, H. Ji, J. B. Potash, S. Sabunciyan, and A. P. Feinberg. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat. Genet.*, 41(2):178–186, Feb 2009.

[306] Jose Roman-Gomez, Xabier Agirre, Antonio Jimenez-Velasco, Victor Arqueros, Amaia Vilas-Zornoza, Paula Rodriguez-Otero, Inaki Martin-Subero, Leire Garate, Lucia Cordeu, Edurne San Jose-Eneriz, Vanesa Martin, Juan Antonio Castillejo, Eva Bandrés, María Jose Calasanz, Reiner Siebert, Anabel Heiniger, Antonio Torres, and Felipe Prosper. Epigenetic regulation of micrornas in acute lymphoblastic leukemia. *J Clin Oncol*, 27(8):1316–1322, Mar 2009.

[307] Tibor A Rauch, Xueyan Zhong, Xiwei Wu, Melody Wang, Kemp H Kernstine, Zunde Wang, Arthur D Riggs, and Gerd P Pfeifer. High-resolution mapping of dna hypermethylation and hypomethylation in lung cancer. *Proc Natl Acad Sci U S A*, 105(1):252–257, Jan 2008.

[308] G. M. Bernardo, G. Bebek, C. L. Ginther, S. T. Sizemore, K. L. Lozada, J. D. Miedler, L. A. Anderson, A. K. Godwin, F. W. Abdul-Karim, D. J. Slamon, and R. A. Keri. Foxa1 represses the molecular phenotype of basal breast cancer cells. *Oncogene*, Mar 2012.

[309] Thomas Dunwell, Luke Hesson, Tibor A Rauch, Lihui Wang, Richard E Clark, Ashraf Dallol, Dean Gentle, Daniel Catchpoole, Eamonn R Maher, Gerd P Pfeifer, and Farida Latif. A genome-wide screen identifies frequently methylated genes in haematological and epithelial cancers. *Mol Cancer*, 9:44, 2010.

[310] Serge Saxonov, Paul Berg, and Douglas L Brutlag. A genome-wide analysis of cpg dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc Natl Acad Sci U S A*, 103(5):1412–1417, Jan 2006.

[311] Ian Campbell. Chi-squared and fisher-irwin tests of two-by-two tables with small sample recommendations. *Stat Med*, 26(19):3661–3675, Aug 2007.

[312] Fabio Mohn, Michael Weber, Dirk Schübeler, and Tim-Christoph Roloff. Methylated dna immunoprecipitation (medip). *Methods Mol Biol*, 507:55–64, 2009.

[313] Hannah Schneider, Annika Braun, Joachim Füllekrug, Wolfgang Stremmel, and Robert Ehehalt. Lipid based therapy for ulcerative colitis-modulation of intestinal mucus membrane phospholipids as a tool to influence inflammation. *Int J Mol Sci*, 11(10):4149–4164, 2010.

[314] A L Henneberry and C R McMaster. Cloning and expression of a human choline/ethanolaminephosphotransferase: synthesis of phosphatidylcholine and phosphatidylethanolamine. *Biochem. J.*, 339:291–298, 1999.

[315] Carsten Büning, Nora Geissler, Matthias Prager, Andreas Sturm, Daniel C Baumgart, Janine Büttner, Sabine Bühner, Verena Haas, and Herbert Lochs. Increased small intestinal permeability in ulcerative colitis: Rather genetic than environmental and a risk factor for extensive disease? *Inflamm Bowel Dis*, Feb 2012.

[316] J. G. Jansen, H. Vrieling, A. A. van Zeeland, and G. R. Mohn. The gene encoding hypoxanthine-guanine phosphoribosyltransferase as target for mutational analysis: Pcr cloning and sequencing of the cdna from the rat. *Mutat Res*, 266(2):105–116, Apr 1992.

[317] M. Suzuki, Q. Jing, D. Lia, M. Pascual, A. McLellan, and J. M. Greally. Optimized design and data analysis of tag-based cytosine methylation assays. *Genome Biol.*, 11(4):R36, 2010.

[318] Hirotugu Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723, 1974.

[319] David P Strachan, Alicja R Rudnicka, Chris Power, Peter Shepherd, Elizabeth Fuller, Adrian Davis, Ian Gibb, Meena Kumari, Ann Rumley, Gary J Macfarlane, Jugnoo Rahi, Bryan Rodgers, and Stephen Stansfeld. Lifecourse influences on health among british adults: effects of region of residence in childhood and adulthood. *Int J Epidemiol*, 36(3):522–531, Jun 2007.

[320] Francisco J Ortega, José M Moreno-Navarrete, Gerard Pardo, Monica Sabater, Manuela Hummel, Anna Ferrer, Jose I Rodriguez-Hermosa, Bartomeu Ruiz, Wifredo Ricart, Belen Peral, and José M Fernández-Real. Mirna expression profile of human subcutaneous adipose and during adipocyte differentiation. *PLoS One*, 5(2):e9022, 2010.

[321] Nora Klöting, Susan Berthold, Peter Kovacs, Michael R Schön, Mathias Fasshauer, Karen Ruschke, Michael Stumvoll, and Matthias Blüher. Microrna expression in human omental and subcutaneous adipose tissue. *PLoS One*, 4(3):e4699, 2009.

[322] Alicia Oshlack and Matthew J Wakefield. Transcript length bias in rna-seq data confounds systems biology. *Biol Direct*, 4:14, 2009.

Appendix A - The regulatory effect of miRNAs is a heritable genetic trait in humans

miRNA name	Subtracted mean RE-score p-val	miRNA expression p-val	No. of targets genes
hsa-miR-142-5p	581.25×10^{-12}	0.39	564
hsa-miR-181a	101.34×10^{-12}	0.30	1039
hsa-miR-183	739.98×10^{-12}	0.54	309
hsa-miR-193a-5p	8.59×10^{-03}	0.32	91
hsa-miR-221	3.07×10^{-12}	0.80	319
hsa-miR-222	2.92×10^{-12}	0.60	319
hsa-miR-30a	5.79×10^{-12}	0.79	1241
hsa-miR-339-5p	73.37×10^{-06}	0.59	103
hsa-miR-551b	833.80×10^{-03}	0.20	5
hsa-miR-574-3p	4.54×10^{-03}	0.99	6
hsa-miR-625	213.80×10^{-06}	0.28	70
hsa-miR-642	61.10×10^{-06}	0.55	172
hsa-miR-665	37.72×10^{-03}	0.20	220
hsa-miR-933	228.42×10^{-06}	0.19	7

Table 6.1: P-values for the associations between the RE-scores of 14 highly variable miRNAs and "subtracted mean RE-score" or the actual expression level of the miRNA (CEU).

miRNA name	Subtracted mean RE-score p-val	miRNA expression p-val	No. of targets genes
hsa-miR-151-5p	96.03×10^{-03}	0.32	6
hsa-miR-17	428.63×10^{-27}	0.06	1114
hsa-miR-27a	500.38×10^{-30}	0.72	1004
hsa-miR-30c	42.39×10^{-27}	0.10	1241
hsa-miR-30e	81.90×10^{-27}	0.06	1241
hsa-miR-340	24.33×10^{-18}	0.65	1094
hsa-miR-361-5p	358.31×10^{-15}	0.88	142
hsa-miR-628-3p	1.40×10^{-18}	0.18	73
hsa-miR-642	608.53×10^{-15}	0.73	172
hsa-miR-768-3p	126.45×10^{-18}	0.07	297
hsa-miR-768-5p	6.27×10^{-12}	0.40	141
hsa-miR-9	73.55×10^{-27}	0.01	1037
hsa-miR-933	26.98×10^{-03}	0.18	7

Table 6.2: P-values for the associations between the RE-scores of 13 highly variable miRNAs and "subtracted mean RE-score" or the actual expression level of the miRNA (YRI).

hsa-miR-30c	hsa-miR-223	hsa-miR-30c-2*	miRPlus_17869	hsa-miR-18a
hsa-miR-542-3p	hsa-miR-519e	hsa-miR-125a-5p	hsa-miR-186	hsa-miR-101
hsa-miR-30a	miRPlus_42487	hsa-miR-921	hsa-miR-9*	hsa-miR-29a
hsa-miR-30e	hsa-miR-19a	hsa-miR-423-3p	hsa-miR-886-3p	hsa-miR-198
hsa-miR-625*	hsa-miR-503	hsa-miR-17	hsa-miR-20b*	hsa-miR-744
hsa-miR-140-3p	hsa-miR-887	hsa-miR-29b	hsa-miR-525-5p	hsa-miR-191
hsa-miR-634	hsa-miR-424	hsa-miR-550	hsa-miR-7	hsa-miR-519d
hsa-miR-30b	hsa-miR-335	hsa-miR-620	hsa-miR-21*	hsa-miR-298
hsa-miR-342-3p	miRPlus_42793	hsa-miR-574-3p	hsa-miR-371-5p	hsa-miR-361-5p
hsa-miR-551b	miRPlus_42745	hsa-miR-33a	hsa-miR-140-5p	hsa-miR-331-3p
hsa-miR-939	hsa-miR-29b-1*	hsa-miR-629*	hsa-miR-20a*	hsa-miR-193b
hsa-miR-302d*	hsa-miR-92b	hsa-miR-374b	miRPlus_42780	hsa-miR-765
hsa-miR-106b	hsa-miR-933	hsa-miR-18b	hsa-let-7a	hsa-miR-518c*
miRPlus_42526	hsa-miR-487b	hsa-miR-516a-5p	hsa-miR-923	hsa-miR-106a
hsa-miR-16	hsa-miR-600	hsa-miR-23a	hsa-miR-665	hsa-miR-590-5p
hsa-miR-34b	hsa-miR-29a*	miRPlus_42856	hsa-miR-185	hsa-miR-549
miRPlus_17858	hsa-let-7f	hsa-miR-141	hsa-miR-181b	hsa-miR-155*
hsa-miR-642	hsa-let-7e	hsa-miR-339-5p	hsa-miR-658	hsa-miR-184
hsa-miR-620	hsa-miR-491-3p	hsa-miR-637	hsa-miR-766	hsa-miR-98
hsa-miR-32*	hsa-miR-29c	hsa-miR-24	hsa-miR-210	hsa-miR-25
miRPlus_28431	hsa-miR-20a	hsa-miR-27a	hsa-miR-27b	hsa-miR-378
hsa-miR-92a	hsa-miR-151-5p	hsa-miR-671-5p	hsa-miR-374b*	hsa-miR-142-5p
hsa-miR-628-3p	miRPlus_17848	hsa-miR-320a	hsa-miR-148b	hsa-miR-155
hsa-miR-130a	hsa-miR-301a	hsa-miR-221	hsa-miR-30b*	hsa-miR-720
hsa-miR-768-5p	hsa-miR-105	hsa-miR-103	hsa-miR-142-3p	hsa-miR-22
hsa-miR-485-3p	hsa-miR-23b	hsa-let-7d	hsa-miR-21	hsa-miR-28-5p
hsa-miR-30e*	hsa-let-7i	hsa-miR-583	hsa-miR-181a	hsa-miR-425
hsa-miR-183	miRPlus_27560	hsa-miR-10a	hsa-miR-146b-5p	hsa-miR-576-3p
hsa-miR-106b*	hsa-miR-185*	hsa-miR-625	hsa-miR-574-5p	hsa-miR-520d-5p
hsa-miR-768-3p	hsa-miR-300	hsa-miR-183*	hsa-miR-93	hsa-miR-19b
hsa-miR-361-3p	hsa-miR-374a	hsa-miR-185	hsa-miR-130b	hsa-miR-25*
hsa-miR-886-5p	hsa-let-7c	miRPlus_17952	hsa-miR-20b	hsa-miR-138
hsa-miR-30d	hsa-miR-193a-5p	hsa-miR-129-5p	hsa-miR-1	hsa-miR-518a-5p/hsa-miR-527
hsa-miR-149*	hsa-miR-510	hsa-miR-148a	hsa-miR-9	hsa-miR-129*
hsa-miR-365	hsa-let-7g	hsa-miR-15a	hsa-miR-34a	hsa-miR-26a
hsa-miR-483-5p	hsa-miR-15b	hsa-miR-107	hsa-miR-885-5p	hsa-miR-146a
miRPlus_42521	hsa-miR-668	hsa-miR-340	hsa-miR-150*	hsa-miR-494
hsa-miR-513a-5p	hsa-miR-363*	hsa-miR-222	hsa-miR-196a*	hsa-miR-17*
hsa-miR-26b	hsa-miR-340*	hsa-miR-675	hsa-miR-519e*	hsa-miR-423-5p
hsa-miR-363	hsa-miR-193b*	hsa-miR-32	hsa-miR-132	hsa-miR-138-1*
				hsa-miR-524-5p

Table 6.3: The complete list of 201 miRNAs whose expression was tested for association with rs17409624.

Appendix B - Improving gene expression estimates from DNA-microarrays using machine learning

		Core	Extended	Full
All Probes	Unadjusted	0.062	0.059	0.058
	With DABG	0.071	068	0.067
	With Zeroing	0.062	0.060	0.060
	With DABG and Zeroing	0.072	0.069	0.068
Only Positive Probes	Unadjusted	0.0778	0.0709	0.0694
	With DABG	0.0886	0.0845	0.0830
	With Zeroing	0.0786	0.0717	0.0701
	With DABG and Zeroing	0.0897	0.0855	0.0840
Only Top 50% Positive Probes	Unadjusted	0.0973	0.0874	0.0856
	With DABG	0.1095	0.1025	0.1006
	With Zeroing	0.0980	0.0879	0.0861
	With DABG and Zeroing	**0.1104**	0.1033	0.1014

Table 6.4: Across sample Spearman correlations (APT=0.154).

		Core	Extended	Full
All Probes	Unadjusted	0.069	0.068	0.067
	With DABG	0.075	0.073	0.073
	With Zeroing	0.067	0.065	0.065
	With DABG and Zeroing	0.080	0.078	0.078
Only Positive Probes	Unadjusted	0.0848	0.0783	0.0771
	With DABG	0.0947	0.0903	0.0892
	With Zeroing	0.0875	0.0809	0.0798
	With DABG and Zeroing	0.1002	0.0957	0.0945
Only Top 50% Positive Probes	Unadjusted	0.1047	0.0950	0.0932
	With DABG	0.1166	0.1097	0.1084
	With Zeroing	0.1073	0.0972	0.0953
	With DABG and Zeroing	**0.1217**	0.1145	0.1131

Table 6.5: Across sample Pearson correlations (ATP=0.163).

		Core	Extended	Full
All Probes	Unadjusted	0.909	0.904	0.903
	With DABG	0.918	0.914	0.913
	With Zeroing	0.918	0.914	0.913
	With DABG and Zeroing	0.923	0.920	0.919
Only Positive Probes	Regular	0.922	0.915	0.914
	With DABG	0.929	0.922	0.921
	With Zeroing	0.925	0.918	0.917
	With DABG and Zeroing	0.930	0.924	0.923
Only Top 50% Positive Probes	Regular	0.930	0.926	0.925
	With DABG	0.935	0.932	0.931
	With Zeroing	0.932	0.928	0.928
	With DABG and Zeroing	**0.936**	0.933	0.932

Table 6.6: Within sample Spearman correlations (ATP=0.845).

		Core	Extended	Full
All Probes	Unadjusted	0.923	0.879	0.878
	DABG	0.926	0.916	0.920
	With Zeroing	0.924	0.883	0.881
	With DABG and Zeroing	**0.927**	0.917	0.920
Only Positive Probes	Unadjusted	0.870	0.853	0.870
	With DABG	0.875	0.861	0.878
	With Zeroing	0.871	0.853	0.870
	With DABG and Zeroing	0.875	0.861	0.879
Only Top 50% Positive Probes	Unadjusted	0.904	0.886	0.883
	With DABG	0.906	0.895	0.893
	With Zeroing	0.905	0.887	0.883
	With DABG and Zeroing	0.907	0.895	0.894

Table 6.7: Within sample Pearson correlations (ATP=0.196).

Appendix C - CpG island hypermethylation is associated with ulcerative colitis

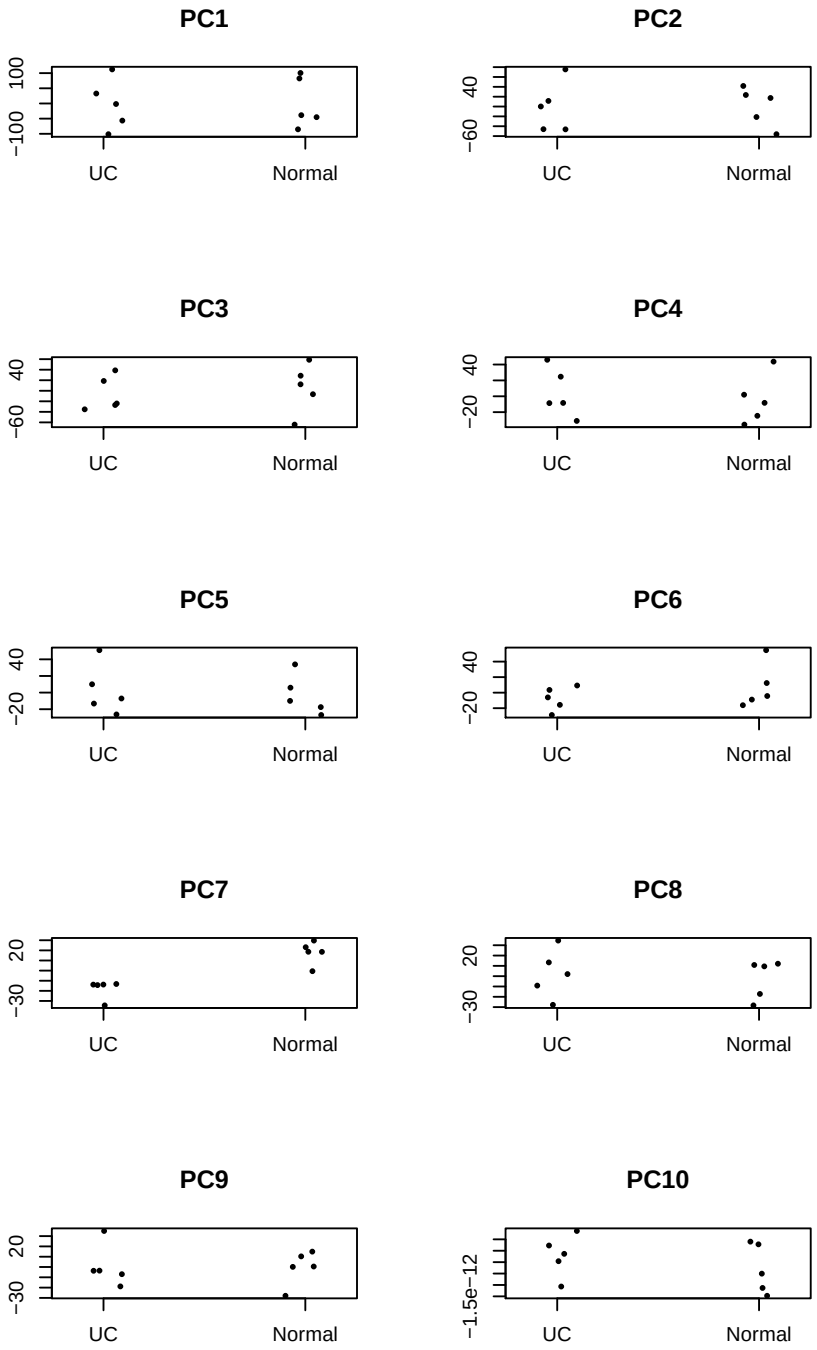

Figure 6.1: Clustering of UC data on principle components 1 to 10.

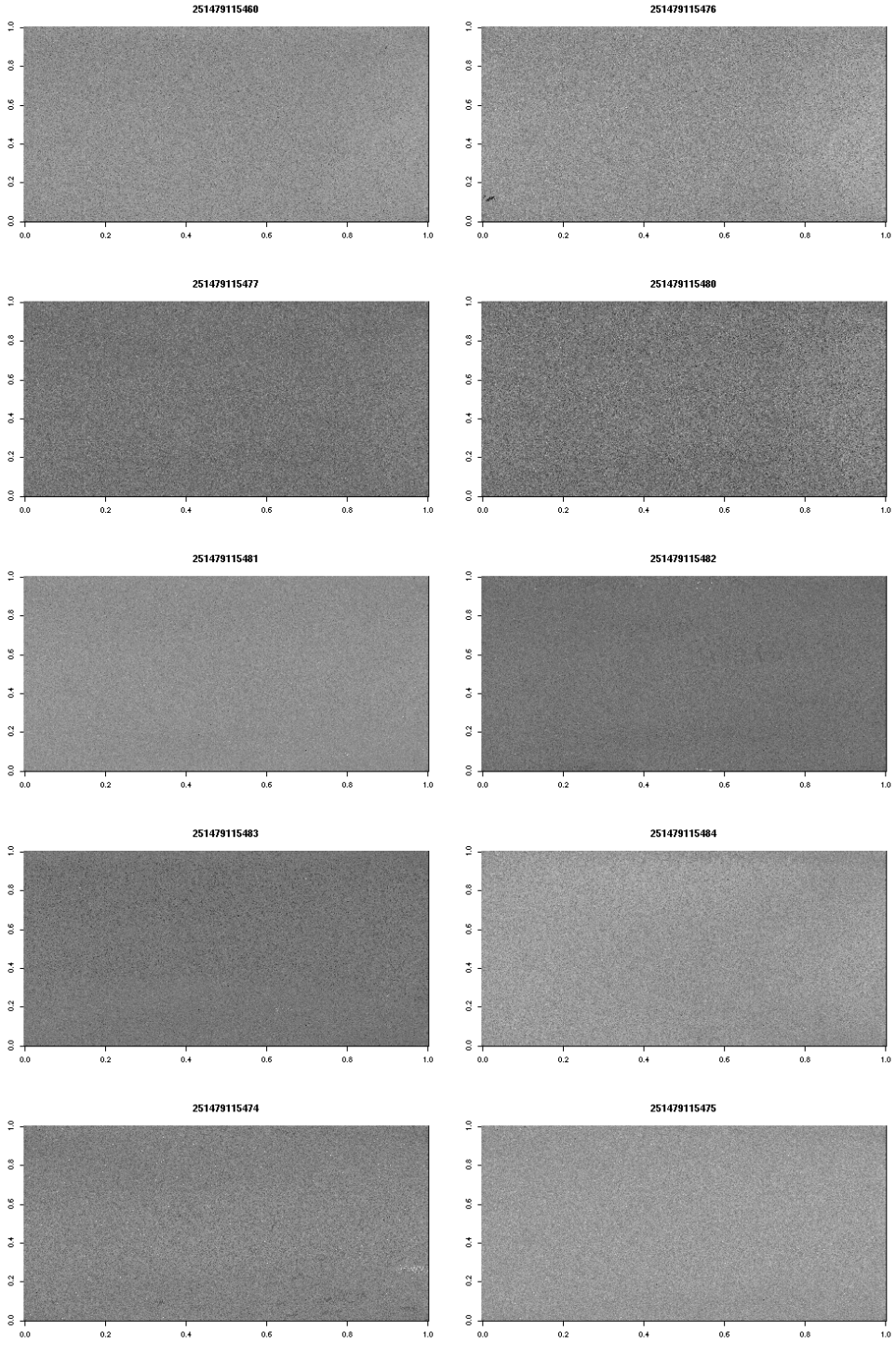

Figure 6.2: Pseudo array images of log intensity ratios, for the 10 Agilent Human CpG Island microarrays.

Appendix D - Severe bias in gene set analysis applied to high-throughput methylation data

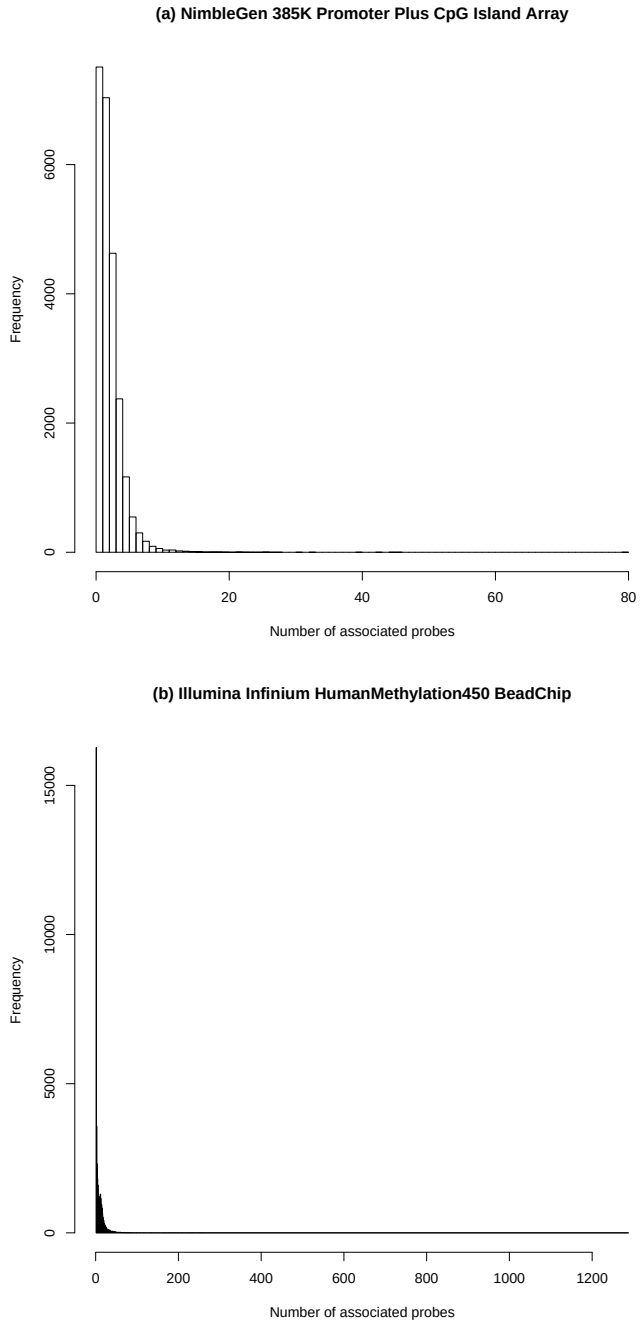

Figure 6.3: A histogram illustrating the distribution of the numbers of microarray probes associated with each gene on (a) the NimbleGen Human DNA Methylation 385K Promoter Plus CpG Island Array and (b) the Illumina Infinium HumanMethylation450 BeadChip.

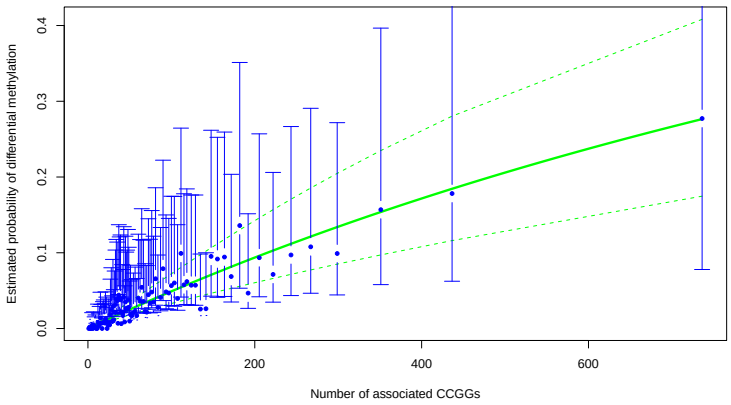

Figure 6.4: Fit of the logistic regression to the HELP-seq data (for gene body hypermethylation).

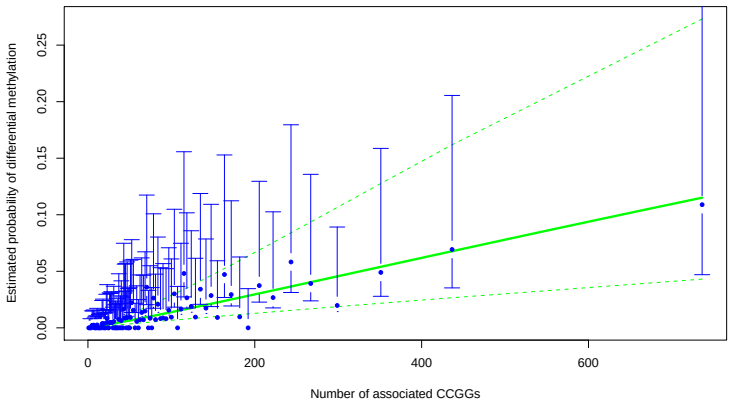

Figure 6.5: Fit of the logistic regression to the HELP-seq data (for gene body hypomethylation).

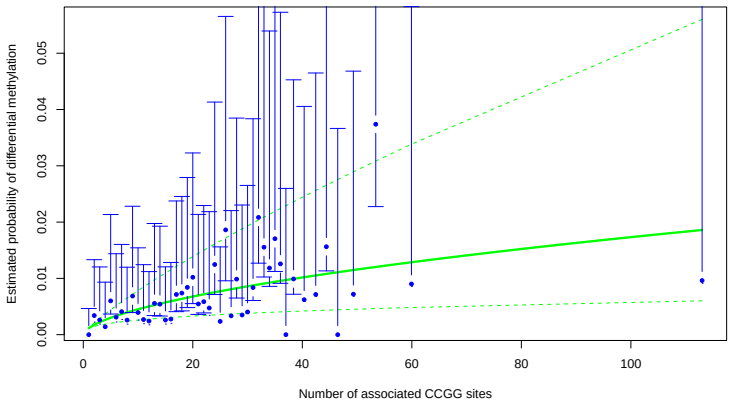

Figure 6.6: Fit of the logistic regression to the HELP-seq data (for promoter hypermethylation).

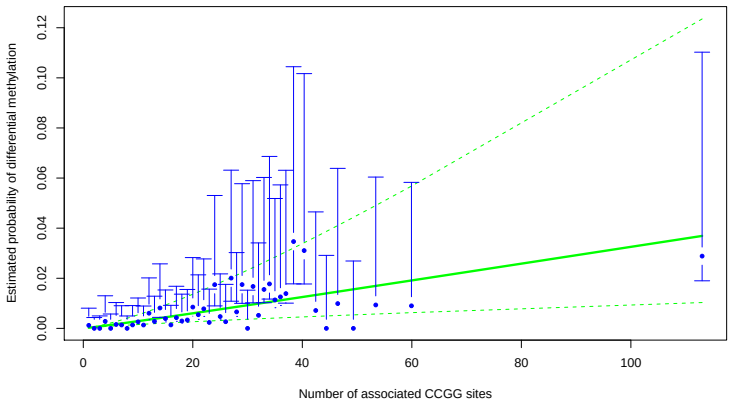

Figure 6.7: Fit of the logistic regression to the HELP-seq data (for promoter hypomethylation).

GOBPID	Term	Count	Size	Expected Count	P-value	Corrected p-value	FDR	Corrected FDR
GO:0007017	Microtubule-Based Process	16	255	6.51	0.00086	0.0014	0.22	0.53
GO:0006913	Nucleocytoplasmic Transport	12	192	4.90	0.0038	0.0028	0.24	0.53
GO:0051169	Nuclear Transport	12	193	4.93	0.0039	0.0029	0.24	0.53
GO:0034613	Cellular Protein Localization	20	434	11.08	0.01	0.01	0.31	0.59
GO:0070727	Cellular Macromolecule Localization	20	437	11.16	0.01	0.01	0.31	0.59
GO:0006886	Intracellular Protein Transport	18	396	10.11	0.01	0.01	0.42	0.69
GO:0000226	Microtubule Cytoskeleton Organization	9	153	3.91	0.02	0.01	0.46	0.69
GO:0034504	Protein Localization In Nucleus	8	124	3.17	0.01	0.01	0.43	0.69
GO:0019953	Sexual Reproduction	16	393	10.03	0.04	0.01	0.64	0.69
GO:0033365	Protein Localization In Organelle	9	167	4.26	0.03	0.01	0.51	0.69

Table 6.8: GSA results from the HELP-seq analysis - gene body hypermethylation.

GOBPID	Term	Count	Size	Expected Count	P-value	Corrected p-value	FDR	Corrected FDR
GO:0048562	Embryonic Organ Morphogenesis	4	107	0.82	0.01	0.0041	0.23	0.72
GO:0042692	Muscle Cell Differentiation	5	132	1.01	0.0034	0.01	0.16	0.72
GO:0051960	Regulation Of Nervous System Development	6	166	1.27	0.0017	0.01	0.16	0.72
GO:0048568	Embryonic Organ Development	4	140	1.07	0.02	0.01	0.32	0.72
GO:0050877	Neurological System Process	14	763	5.84	0.0019	0.01	0.16	0.72
GO:0045595	Regulation Of Cell Differentiation	9	434	3.32	0.01	0.01	0.20	0.72
GO:0060284	Regulation Of Cell Development	6	178	1.36	0.0024	0.01	0.16	0.72
GO:0007600	Sensory Perception	7	363	2.78	0.02	0.02	0.32	0.85
GO:0050767	Regulation Of Neurogenesis	5	146	1.12	0.01	0.02	0.19	0.88
GO:0061061	Muscle Structure Development	6	261	2.00	0.01	0.03	0.29	0.88

Table 6.9: GSA results from the HELP-seq analysis - gene body hypomethylation.

GOBPID	Term	Count	Size	Expected Count	P-value	Corrected p-value	FDR	Corrected FDR
GO:0009308	Amine Metabolic Process	7	391	2.26	0.01	0.0049	0.96	0.79
GO:0006575	Cellular Amino Acid Derivative Metabolic Process	4	159	0.92	0.01	0.01	0.96	0.79
GO:0032269	Negative Regulation Of Cellular Protein Metabolic	4	171	0.99	0.02	0.01	0.96	0.79
GO:0051248	Negative Regulation Of Protein Metabolic Process	4	177	1.02	0.02	0.01	0.96	0.79
GO:0044283	Small Molecule Biosynthetic Process	5	305	1.76	0.03	0.02	0.96	0.79
GO:0044106	Cellular Amine Metabolic Process	5	302	1.75	0.03	0.02	0.96	0.79
GO:0045926	Negative Regulation Of Growth	3	111	0.64	0.03	0.02	0.96	0.79
GO:0001558	Regulation Of Cell Growth	4	192	1.11	0.03	0.02	0.96	0.79
GO:0034641	Cellular Nitrogen Compound Metabolic Process	6	429	2.48	0.04	0.02	0.96	0.79
GO:0006519	Cellular Amino Acid And Derivative Metabolic	5	345	1.99	0.05	0.03	0.96	0.79

Table 6.10: GSA results from the HELP-seq analysis - promoter hypermethylation.

GOBPID	Term	Count	Size	Expected Count	P-value	Corrected p-value	FDR	Corrected FDR
GO:0009966	Regulation Of Signal Transduction	13	890	4.69	0.00069	0.0027	0.16	0.32
GO:0023051	Regulation Of Signaling Process	13	890	4.69	0.00069	0.0027	0.16	0.32
GO:0007178	Transmembrane Receptor Protein Serine/ Pathway	5	145	0.76	0.00099	0.0029	0.16	0.32
GO:0010648	Negative Regulation Of Cell Communication	6	249	1.31	0.0019	0.0036	0.24	0.32
GO:0019226	Transmission Of Nerve Impulse	7	400	2.11	0.01	0.0046	0.32	0.32
GO:0030099	Myeloid Cell Differentiation	4	143	0.75	0.01	0.01	0.32	0.32
GO:0051047	Positive Regulation Of Secretion	3	104	0.55	0.02	0.01	0.32	0.32
GO:0006310	Dna Recombination	3	113	0.60	0.02	0.01	0.32	0.32
GO:0006917	Induction Of Apoptosis	6	323	1.70	0.01	0.01	0.32	0.32
GO:0012502	Induction Of Programmed Cell Death	6	324	1.71	0.01	0.01	0.32	0.32

Table 6.11: GSA results from the HELP-seq analysis - promoter hypomethylation.