ABSTRACT

HOWARD, BRIAN EDWARD. Methods for Accurate Analysis of High-Throughput Transcriptome Data. (Under the direction of Dr. Steffen Heber).

A detailed understanding of the transcriptome is a prerequisite for deciphering the flow of information from genotype to phenotype. Fortunately, modern high-throughput technologies now provide an unprecedented ability to observe the full complement of transcriptional events, which extend far beyond the classical "one gene, one protein" hypothesis to include alternatively spliced genes, microRNAs, RNA interference, anti-sense transcription, and a variety of other, until recently, unknown phenomena. However, in order to accurately interpret the results of these assays, new statistical and bioinformatic methods must be developed in parallel to biotechnological advances. In this thesis, we present several methods for improving the accuracy of inferences obtained from the high-throughput transcriptome data generated by these new technologies.

First, we present a novel method for microarray quality assessment. Since accurate inference is dependent on the quality of the underlying data, quality assessment is a critical component in any microarray data analysis. Our method, which uses an unsupervised classifier to discriminate between high and low quality microarray datasets, exhibits performance comparable to supervised learners constructed using the same training data. However, because our approach requires only unnannotated data, it is easy to customize and to keep up-to-date as technology evolves.

Next, we present an alternative method for microarray quality assessment, which identifies low quality microarrays by simulating a set of differentially expressed genes. This method directly measures the ability of a planned statistical analysis to identify differential

gene expression when suspected low quality arrays are included in the dataset. A key advantage of this approach is that, unlike other methods, this method provides a specific recommendation about whether to retain or discard low quality chips in the context of a particular experimental setting.

Finally, we introduce a procedure for accurately quantifying alternative splicing using RNA-Seq data. Our method uses a familiar linear models approach, but improves upon similar methods that assume uniform coverage of RNA-Seq reads along the targeted transcripts. We first show, through simulation, that using an incorrect read sampling distribution can lead to incorrect conclusions about the expression of isoforms in a mixture. Applying our method to an example dataset, we identify 438 differentially spliced genes, exhibiting a range of expression patterns including genes with switch-like differential splicing between two tissues, as well as genes with more subtle variations in isoform expression.

Taken together, we expect that these methods can serve to increase the accuracy of inferences drawn from high-throughput transcriptome data, and in doing so, lead to an advancement of our understanding of the biology of genome expression.

Methods for Accurate Analysis of High-Throughput Transcriptome Data

by
Brian Edward Howard

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Bioinformatics

Raleigh, North Carolina

2009

APPROVED BY:

_____	_____
Dr. Steffen Heber Committee Chair	Dr. David Bird
_____	_____
Dr. Dahlia Nielsen	Dr. Heike Winter-Sederoff

_____
Dr. Hao Helen Zhang

BIOGRAPHY

Brian E. Howard was born in scenic Frederick, Maryland in 1972, where he was raised as an only child until the age of 5, when the first of two younger sisters arrived. Surviving this early trauma eventually translated into an increased appreciation for quiet reflection and may have helped to steer Brian towards a career in science. In 1995, Brian graduated from UMBC, in Baltimore, Maryland, with Bachelor of Science degrees in both Computer Science and Biology. After working as a software engineer for 9 years and obtaining a Masters degree in Computer Science from Johns Hopkins University in 2003, Brian decided to pursue a research career in the genomic sciences. In 2004, Brian and his wife, Grace, moved to Raleigh where Brian enrolled in the excellent graduate program at North Carolina State University. When he is not at work researching important questions in Bioinformatics and Computational Biology, Brian enjoys spending time with his wife, family, and dog, hiking and running, traveling, making music, and playing the occasional video game.

ACKNOWLEDGMENTS

I would like to thank my wife, Grace, for her patience and support, not only during the time I have spent working on this thesis, but also for everything we have experienced together that has brought me to this place. The same thanks I also extend to my parents and to my sisters. Thanks, also, to the staff, faculty and students at the BRC who have helped to make this a successful journey – especially to Dr. Steffen Heber, who has provided invaluable insights and support along the way. I would also like to express my gratitude to Dr. John Sheppard, who introduced me to academic research and started me out on this path when I was a student at JHU. And finally, thanks to my dog, Sandy, who has always been a loyal and affectionate master.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

Study of the Transcriptome in the (Post-) Genomic Era

Genome size alone is a poor predictor of phenotypic complexity. Ever since the first crude, genome-level investigations in the 1950's, biologists have been puzzled by a lack of correlation between the raw count of genomic nucleotides and organism morphology (Pray, 2008; see also Figure 1-1.) Today, with the availability of many whole genome sequences, it is clear that there is also no meaningful association between gene count and organism complexity. For example, the organism with the largest known number of genes (60,000) is the single-celled parasite *Trichomonas vaginalis*. In contrast, the current gene count estimate for *Homo sapiens* is about 25,000 — approximately the same as the plant, *Arabidopsis thaliana*, and less than both mouse (30,000 genes) and rice (51,000 genes) (Pray, 2008).

The key to resolving this information gap is widely believed to hinge on a detailed understanding of the spatial and temporal gene expression patterns that confer phenotype from genotype (Sharp, 2009). In between the latent information stored in raw DNA and outwardly observable biological forms, lies a complex, and often mysterious, world of RNA. As a first step towards understanding the flow of information from genes to proteins, it is imperative to study the entire *transcriptome*, or "the complete set of transcripts in a cell, and their quantity, for a specific developmental stage or physiological condition." (Wang *et al*., 2009) In recent years, our understanding has advanced far beyond the classic "one gene, one protein" hypothesis, to include a rich transcriptional landscape in which a single gene can, in fact, encode multiple protein products, and in which many important regulatory transcripts do not result in functional proteins, at all. The last decade has witnessed an explosion of

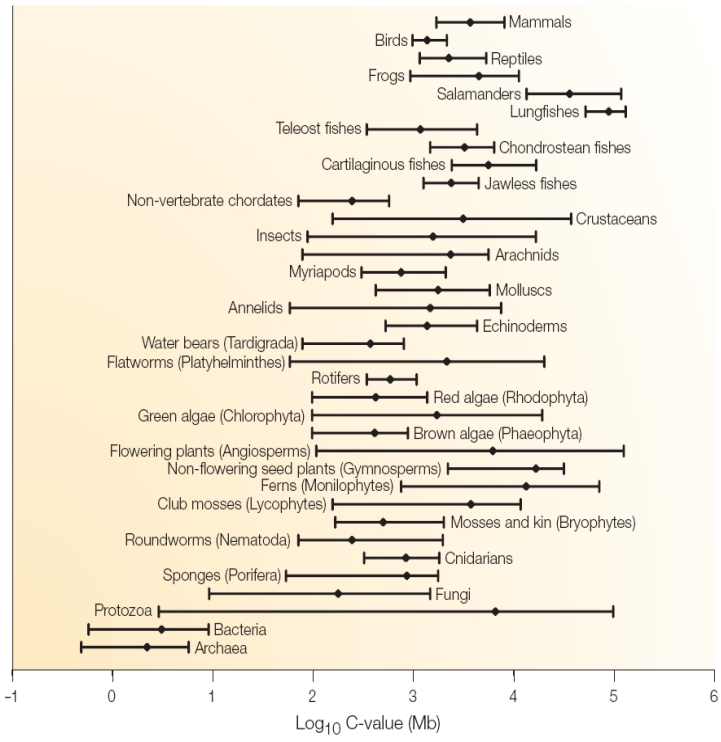


Figure 1-1. Extensive variation in genome size within and among the main groups of life (Pray, 2008).

As an example, some protozoa have larger genomes than mammals, while other protozoa have genome sizes comparable to bacteria and archaea. There is a striking lack of association between genome size and organism complexity.

research in this area, thanks, in part, to a variety of new high-throughput technologies that allow researchers to study the detailed expression of thousands, or tens of thousands, of RNA transcripts at once.

This introductory chapter will begin by reviewing several of the important components of the transcriptome, which includes not only the canonical, "carbon-copy" mRNAs fundamental to the "Central Dogma" of molecular biology, but also alternatively-

spliced transcripts, non-protein coding microRNAs, and anti-sense RNAs. Then, we will review the various high-throughput technologies that are currently available for measuring transcriptome expression. Special attention will be given to both microarrays and high-throughput RNA sequencing (RNA-Seq), both of which are central to this thesis. Finally, relevant background information regarding current methods for microarray quality assessment is discussed, in an effort to lay the groundwork for subsequent chapters.

The focus of our research is the development of new methods that can be used to maximize the accuracy of inferences obtained through analysis of transcriptome datasets. In the case of microarrays, which are often subject to high levels of experimental noise (Wilkes, *et al*., 2007), data quality assessment is a critical component of any analysis. In chapter 2, "Unsupervised assessment of microarray data quality using a Gaussian mixture model," we will introduce a new method for interpreting a variety of widely-used quality indicators in the effort to identify low quality microarrays. This chapter was previously published in the journal *BMC Bioinformatics* (Howard, *et al*., 2009a). In chapter 3, "Practical quality assessment of microarray data by simulation of differential gene expression," we take a different approach to quality assessment. In small, real-world experiments, it is often not feasible to re-run a defective hybridization; instead the relevant choice is whether or not to completely discard imperfect data. To address this question, we introduce a simulation-based method that determines the effect of discarding a particular array in the context of a test for differential gene expression. This chapter was published in the conference proceedings for the 5th International Symposium on Bioinformatics Research and Applications (ISBRA 2009) (Howard, *et al*., 2009b). In Chapter 4, "Towards reliable isoform quantification using RNA-

Seq data," we turn our attention to an alternative high throughput technology. In particular, our interest is in developing a method for quantifying the relative transcription levels for the various alternatively spliced isoforms for mutli-isoform genes, using RNA-Seq data. This chapter is an accepted submission to the upcoming International Conference on Bioinformatics and Biomedicine (BIBM 2009).

Components of the Transcriptome

Early large scale gene expression studies often assume a limited definition of the transcriptome as the "identity of each expressed gene and its level of expression for a defined population of cells" (Velculescu, *et al*., 1997). In these studies, gene expression levels are inferred from the measured quantities of *messenger RNA* (*mRNA*) found in a sample. This approach is in correspondence with the "Central Dogma" of biology which describes how DNA genes are first copied to mRNA transcripts, and then delivered to ribosomes where they are decoded and used as blueprints to assemble functional proteins. Indeed, most studies of the transcriptome have involved isolation of mRNA molecules followed by a quantification step using microarrays or, more recently, high-throughput RNA sequencing. Much has been learned from this approach, with applications in drug development (Marton, *et al.*, 1998), toxicology (Nuwaysir, 1999), biomarker discovery (Golub, *et al.*, 1999; van't Veer, *et al*., 2002; Singh, *et al*., 2002; Wang, *et al.*, 2000; Alon, *et al.*, 1999; Ramaswamy, *et al.*, 2001), evolutionary biology (Kant and Baldwin, 2007), functional genomics (Jares, 2006) and clinical practice (Li, *et al.*, 2008). Nevertheless, as the science has evolved, we have

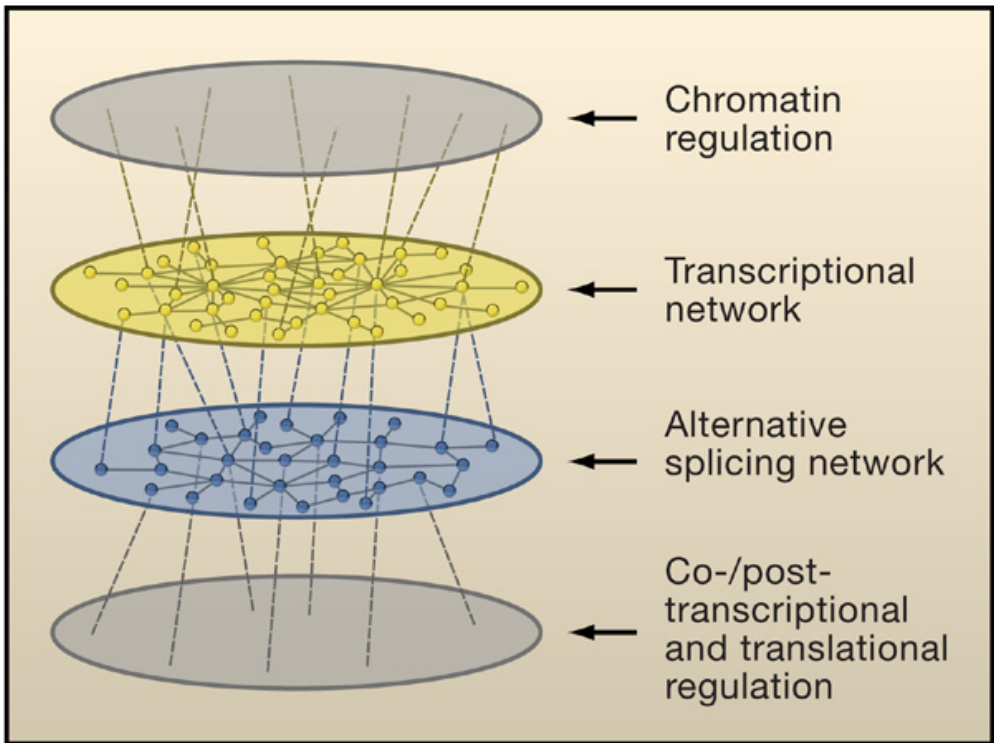


Figure 1-2. Layers of coordinated gene regulation (Blencowe, 2006).

Recent evidence suggests that, for a given tissue and treatment, the set of differentially expressed genes is independent from the set of differentially spliced genes. This contributes to a model for genetic regulation that involves several distinct layers acting independently to determine phenotype. In the first layer, the physical DNA structure and packaging, which can be influenced dynamically by histone modifications, controls the availability of genes to the transcriptional machinery. Subsequently, various *cis* and *trans* regulatory signals, including promoters, enhancers and transcription factor binding sites, further modulate gene expression. Likewise, at the level of alternative splicing, a variety of signals, including canonical splice recognition sites, along with splicing enhancers and silencers, determine the observed patterns of alternative splicing. Finally, a variety of post-transcriptional and post-translational mechanisms, including RNA degradation, RNA silencing, and post-translational modifications, can further modify expression.

developed an increasing appreciation for complexities of the transcriptome that extend well beyond the simple one gene, one protein model of Beadle and Tatum (1941).

In humans and other metazoans, most genes are divided into multiple protein coding regions (*exons*), separated by long intervening non-coding *intronic* regions. After transcription, but before being transported out of the nucleus for conversion to protein, the *introns* are spliced out of the primary transcripts. In many cases, different patterns of splicing can lead to multiple transcripts and proteins from the same gene. It is now estimated that 70% or more of human genes undergo *alternative splicing*, and mutations in splicing regions are increasingly recognized as playing an important role in inherited diseases (Hertel 2008). Furthermore, microarray experiments have revealed that transcription and splicing appear to be regulated in an orthogonal manner (Figure 1-2); the set of genes that is alternatively spliced is often very different than the set of genes that is differentially expressed under the same conditions (Blencowe 2006). Technological advances that increase our ability to detect and measure this phenomenon are expected to lead to an increased appreciation for the role played by alternative splicing in the domains of evolution, tissue differentiation, phenotypic response to environmental stimulus and disease pathology (Gravely 2001).

RNA interference (*RNAi*) is an additional complication that must be included in any modern definition of the transcriptome (Fire, *et al*. 1998). RNAi is a complex, conserved, innate genetic mechanism that can be exploited to post-transcriptionally "silence" targeted genes with high specificity. The primary endogenous variety of RNAi occurs in the form of *micro-RNAs* (*miRNAs*) — short, non-protein coding, functional RNAs that occur in the genomes of many organisms. Currently, there are 678 predicted human miRNAs cataloged in the miRBase database (Griffith-Jones 2004). Micro-RNA genes are transcribed in the nucleus into primary miRNA (pri-miRNA) transcripts, which are then transported to the cytoplasm,

cleaved into mature 21-23 nucleotide miRNAs by an enzyme called DICER, and then incorporated into an RNA-induced silencing complex (RISC). When a RISC encounters a target RNA with base pair complementarity, the targeted mRNAs are either degraded or sequestered. Because base-pairing is imperfect, each miRNA appears to target on average 5-10 mRNAs in humans (Sarnow 2006). Recent research suggests that microRNAs may be nearly as important as transcription factors for the purpose of gene regulation in vertebrate genomes, and that the two regulatory mechanisms may work in concert to control gene expression (Hobert 2008). For example, it has been shown that, in many cases, up-regulation of transcription factors confers wide-spread up-regulation of targeted genes, and that miRNAs may then modulate this response through post-transcriptional silencing in a tissue-specific manner (Hobert 2004). In contrast to transcription factors, however, it has been noted that the "speed, reversibility, and compartmentalization of miRNA-mediated control mechanisms predestine miRNAs to be involved in rapid, adaptive changes in gene expression to maintain homeostasis and to respond to specific environmental, nutrient, or neuronal signals." (Hobert 2008).

In addition to microRNAs, there are also a number of other important non-coding RNA species now recognized as components of the transcriptome. *Piwi-interacting RNA*s (*piRNA*s), for example, are short transcripts with lengths between 26 to 32 nucleotides. Unlike miRNAs, they lack secondary structure and are not highly conserved (Carmi, 2006). They are thought to function in germline development, possibly in the capacity of silencing movable genetic elements, such as retrotransposons, during spermatogenesis (Watanabe, *et al.*, 2006). *Repeat-associated small interfering RNAs* (*rasiRNA*s) are also thought to play a

role in silencing germline retrotransposons, but their structure and properties are different (Watanabe, *et al*., 2006). In addition, endogenous, a*nti-sense* RNA transcripts are also known to play an important regulatory role in some organisms by causing the degradation of their targeted sense transcripts (e.g. Dühring, *et al.*, 2006). At this time, many of the mechanisms and properties of these functional non-coding RNAs remain unknown. In addition, there are probably additional types of non-coding RNA, that are yet undiscovered — especially longer transcripts that have been more difficult to characterize with high-throughput RNA sequencing technologies (Sharp, 2009). In fact, the recent ENCODE project (The ENCODE Project Consortium 2007) has revealed that the majority of the human genome is actually transcribed into RNA, with only a tiny fraction of this corresponding to protein-coding genes.

High-Throughput Technologies for Transcriptome Analysis

In this section we will provide a brief description of the two main technologies used to assay the transcriptome in a high-throughput manner. There are also several other important technologies with both historical and modern significance, including SAGE, EST libraries, and MPSS. However, since our research is focused on data generated from microarrays and high-throughput RNA sequencing, we will limit our discussion to these two technologies.

Microarrays

The two main types of microarrays are spotted cDNA arrays and oligonucleotide arrays. Both array types rely on hybridizing labeled RNA or cDNA from a sample to complementary nucleotides fixed to a solid surface. The amount of RNA hybridizing to the array for each species is assumed to be proportional to its prevalence in the sample and can therefore be used as a proxy for quantification of the sample's RNA content.

Spotted cDNA arrays are usually manufactured in-house by an individual lab. The procedure typically involves isolating RNA from a targeted transcriptome, and then making reverse transcribed cDNAs. These cDNAs are then affixed to a solid surface in a grid-like pattern, with a distinct "spot" for each cDNA species. Each spot is complementary to a specific RNA transcript from the original sample, and can be used to quantify the RNA content of future samples from the same organism, tissue, and/or treatment. Because cDNA arrays are typically made by individual labs, the resulting measurements can be more difficult to compare across labs than results obtained from industrially manufactured oligonucleotide arrays (Bammler, *et al*., 2005). However, the degree of customization is attractive for many researchers, and, for non-model organisms, spotted arrays are often the only microarray platform available.

Oligonucleotide arrays such as those available from Agilent (http://www.chem.agilent.com), Affymetrix (http://www.affymetrix.com), Illumina (www.illumina.com), NimbleGen (http://www.nimblegen.com), and other biotechnology companies, rely on a set of short standardized oligonucleotide probes. Probes are carefully chosen according to their complementarity to selected targets, and so as to minimize cross-

hybridization to homologous regions of similar genes. Many types of arrays are available, varying according to the targeted molecules. *Gene expression arrays* attempt to measure the expression of mRNAs from protein coding genes, and typically target the least variable gene regions at the 3' and 5' termini. For this reason, this platform is not usually capable of discriminating between alternatively spliced transcripts from the same gene. *Exon arrays* or *exon junction arrays* have probes that are designed to query individual exons, or regions spanning adjacent exons, and can be used to quantify the occurrence of alternative splice isoforms. *Genome tiling arrays* are designed to measure transcription across the entire genome, with probes spaced equally at regular intervals, regardless of known locations of protein coding genes. Other microarrays include probes for known miRNAs and other non-coding RNAs. In this thesis, we make use of example datasets that originate from the Affymetrix GeneChip Array and Affymetrix Exon Array platforms. Accordingly, the following sections provide additional details regarding these arrays.

Affymetrix 3' GeneChip Arrays

An Affymetrix GeneChip® microarray consists of an approximately one-half square inch quartz surface divided into a grid containing hundreds of thousands of "probe cells." Each probe cell contains hundreds of thousands, or even millions of copies, of a single short oligonucleotide probe. Each probe on the GeneChip "3' Expression" arrays is typically 25 nucleotides in length. These probes, which are synthesized directly on the quartz surface using a photolithographic process, are designed to be complimentary to a particular gene in the target organism's genome. Each probe exists as a member of a probe pair. Within a

probe pair, the "perfect match" probe is perfectly complimentary to its intended target; the "mismatch probe," on the other hand, contains a single mismatched nucleotide, usually at the center base within the probe sequence. Although the mismatch probe is intended to measure non-specific, "background" hybridization, some pre-processing algorithms actually ignore the signal from this probe. Probe pairs are further organized into larger "probe sets." Each probe pair within a probeset is intended to hybridize to a different subsection of the intended target sequence. The 3' expression arrays typically have about 11 probe pairs per targeted gene. In order to minimize the effects of various spatial artifacts that can occur during hybridization, the probes for each probeset are not positioned contiguously, but rather are scattered across the entire array.

In the case of 3' expression arrays, the goal is to measure the amount of mRNA in a sample and thereby gain information concerning the relative transcription levels of the targeted genes. Once mRNA has been isolated from the biological sample, oligo-dT primers, which are complementary to the poly-A tails on the mRNAs, are used to prime reverse transcription reactions that convert the mRNA into double-stranded cDNA. Next, these cDNAs are re-transcribed back into cRNA using biotinylated nucleotides. In addition, an amplification procedure can be used to increase the amount of RNA available for hybridization. The resulting cRNAs are then fragmented and hybridized to the array. Any cRNA fragments that are complimentary to the probes on the array will base pair and anneal to the probes. Next, the array is exposed to a Streptavidin-phycoerythrin (SAPE) solution. The Streptavidein adheres to the biotinylated cRNA fragments bound to the array. Afterword, a laser is used to excite the fluorescent dye (phycoerythrin), which then emits light. The light

can be read using an optical scanner, and the amount of light emitted from each cell is roughly proportional to the amount of cRNA hybridized to the probes in that cell.

After a microarray experiment has been performed, a variety of data processing steps need to be performed before an analysis of differential gene expression can proceed. Affymetrix provides software to assist with these steps, but there are also a variety of free, public domain software packages available, as well. The first step is to convert the digitized image produced by the optical scanner into a raw intensity measurement for each probe. This involves determining the borders of each "spot" and computing an average intensity measurement. Once raw intensity values have been read for each probe, it is next necessary to perform "background subtraction," which is an attempt to remove noise caused by a variety of factors including cross-hybridization, and other local and global artifacts. The mismatch probes may or may not be used for this purpose. After background subtraction, a normalization step is performed to ensure that each of the microarrays in an experiment has a similar distribution of intensity values. In addition, the signals from all probes within a single probeset are typically summarized using a single statistic intended to represent the intensity of the target gene. Once these values have been computed, it is possible to use a variety of statistical procedures, including t-tests, ANOVA, and other more sophisticated techniques, to compare the gene expression levels. In general, the outcome of a microarray experiment is highly dependent on the pre-processing steps performed and the statistical analysis employed.

Affymetrix Exon Arrays

The Affymetrix GeneChip® Exon Array platform, currently available for mouse, rat and human, is an oligonucleotide microarray platform designed to interrogate expression levels of individual exons, rather than genes. The Human Exon 1.0 ST Array, for example, contains 1.4 million probesets querying more than one million exon clusters. On average, each potential exon or splice region is covered by a single probeset, with about 4 probes per probeset, for a total of more than 5,500,000 probes. These probesets are divided into annotation categories which quantify the degree of evidence supporting the corresponding target features. "Core" probesets, for example, are designed to measure expression of well-annotated exons, while the "Extended" and "Full" probeset lists contain probesets designed to measure expression of more speculative transcriptional loci, including computationally predicted exons. While the platform is relatively new, a sample of several recent publications have reported reasonable rates of validation for detected differentially expressed exons (Gardinia, *et al*., 2006; Kwan, *et al*., 2007; French, *et al*., 2007; Chueng, *et al*., 2008; Clark, *et al*., 2007). In general, better validation rates tend to occur when more stringent filters are applied to the candidate list - for example, by only accepting predictions from probesets with signal levels that are clearly above background, and by using only those probesets from the "Core" list. The fact that each probeset is comprised of, on average, only four probesets, may make the results somewhat noisier than for traditional 3' arrays. Additionally, algorithms used to interpret the data are still being researched, with fewer public domain options available than for gene-level array platforms.

One important design limitation inherent in this platform, is that exon arrays measure *exon* expression and not *transcript* expression. For a given splicing isoform, exon arrays can only provide measurements of the component exons individually; they do not directly measure what combinations of exons are being co-expressed within the same transcript. Consequently, published analyses made using this platform have necessarily emphasized differential exon expression rather than differential splicing products.

High-Throughput RNA Sequencing

High-throughput RNA sequencing (*RNA-Seq*) includes sequencing by synthesis according to a variety of technologies, including those from Solexa / Illumina, Applied Biosystems / SOLID, Roche / 454 Life Sciences, as well as several other biotechnology companies. In these procedures, either RNA or reverse transcribed cDNA is fragmented and then used as a template for massively parallel polymerization reactions while various techniques are used to record each nucleotide added to the growing oligonucleotide chains. Various platforms are capable of sequencing millions of short reads in parallel, which can then be mapped to a reference genome or assembled *de novo*. Chapter 4 of this thesis presents an analysis of datasets generated from both the Roche 454 and Illumina platforms. In the next two sections, we provide additional background information specific to these technologies.

Rouche / 454 Life Sciences Pyrosequencing

Using the 454 Life Sciences pyrosequencing method, nucleotide sequencing can be performed directly from a DNA sample, with no cloning required. The details of the

procedure are described in (Margulies et al. 2005). Although originally designed for the purpose of genome sequencing, the same technology can be used to measure gene expression by simply converting expressed RNA into cDNA. To begin, the cDNA is broken into small double-stranded fragments using a nebulizer (Figure 1-3, Panel 1). The resulting fragments typically range from 50-900 base pairs in length, with a mean size of about 325 base pairs. Next, "adaptors" are attached to both ends of the DNA fragments (Figure 1-3, Panel 2; Figure 1-4). Once the adaptors have been attached, the DNA fragments are electrophoresed through an agarose gel to separate the fragments according to length. Fragments in the length range of 250-500 base pairs are retained. The nicks at the 3' junctions between the adaptors and the fragments are then repaired, and the overhangs are filled in.

In order to isolate single stranded DNA fragments having an A adaptor on one end and a B adaptor on the opposite end (configuration A-dsDNA-B in Figure 1-4), the following procedure is utilized: first the DNA are washed through a column of Streptavidin beads allowing the A-dsDNA-B and the B-dsDNA-B fragments to adhere to the column, while the A-dsDNA-A fragments are eluted. Next, the column is washed with a melting solution which causes the remaining double stranded A-dsDNA-B and B-dsDNA-B DNA fragments to separate into single stranded fragments. Furthermore, one of the A-ssDNA-B fragments will lack a biotin tag, and therefore elute through the column while the B-ssDNA-B fragments and the other single A-ssDNA-B fragments remain attached to the beads.

In the next step, the eluted single-stranded DNA library is attached to DNA capture beads. Each of the DNA capture beads is joined to oligonucleotide primers that are complimentary to the A adaptor. The DNA library is then bound to the capture beads using a

process of limiting dilution: a very dilute solution of the DNA library is annealed to a vast excess of the beads such that it is highly improbable that any bead will anneal to more than one DNA molecule. The beads are then suspended in a water-in-oil emulsion along with a PCR reaction mixture (Figure 1-3, Panel 3). Each tiny droplet serves as a microreactor, in which the single DNA strands are amplified into millions of copies per bead. When the PCR reactions have completed, the emulsion is broken, and the millions of amplified, double stranded DNAs on each bead are converted to single stranded DNAs using a melting solution. While approximately zero beads will contain more than one species of DNA as a result of the limiting dilution, many beads will have no DNA attached. In order to remove the majority of these "null beads," an enrichment procedure is performed in which a 40 base pair primer is mixed in with the beads. This primer, which is partially complementary to the 3' adaptor and which has a biotin tag attached, binds to the DNA-containing beads. Next, Streptavidin beads are added to the mixture, and the biotin-attached capture beads adhere to them. A magnet is used to separate these Streptavidin beads from the null DNA capture beads, which are discarded, and then the Streptavidin beads and biotin tagged primers are released from the capture beads with an application of melting solution.

Now, the beads are placed on a fiber optic plate (Figure 1-3, Panel 4), which is made of individual fiber optic core threads packed tightly together, side-by-side. Etched into each of the fiber optic threads is a tiny 75 pico-liter well. The size of the wells is such that each one can accommodate exactly one of the DNA capture beads, which are loaded onto the plate along with luciferase, ATP sulfurylase, sequencing primer and DNA polymerase. After the beads have been loaded onto the fiber optic plate, the DNA fragments attached to the beads

are ready to undergo a series of massively parallel, sequencing-by-synthesis reactions. In a series of cycles, reaction mixture containing a single species of nucleotide (A,C,G or T) is washed across the fiber optic plate (Figure 1-3, Panel 5). If this particular nucleotide is complimentary to the next base pair on one of the single stranded DNA's, it is incorporated into the growing double-stranded chains by DNA polymerase, and a series of chemical reactions occurs culminating in the conversion of ATP into light by luciferase (Figure 1-3, Panel 7). This light is then conveyed through the fiber optic tubules and automatically measured, with the amount of light from each well roughly proportional to number of sequential nucleotides being incorporated. This linearity is maintained for strings of up to eight consecutive repeated nucleotides; above that, it becomes difficult to accurately quantify the number of nucleotides from the measured light. After each nucleotide is added, a solution containing apyrase is washed over the wells to quench any residual luciferase reactions and complete the cycle. Then the next nucleotide is added and a new cycle begins. Nucleotides are added in this manner in a sequential fashion, e.g. A-C-G-T-A-C-G-T-A-C-G-T, etc. By measuring the amount of light emitted from each well at each step, in correlation with the order in which the nucleotides are added, it is possible to accurately determine the sequence of the single stranded DNA fragment attached to the bead in each well (Fig 1.3, Panel 6). During the incorporation of each nucleotide, a small fraction of the amplified strands on each bead can become out of synch. The effect, however, is cumulative, and this places an upper limit to the length of accurate reads. The first generation of sequencing machines were capable of average read lengths of around 110 base pairs. The more recent hardware has apparently increased the average length of reads to around 400 base pairs.

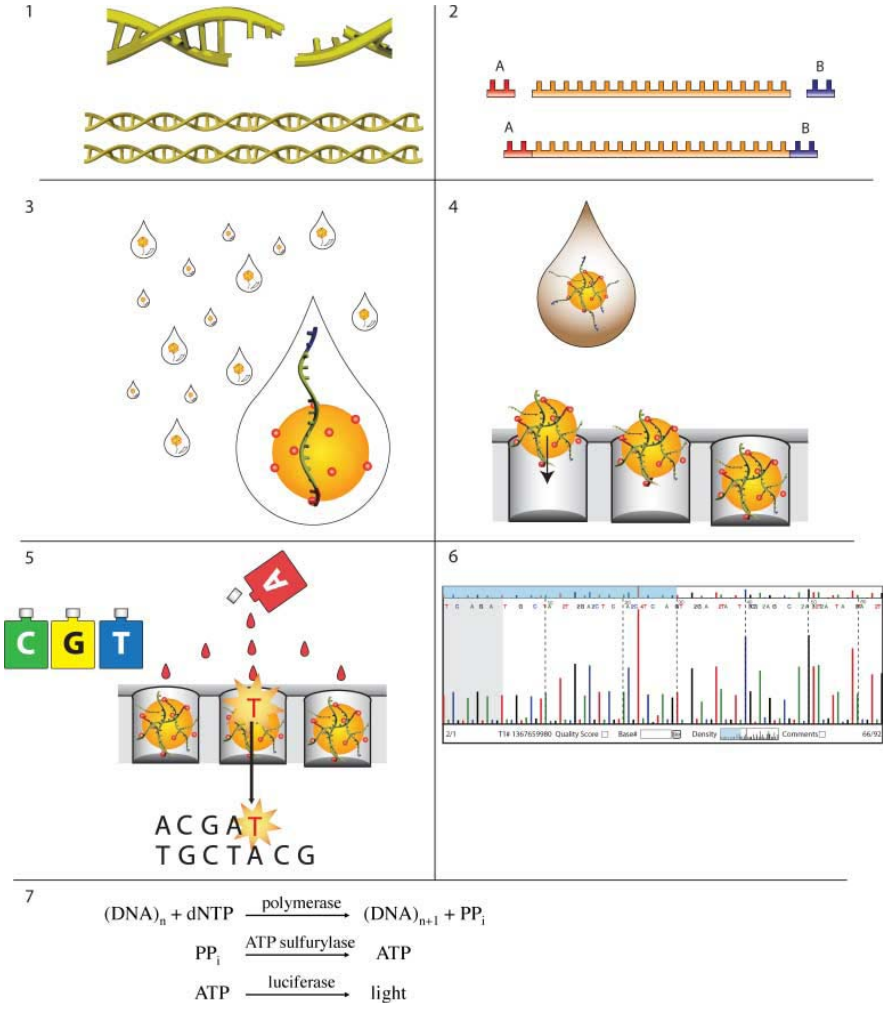


Figure 1-3. Steps in Roche/454 Life Sciences pyrosequencing (Ellengren, 2008).

In the final step of the analysis, if no reference genome is available, the fragments need to be combined in a manner analogous to sequence assembly from traditional shotgun sequencing. However, because the read length is currently much shorter than the length of Sanger sequencing reads, greater coverage of the genome may be required to achieve the

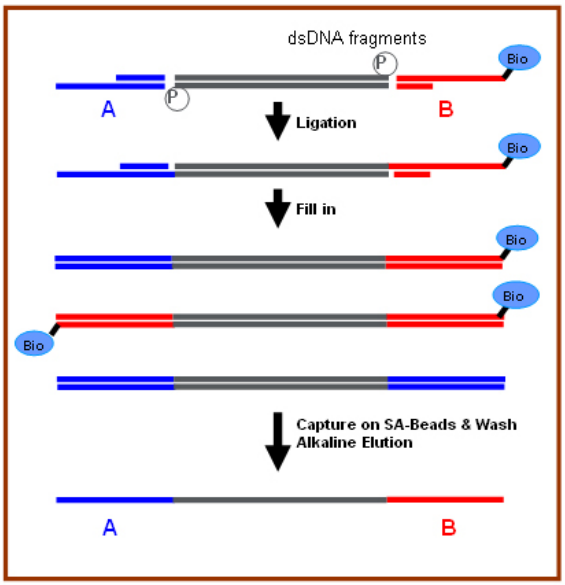


Figure 1-4. Pyrosequencing adaptors (Marguiles *et al.*, 2005).

There are two double-stranded adaptors, the "A" and "B" adaptors. Each adaptor consists of a 20 base pair amplification primer, followed by another 20 base pair sequencing primer, and a 4 base pair "key sequence," which is used to identify the beginning of each sequenced fragment. Both adaptors have a blunt end on one side, and a 5' overhang on the opposite side. In addition, the B adaptor also has a biotin tag attached to the 5' overhang. Adaptors can only attach to the DNA fragments blunt end to blunt end, in 3 possible configurations: A-dsDNA-A, A-dsDNA-B, B-dsDNA-B.

same level of assembly completeness. If a reference genome is available, the reads can instead simply be aligned to the reference using a program such as BLAT (Kent, 2002).

Illumina / Solexa Sequencing

Transcriptome sequencing with the Illumina platform typically begins by using sonication to shear RNAs into fragments that are then used to construct complimentary cDNA. Generic work flows are described in (Mardis, 2008; Metzker, 2008; see also, Figure

1-5). Adaptors are attached to both ends of the fragments, and single-stranded fragments are randomly attached to the solid surface of a flow cell. The flow cell surface also contains a dense lawn of primers that are complementary to the adaptors.

Next, a technique known as "bridged amplification" is used to make many copies of each of the individual cDNA fragments adhering to the surface. This step is important because it ensures that the subsequent sequencing-by-synthesis reactions will produce adequate signal for detection by the optical instrumentation. During bridge amplification, the attached cDNA fragments bend towards the surface so that their free ends also come into contact with the surface. Since the free ends are attached to adaptors that are complementary to the primers on the surface, they bind to the surface of the flow cell, forming a "bridge." At this point, the primers are used for a round of DNA synthesis by DNA polymerase. After several rounds of synthesis and denaturization, the result is millions of individual, homogeneous cDNA clusters, each containing about 1 million copies of a single original fragment.

After this preliminary step, the stage is set for the subsequent sequencing by synthesis reactions. First, primer is added to the attached fragments, and then, in each sequencing round, all four nucleotide bases are added simultaneously. Each of the four nucleotide bases is given a unique fluorescent label. In addition, the 3' hydroxyl group is chemically blocked so that only one base can be added at a time. During each synthesis cycle, a single base is incorporated into each growing oligonucleotide chain. Then, the fluorescent labels are excited and the resulting light is detected and recorded by a specialized optical instrument. After each cycle, the fluorescent label and the 3' chemical block are removed so that the

sequencing reaction can continue in the next round. By keeping track of the colors of light originating from different locations on the surface, the sequencing machinery is able to accumulate sequence information for millions of short cDNA fragments in parallel. In the first generation of equipment, these fragments ranged in length from about 25-35 base pairs. Each flow cell has 8 separate lanes, each of which can be loaded with a different sample.

Following the sequencing reactions, it is necessary to perform the same basic data analysis steps required with 454 sequencing. However, because the number of reads is so much greater, and the length of those reads, so much shorter, it is generally necessary to use a completely separate set of software tools to do the alignment and assembly steps.

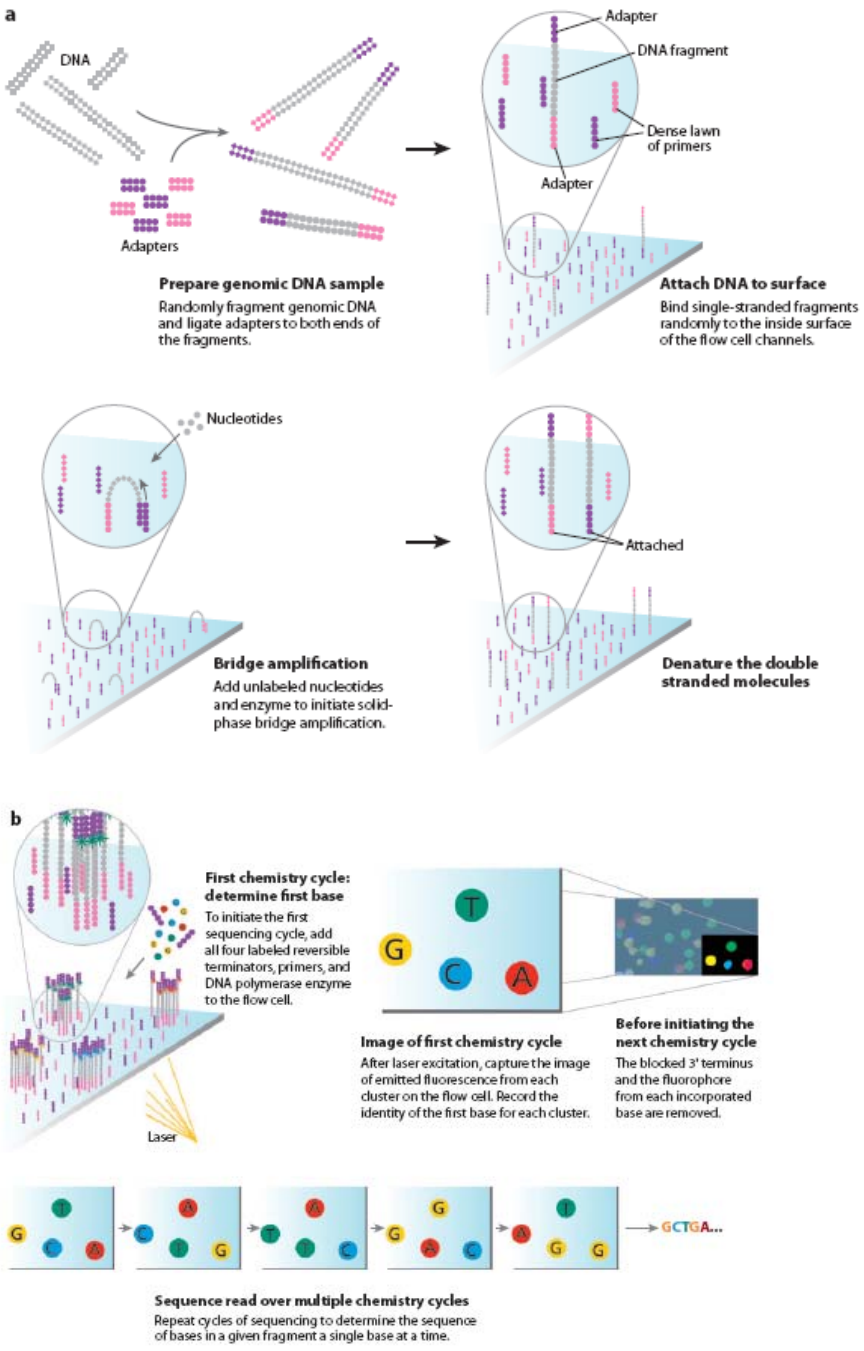


Figure 1-5. Illumina sequencing (Mardis, 2008).

A) Bridge amplification. B) Sequencing by synthesis.

Microarray Quality Assessment

Sources of variability

As with any biological experiment, the goal of a typical microarray study is to identify meaningful variation between treatments while controlling for random technical and biological noise. However, the generation of reliable microarray data involves many steps, and often things go wrong. Because of this, recent efforts have emphasized the need for rigorous quality control, as well as procedures and metrics designed to assist researchers in identifying low quality data. Here we consider some of the various sources of error that may occur during a typical microarray experiment. The following list is not exhaustive, but does highlight several important factors influencing the outcome for a typical microarray experiment. For concreteness, the Affymetrix GeneChip platform is assumed. The ordering of the list roughly follows the sequence of steps in an array analysis, from design and manufacturing of the chips through analysis of the data:

- *Sub-optimal probes and cross hybridizations.* Probe selection is a very important aspect of microarray design. Unfortunately, we do not yet completely understand how probe sequence translates to binding affinity. Microarrays are subject to cross-hybridization from partially complimentary sequences, and this effect is difficult to predict or quantify. Even for perfectly complimentary target and probe pairs, there are differences in binding affinity that are difficult to model.
- *Damaged or Scratched slide*. Damaged slides are obviously problematic. Generally speaking, commercially produced microarray slides are of exceptional quality, but

mistakes can occur. More commonly, slides can be damaged at any point during their lifecycle, including during shipping and storage.

- *RNA degradation.* All biological samples contain varying amounts of enzymes (nucleases) which function to digest RNA. These continue to be active even after the death of the cell or tissue from which the RNA is extracted. Furthermore, nucleases are prevalent in the environment, and extreme caution must be exercised not to introduce exogenous nucleases into a sample during an experiment.
- *Sample contamination.* Depending on the sample preparation procedure, numerous opportunities for sample contamination may occur. For example, in some cases RNA is extracted from specific organs or tissues, which may be difficult to properly isolate under certain conditions. It is also possible to contaminate samples with foreign RNA or DNA, for example from bacteria or viruses present in the environment. It is even possible (but hopefully rare) to inadvertently mislabel or switch samples during an analysis.
- *Improper storage of reagents.* The Affymetrix microarray preparation protocol is fundamentally a sequential series of chemical reactions. In order for these reactions to proceed efficiently according to specifications, proper storage of reagents is necessary.
- *Transcription efficiency.* The first steps in preparation of an RNA sample for hybridization involve reverse transcription reactions that copy RNA into cDNA; these cDNAs are then transcribed back into RNA, incorporating biotin-labeled nucleotides. If these reactions do not occur with the same efficiency across all samples, and for all

transcripts, a bias has been introduced.

- *PCR amplification.* Similarly, during sample preparation, it is usually necessary to perform a PCR step in order to amplify the small amounts of RNA present in the original sample. While the goal is to have the resulting amount of RNA proportionate to the starting amount of each transcript, it is possible that bias can be introduced if the PCR reactions proceed with differing efficiency for different transcripts or samples.
- *Spot identification.* Quantification of the relative hybridization occurring at each probe involves averaging the signal observed at each of the microarray features. This requires identifying the edges of each feature "spot," subtracting any contaminating background intensity and producing a composite average raw intensity to serve as a proxy for hybridization at the underlying probes. Numerous opportunities for introducing error exist during this process.
- *Spatial artifacts.* If the target solution is not evenly applied during hybridization, numerous bubble effects and edge effects can occur. Unfortunately, these spatial artifacts are not uncommon.
- *Paralogous sequences.* In most organisms there are many sets of paralogous genes having highly similar sequences. These genes are one source of cross-hybridization effects that can influence observed intensities.
- *Alternatively spliced isoforms.* For most genes, the Affymetrix 3' expression arrays only contain probesets corresponding to the 3' end of the transcript. For this reason, these arrays are unable to measure relative levels of alternative transcripts, and the

presence of alternatively spliced gene variants can confound the observed signal for the constitutive isoforms.

- *Normalization method.* The goal of normalization procedures is to remove systematic biases among a set of microarray slides. However, there are many different normalization procedures, and each method is based on a set of specific assumptions. Unfortunately, the method of normalization chosen can greatly influence the outcome of an experiment, and in some cases normalization can introduce unwanted biases into the measured gene expression levels.
- *Above background detection.* Microarrays have a finite dynamic range, and features having low signal levels often originate due to background signal and cross-hybridization. It is important to exclude low intensity signals from an analysis, but it is not always clear what threshold to use for this determination, and this decision can have an impact on the analysis results.
- *Saturation effects.* At the other extreme of the spectrum, genes expressed at very high levels may saturate their corresponding probes. In these cases, it will not be possible to obtain accurate measurements for expression differences between transcription levels of these genes.
- *Probeset averaging.* For most applications, the multiple probe pairs in a probeset are combined to produce a composite signal intended to represent overall expression for the target gene. However, this averaging approach can, unfortunately, obscure the underlying variation among probe measurements and conceal this variance from higher level analysis steps.

- *Analysis methods.* A whole spectrum of analysis methods exists for comparing microarray expression levels. These range from simple fold change comparisons to detailed linear models and sophisticated Bayesian estimation procedures. The statistical analysis method employed can have large impact on the resulting lists of differentially expressed genes.
- *Sample size.* Although microarrays are a very cheap way to get high-throughput gene expression measurements, as compared to other alternative methods like SAGE, EST libraries, rt-PCR, and, more recently, high-throughput sequencing technologies, most microarray experiments have, nevertheless, been limited by small sample sizes. It is not uncommon for researchers to allocate only 2 or 3 slides per individual treatment. Because of this, it is often very hard to get reliable estimates of the means and variances of gene expression levels for these treatments, and robust comparisons can be difficult even with quality data.
- *Biological noise.* Even when everything else goes well, it is still possible for results to be confounded by uncontrolled biological variability. Biological systems are inherently complex, and it is always possible that a subset of the experimental units is in a completely different "state" than the others. A typical mammalian organ is a mixture of a wide variety of distinct cell types, each of which may be reacting to a distinct set of environmental signals; a study of gene expression in the liver, for example, would be comprised of a complex mixture of distinct expression patterns. Similarly, an unhealthy organism could react very differently to a treatment than a healthy subject.

For array types other than Affymetrix GeneChips, various additional sources of error are also possible. For example, two color arrays are subject to "dye bias" effects, i.e. differential affinity of the dyes used to label the two samples compared. It has even been observed that local ozone concentrations can have a noticeable affect on dye affinities ("Making the most of Microarrays," 2006, p.1039). The bottom line is that there is considerable variance inherent in microarray measurements. Given the fact that biological systems are already highly complex and variable, and the fact that small sample sizes are often necessary, it is paramount to make efforts to identify and manage controllable sources of variance.

Review of Popular Quality Metrics

Having established the importance of quality control metrics, our discussion now turns to a few of the more popular metrics used in practice. These metrics can be divided into two broad categories: *pre-hybridization metrics* and *post-hybridization metrics*.

Pre-hybridization quality metrics are used to measure the quality of an RNA sample before the sample is actually hybridized to a microarray slide. Ideally, if there is some quality problem with an RNA sample, this problem would be detected prior to hybridization, before continuing the analysis and incurring the expense and wasted effort of following through with corrupted data. Popular pre-hybridization metrics include the 28S/18S rRNA ratio (Sambrook, *et al.*, 2001), the RNA integrity number (RIN) (Schroeder, *et al.*, 2006), the RQS Score (Copois, *et al.*, 2007), and the RNA degradation factor (Auer, *et al.*, 2003).

Even in the cases where pre-hybridization quality checks suggest that the initial RNA sample is of high quality, there are still many sources of error that can occur during the course of an RNA experiment. For this reason, it is also critical to measure microarray data quality after the sample has been hybridized to the chip, scanned and converted into raw intensity measurements. For example, for the Affymetrix microarray platform, several *post-hybridization quality assessment* metrics are included with the manufacturer's analysis software (Affymetrix, 2003); it is also possible to compute these metrics using the free R "simpleaffy" package (Wilson and Miller, 2005). In addition, the R BioConductor package (Gentleman, *et al.*, 2004) contains additional facilities for assessing the post-hybridization quality of Affymetrix microarray data. The software contains a variety of useful diagnostic plots that are often applied in conjunction with the standard Affymetrix quality control metrics discussed above (see Gentleman, *et al.*, 2005.) It is also possible to derive numerical statistics from these plots, and this approach has been used to produce software for automatically assessing microarray quality (e.g. Heber and Sick 2006). Because these BioConductor metrics are used extensively in the examples discussed in chapters 2 and 3 of this thesis, brief descriptions are provided below.

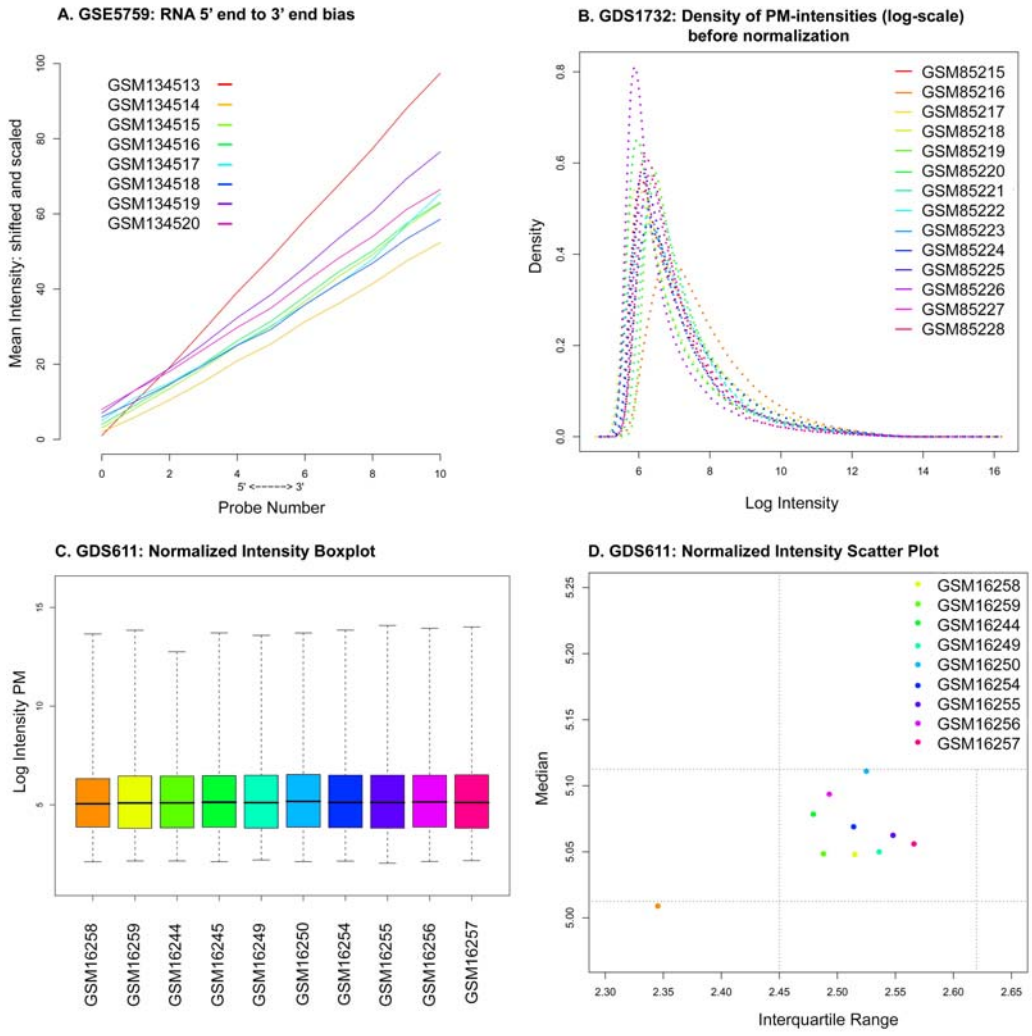


Figure 1-6. Examples of microarray quality assessment diagnostics.

(A) RNA degradation plot. The graph indicates that RNA degradation may be occurring at a different rate in sample GSM134513 compared to the other samples in the experiment. Inclusion of this sample in the analysis may adversely affect the resulting inferences. (B) Raw intensity distribution. A comparison of the un-normalized PM probe intensity distributions of the chips in an experiment can be used to identify outliers. (C) Normalized intensity box plot. After quantile normalization, the intensity distribution for all chips in an experiment is expected to be similar. Chips that deviate significantly from this expectation should be considered for exclusion. (D) Normalized intensity scatter plot. This plot is constructed from the same underlying data as C. Differences among the chips suggest that sample GSM16258 may be an outlier.

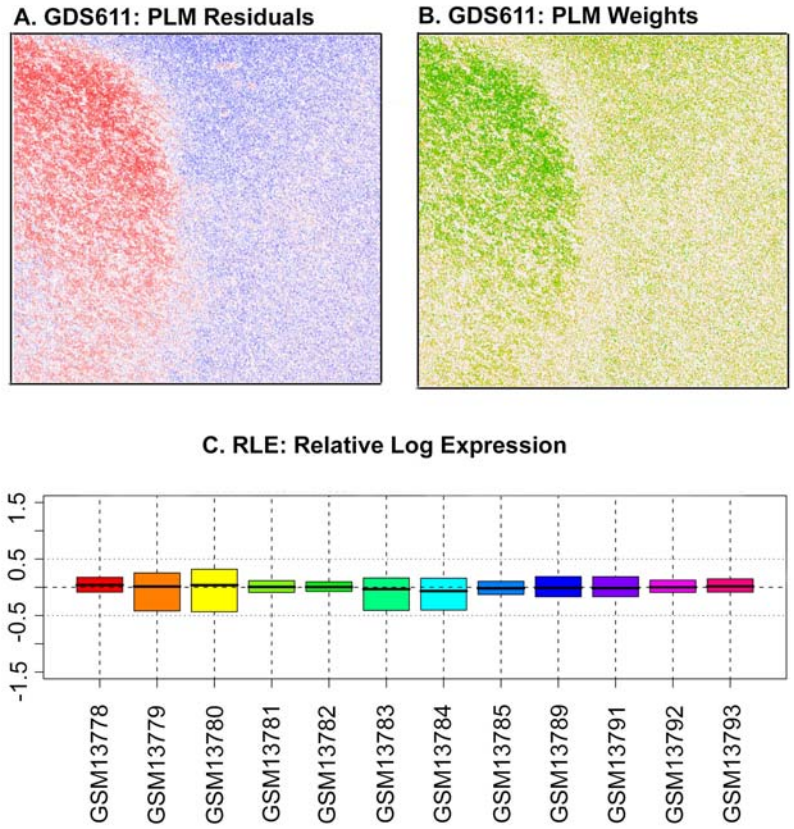


Figure 1-7. Additional microarray quality assessment diagnostics.

(A) Probe level model (PLM) residuals. The residuals in the probe level summarization model are expected to have a mean of zero. A large field of positive residuals (red) separated from a large field of negative residuals (blue) indicates a possible quality problem for sample GSM16258 from GEO dataset GDS611. (B) Probe level model weights. An excess of down-weighted probes (dark green area) in the probe level model for chip GSM16258 is an additional indicator of potential quality problems. (C) Relative Log Expression (RLE) box plot. Since most genes are not expected to be differentially expressed, a non-zero median and a large inter-quartile range in the RLE box plot can serve as indicators of low quality data.

5' to 3' RNA degradation plot

The 5' to 3' RNA degradation plot (Figure 1-6a), which can be produced using the BioConductor "affy" package (Gautier, *et al.*, 2004), shows the relationship between signal

intensity and probe position within each multi-probe probeset. To produce this graph, the individual probes within each probeset are ordered according to their position within the probeset, in the 5' to 3' direction. For each chip, the average signal at each probe position is computed across all probesets, and plotted as a function of probe position. In addition, a linear regression may be performed to model signal intensity as a function of probe position (assuming equal spacing of probes within the probesets). An upward slope is expected, since in almost all samples there is some RNA degradation, which occurs in the 5' to 3' direction. In general, outlier chips can be identified by looking for chips having slopes that are very different than the others; this may indicate, for example, an inconsistency in sample handling that has introduced additional variance for this chip.

Raw intensity distribution

Another useful diagnostic approach is to plot the raw intensity distribution for each chip prior to normalization. While it is reasonable to expect small differences among the chips, too much variation may indicate data quality problems. For example, Figure 1-6b shows the raw intensity plot for GEO experiment GDS1732. The intensity distribution for chip GSM85216 seems to be skewed to the right with more density at higher intensities.

Normalized intensity distribution

A comparison of signal distributions after normalization is probably an even better quality indicator. After normalization, it is expected that the intensity distributions should be similar for all chips. Figure 1-6c and Figure 1-6d show diagnostic plots for GEO dataset GDS611. In this case, both the median and the interquartile range of the normalized intensity

values for chip GSM16258 appear to be different from the other chips in the experiment, with the apparent magnitude of this difference depending on the nature of the visualization method employed.

PLM weights and residuals

The probe level summarization model is another often-used quality control indicator. As described in (Gentleman, *et al.*, 2005), the BioConductor affyPLM package contains functionality to fit (using a median polish procedure) a probeset summarization model for each probeset:

$$\log(Y_{gji}) = \theta_{gi} + \phi_{gj} + \varepsilon_{gji}$$

where θ_{gi} is the log-scale expression level for gene g on array *i*, ϕ_{gj} is the effect of the j^{th} probe for gene *g*, and ε_{gij} is the error measurement. An important assumption of this model is that the expected value of ε_{gij} is zero. A plot of the residuals from this model, as well as the weights used by the regression procedure to down-weight outlier probes, may reveal irregularities in the data. For example, Figure 1-7a shows that the residuals plot for sample GSM16258 from GEO dataset GDS611 indicates a large field of highly positive residuals (red) and another field of highly negative residuals (blue). In addition, the PLM weights plot (Figure 1-7b) indicates a large set of highly down-weighed probes (dark-green area).

RLE box plot

Given the normalized (log-scale) expression values, the RLE box plot displays the distribution of the quantity $M_{gi} = \hat{\theta}_{gi} - m_g$ for each chip, where $\hat{\theta}_{gi}$ is the log expression

measurement for probeset g, on chip i, and m_g is the median expression of probeset g across all arrays. In general, since it is normally assumed that the majority of genes are not differentially expressed across chips, the quantity M_{gi} is ordinarily expected to be distributed with median 0. Figure 1-7c shows the RLE box plot from GEO dataset GDS515. Several of the chips seem to have larger variance in the RLE distribution, and a few of the medians appear to deviate significantly from zero.

Microarray Quality Assessment Summary

The previous discussion has emphasized the importance of quality control in the analysis of microarray data. In addition, a variety of popular quality control methods were described. However, due to space constraints, many other important metrics were not covered here. For example, spike-in controls are available to assist with the determination of Affymetrix GeneChip data quality (Affymetrix, 2003), while NUSE plots, hierarchical clustering, and pairwise correlations are also available as part of the BioConductor library (Gentleman, *et al.*, 2005). Furthermore, recent research has focused specifically on detection of spatial artifacts (e.g. Reimer and Weinstein, 2005; Stokes, *et al.*, 2007). Recent research has also explored the development of similar metrics for RNA-Seq data (e.g. Morgan, *et al.*, 2009).

One difficulty for all of the metrics explored is that it is often not clear how to separate "good" quality scores from "bad" quality scores – the use of an arbitrary threshold is generally required. For example, Affymetrix supplies suggested score ranges for several of its post-hybridization quality scores, but with little or no justification for these

recommendations. Furthermore, since many of the metrics measure different aspects of quality, it is usual practice to use one or more metrics simultaneously, but it is not always clear what quality decision to make when these metrics give conflicting scores. Nevertheless, it is generally assumed that even a relatively naive use of some of these metrics will detect the most adverse cases of low quality data. Ongoing research is required to identify the best metrics to use in various scenarios, the distributional properties of these metrics, and efficient ways to combine them to generate useful composite quality scores. The topic of the next chapter is a new method for identifying low quality microarray data using quality metrics like the ones reviewed here.

References

Affymetrix, Inc. **GeneChip expression analysis, data analysis fundamentals.** Affymetrix 2003, Santa Clara, CA. Retrieved June 1, 2007 from http://www.affymetrix.com/support/downloads/manuals/data_analysis_fundamentals_manual.pdf.

Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ. **Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays**. *Proc Natl Acad Sci USA* 1999, **96**:6745–6750.

Auer H, Lyianararchchi S, Newsom D, Klisovic MI, Marcucci G, *et al*. **Chipping away at the chip bias: RNA degradation in microarray Analysis**. *Nature Genetics* 2003, **35**:292-293.

Bammler T, Beyer RP, Bhattacharya S, Boorman GA, Boyles A, Bradford BU, *et al.* **Standardizing global gene expression analysis between laboratories and across platforms**. *Nature Methods* 2005, **2**: 351–356

Beadle GW, Tatum EL. **Genetic control of biochemical reactions in Neurospora.** *Proc Natl Acad Sci USA* 1941, **27(11)**:499-506.

Blencowe BJ. **Alternative Splicing: New insights from global analysis.** *Cell* 2006, **126:**37-47.

Carmi I. **Molecular Biology Select: Taking a Peak at Piwi RNAs**. *Cell* 2006 **126(2)**:223

Cheung HC, Baggerly KA, Tsavachidis S, Bachinski LL, Neubauer VL, Nixon TJ, Aldape KD, Cote GJ, Krahe R. **Global analysis of aberrant pre-mRNA splicing in glioblastoma using exon expression arrays.** *BMC Genomics* 2008, **9**:216.

Clark TA, Schweitzer AC, Chen TX, Staples MK, Lu G, Wang H, Williams A, Blume JE. **Discovery of tissue-specific exons using comprehensive human exon microarrays.** *Genome Biology* 2007, **8(4)**:R64.

Copois V, Bibeau F, Bascoul-Mollevi C, Salvetat N, Chalbos P, *et al*.: **Impact of RNA degradation on gene expression profiles: assessment of different methods to reliably determine RNA quality**. *Journal of Biotechnology* 2007, **127(4)**:549-59.

Dühring U, Axmann IM, Hess WR, Wilde A. **An internal antisense RNA regulates expression of the photosynthesis gene isiA.** PNAs 2006 103(18):7054-58.

Ellegren H. **Sequencing goes 454 and takes large-scale genomics into the wild.** *Molecular Ecology* 2008, **17**:1629-35.

The ENCODE Project Consortium. **Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.** *Nature* 2007, **447**:799-816.

Fire A, Xu S, Montgomery MK, Kostas SA,Driver SE, Mello CC. **Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans.** *Nature* 1998, **391**:806-811.

French PJ, Peeters J, Horsman S, Duijm E, Siccama I, van den Bent MJ, Luider TM, Kros JM, van der Spek P, Sillevis Smitt PA. **Identification of differentially regulated splice variants and novel exons in glial brain tumors using exon expression arrays.** *Cancer Research* 2007, **67(12)**:5635-42.

Gardina PJ, Clark TA, Shimada B, Staples MK, Yang Q, Veitch J, Schweitzer A, Awad T, Sugnet C, Dee S, Davies C, Williams A, Turpaz Y. **Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array.** *BMC Genomics* 2006, **7**:325.

Gautier L, Cope L, Bolstad BM, Irizarry RA: **affy: analysis of Affymetrix GeneChip data at the probe level.** *Bioinformatics* 2004, 20(3):307-15.

Gentleman RC, Carey VJ, Bates BM, Bolstad B, Dettling M, *et al*.: **Bioconductor: open software development for computational biology and bioinformatics.** *Genome Biology* 2004, **5(10)**:R80.

Gentleman R, Carey V, Huber W, Irizarry R, Dudoit S: *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. New York: Springer; 2005.

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. **Molecular classification of cancer: class discovery and class prediction by gene expression monitoring**. *Science* 1999, **286**:531–537.

Graveley BR. **Alternative splicing: increasing diversity in the proteomic world.** *Trends in Genetics* 2001, **17(2):**100-107.

Griffiths-Jones S. **The microRNA Registry.** *NAR 32(Database Issue)* 2004:D109-D111.

Heber S, Sick B: **Quality assessment of Affymetrix GeneChip data.** *OMICS: A Journal of Integrative Biology* 2006, **10(3)**:358-68.

Hertel KJ. **Combinatorial control of exon recognition.** *Journal of Biological Chemistry* 2008, **283(3)**:1211-5.

Hobert O. **Common logic of transcription factor and microRNA action.** *Trends in Biochemical Sciences* 2004, 29(9):462-8.

Hobert O. **Gene regulation by transcription factors and microRNAs.** *Science* 2008, 319:1785-1786.

Howard BE, Sick B, Heber S. **Practical Quality Assessment of Microarray Data by Simulation of Differential Gene Expression**. In Mandoiu I, Narasimhan G, Zhang Y (Eds.), *Proceedings of the 5th International Symposium on Bioinformatics Research and Applications 2009 (ISBRA 2009)*, Ft. Lauderdale, FL, p 18-27.

Howard BE, Sick B, Heber S. **Unsupervised Assessment of Microarray Data Quality Using a Gaussian Mixture Model.** *BMC Bioinformatics* 2009, **10:191**.

Jares P. **DNA microarray applications in functiona genomics.** *Ultrastructural Pathology* 2006, **30(3)**:209-19.

Kant MR, Baldwin IT. **The ecogenetics and ecogenomics of plant-herbivore interactions: rapid progress on a slippery road**. *Current Opinion in Genetics and Development*, **17(6)**:519-24.

Kent WJ. **BLAT: The BLAST-Like Alignment Tool.** *Genome Research* 2002, **12**:656-664.

Kwan T, Benovoy D, Dias C, Gurd S, Serre D, Zuzan H, Clark TA, Schweitzer A, Staples MK, Wang H, Blume JE, Hudson TJ, Sladek R, Majewski J. **Heritability of alternative splicing in the human genome.** *Genome Research* 2007, **17(8)**:1210-18.

Li X, Quigg RJ, Zhou J, Gu W, Rao PN, Reed EF. **Clinical Utility of Microarrays: Current Status, Existing Challenges and Future Outlook**. *Current Genomics* 2008, 9: 466-474.

Making the most of microarrays. *Nature Biotechnology* 2006, **24**:1039.

Mardis ER. **Next-generation DNA sequencing methods.** *Annual Review of Genomics and Human Genetics* 2008, **9**:387-402.

Margulies M, *et al*. **Genome sequencing in microfabricated high-density picolitre reactors.** *Nature* 2005, **437**:376-80.

Marton MJ, DeRisi JL, Bennet HA, Iyer VR, Meyer MR, *et al.* **Drug target validation and identification of secondary drug target effects using DNA microarrays.** *Nature Medicine* 1998, **4(11)**:1293-301.

Metsker ML. **Sequencing technologies – the next generation.** *Nature Reviews Genetics* 2008, **9(11)**, poster.

Morgan M, Anders S, Lawrence M, Aboyoun P, Pages H, Gentleman R. **ShortRead: a Bioconductor package for input, quality assessment, and exploration of high throughput sequence data**. *Bioinformatics* 2009, August 3 [Epub ahead of print].

Nuwaysir EF, Bittner M, Trent J, Barrett JC, Afshari CA. **Microarrays and toxicology: the advent of toxicogenomics.** *Molecular Carcinogenesis* 1999, **24(3)**: 153-9.

Pray LA. **Eukaryotic genome complexity**. *Nature Education* 2008, **1(1)**.

Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang CH, Angelo M, Ladd C, Reich M, Latulippe E, Mesirov JP, Poggio T, Gerald W, Loda M, Lander ES, Golub TR.

Multiclass cancer diagnosis using tumor gene expression signatures. *Proc Natl Acad Sci USA* 2001, **98**:15149–15154.

Reimer M, Weinstein JN: **Quality assessment of microarrays: visualization of spatial artifacts and quantitation of regional biases.** *BMC Bioinformatics* 2005, **6**:166.

Sambrook J, Russel DW. *Molecular Cloning: A Laboratory Manual, 34d ed*. Cold Spring Harbor Laboratory Press 2001, Cold Spring Harbor, NY.

Sarnow P, Jopling CL, Norman KL, Schutz S, Wehner KA. **MicroRNAs: expression, avoidance and subversion by vertebrate viruses.** *Nature Reviews, Microbiology* 2006, 4:651-59.

Schroeder A, Mueller O, Stocker S, Salowsky R, Leiber M, *et al*. **The RIN: an RNA integrity number for assigning integrity values to RNA measurements.** *BMC Molecular Biology* 2006, **7:3**.

Sharp PA. **The centrality of RNA.** *Cell* 2009, **136(4)**:577-580.

Stokes TH, Moffitt RA, Phan JH, Wang MD: **chip artifact CORRECTion (caCORRECT): a bioinformatics system for quality assurance of genomics and proteomics array data.** *Annals of Biomedical Engineering* 2007, **35(6)**:1068-80.

Singh D, Febbo PG, Ross K, Jackson DG, Manola J, Ladd C, Tamayo P, Renshaw AA, D'Amico AV, Richie JP, Lander ES, Loda M, Kantoff PW, Golub TR, Sellers WR. **Gene expression correlates of clinical prostate cancer behavior.** *Cancer Cell* 2002, **1**:203–209.

van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH. **Gene expression profiling predicts clinical outcome of breast cancer.** *Nature* 2002, **415**:530–536.

Velculescu VE, Zhang L, Zhou W, Vogelstein, Basrai MA, *et al.* **Characterization of the Yeast Transcriptome.** *Cell* 1997, **88(2):**243-251.

Wang T, Hopkins D, Schmidt C, Silva S, Houghton R, Takita H, Repasky E, Reed SG. **Identification of genes differentially over-expressed in lung squamous cell carcinoma using combination of cDNA subtraction and microarray analysis**. *Oncogene* 2000, **19**:1519–1528.

Wang Z, Gerstein M, Synder M. **RNA-Seq: a revolutionary tool for transcriptomics**. *Nature Reviews Genetics* 2009, **10(1)**:57-63.

Watanabe T, Taked A, Tsukiyam T, Mise K, Okuno T, *et al.* **Identification and characterization of two novel classes of small RNAs in the mouse germline: retrotransposon – derived siRNAs in oocytes and germline small RNAs in testes**. *Genes and Development* 2006, **20(13)**:1732-1743.

Wilson CL, Miller CJ. Simpleaffy: a BioConductor package for Affymetrix Quality Control and data analysis. *Bioinformatics* 2005, 21(18):3683-5.

Chapter 2

UNSUPERVISED ASSESSMENT OF MICROARRAY DATA QUALITY USING A GAUSSIAN MIXTURE MODEL

Brian E Howard, Beate Sick, Steffen Heber

***BMC Bioinformatics* 2009, 10:191**

Abstract

Quality assessment of microarray data is an important and often challenging aspect of gene expression analysis. This task frequently involves the examination of a variety of summary statistics and diagnostic plots. The interpretation of these diagnostics is often subjective, and generally requires careful expert scrutiny. We show how an unsupervised classification technique based on the Expectation-Maximization (EM) algorithm and the naïve Bayes model can be used to automate microarray quality assessment. The method is flexible and can be easily adapted to accommodate alternate quality statistics and platforms. We evaluate our approach using Affymetrix 3' gene expression and exon arrays and compare the performance of this method to a similar supervised approach. This research illustrates the efficacy of an unsupervised classification approach for the purpose of automated microarray data quality assessment. Since our approach requires only unannotated training data, it is easy to customize and to keep up-to-date as technology evolves. In contrast to other "black box" classification systems, this method also allows for intuitive explanations.

Background

Recently, the MicroArray Quality Control (MAQC) consortium found that most microarray platforms will generate reproducible data when used correctly by experienced researchers (Shi, *et al.*, 2006). Despite this positive result, it has been suggested that 20% or more of the data available in public microarray data repositories may be of questionable quality (Larsson *et al.*, 2006). For this reason, discriminating between high and low quality

microarray data is of the highest importance, and several recent publications have dealt with this problem; detailed reviews are provided by Wilkes *et al.* (2007) and Eads *et al.* (2006).

Several approaches have emphasized the importance of measuring, either directly or indirectly, the integrity of the RNA samples used in the experiment (e.g. Copois *et al.*, 2007; Archer *et al.,* 2006; Jones *et al., 2006*). Other research has focused on spatial artifacts: problems that typically arise during hybridization due to bubbling, scratches and edge effects (Reimer *et al*., 2005; Stokes *et al*., 2007).

In the case of Affymetrix GeneChips, which we will use to demonstrate our method, there are standard benchmark tests provided by the manufacturer (Affymetrix, Inc., 2003). A standard complementary approach is to use the R statistical software, along with the BioConductor (Gentleman *et al*., 2004) "affy" (Gautier *et al*., 2004) and "affyPLM" (Bolstad, 2007) packages, to produce a series of diagnostic plots for the assessment of GeneChip quality (see Figure 1-6 and Figure 1-7). A review of the quality control features available in BioConductor can be found in (Gentleman *et al.*, 2005), and a variety of software packages are now available to assist in the automation of this process (Heber and Sick, 2006; Psarros, *et al*. 2005; Howard, *et al.* 2007; Lee, *et al.*2006; Lozano and Kalko, 2006).

In general, the goal of these approaches is to identify chips that are outliers - either in relation to other chips in the same experiment or the entire theoretical population of similar chips. Often, it is assumed that a rational decision regarding data quality is made only after considering several quasi-orthogonal dimensions of quality. Chips are typically rejected only

after a preponderance of the evidence indicates poor quality; a slightly unusual score on a single metric is frequently ignored, while a number of moderately or highly unusual scores on a variety of quality metrics is often grounds for exclusion of a particular chip from further analysis. However, there are no universal, robust thresholds available for the identification of outliers according to the various quality variables. Instead, decisions are necessarily made using historical data, either implicitly or explicitly.

Therefore, recent efforts have focused on providing a "holistic", accurate, and automatic interpretation of diagnostic plots and quality metrics. Burgoon *et al.* (2005) describe a custom, in-house protocol for assessing data quality of two-color spotted cDNA arrays. The authors advocate an integrated "Quality Assurance Plan" which attempts to integrate quality control at every level of the experimental procedure. Another example is the RACE system (Heber and Sick, 2006; Psarros, *et al*. 2005). This system utilizes various statistics extracted from the BioConductor diagnostic plots, along with a random forest classifier, to automatically identify low quality data. However, like the quality assurance protocol described by Burgoon *et al.*, the RACE system relies on a large expert-annotated data set. For this reason, it is difficult to keep the system up-to-date in the face of rapidly changing technology, with new chip types continually being introduced into the market. A further challenge is to adapt such a system to similar, but slightly different, types of data such as Affymetrix SNP arrays, exon arrays, or arrays produced by other manufacturers such as Illumina and Agilent.

In this paper we investigate a method for unsupervised classification that was designed with these considerations in mind. First, we describe how to frame the

interpretation of microarray quality indicators as an unsupervised classification problem using a Gaussian mixture model. We show how the model parameters can be estimated using the Expectation-Maximization (EM) algorithm (Dempster *et al*., 1977), and how they can be used to construct a Naïve Bayes classifier for identifying low quality data.

Previous work has demonstrated that naïve Bayes classifiers perform well with labeled training sets in the supervised version of the problem discussed in this paper (Heber and Sick, 2006). The combination of Naïve Bayes together with EM has been used with considerable success in other problem domains, including text classification (Nigam, *et al*., 2000). Gaussian mixture models have been applied to automatic quality assessment of phone signal clarity (Falk and Chan, 2004) and mass spectrometry data (Wong, *et al*., 2007), and in other stages of the microarray processing pipe-line, including identification of differentially expressed genes (Najarian, *et al.*, 2004), assessment of the concordance between sets of similar microarray data sets (Lai, *et al.*, 2007), and even quality control at the spot detection and image fluorescence analysis level (Asyali and Alci, 2005). However, this is the first research we are aware of that employs this estimation approach, in conjunction with a naïve Bayes classifier, for the purpose of array-level quality control of microarray data.

In the following sections, we describe the datasets used in this research, and explain the implementation of both the supervised and unsupervised versions of the quality classifier. We demonstrate that the performance of the unsupervised classifier is comparable to a supervised classifier constructed from expert-labeled data. We also apply the algorithm to Affymetrix exon array data, and compare the observed quality indicator distributions with those obtained from 3' expression arrays.

Methods

Datasets

Our first dataset is a set of 603 Affymetrix raw intensity microarray data files, from 32 distinct experiments downloaded from the NCBI GEO database (Edgar, *et al*., 2002). A variety of Affymetrix GeneChip 3' Expression array types are represented in the dataset, including:

- **ath1121501** (Arabidopsis, 248 chips); GEO accession numbers: GSE5770, GSE5759, GSE911 (William *et al*., 2004), GSE2538 (Ramonell *et al*., 2005), GSE3350 (Vanneste *et al*., 2005), GSE3416 (Bläsing *et al*., 2005), GSE5534, GSE5535, GSE5530, GSE5529, GSE5522, GSE5520, GSE1491 (Armstring *et al*., 2004), GSE2169, GSE2473
- **hgu133a** (human, 72 chips); GSE1420 (Kimchi *et al*., 2005), GSE1922
- **hgu95av2** (human, 51 chips); GSE1563 (Flechner *et al*., 2004)
- **hgu95d** (human, 22 chips); GSE1007 (Haslett *et al*., 2003)
- **hgu95e** (human, 21 chips); GSE1007
- **mgu74a** (mouse, 60 chips); GSE76, GSE1912 (Lin *et al*., 2004)
- **mgu74av2** (mouse, 29 chips); GSE1947 (Giambonini-Brugnoli *et al*., 2005), GSE1419 (Chen *et al*., 2005; Herman *et al*., 2004)
- **moe430a** (mouse, 10 chips); GSE1873 (Li *et al*., 2005)

- **mouse4302** (mouse, 20 chips); GSE5338 (Cheng *et al*., 2006), GSE1871 (Jacobson *et al*., 2005)
- **rae230a** (rat, 26 chips); GSE1918, GSE2470
- **rgu34a** (rat, 44 chips); GSE5789 (Ovanod *et al*., 2006), GSE1567 (Gonzalez *et al*., 2005), GSE471 (Fischer *et al*., 2002).

These experiments cover many of the species commonly analyzed using the GeneChip platform, and were selected to represent a variety of tissue types and experimental treatments.

The BioConductor rma() function was used to perform probeset summarization, background subtraction and quantile normalization, with each raw intensity (.CEL) file preprocessed together with the other chips from the same GEO experiment. A variety of quality control indicators, listed in Table 2-1, were then computed for each chip. (For a list of all the .CEL files and their GEO identifiers, along with quality control feature scores and expert annotations, see http://www.biomedcentral.com/1471-2105/10/191/additional, "Additional file 2.") Also included in the file are descriptions explaining how each of the 29 quality control feature scores is computed from the raw expression data.

The second dataset consists of all of the exon array .CEL files available in the GEO database at the time of this analysis (540 .CEL files). Fourteen different experiments are represented: GSE10599 (Zhang *et al*., 2008), GSE10666 (Sandberg *et al*., 2008), GSE11150 (Chahrour *et al*., 2008), GSE11344 (Xing *et al*., 2008), GSE11967 (Soreq *et al*., 2008), GSE12064 (Douglas *et al*., 2008), GSE6976 (Platts *et al*., 2007), GSE7760 (Hu *et al*., 2008),

Table 2-1. BioConductor quality control statistics

Quality Statistic [1]	**Description**
mean.raw.int, sd.raw.int, median.raw.int, interQuartile.raw.int	mean, standard deviation, median and inter-quartile range of raw log intensity distribution.
q.5.raw.int, q.95.raw.int	5th and 95th percentile of raw log intensity distribution.
slope.bias, p.bias	slope parameter and associated p-value of linear regression of log expression level versus probe number, as computed by R affy library function AffyRNAdeg().
mean.norm.int, sd.norm.int, median.norm.int, interQuartile.norm.int, q.5.norm.int, q.95.norm.int	mean, standard deviation, median, inter-quartile range, and 5th and 95th percentiles of normalized log intensity distribution.
PLM.w.q.0.001, PLM.w.q.0.01, PLM.w.q.0.1, PLM.w.q.0.2	0.1th, 1st, 10th and 20th percentile of the probe-level model weights, computed using affyPLM library functionality.
PLM.res.q.0.01, PLM.res.q.0.1, PLM.res.q.0.25, PLM.res.q.0.75, PLM.res.q.0.9, PLM.res.q.0.99	1st, 10th, 25th, 75th, 90th, and 99th percentile of probe-level model residuals, computed using affyPLM library functionality.
RLE.median, RLE.interQuartile, RLE.lower.whisker, RLE.upper.whisker	median, inter-quartile range, lower tail and upper tail of "relative log intensity", computed using affyPLM library functionality.

[1.] The "SCORE" function was used to normalize values for each statistic, *t*, for each chip, *i*, relative to the values observed in other chips from the same experiment:

$SCORE(t_i) = \frac{t_i - median(t)}{mad(t)}$; with *median*() and *mad*() computed across all chips in the experiment.

GSE7761 (Huang *et al*., 2007), GSE8945 (Hung *et al*., 2008), GSE9342, GSE9372 (Kwan *et al*., 2008), GSE9385 (French *et al*., 2007), GSE9566 (Cahoy *et al*., 2008). The dataset includes examples of the Mouse Exon 1.0 ST array and several versions of the Human Exon 1.0 ST array. This dataset was processed using two different methods. First, the same set of quality indicators described above for the 3' expression dataset was prepared using the BioConductor packages in R. The "aroma" .cdf annotation files (Bengtsson *et al*., 2008) were used to read in expression values for the core probes on the arrays. In addition, this second dataset was also processed using the Affymetrix Expression Console software. Only the "core" probesets were considered and the software was used to perform "gene-level" probeset summarization, background subtraction and quantile normalization using the "RMA sketch" option in the software. Several alternative quality indicators were then computed (Table 2-2). A list of the .CEL files and their GEO identifiers and also the various quality control feature scores is can be found in "Additional file 3," available at http://www.biomedcentral.com/1471-2105/10/191/additional. Detailed descriptions of the Affymetrix Expression Console quality control features can be found in (Affymetrix, Inc, 2007).

Expert Annotation

A domain expert analyzed the 3' expression dataset (dataset 1) and assigned quality scores according to a procedure which is based on experience gained during almost three years of bioinformatics support within the Lausanne DNA Array Facility (DAFL). This quality control procedure is described in (Heber and Sick, 2006). Briefly, the chip scan

Table 2-2. Affymetrix Expression Console quality control statistics (exon arrays)

Quality Statistic [1]	**Description**
pm.mean	mean of the raw intensity for all PM probes, prior to any normalizations.
bgrd.mean	mean of the raw intensity for all probes used to compute background intensity. (Note: may be higher than pm.mean because GC compositions of probes used to compute background and PM probes can be quite different.)
pos.vs.neg.auc	area under ROC curve discriminating between positive control probesets and negative control probesets.
probeset.mean, probeset.stdev	mean and standard deviation of probeset signals after normalization. [2]
probeset.mad.residual.mean, probeset.mad.residual.stdev	mean and standard deviation of the absolute deviations of the RMA probe level model residuals from the median across chips . [2]
probeset.rle.mean, probeset.rle.stdev	mean and standard deviation of the absolute values of the relative log expression (RLE) for all probesets. [2]

[1.] The "SCORE" function was used to normalize values for each statistic, *t*, for each chip, *i*, relative to the values observed in other chips from the same experiment:

$$SCORE(t_i) = \frac{t_i - median(t)}{mad(t)};$$ with *median*() and *mad*() computed across all chips in the experiment.

[2.] Separate statistics are computed for a) all probesets, b) negative control probesets, and c) positive control probesets.

images and the distributions of the log scale raw PM intensities are visualized. Smaller discrepancies between chips are common and can often be removed by normalization. Remaining discrepancies usually indicate low quality data, possibly caused by problems in the amplification or labeling step. The general 5' to 3' probe intensity gradient averaged over all probe sets on a chip is also examined. The slope and shape of the resulting intensity curves depend on the RNA sample source, the amplification method, and the array type. In general, the specific shape of the curves is less important for the quality check than their agreement across the experiment. Pseudo-images representing the spatial distribution of residuals and weights derived from the probeset summarization model are very important diagnostics. Small artifacts are not critical when using robust analysis methods; however, extended anomalies are taken as an indication of low quality. In addition, box plot representations of the Normalized Un-scaled Standard Error (NUSE) from the probe level model fit and the Relative Log Expression (RLE) between each chip and a median chip are examined. These plots are used to identify problematic chips showing an overall deviation of gene expression levels from the majority of all measured chips. A chip may be judged as having poor quality if it is an apparent outlier in the experiment-wide comparison of several quality measures. Each array was given a quality score of 0, 1 or 2, with 0 being “acceptable quality” (519 chips), 1 being “suspicious quality” (56 chips) and 2 being “unacceptable quality” (28 chips). For the purposes of classification, chips with scores of 1 or 2 were combined into the composite “low quality” class.

Supervised Naïve Bayes Classifier

Previous research has demonstrated that quality assessment of microarray data can be successfully automated with the use of a supervised classifier (Heber and Sick 2006; Burgoon *et al*., 2005). The goal of supervised classification is to utilize an annotated training dataset to learn a function that can be used to correctly classify unlabeled instances. In the case of microarray quality assessment, the training dataset consists of the quality control features computed for each chip, combined with the quality annotation for each chip.

By making the simplifying assumption that all features are conditionally independent, naïve Bayes classifiers attempt to directly model the probability that a particular data point belongs to each class. Given the class label, each feature is assumed to follow an independent, univariate distribution. These distributions are, of course, unknown, but the maximum likelihood parameter estimates can be determined from a labeled training set. Then, for each unlabeled instance, Bayes' rule can be applied to compute the conditional probability that the instance belongs to each of the possible classes. Because we had prior success performing classification on a similar data set using Naïve Bayes with Gaussian feature distributions (Heber and Sick, 2006), we again chose to model the features using independent normal distributions. However, the approach could easily be adapted to use alternative distributions, for example, Student's t-distribution or the skew-normal distribution.

Under this framework, the probability that an unlabeled instance belongs to the low quality class is estimated as follows:

$$\Pr\{c=1 \mid \vec{x}\} = \frac{\left(\prod_{i=1}^{p} f(\vec{x}^{(i)} \mid c=1)\right)\Pr\{c=1\}}{f(\vec{x})} \quad (2.1)$$

where $c \in \{0,1\}$ signifies the class label, with 0 denoting "high quality" and 1 denoting "low quality," $\vec{x}$ is a length p vector of features describing the unlabeled instance., and $f(\vec{x}^{(i)} \mid c=1)$ is the Gaussian density for the i^{th} feature, among low quality chips.

The marginal probability of observing a low quality chip, $\Pr\{c=1\}$, can be estimated from the proportion of low quality chips in the training set. Furthermore, the marginal density for a particular combination of feature values, $f(\vec{x})$, independent of the class label, is equal to:

$$f(\vec{x}) = \left[\left(\prod_{i=1}^{p} f(\vec{x}^{(i)} \mid c=0)\right)\Pr\{c=0\}\right] + \left[\left(\prod_{i=1}^{p} f(\vec{x}^{(i)} \mid c=1)\right)\Pr\{c=1\}\right] \quad (2.2)$$

For the purposes of classification, this algorithm assigns class 1 to an unlabeled instance $\vec{x}$, if $\Pr(c=1 \mid \vec{x}\} > t$, where t is a threshold parameter, ordinarily set to 0.5 in order to approximate the Bayes optimal decision rule. By varying this parameter, it is also possible to construct ROC curves which display the tradeoff between sensitivity and specificity for various decision thresholds.

Unsupervised Naïve Bayes Classifier

The standard (supervised) approach to constructing a naïve Bayes classifier employs maximum likelihood estimation to infer the distribution parameters of each classification feature from an expert-annotated training set. It is, however, also possible to construct an "unsupervised" naïve Bayes classifier by using an unannotated dataset as input. In this case,

the EM algorithm is used to infer the feature distributions, assuming an appropriate Gaussian mixture model, as described in the following section.

Gaussian Mixture Model and the EM Algorithm

The naïve Bayes classification model described above requires parameter estimates for the quality control metrics, conditional on each quality class. In the absence of annotated data, however, the quality classes of the unannotated training instances are additional unknowns that must be estimated along with the distributional parameters. We model the unannotated dataset using a Gaussian Mixture Model, under the assumption that microarray data can be reasonably classified into the dichotomy of "high quality" and "low quality" chips, and that the unlabeled training set contains examples of each.

Given a large set of microarray data files, the first step is to compute values for each of the various quality control features. Then, for each feature, we assume that the observed distribution of scores is generated by an underlying Gaussian mixture model with two components: 1) chips having high quality and 2) chips having low quality. Given the mixture component, $c \in \{0,1\}$, each feature is assumed to follow a Normal $\left(\mu_c, \sigma_c^2\right)$ distribution. However, in the case of an unlabeled dataset, the true mixture component is unknown. We further assume that, marginally, the class label for each instance is a simple Bernoulli random variable with probability ϕ of indicating a low quality chip. Under this model, the (log) likelihood of the dataset is:

$$\log L(\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \vec{\phi} \mid \boldsymbol{x}) = \sum_{i=1}^{N} \log(f(\vec{x}_i; \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \vec{\phi}) = \sum_{i=1}^{N} \log \sum_{j=0}^{1} f(\vec{x}_i \mid \vec{c}^{(i)} = j; \vec{\mu}_j, \vec{\sigma}_j^2) \Pr\{\vec{c}^{(i)} = j; \vec{\phi}\} \quad (2.3)$$

where:

- $\boldsymbol{x}$ is an $N \times p$ matrix containing the p feature values for the N items in the dataset, with $\vec{x}_i$ denoting the length p feature vector for the i^{th} data point.
- $\boldsymbol{\mu}$ is a $2 \times p$ parameter matrix containing, in each column, μ_0 and μ_1 for the p^{th} feature; $\vec{\mu}_j$ is the length p parameter vector for the j^{th} Gaussian mixture component ($j \in \{0,1\}$).
- $\boldsymbol{\sigma}^2$ is a $2 \times p$ parameter matrix containing, in each column, σ_0^2 and σ_1^2 for the p^{th} feature; $\vec{\sigma}_j^2$ is the length p parameter vector for the j^{th} Gaussian mixture component.
- $\vec{c}$ is a length N vector containing the (unknown) class labels for each of the N data points.
- $\vec{\phi}$ is a length 2 probability vector containing the probability that a randomly chosen data point belongs to each class.

The likelihood function in equation 2.3 can be maximized using the EM algorithm (Dempster *et al*., 1977). The EM algorithm is a well-known method for maximizing mixture model likelihood functions by iteratively performing two steps:

- **E Step**: Estimate the unknown class labels, based on the current estimates for the other parameters.

- **M Step**: Given current class labels, compute the maximum likelihood estimators for the parameters $\boldsymbol{\mu}$, $\boldsymbol{\sigma}^2$, and $\vec{\phi}$.

To implement the EM algorithm, we introduce an additional $N \times 2$ matrix, $\boldsymbol{w}$, which contains, for each data point, i, the current guesses for $p(\vec{c}^{(i)}=0)$ and $p(\vec{c}^{(i)}=1)$. After initializing all parameters and the weight matrix, w, to random values, the EM algorithm proceeds as follows:

M step: For $j \in \{0,1\}, k \in \{1...p\}$

$$\vec{\phi}^{(j)} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{w}_{ij} \quad (2.4)$$

$$\boldsymbol{\mu}_{jk} = \frac{\sum_{i=1}^{N} \boldsymbol{w}_{ij}\boldsymbol{x}_{ik}}{\sum_{i=1}^{N} \boldsymbol{w}_{ij}} \quad (2.5)$$

$$\boldsymbol{\sigma}_{jk}^2 = \frac{\sum_{i=1}^{N} \boldsymbol{w}_{ij}\left(\boldsymbol{x}_{ik} - \boldsymbol{\mu}_{jk}\right)^2}{\sum_{i=1}^{N} \boldsymbol{w}_{ij}} \quad (2.6)$$

E Step: For $i \in \{1...N\}, j \in \{0,1\}$

$$\begin{aligned} \boldsymbol{w}_{ij} &= \Pr(c_i = j \mid \vec{x}_i; \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \vec{\phi}) \\ &= \frac{\left(\prod_{k=1}^{p} normpdf(\boldsymbol{x}_{ik}, \boldsymbol{\mu}_{jk}, \boldsymbol{\sigma}_{jk}^2)\right)\vec{\phi}^{(j)}}{\sum_{l=0}^{1}\left(\left(\prod_{k=1}^{p} normpdf(\boldsymbol{x}_{ik}, \boldsymbol{\mu}_{lk}, \boldsymbol{\sigma}_{lk}^2)\right)\vec{\phi}^{(l)}\right)} \end{aligned} \quad (2.7)$$

where $normpdf(x, \mu, \sigma^2)$ denotes the probability density of a normal distribution evaluated at x. Because the algorithm can possibly converge to local optima, it is prudent to run the algorithm several times after random restarts. Additionally, each $\boldsymbol{\sigma}_{jk}^2$ was constrained to be

>= .001 to avoid convergence to a trivial solution. Further details concerning this implementation of the EM algorithm and the associated Gaussian mixture model can be found in (Ng 2006). Once estimates have been obtained for μ, σ^2 and $\vec{\phi}$, any unlabeled instance can be classified according to these mixture components using naïve Bayes, according to equation 2.1 (or equivalently, equation 2.7, in the case of the original unlabeled dataset). Since our assumption is that low quality chips are outliers with respect to these quality features, we use the mixture component corresponding to the smallest value from $\vec{\phi}$ to identify the low quality class.

Feature Selection

In order to achieve optimal classification performance, it is important to select an appropriate subset of the classification features. Ideally, this subset should include independent features that are each individually predictive of the class label.

To measure the ability of each feature to predict the correct class label in a training set (where "correct" label is defined as either the expert annotation in the supervised case, or the estimated $\boldsymbol{w}$ matrix in the unsupervised case), we first constructed an $N \times p$ score matrix, S, where each cell S_{ij} contains a distance measuring the discrepancy between the true and predicted class for data point $\vec{x}_i$, given the j^{th} feature and the parameter estimates for that feature:

$$\boldsymbol{S}_{ij} = \left| \vec{c}^{(i)} - \frac{normpdf(\boldsymbol{x}_{ij}, \boldsymbol{\mu}_{1j}, \boldsymbol{\sigma}_{1j}^2)\vec{\phi}^{(1)}}{normpdf(\boldsymbol{x}_{ij}, \boldsymbol{\mu}_{0j}, \boldsymbol{\sigma}_{0j}^2)\vec{\phi}^{(0)} + normpdf(\boldsymbol{x}_{ij}, \boldsymbol{\mu}_{1j}, \boldsymbol{\sigma}_{1j}^2)\vec{\phi}^{(1)}} \right| \quad (2.8)$$

Then for each feature, j, these scores were totaled across all N data points

$$\boldsymbol{S}_{+j} = \sum_{i=1}^{N} \boldsymbol{S}_{ij} \qquad (2.9)$$

Finally, the p scores were sorted in ascending order, to rank the features by their ability to predict the correct class label. Denote the rank of feature j according to the value of this score as $S_{[j]}$.

To identify correlations among the quality control features, we next computed the $p \times p$ Pearson correlation matrix. Let ρ_{jk} denote the correlation between features j and k, and $\rho_{[j]k}$ represent the *rank* of the correlation of feature j with feature k among all other features correlated with k, with features ranked in order of *descending* correlation. To select a subset of n features, we used the following forward selection algorithm:

- First, select the single feature that is most predictive of the class labels, i.e. the feature with $S_{[j]} = 1$.
- Then, sequentially, for the remaining n-1 features, select the feature j to satisfy:

$$\arg\min_{j} \left(c_1 S_{[j]} + \frac{c_2 \sum_{i \in F} \rho_{[j]i}}{|F|} \right) \qquad (2.10)$$

where F denotes the set of previously selected features. The constants c_1 and c_2 in this expression are weighting factors that can be modified to control the tradeoff between selection for independent features and features that are highly correlated with the class label. We used 0.5 for each.

Results and Discussion

Parameter Estimates

3' Expression Arrays

We applied the unsupervised mixture model described above to the 3' expression array data (hiding the expert quality labels). For nearly all of the 29 quality control features considered, the unsupervised EM parameter estimates very closely approximate the corresponding supervised MLE estimates, a result which indicates that the unsupervised approach was able to discover patterns in the data that are in agreement with the expert annotations. "Additional file 4" (available at http://www.biomedcentral.com/1471-2105/10/191/additional) contains the mixture model parameter estimates for $\vec{\phi}$, μ_0, μ_1, σ_0^2 and σ_1^2 for each of the quality control features. These estimates were obtained by applying the EM algorithm to the entire unlabeled dataset. For comparison, the table also includes the maximum likelihood estimates obtained using the expert-annotated class labels. Figure 2-1 shows some representative examples. Plots of this nature reveal that, in most cases, the EM and (supervised) MLE estimates exhibit only minor differences, generally with magnitudes analogous to the discrepancies shown in Figures 2.1a-d.

The EM estimates appear to be reasonable in all cases, given the original intent of each quality metric. For example, given the normalized (log-scale) expression values, the RLE metric measures the distribution of the quantity $M_{gi} = \hat{\theta}_{gi} - m_g$ for each chip, where $\hat{\theta}_{gi}$ is the log expression measurement for probeset g, on chip i, and m_g is the median expression

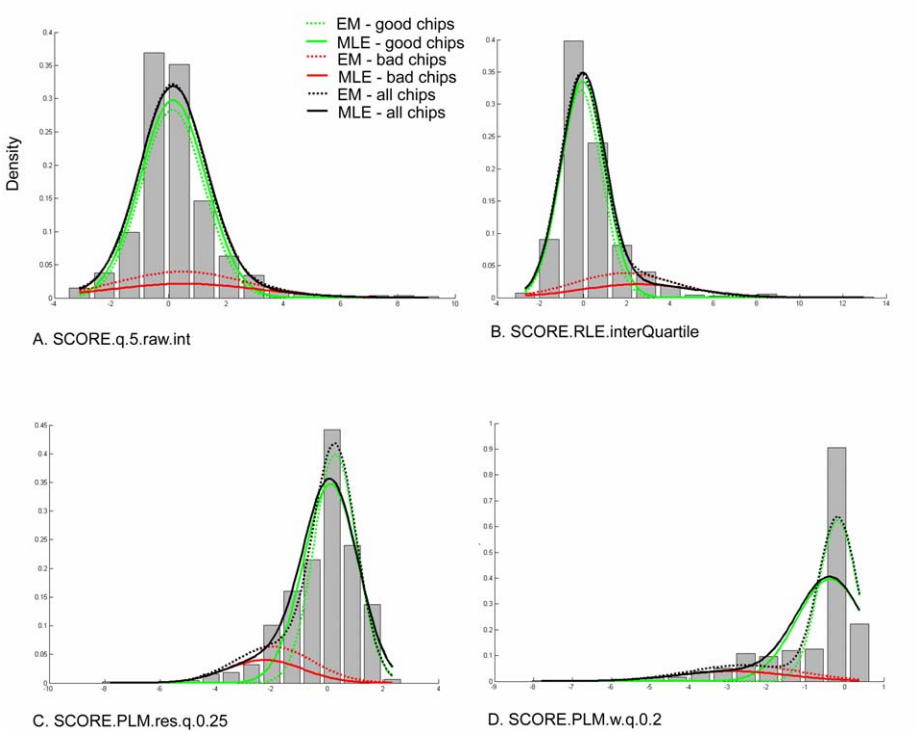


Figure 2-1. Mixture model parameter estimates.

Supervised (MLE) and Unsupervised (EM) estimates shown are for the following features from the 3' expression arrays: (A) 5th percentile of raw intensities, (B) inter-quartile range of the Relative Log Intensity (RLE), (C) 25th percentile of the probe-level model residuals, and (D) the 20th percentile of the probe-level model weights. All features were normalized relative to other chips in the same experiment, using the SCORE function (see Table 2-1).

of probeset g across all arrays. In general, since it is ordinarily assumed that the majority of genes are not differentially expressed across chips, the quantity M_{gi} is expected to be distributed with median 0. In addition, chips that more frequently have extreme expression values will have a large inter-quartile range for this statistic. Figure 2-1b indicates that, as expected, low quality chips were indeed more likely to have a large inter-quartile range for the RLE statistic.

Parameter estimates for the other metrics also agree with our expectations. For example, the estimates for metrics relating to probe-level model weights and residuals reflect the expectation that low quality chips should have larger residuals and more down-weighted probesets (Figure 2-1c,d). Similarly, the estimates indicate that low quality chips are more likely to have RNA degradation plots that are different from other chips in the same experiment. The low quality chips also tend to have both mean raw and mean normalized intensities that are either significantly higher or lower than other chips in the same experiment.

Exon Arrays

The Affymetrix exon array platform is different from the 3' expression array platform in several important ways (Robinson and Speed, 2007). For example, the 3' expression array targeting the human genome (Hgu133) has, on average, 1 probeset pair for each well-annotated gene; each probeset consists of 11 individual 25-mer probes, which primarily target the 3' region of the gene. In contrast, the Human Exon 1.0 ST array has 1 probeset for each exon for each gene in the target genome. Each probeset contains, in general, 4 (rather

than 11) 25-mer probes. Unlike 3' expression arrays, exon arrays lack mismatch probes. Instead, the background expression level for each probe is estimated by averaging the intensities of approximately 1000 surrogate genomic and anti-genomic background probes having the same GC content as the target probe. Because most genes consist of several exons, the median number of probes per gene is increased on the exon array from 11 on the 3' array to between 30-40 (Gardina *et al*., 2006). However, genes with fewer exons are covered by fewer probes. In fact, there are a few thousand well-annotated single exon genes covered by only 4 probes (Robinson and Speed, 2007). Furthermore, the feature size on the exon arrays has been reduced from 11x11 microns on the HGU133 array to 5x5 microns on the Human Exon 1.0 ST array (about 1/5 the area). This change may increase the expression variance, at least at the probeset level (Robinson and Speed, 2007). Exon arrays also utilize a different hybridization protocol which uses sense-strand labeled targets, and results in DNA-DNA hybridizations rather than the DNA-RNA hybridizations used with traditional 3' arrays (Abdueva *et al*., 2007). These differences suggest that the distributions of key quality control indicators may differ between the two platforms.

For the exon arrays, the resulting probability estimate for low quality chips was .397 - nearly twice what was obtained for the 3' arrays. This is reflected in Figure 2-2 as the larger areas under the red curves for exon arrays compared to 3' arrays, and as the smaller areas under the green curves for exon arrays compared to 3' arrays. For the majority of the indicators, the estimated distributions were qualitatively similar to those estimated for the 3' arrays (Figure 2-2a,c,d). One interesting difference is that in the exon arrays, the low quality chips appear to be more likely to have median raw intensity values that are lower than other

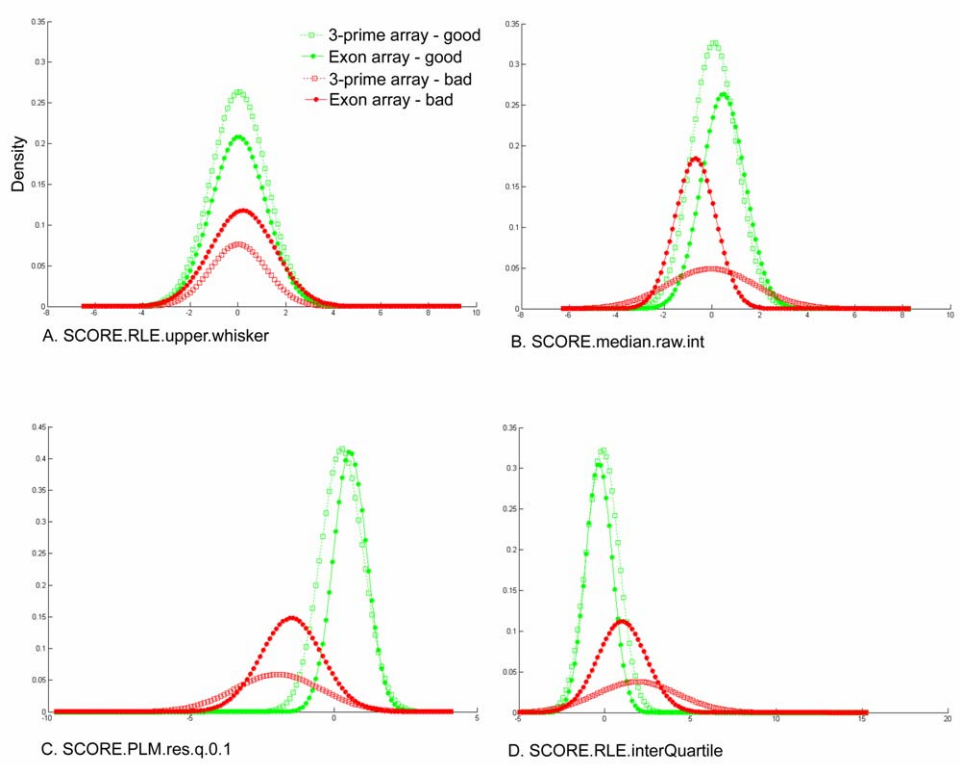


Figure 2-2. Comparison of parameter estimates for 3' expression arrays and exon arrays.

Each diagram illustrates the unsupervised Gaussian parameter estimates for one of the quality control features, for each of the two chip types. Estimates shown are for the following features: (A) Upper tail of the Relative Log Intensity (RLE), computed using the affyPLM functionality, (B) median of the raw intensity distribution, (C) 10th percentile of the probe-level model residuals, and (D) inter-quartile range of the RLE.

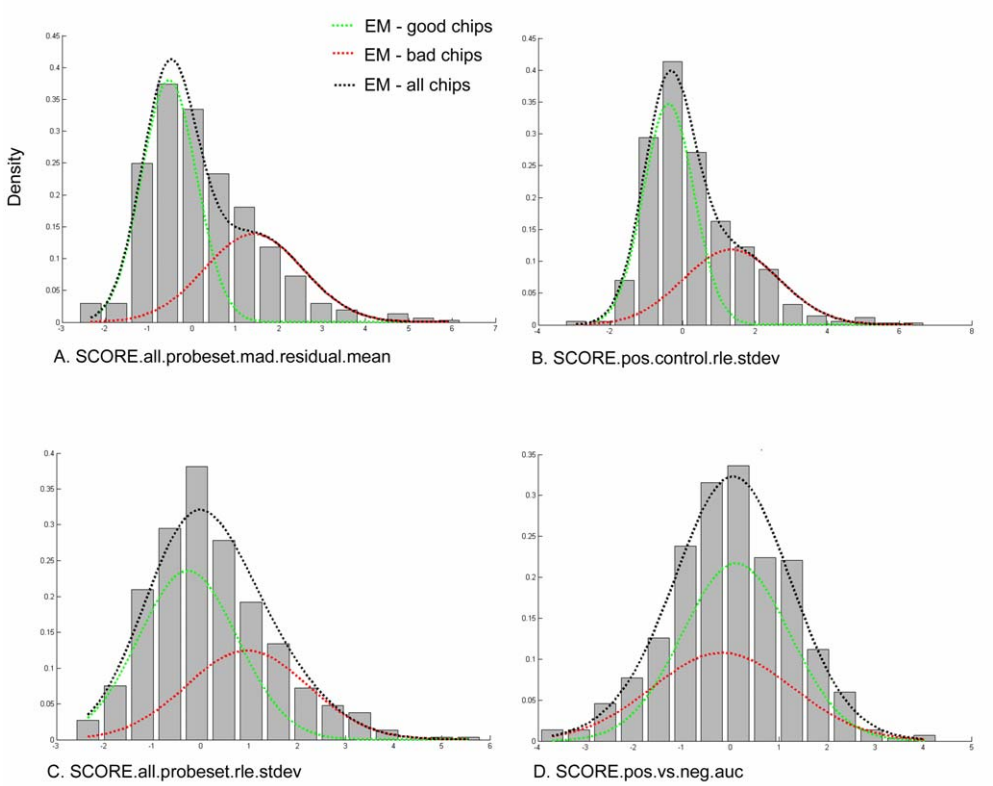
chips in the same experiment (Figure 2-2b), whereas for the 3' arrays, both abnormally high and low median raw intensities appear to be indicative of bad chips.

To check the robustness of our estimates, we also analyzed a separate set of quality control indicators (Table 2-2) computed using the Affymetrix Expression Console software. In agreement with the estimate obtained using the first set of quality metrics, the inferred probability for low quality chips was .394 using the Expression Console quality indicators. At a qualitative level, the estimates for the Expression Console quality indicators generally agreed with our expectations. For example, Figure 2-3a shows that, as expected, lower quality chips tend to have larger residuals when fitting the RMA probe-level summarization model. Similarly, Figure 2-3b and Figure 2-3c show that low quality chips are more likely to have higher variability in the RLE metric. Interestingly, the SCORE.pos.vs.neg.auc metric, which measures the area under an ROC curve discriminating between positive and negative controls, did not indicate a major difference between high and low quality chips. This seems to be in conflict with the recommendation by Affymetrix that this is potentially one of the most useful quality control indicators for exon arrays (Affymetrix, Inc, 2007). This observation could reflect the fact that labs detecting unusual values for this metric may have been more likely to exclude the corresponding chips from further analysis.

Classifier Performance Evaluation

3' Expression Arrays

After obtaining parameter estimates for various quality control features for the 3' expression arrays, we next sought to compare the performance of the unsupervised



and

Figure 2-3. Parameter estimates for exonarrya expression console QC features.

Shown are the parameter estimates obtained using the EM algorithm for various exon array quality control features available in the Affymetrix Expression Console software. Estimates shown are for the following features: (A) mean of the absolute deviation of the RMA probe level model residuals from the median across chips, (B) standard deviation of signal from positive control probesets after normalization, (C) standard deviation of signal from all probesets after normalization, and (D) area under ROC curve discriminating between positive control probesets and negative controls.

supervised classifiers. A 10-fold cross-validation procedure was used to compare the performance of naïve Bayes classifiers constructed using distribution parameters estimated using either the standard maximum likelihood method or, alternatively, the unsupervised mixture model approach. For each of 10 iterations, 9/10 of the 603 data instances were used as a training set, for both parameter estimation and also the selection of 5 classification features. For classifiers built using supervised MLE estimation ("MLE + Naïve Bayes"), the expert generated labels were used to distinguish between high and low quality chips in the training set. For the unsupervised classifier ("EM + Naïve Bayes"), the expert labels in the training set were ignored and the EM algorithm was used to estimate parameters of a Gaussian mixture model. The remaining unused 10^{th} of the data was used to assess the performance of the classifier, using the expert labels as the standard of truth. The performance of the two algorithms was nearly identical. The confusion matrices (Table 2-3) show the classification results for the two algorithms using a classification threshold of 0.5. The accuracy of the MLE + Naïve Bayes method was .907 with a false positive rate of .058, while the accuracy of the EM + Naïve Bayes method was .910 with a false positive rate of .079. An ROC curve, constructed by varying the classification threshold, is shown in Figure 2-4. The area under the ROC curve (AUC) was .9455 for the unsupervised method and .9402 for the supervised method. Although this performance is good, it is possible that these results could be improved even more by identifying and using alternative (other than normal) distributions to model one or more of the classification features.

In many real world scenarios the amount of unlabeled data available greatly exceeds the amount of expert-labeled data. To test the performance of the two classifiers under these

Table 2-3. Confusion matrices, full training set

		MLE + Naïve Bayes (Supervised)		
		0	1	◄ Expert Label
Classified As►	0	489	26	
	1	30	58	
		EM + Naïve Bayes (Unsupervised)		
		0	1	◄ Expert Label
Classified As►	0	478	13	
	1	41	71	

Table 2-4. Confusion matrices, given 30 labeled instances for supervised method

		MLE + Naïve Bayes (Supervised)		
		0	1	◄ Expert Label
Classified As►	0	492	44	
	1	27	40	
		EM + Naïve Bayes (Unsupervised)		
		0	1	◄ Expert Label
Classified As►	0	478	13	
	1	41	71	

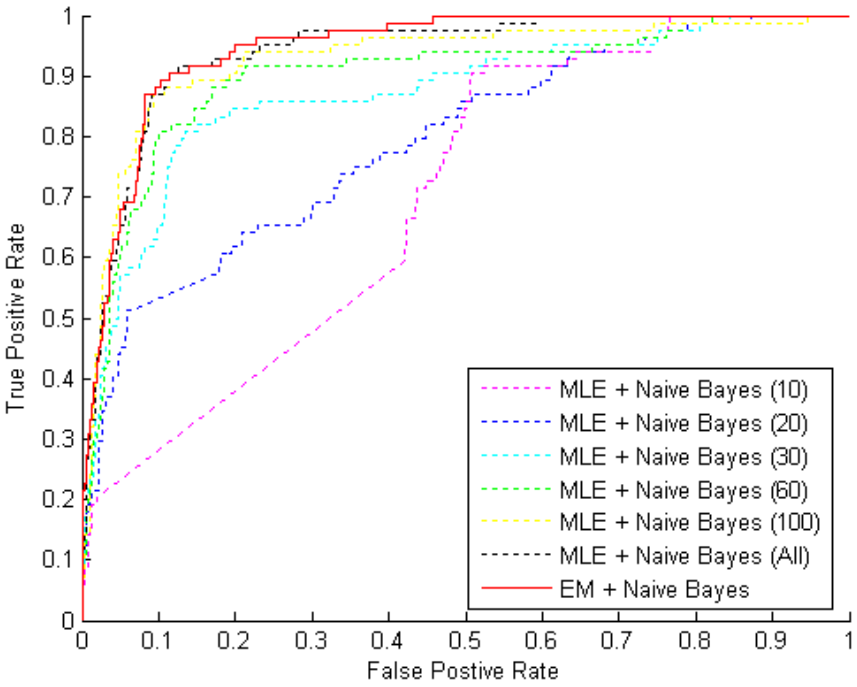


Figure 2-4. Classifier performance.

Unsupervised versus supervised classifier using labeled data sets of various sizes. When the full labeled training dataset (~540 labeled instances per fold) is available, the performance of the unsupervised classification method (EM+Naïve Bayes) and the supervised classification method (MLE+Naïve Bayes) are equivalent on the test dataset. When the amount of labeled data is limited, but unlabeled data is abundant, the unsupervised method outperforms the supervised method.

conditions, we performed additional 10-fold cross-validation experiments similar to the previous test. However, in this case, the supervised MLE + Naïve Bayes classifier was trained using random subsets of instances from each labeled training fold, while the EM + Naïve Bayes classifier was constructed using the entire unlabeled training fold. Subsets containing 10, 20, 30, 60, 75, and 100 instances were used to train the supervised classifier. The ROC curves in Figure 2-4 indicate that the EM + Naïve Bayes classifier appears to have an advantage when the amount of unlabeled training data available greatly exceeds the amount of expert-labeled data. For example, the unsupervised method clearly outperforms the supervised method when 30 or fewer labeled instances were available. Table 2-4 contains the resulting confusion matrix for the case in which 30 labeled training instances were used, with a classification threshold of 0.5.

Exon Arrays

To demonstrate the general applicability of our method, we constructed unsupervised classifiers using the two sets of quality control variables and the entire unlabeled training set. These classifiers were then used to predict classification labels for each data point. Figure 2-5 shows a Venn diagram comparing the classification results for classifiers constructed using the BioConductor quality features and the Expression Console quality features. In most, but not all, cases, the classifiers agree on the characterization of each chip with regard to quality. In addition, both classifiers agree that approximately 39% of the data is low quality. "Additional file 3" (available at http://www.biomedcentral.com/1471-

2105/10/191/additional) contains the classification labels obtained using unsupervised classifiers constructed using each set of quality variables.

Simulation Results

The agreement between the quality control feature distribution parameters estimated using the supervised maximum likelihood method and the estimates obtained with the unsupervised Gaussian mixture model suggests that our domain expert has uncovered a plausible dichotomy of chips within our dataset. To further confirm that the chips classified as having low quality were indeed more likely to negatively impact tests for differential expression, we performed a simple simulation. The procedure involved adding an offset to the observed expression measurements for a subset of the probesets on a set of "treatment" arrays, and then comparing these arrays with a set of unmodified "control" arrays sampled from the same experiment (details not shown). Among those chips designated by the expert as low quality, the majority (approximately 70%) impaired the ability to detect simulated differential expression when included in an analysis, compared to only about 10% of the chips classified as having high quality.

Conclusions

In this paper we have illustrated the efficacy of an unsupervised classification approach to assessing microarray data quality. Our method uses unlabeled training data to identify apparent distinctions between "good" and "bad" quality chips within the dataset. The method then integrates measurements obtained across a variety of quality dimensions into a

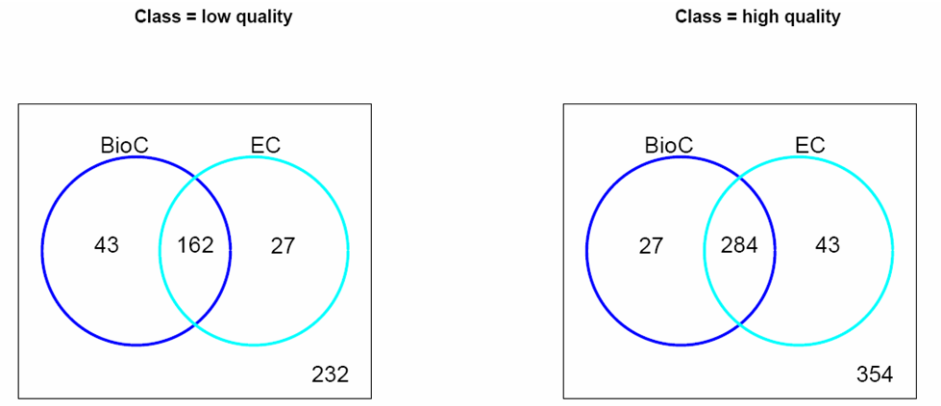


Figure 2-5. Exon arrays identified as high and low quality using two sets of QC indicators.

The Venn diagram displays the number of exon arrays classified as low quality using the BioConductor QC variables and the Expression Console QC variables (left panel), as well as the number of exon arrays classified as high quality using the same two sets of QC indicators (right panel).

single composite quality score which can be used to accurately identify low quality data.

Our method is flexible and can be easily adapted to accommodate alternate quality statistics and platforms. Because this technique requires only unannotated training data, it is easy to keep the resulting classifier up-to-date as technology evolves, and the adaptable nature of the system makes arbitrary, universal quality score thresholds unnecessary. Moreover, since a naïve Bayes classification approach involves the estimation of the underlying, univariate distributions for each of the classification parameters, this method allows for intuitive explanations that offer an advantage over other "black box" classification systems (Mozina *et al*., 2004; Pulin *et al*., 2006). For example, under this framework, it is possible to infer which diagnostic plots and features are most relevant for the classification of a particular chip. These plots can then be presented to the user in order to explain the

classification. A quality control method that incorporates an interpretation of standard diagnostic plots is an extension of a familiar process already used by many labs, and good diagnostic plots can provide powerful and convincing evidence of data quality artifacts.

An important caveat for this, and any quality control methodology, is that the decision about what to do with the detected low quality chip(s) is dependent on the experimental design, the number of low quality chips detected, and the magnitude of the defects encountered. In many cases, low quality chips still contain valuable information, and in some cases the most effective strategy may be to simply down-weight these chips rather than discarding them entirely (Ritchie *et al*., 2006).

Nevertheless, with the availability of a variety of rapidly growing public repositories for microarray data, the continual appearance of new microarray chip types, and the increasing usage of genomics data by research organizations worldwide, the development of robust and flexible methods for microarray quality assessment is now more important than ever. An advantage of the approach described in this paper is that, once a classifier has been constructed, the run-time required to automatically classify new instances is minimal. This makes the method ideal for use as a component of a batch processing system, such as a screening tool for use with public databases, or as a step in a meta-analysis pipeline.

Software Availability and Requirements

- ***Project name***: Unsupervised Assessment of Microarray Data Quality Using a Gaussian Mixture Model.

- ***Availability***: A Matlab implementation of these algorithms and the corresponding analyses is available in "Additional file 5" at http://www.biomedcentral.com/1471-2105/10/191/additional.
- ***Operating system***: Implemented and tested under Windows XP.
- ***Programming language***: Matlab 7.0.1.15, service pack 1.
- ***Other requirements***: Matlab Statistics Toolbox version 6.1.
- ***License***: Brian E. Howard. Free for non-commercial use.
- ***Any restrictions to use by non-academics***: Contact corresponding author.

Acknowledgements

Funded by the NCSU/EPA Cooperative Training Program in Environmental Sciences Research, Training Agreement CT833235-01-0 with North Carolina State University.

References

Abdueva D, Wing MR, Schaub B, Triche TJ: **Experimental comparison and evaluation of the Affymetrix exon and U133Plus2 GeneChip arrays**. *PLoS ONE* 2007, **2(9)**:e913.

Affymetrix, Inc.: *GeneChip expression analysis, data analysis fundamentals.* Affymetrix, Santa Clara, CA, 2003. Retrieved June 1, 2007 from http://www.affymetrix.com/support/downloads/manuals/data_analysis_fundamentals_manual.pdf.

Affymetrix, Inc.: *Quality assessment of exon and gene arrays*. Affymetrix, Santa Clara, CA, 2007. Retrieved October 3, 2008 from http://www.affymetrix.com/support/technical/whitepapers/exon_gene_arrays_qa_whitepaper.pdf

Archer KJ, Dumur CI, Joel SE, Ramakrishnan V: **Assessing quality of hybridized RNA in Affymetrix GeneChip experiments using mixed-effects models.** *Biostatistics* 2006, **7(2)**:198-212.

Armstrong JI, Yuan S, Dale JM, Tanner VN, Theologis A: **Identification of inhibitors of auxin transcriptional activation by means of chemical genetics in Arabidopsis**. *PNAS* 2004, **101(41)**:14978-83.

Asyali MH, Alci M: **Reliability analysis of microarray data using fuzzy c-means and normal mixture modeling based classification methods.** *Bioinformatics* 2005, **21(5)**:644-9.

Bengtsson H, Simpson K, Bullard J, Hansen K: **aroma.affymetrix: A generic framework in R for analyzing small to very large Affymetrix data sets in bounded memory.** Tech Report #745, Department of Statistics, University of California, Berkley, February 2008.

Bläsing OE, Gibon Y, Günther M, Höhne M, Morcuende R, *et al*.: **Sugars and circadian regulation make major contributions to the global regulation of diurnal gene expression in Arabidopsis**. *Plant Cell* 2005, **17(12)**:3257-81.

Bolstad, B: *affyPLM: methods for fitting probe-level models.* BioConductor version 2.0 package. Retrieved July 2, 2007 from http://bioconductor.org/packages/2.0/bioc/html/affyPLM.html

Burgoon LD, Eckel-Passow JE, Gennings C, Boverhof DR, Burt JW, *et al.*: **Protocols for the assurance of microarray data quality and process control.** *Nucleic Acids Research* 2005, 33(19):e172.

Cahoy JD, Emery B, Kaushal A, Foo LC, Zamanian JL, *et al.*: **A transcriptome database for astrocytes, neurons, and oligodendrocytes: a new resource for understanding brain development and function.** *Journal of Neuroscience* 2008, **28(1)**:264-78.

Chahrour M, Jung SY, Shaw C, Zhou X, Wong ST, *et al.*: **MeCP2, a key contributor to neurological disease, activates and represses transcription**. *Science* 2008, **320(5880)**:1224-9.

Chen Z, Herman AE, Matos M, Mathis D, Benoist C: **Where CD4+CD25+ T reg cells impinge on autoimmune diabetes.** *Journal of Experimental Medicine* 2005, **202(10)**:1387-97.

Cheng H, Aleman TS, Cideciyan AV, Khanna R, Jacobson SG, Swaroop A: **In vivo function of the orphan nuclear receptor NR2E3 in establishing photoreceptor identity during mammalian retinal development.** *Human Molecular Genetics* 2006, **15(17)**:2588-602.

Copois V, Bibeau F, Bascoul-Mollevi C, Salvetat N, Chalbos P, *et al.*: **Impact of RNA degradation on gene expression profiles: assessment of different methods to reliably determine RNA quality**. *Journal of Biotechnology* 2007, **127(4)**:549-59.

Dempster AP, Laird NM, and Rubin DB: **Maximum likelihood from incomplete data via the EM algorithm.** *Journal of the Royal Statistical Society, Series B* 1977, **39(1)**:1-38.

Douglas D, Hsu JH, Hung L, Cooper A, Abdueva D, *et al.*: **BMI-1 promotes ewing sarcoma tumorigenicity independent of CDKN2A repression**. *Cancer Research* 2008, **68(16)**:6507-15.

Eads B, Cash A, Bogart K, Costello J, Andrews J: **Troubleshooting microarray hybridizations.** *Methods in Enzymology* 2006, **411**:34-49.

Edgar R, Domrachev M, Lash AE: **Gene Expression Omnibus: NCBI gene expression and hybridization array data repository**. *Nucleic Acids Research* 2002, **30(1)**:207-10.

Falk TH, Chan W-Y: **Objective speech quality assessment using Gaussian mixture models**. In *22nd Biennial Symposium on Communications*, 31 *May – 3 June 2004; Ontario, Canada*: 2004.

Fischer MD, Gorospe JR, Felder E, Bogdanovich S, Pedrosa-Domellöf F, *et al*. **Expression profiling reveals metabolic and structural components of extraocular muscles**. *Physiological Genomics* 2002,**9(2)**:71-84.

Flechner SM, Kurian SM, Head SR, Sharp SM, Whisenant TC, *et al*.: **Kidney transplant rejection and tissue injury by gene profiling of biopsies and peripheral blood lymphocytes**. *American Journal of Transplantation* 2004, **4(9)**:1475-89.

French PJ, Peeters J, Horsman S, Duijm E, Siccama I, *et al*.: **Identification of differentially regulated splice variants and novel exons in glial brain tumors using exon expression arrays.** *Cancer Research* 2007, **67(12)**:5635-42.

Gardina PJ, Clark TA, Shimada B, Staples MK, Yang Q, Veitch J, Schwitzer A, Awad T, Sugnet C, Dee S, Davies C, Williams A, Turpaz Y: **Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array.** *BMC Genomics* 2006, **7**:325.

Gautier L, Cope L, Bolstad BM, Irizarry RA: **affy: analysis of Affymetrix GeneChip data at the probe level.** *Bioinformatics* 2004, 20(3):307-15.

Gentleman RC, Carey VJ, Bates BM, Bolstad B, Dettling M, *et al*.: **Bioconductor: open software development for computational biology and bioinformatics.** *Genome Biology* 2004, **5(10)**:R80.

Gentleman R, Carey V, Huber W, Irizarry R, Dudoit S: *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. New York: Springer; 2005.

Giambonini-Brugnoli G, Buchstaller J, Sommer L, Suter U, Mantei N: **Distinct disease mechanisms in peripheral neuropathies due to altered peripheral myelin protein 22 gene dosage or a Pmp22 point mutation.** *Neurobiology of Disease* 2005, **18(3)**:656-68.

Gonzalez R, Yang YH, Griffin C, Allen L, Tique Z, Dobbs L. **Freshly isolated rat alveolar type I cells, type II cells, and cultured type II cells have distinct molecular phenotypes.** *American Journal of Physiology. Lung Cellular and Molecular Physiology* 2005 **288(1)**:L179-89.

Jacobson JR, Barnard JW, Grigoryev DN, Ma SF, Tuder RM, Garcia JG: **Simvastatin attenuates vascular leak and inflammation in murine inflammatory lung injury**. *American Journal of Physiology. Lung Cellular and Molecular Physiology* 2005 **288(6)**:L1026-32.

Jones L, Goldstein DR, Hughes G, Strand AD, Collin F, *et al*.: **Assessment of the relationship between pre-chip and post-chip quality measures for Affymetrix GeneChip expression data.** *BMC Bioinformatics* 2006, **7**:211.

Haslett JN, Sanoudou D, Kho AT, Han M, Bennett RR, *et al*.: **Gene expression profiling of Duchenne muscular dystrophy skeletal muscle.** *Neurogenetics* 2003, **4(4)**:163-71.

Heber S, Sick B: **Quality assessment of Affymetrix GeneChip data.** *OMICS: A Journal of Integrative Biology* 2006, **10(3)**:358-68.

Herman AE, Freeman GJ, Mathis D, Benoist C.: **CD4+CD25+ T regulatory cells dependent on ICOS promote regulation of effector cells in the prediabetic lesion.** *Journal of Experimental Medicine* 2004, **199(11)**:1479-89.

Howard BE, Perera I, Im YJ, Winter-Sederoff H, Sick B, Heber S: **Quality assessment of Affymetrix GeneChip data using the EM algorithm and a naïve Bayes classifier.** In *Proceedings of the IEEE 7th International Symposium on Bioinformatics & Bioengineering (BIBE 2007): 14-17 October 2007; Cambridge, MA*. Edited by Jack Y Yang, Mary Qu Yang, Michelle M Zhu, et al: IEEE; 2007:145-150.

Hu Z, Zimmermann BG, Zhou H, Wang J, Henson BS, *et al*.: Exon-level expression profiling: a comprehensive transcriptome analysis of oral fluids. *Clinical Chemistry* 2008, **54(5)**:824-32.

Huang RS, Duan S, Shukla SJ, Kistner EO *et al*. **Identification of genetic variants contributing to cisplatin-induced cytotoxicity by use of a genomewide approach.** *American Journal of Human Genetics* 2007, **81(3)**:427-37.

Hung LH, Heiner M, Hui J, Schreiner S, Benes V, Bindereif A: **Diverse roles of hnRNP L in mammalian mRNA processing: a combined microarray and RNAi analysis.** *RNA* 2008, **14(2)**:284-96.

Kimchi ET, Posner MC, Park JO, Darga TE, Kocherginsky M, *et al*.: **Progression of Barrett's metaplasia to adenocarcinoma is associated with the suppression of the transcriptional programs of epidermal differentiation.** *Cancer Research* 2005, **65(8)**:3146-54.

Kwan T, Benovoy D, Dias C, Gurd S, Provencher C, *et al*. **Genome-wide analysis of transcript isoform variation in humans.** *Nature Genetics* 2008, **40(2)**:225-31.

Lai Y, Adam B, Podolsky R, She J: **A mixture model approach to the tests of concordance and discordance between two large-scale experiments with two-sample groups**. *Bioinformatics* 2007, **23(10)**:1243-1250.

Larsson O, Wennmalm K, Sandberg R: **Comparative microarray analysis**. *OMICS: A Journal of Integrative Biology* 2006, **10(3)**: 381-397.

Lee EK, Yi SG, Park T: **arrayQCplot: software for checking the quality of microarray data**. *Bioinformatics* 2006, **22(18)**:2305-7.

Li J, Grigoryev DN, Ye SQ, Thorne L, Schwartz AR *et al.*: **Chronic intermittent hypoxia upregulates genes of lipid biosynthesis in obese mice.** *Journal of Applied Physiology* 2005, **99(5)**:1643-8.

Lin KK, Chudova D, Hatfield GW, Smyth P, Andersen B: **Identification of hair cycle-associated genes from time-course gene expression profile data by using replicate variance**. *PNAS* 2004, **101(45)**:15955-60.

Lozano JJ, Kalko SG: **AMarge: Automated Extensive Quality Assessment of Affymetrix chips.** *Applied Bioinformatics* 2006, **5(1)**:45-47.

Mozina M, Demsar J, Kattan M, Zupan B: (2004) **Nomograms for visualization of naïve Bayesian classifiers.** In *Proc. of Principles and Practice of Knowledge Discovery in Databases (PKDD-2004): 20-24 Sept. 2004; Pisa, Italy.* Edited by Jean-Francois Boulicaut, Floriana Esposito, Fosca Giannotti, Dino Pedreshci: ACM; 2004:337-348.

Najarian K, Zaheri M, Rad AA, Najarian S, Dargahi J: **A novel mixture model method for identification of differentially expressed genes from DNA microarray data.** *BMC Bioinformatics* 2004, **5:201**

Ng A: **Mixtures of Gaussians and the EM algorithm. CS229 Lecture notes.** 2006 [http://www.stanford.edu/class/cs229/notes/cs229-notes8.pdf]. Standford University, Palo Alto, CA

Nigam K, McCallum A, Thrun S, Mitchell T: **Text classification from labeled and unlabeled documents using EM**. *Machine Learning* 2000, **39(2/3)**: 103-134.

Ovando BJ, Vezina CM, McGarrigle BP, Olson JR. **Hepatic gene downregulation following acute and subchronic exposure to 2,3,7,8-tetrachlorodibenzo-p-dioxin**. *Toxicological Sciences* 2006 **94(2)**:428-38.

Platts AE, Dix DJ, Chemes HE, Thompson KE, Goodrich R, *et al.*: **Success and failure in human spermatogenesis as revealed by teratozoospermic RNAs**. *Human Molecular Genetics* 2007,**16(7)**:763-73.

Poulin B, Eisner R, Szafron D, Lu P, Greiner R, *et al*.: (2006) **Visual explanation of evidence in additive classifiers.** In *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI): 16-20 July, Boston, MA*.

Psarros M, Heber S, Sick M, Thoppae G, Harshman K, Sick B: **RACE: Remote Analysis Computation for gene Expression data.** *Nucleic Acids Research* 2005, 33:W638-43.

Ramonell K, Berrocal-Lobo M, Koh S, Wan J, Edwards H, Stacey G, Somerville S: **Loss-of-function mutations in chitin responsive genes show increased susceptibility to the powdery mildew pathogen Erysiphe cichoracearum.** *Plant Physiology* 2005, **138(2)**:1027-36.

Reimer M, Weinstein JN: **Quality assessment of microarrays: visualization of spatial artifacts and quantitation of regional biases.** *BMC Bioinformatics* 2005, **6**:166.

Ritchie ME, Diyagama D, Neilson J, van Laar R, Dobrovic A, Holloway A, Smyth G: **Empirical array quality weights in the analysis of microarray data.** *BMC Bioinformatics* 2006, 7:261.

Robinson MD, Speed TP: **A comparison of Affymetrix gene expression arrays.** *BMC Bioinformatics* 2007, **8(1)**:449.

Sandberg R, Neilson JR, Sarma A, Sharp PA, Burge CB: **Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites.** *Science* 2008, **320(5883)**:1643-7.

Shi L, Reid LH, Jones WD, Shippy R, Warrington JA, *et al.*: **The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements**. *Nature Biotechnology* 2006, **24(9)**: 1151-1161.

Soreq L, Gilboa-Geffen A, Berrih-Aknin S, Lacoste P, Darvasi A, *et al*.: **Identifying alternative hyper-splicing signatures in MG-thymoma by exon arrays.** *PLoS ONE* 2008, **3(6)**:e2392.

Stokes TH, Moffitt RA, Phan JH, Wang MD: **chip artifact CORRECTion (caCORRECT): a bioinformatics system for quality assurance of genomics and proteomics array data.** *Annals of Biomedical Engineering* 2007, **35(6)**:1068-80.

Vanneste S, De Rybel B, Beemster GT, Ljung K, De Smet I, *et al*.: **Cell cycle progression in the pericycle is not sufficient for SOLITARY ROOT/IAA14-mediated lateral root initiation in Arabidopsis thaliana.** *Plant Cell* 2005, **17(11)**:3035-50.

Wilkes T, Laux H, Foy CA: **Microarray data quality – review of current developments.** *OMICS: A Journal of Integrative Biology* 2007:**11(1)**:1-13.

William DA, Su Y, Smith MR, Lu M, Baldwin DA, Wagner D: **Genomic identification of direct target genes of LEAFY.** *PNAS* 2004, **101(6):**1775-80.

Wong JWH, Sullivan MJ, Cartwright HM, Cagney G: **msmsEval: tandem mass spectral quality assignment for high-throughput proteomics**. *BMC Bioinformatics* 2007, **8:51**.

Xing Y, Stoilov P, Kapur K, Han A, Jiang H, *et al*.: **MADS: a new and improved method for analysis of differential alternative splicing by exon-tiling microarrays.** *RNA* 2008, **14(8)**:1470-9.

Zhang Z, Lotti F, Dittmar K, Younis I, Wan L, *et al*.: **SMN deficiency causes tissue-specific perturbations in the repertoire of snRNAs and widespread defects in splicing**. *Cell* 2008, **133(4)**:585-600.

Chapter 3

PRACTICAL QUALITY ASSESSMENT OF MICROARRAY DATA BY SIMULATION OF DIFFERENTAL GENE EXPRESSION

Brian E Howard, Beate Sick, Steffen Heber

***ISBRA* 2009, Ft. Lauderdale, FL**

Abstract

There are many methods for assessing the quality of microarray data, but little guidance regarding what to do when defective data is identified. Depending on the scientific question asked, discarding flawed data from a small experiment may be detrimental. Here we describe a novel quality assessment method that is designed to identify chips that should be discarded from an experiment. This technique simulates a set of differentially expressed genes and then assesses whether discarding each chip enhances or obscures the recovery of this known set. We compare our method to expert annotations derived using popular quality diagnostics and show, with examples, that the decision to discard a chip depends on the details of the particular experiment.

Introduction

Considerable attention has been paid to methods and metrics that can be used to measure the quality of microarray data (for recent reviews, see Larsson, *et al*., 2006; Wilkes *et al*., 2007). For example, a common approach employs a routine set of diagnostic plots and statistics to identify arrays having low quality relative to the other chips in an experiment (Archer *et al*., 2006; Reimer and Weinstein, 2005; Stokes *et al.*, 2007; Gentleman *et al.*, 2005; Heber and Sick, 2006; Howard *et al.*, 2009). In the majority of cases, these methods are used as a filtering step, with the assumption that discarding low quality arrays should increase both the sensitivity and specificity of tests for differentially expressed genes (Wilkes

et al., 2007); however, in reality, many of these chips still contain valuable signal, even if that signal is obscured by extensive statistical noise. For a given FDR level, increasing sample size can increase the power to identify differentially expressed genes with decreased probability of declaring false positives (Pawitan *et al.*, 2005). Hence, as demonstrated in (Ritchie *et al.*, 2006), discarding moderately noisy chips can actually be detrimental in many cases. Unfortunately, no clear guidelines currently exist for differentiating scenarios in which it is advantageous to discard low quality data from situations where that data should be retained.

Here we present a simple procedure that can be used to assess the quality of microarray data. In contrast to other methods, however, this procedure also provides practical advice about what to do when low quality chips are identified. The method works by first simulating a set of differentially expressed genes, using gene expression distributions estimated from the dataset. Then, the procedure identifies arrays whose inclusion impairs the recovery of this known set of genes. This method is intended not to merely categorize arrays into binary "high quality" and "low quality" categories, but to identify arrays that should actually be excluded from a particular analysis. Because this approach to quality assessment depends on the details of the particular microarray experiment considered, the assessment framework we describe is easily adaptable to a variety of analysis protocols and experimental frameworks.

In the first section, we will describe the dataset used in this paper, and explain the simulation algorithm. Then, we will compare the results obtained from this approach with previous expert annotations created with the aid of a set of popular quality diagnostics. We

will illustrate the observation that any decision about whether to include a given array should be dependent not only on the noise profile of the array itself, but also on the details of the specific experiment being performed, including the number of replicates in each sample and the analysis method used to interpret the results.

Methods

Datasets

The dataset for this research consists of a set of 531 Affymetrix raw intensity (.CEL) files obtained from the NCBI GEO database (Edgar *et al.*, 2002). These data are a subset of the dataset described in (Howard *et al.*, 2009) and consist of all the experiments having at least three samples per treatment. Several of the most commonly used Affymetrix GeneChip 3' expression array types are represented and were chosen to include a variety of frequently investigated tissue types, experimental treatments and species, including Arabidopsis (ath1121501 array), mouse (mgu74a, mgu74av2, mo3430a, and mouse4302 arrays), rat (rae230a and rgu34a arrays), and human (hgu133a, hgu95av2, hgu95d, and hgu95e arrays).

Expert Annotations

Quality scores were assigned to each chip by a domain expert, according to a procedure previously established and applied in the Lausanne DNA Array Facility (DAFL) (Heber and Sick, 2006). Briefly, this procedure involves the systematic analysis of a variety of common predictive quality assessment metrics including: chip scan images, distributions of the log scale raw and normalized PM probe intensities, plots of the 5' to 3' probe intensity

gradients, pseudo-images of the PLM weights and residuals, and boxplots of the Normalized Unscaled Standard Error (NUSE) and Relative Log Expression (RLE) scores for each chip. After consideration of each of these quality features, the expert identified arrays that appeared to be outliers with respect to other chips in the same experiment, and each array was assigned a quality score of 0, 1, or 2, with 0 being "acceptable quality" (462 chips), 1 being "suspicious quality" (45 chips) and 2 being "unacceptable quality" (24 chips). These scores were then used as a basis of comparison to quality assessments made using the empirical quality approach described in this paper.

Quality Assessment Algorithm

Our approach takes a very practical definition of microarray data "quality": a low quality microarray is an array that diminishes the chances of accurately detecting differentially expressed genes, given a particular experimental design, dataset, and analysis methodology. To make this determination, our algorithm uses simulated data to find out if excluding a particular chip is likely to improve the ability to detect differentially expressed genes in an experiment similar to the one intended by the investigator. The simulated dataset is constructed using the observed gene expression distributions from the original experiment. Within this simulated dataset, which includes both a "treatment" group and a "control" group, some of the genes are differentially expressed. The quality assessment procedure operates by performing a statistical test for differential expression under two different scenarios: (1) using only the simulated data, and (2) using the simulated data plus the actual expression measurements for one of the chips. If excluding the actual expression

measurements for this chip enhances the recovery of the known set of simulated differentially expressed genes, then that chip is flagged as "low quality".

Note that this definition of quality depends on the details of the experiment examined, and, accordingly, our quality assessment framework is adaptable to a variety of microarray platforms and statistical procedures. For concreteness, we will describe the algorithm as it might be applied to a set of one-color microarray data of the sort that comprises our previously described test dataset. However, the details of this approach, including the normalization procedure, gene expression parameters, and choice of statistical test, are flexible. These can, and should, be adapted to match the analysis approach used for the actual experimental data.

Goal

- To determine whether or not a particular microarray chip should be excluded from an experiment designed to test for differential expression between two treatment groups.

Input

- A set of microarray expression values from treatment Group 1, which contains N_1 ($\geq$2) replicate chips.
- A set of microarray expression values from treatment Group 2, which contains N_2 replicate chips.
- A suspected low quality chip, *c*, from Group 1.

Output

- A decision whether or not to exclude chip c from the test for differential expression between Group 1 and Group 2.

Procedure

1. Normalize the complete dataset using whatever procedure would normally be used in the final analysis (e.g. quantile/RMA in Irizarry *et al.*, 2003, etc.).
2. Exclude the suspect chip, c, and use the N_1-1 remaining chips from Group 1 to estimate the mean, $\hat{\mu}_g$, and sample variance, s_g^2, for every probeset, g, on the chip.

Repeat 30 times:

3. Simulate a set of G_1 consistently expressed genes (CEGs) as follows:
    - Randomly select G_1 probesets from the set of all probesets on the chip.
    - For each selected probeset, sample N_1+N_2-1 values from a Normal($\hat{\mu}_g$, s_g^2) distribution.
    - Append the actual expression values from chip c to the simulated data for Group 1.

    The result is a $G_1 \times (N_1+N_2)$ expression matrix, where the first N_1 columns correspond to "treatment 1" and the second N_2 columns are "treatment 2".
4. Use the same procedure to simulate a set of G_2 differentially expressed genes (DEGs), with the following additional step:

- Add a small multiple of the probeset-specific standard deviations, s_g, to the N_2 "treatment 2" expression values, shifting the mean of the second treatment group relative to the first.

5. Perform a test for differential expression between the two treatments (e.g. using LIMMA (Smyth, 2004)) in each of the G_1+G_2 rows.
6. Evaluate the performance of this test by computing an ROC curve, which can be constructed from the sorted p-values from the tests in step 5. Using this ROC curve, compute the corresponding area under the curve (AUC). (A detailed guide to ROC curves can be found in Fawcett, 2006).
7. Discard the expression values from the suspect chip, and re-compute the ROC curve and AUC (i.e. repeat steps 5 and 6).
8. Record the difference between the AUC scores computed in steps 6 and 7.

9. Discard chip c if the AUC without chip *c* is significantly higher than with chip *c*.

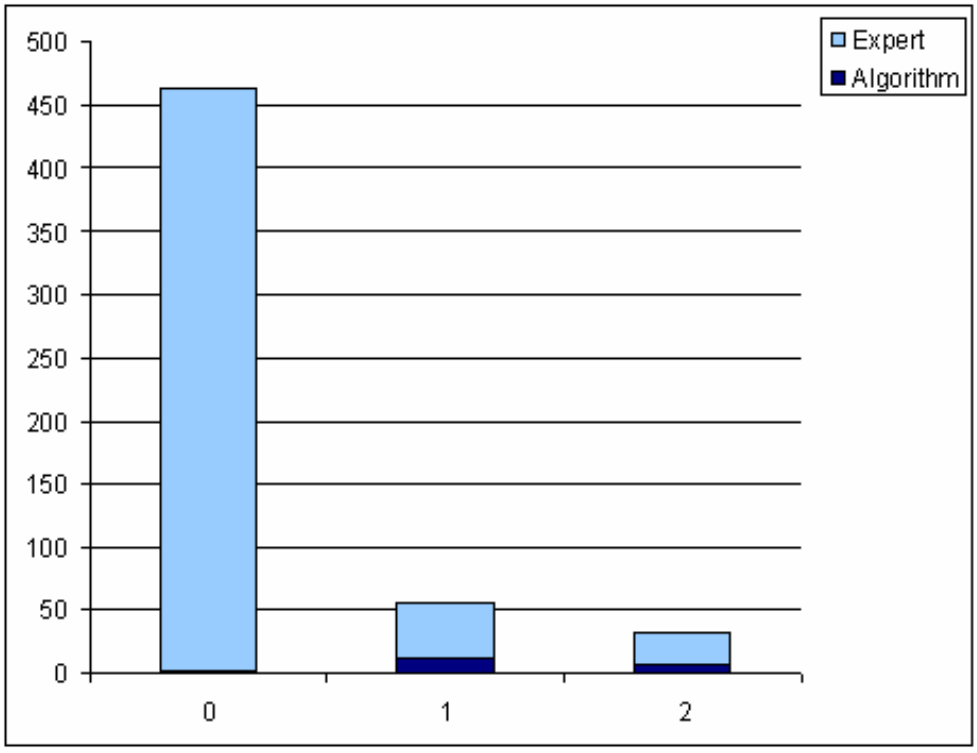


Figure 3-1. Low-quality calls by expert quality group.

Expert quality score is shown on the x-axis. Light blue indicates frequency of this category among expert annotations. Dark blue shows proportion of this category flagged for exclusion using the simulation approach.

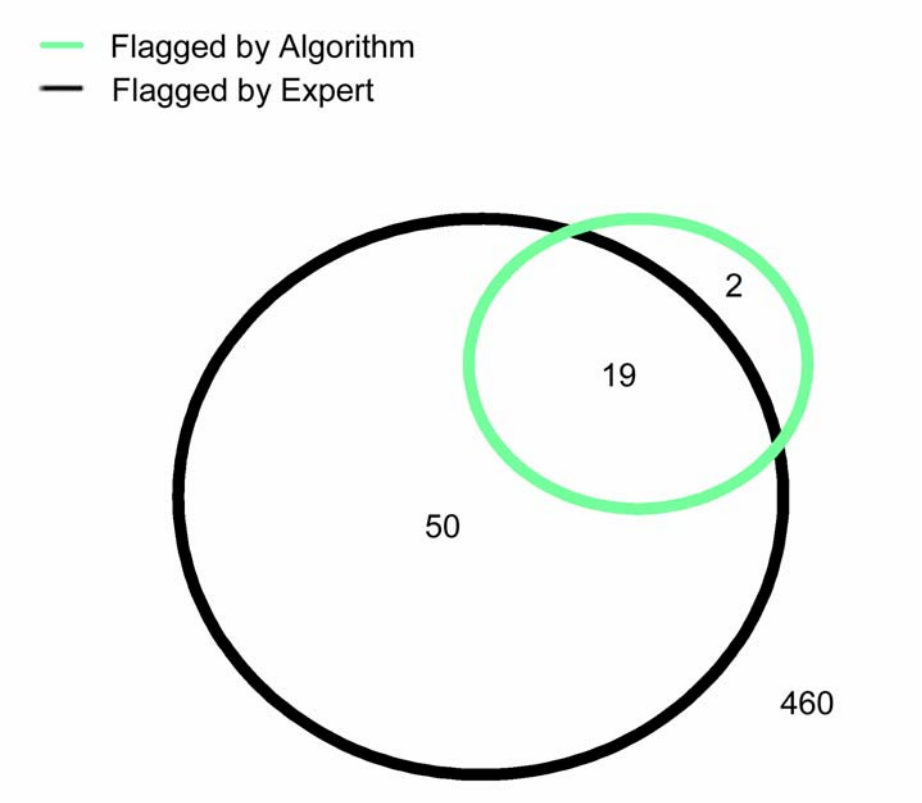


Figure 3-2. Comparison of expert and simulation determined quality scores.

Chips with expert quality scores of 1 or 2 are included in the "Flagged by Expert" set. Chips with simulation p-values < .001 are included in the "Flagged by Algorithm" set.

Results

Comparsion with Expert Annotations

After normalizing all arrays from each experiment using RMA (Irizarry *et al.* 2003), we applied the previously described simulation-based quality assessment procedure to each of the chips in our dataset. For each chip, we simulated 30 N × N experiments, where N is the number of replicates for that chip's treatment in the original dataset. Each experiment contained 500 consistently expressed genes (CEGs) and 500 differentially expressed genes (DEGs). Differential expression was simulated by adding ± 1 standard deviation to the second treatment group (odd probesets were given positive deltas, and even probesets were given negative deltas). The R LIMMA (Smyth, 2004) package was then used to identify differentially expressed genes, both with and without the suspect chip, and the resulting ROC curves were computed in each case. Chips whose inclusion significantly lowered the AUC according to a paired t-test (p-value < .001) were identified as having low quality. The entire analysis was performed using the R statistical programming language (code available from the author by request.)

We then compared the chips identified using this procedure with those identified previously by the domain expert. Figure 3-1 shows that, for the 24 chips identified by the expert as having the lowest quality (i.e. scored as 2's), the simulation identified 8 chips as being candidates for exclusion (33.3%). Among the 45 chips flagged by the expert as suspicious (1's), 11 were identified by the simulation procedure as candidates for exclusion (24.4%). For the 462 chips regarded by the expert as having acceptable quality, only 2 were

identified by the algorithm as candidates for exclusion (0.43%). Figure 3-2 summarizes the chips flagged as low quality by the two methods.

Practical Quality Judgment Depends on the Details of the Experiment

Quality assessment procedures based on predictive quality metrics sometimes have difficulty determining the utility of excluding suspicious chips because this decision is inextricably tied to the details of the particular experiment and the analysis method used. Unfortunately, the values for most quality metrics do not explicitly incorporate the sample size, target effect magnitude or analysis method employed. However, these experimental details are critical for making a rational decision regarding the inclusion or exclusion of low quality data. This scenario is illustrated in the following examples.

Example 1. GEO dataset GSE1873 (Li, *et al.* 2005) contains gene expression measurements taken from liver tissue of obese mice. The experiment used 5 Affymetrix microarrays to measure gene expression of obese mice exposed to intermittent hypoxic conditions and 5 microarrays to measure gene expression of obese mice used as controls. Using the protocol described in section 2.2, our domain expert examined this dataset and identified 3 chips as having low or suspicious quality (GSM32860, GSM32861 and GSM32866). However, in a simulation using 5 chips in each treatment group, only GSM32860 and GSM32866 were found to be worthy of exclusion (when considered individually). On the other hand, in simulated 3x3 and 4x4 experiments, exclusion of chip GSM32866 is no longer recommended by our procedure. Conversely, as simulated

experiment size increases, the p-value for chip GSM32861 approaches the threshold for exclusion, with a p-value of less than 0.01 for experiments of size 9x9 or greater.

Example 2. Recent research has demonstrated that many of the common quality problems observed in a typical microarray experiment can be mitigated with the use of robust analysis methods. For example, many typical quality problems can be captured with a heteroscedastic variance model which allows each chip to have different levels of random noise (Ritchie, *et al*. 2006). Smyth showed in simulation that, in many cases, procedures that simply down-weight noisier chips perform better than methods that attempt to identify and exclude these low quality chips.

Again, consider experiment GSE1873. Figure 3-3 shows the expression values for a few representative probesets (expert-identified low quality probesets are shown in colored dots). The diagram illustrates the fact that there is greater variance between probesets than among chips within each probeset. On the other hand, the expression values for the low quality chips appear to more often have extreme values than the other chips, although not consistently in one direction. The RLE boxplot also reflects this observation (Figure 3-4); the interquartile ranges for the low quality chips are larger than for the high quality chips. These observations suggest that the heteroscedastic variance model may indeed be useful in the analysis of this data set. To test this hypothesis, we repeated the quality simulation with one modification: we used the "arrayWeights" functionality of the LIMMA package to identify and downweight noisy chips. Under this analysis framework, the quality simulation showed that excluding these chips is no longer recommended.

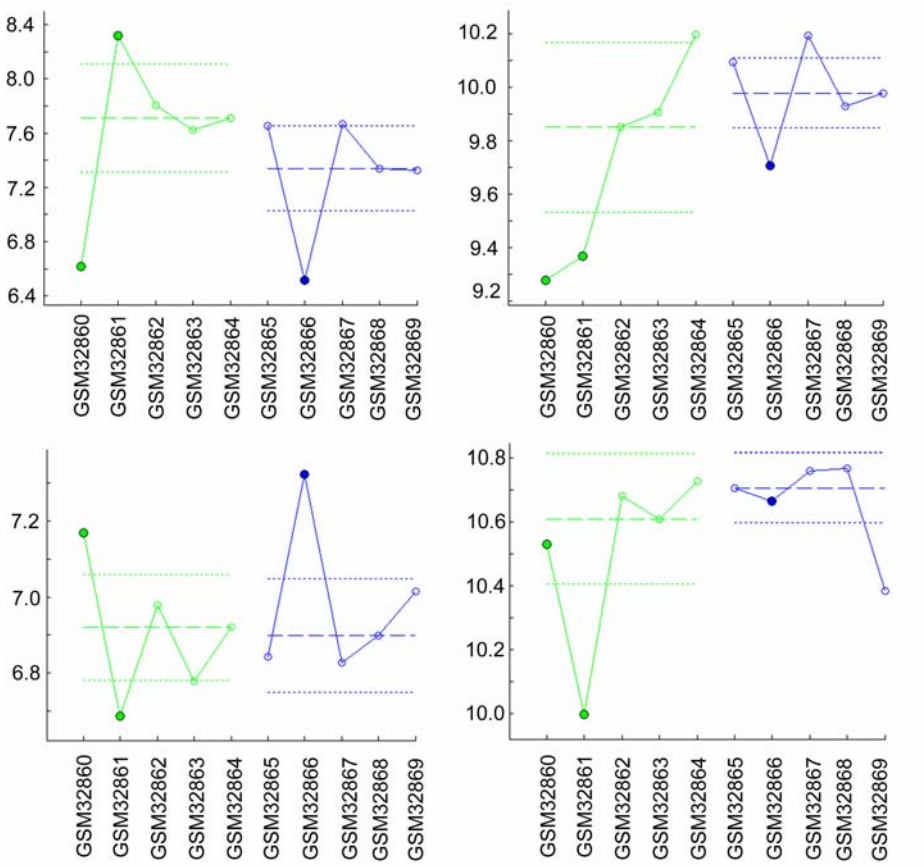


Figure 3-3. Normalized expression levels for 4 probesets from experiment GSE1873.

Green circles correspond to 'treatment 1' and blue circles to 'treatment 2.' The colored circles represent chips flagged by the expert as low quality. The dashed lines indicate the median expression level for each treatment, while the dotted lines correspond to treatment median ± 1 MAD (median absolute deviation). X-axis is chip name; y-axis is normalized expression.

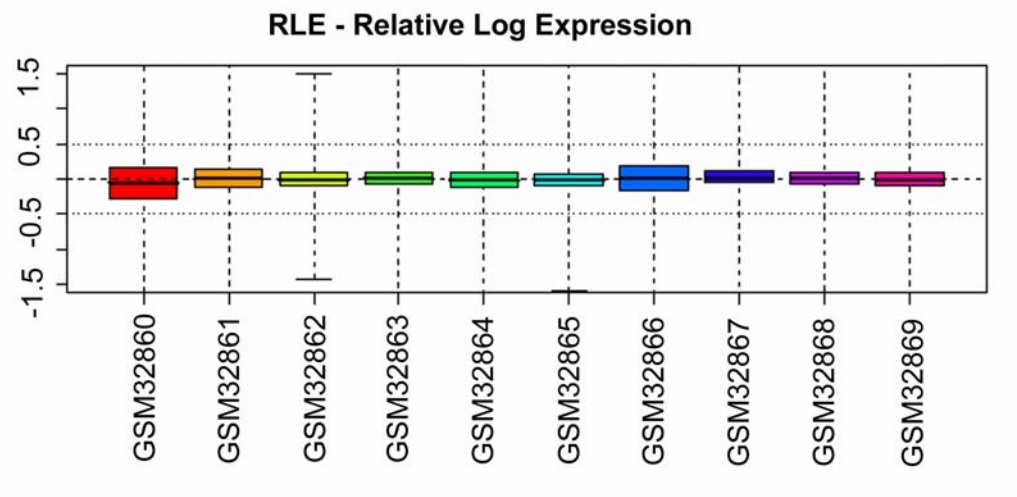


Figure 3-4. Log expression for experiment GSE1873.

Height of box corresponds to interquartile range of the RLE, and midline indicates RLE median.

On the other hand, even robust methods can not be expected to correct the most extreme types of errors. For example, we simulated mislabeled samples by interchanging data from different GEO datasets and observed that in these cases it was often still better to remove the foreign arrays than to apply the downweighting procedure (data not shown).

Discussion

The quality assessment method described here addresses an important question not often considered by other procedures: what to do with the low-quality chips that are identified. In many real-world scenarios, better results can be obtained by retaining slightly flawed data, instead of discarding it completely. Unfortunately, there is currently little guidance available with regard to this decision. Our method takes an empirical approach to this problem by simulating a set of differentially expressed genes and then evaluating the

contribution of each suspected chip with regard to identifying these genes. For the datasets examined in our research, the chips identified by the simulation algorithm as "excludable" were roughly a subset of the chips identified by the domain expert as having low quality (Figure 3-2). This may imply that although the expert is correctly identifying the chips with higher noise levels, many of those chips still retain useful signal, especially within the context of the small experiments considered.

This approach is easily adapted to other analysis settings, and, in general, it is recommended that the analysis method and parameter settings chosen for the simulation should match the protocol intended for the real data set. For example, here we have used the LIMMA library for statistical analysis, but other methods, such as SAM (Tusher *et al*., 2001) or Cyber-T (Baldi and Long, 2001) could just as easily be applied instead. Alternatively, if the researcher is interested in controlling false discoveries at a specific rate, then one could apply an FDR control procedure and compare the number of true discoveries made instead of the area under ROC curves. In future work we intend to explore more thoroughly the influence of these parameters on the resulting quality decisions. It would also be interesting to enhance our simulation approach to emulate more complex gene expression models, possibly allowing for correlated genes, non-normal distributions and variable effect sizes. It should be noted that when applying the procedure as described here, it is important to look not only at the resulting p-value, but also the magnitude of the observed difference in AUC obtained with and without each chip. Very small differences can sometimes accompany significant p-values, especially if enough replications are performed; in these cases it is probably prudent to retain the chip anyway.

Like other quality assessment procedures that attempt to identify outliers among a particular set of microarrays, our method is susceptible to scenarios where the dataset is corrupted by a majority of chips with systematic error. For example, in a dataset where one of the arrays is mislabeled with regard to the experimental treatment applied, our method would likely identify the mislabeled array as an outlier; however, if all of the arrays except one particular array were mislabeled, our algorithm may erroneously identify the correctly labeled array as the outlier.

Robust analysis methods such as the approach described in Ritchie (2006) can potentially mitigate many of the common problems observed in microarray datasets. On the other hand, there are still scenarios where even the most robust methods cannot recover useful signal from a particular low quality array. Arrays showing evidence of large spatial artifacts, contamination or other gross errors such as mislabeled samples can rarely be salvaged. Our method can be used to identify these scenarios. In addition, while the method we have described can be used on its own for quality assessment, this technique can also be used in conjunction with other traditional quality diagnostics, which may provide additional clues as to what sorts of errors are present in a batch of arrays and thereby assist in avoiding these problems in the future.

References

Archer KJ, Dumur CI, Joel SE, Ramakrishnan V: **Assessing quality of hybridized RNA in Affymetrix GeneChip experiments using mixed-effects models**. *Biostatistics* 2006, **7(2)**:198-212

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

Edgar R, Domrachev M, Lash, AE: **Gene Expression Omnibus: NCBI gene expression and hybridization array data repository**. *Nucleic Acids Research* 2002, **30(1)**:207-10 (2002)

Fawcett T: **An introduction to ROC analysis**. *Pattern Recognition Letters* 2006, **27**:861-874.

Gentleman R, Carey V, Huber W, Irizarry R, Dudoit S.: *Bioinformatics and computational biology solutions using R and Bioconductor*. New York: Springer; 2005.

Heber S, Sick B: **Quality assessment of Affymetrix GeneChip data**. *OMICS: A Journal of Integrative Biology* 2006, **10(3)**:358-68.

Howard BE, Sick B, Heber S: **Unsupervised assessment of microarray data quality using a Gaussian mixture model**. *BMC Bioinformatics* 2009, **10**:191.

Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP: **Summaries of Affymetrix GeneChip probe level data**. *Nucleic Acids Research* 2003, **31(4)**:e15 .

Larsson O, Wennmalm K, Sandberg R: **Comparative microarray analysis**. *OMICS: A Journal of Integrative Biology* 2006, **10(3)**: 381-397.

Li J, Grigoryev DN, Ye SQ, Thorne L *et al*.: **Chronic intermittent hypoxia upregulates genes of lipid biosynthesis in obese mice**. *Journal of Applied Physiology* 2005, **99(5)**:1643-1648.

Pawitan Y *et al*.: **False discovery rate, sensitivity and sample size for microarray studies.** *Bioinformatics* 2005, **21:** 3017-3024.

Reimer M, Weinstein JN: **Quality assessment of microarrays: visualization of spatial artifacts and quantitation of regional biases**. *BMC Bioinformatics* 2005, **6**:166.

Ritchie ME, Diyagama D, Neilson J, van Laar R, Dobrovic A, Holloway A, Smyth G: **Empirical array quality weights in the analysis of microarray data**. *BMC Bioinformatics* 2006, **7**:261.

Smyth GK: **Linear models and empirical Bayes methods for assessing differential expression in microarray experiments**. *Statistical Applications in Genetics and Molecular Biology* 2004, **3(1)**.

Stokes TH, Moffitt RA, Phan JH, Wang MD: **chip artifact CORRECTion (caCORRECT): a bioinformatics system for quality assurance of genomics and proteomics array data**. *Annals of Biomedical Engineering* 2007, **35(6)**:1068-80.

Tusher VG, Tibshirani R, Chu G: **Significance analysis of microarrays applied to the ionizing radiation response**. *PNAS* 2001, **98(9)**:5116-5121.

Wilkes T, Laux H, Foy CA: **Microarray data quality – review of current developments.** *OMICS: A Journal of Integrative Biology* 2007, **11(1)**:1-13.

Chapter 4

TOWARDS RELIABLE ISOFORM QUANTIFICATION USING RNA-SEQ DATA

Brian E Howard and Steffen Heber

***BIBM* 2009, Washington, DC**

Abstract

In eukaryotes, alternative splicing often generates multiple splice variants from a single gene. Here we explore the use of RNA sequencing (RNA-Seq) datasets to address the isoform quantification problem. Given a set of known splice variants, the goal is to estimate the relative abundance of the individual variants. Our method employs a linear models framework to estimate the ratios of known isoforms in a sample. A key feature of our method is that it takes into account the non-uniformity of RNA-Seq read positions along the targeted transcripts. Preliminary tests indicate that the model performs well on both simulated and real data.

Introduction

In higher organisms many multi-exon genes undergo alternative splicing (AS) reactions that produce multiple splice variants, often encoding distinct, but related, protein products. In contrast to the traditional "one gene, one protein" hypothesis, more than 70% of human genes are now believed to be subject to alternative splicing (Hertel, 2008), with AS isoforms apparently responsible for many of the salient differences between diverse tissue types. A significant degree of AS has also been observed in various plant species, although the precise magnitude and functional relevance of these events is unknown (Wang and Brendel, 2006).

While it has long been assumed that differential gene splicing plays an important role in determining the phenotypes of organisms, it has been difficult to quantify AS using

available high throughput methods. However, recently developed massively parallel sequencing-by-synthesis technologies from Illumina, Applied Biosystems and Roche 454 Life Sciences now have the potential to revolutionize the study of the transcriptome (Wang *et al.*, 2009). It is now possible to produce enough high quality reads in a single run to rival traditional EST libraries that have accumulated over a span of decades. Furthermore, the resulting digital counts are more comparable to the "gold standard" quantification method, RT-PCR, and may overcome many of the shortcomings inherent in hybridization-based microarray gene expression studies.

Because the technological potential of these "RNA-Seq" protocols is well appreciated, and rapidly advancing, methods for accurate estimation of isoform expression levels are an active area of research. To compute an isoform expression score, the reads that map to each isoform must be converted into a quantitative expression value. One approach is to count the number of reads that map to an isoform, normalizing against the feature length and sequencing depth (Mortazavi *et al.*, 2008). Unfortunately, this technique is often infeasible for AS variants because many reads can map to multiple isoforms simultaneously. Recently, Lacroix *et al.* (2008) investigated the theoretical limitations of transcriptome reconstruction and quantification from a combinatorial perspective. Their analysis operated under an "exact information hypothesis" whereby the exact abundances of all relevant transcribed regions is provided error-free. However, this approach ignores the sampling process that actually generates observed data along with the associated measurement error; in practice, statistical approaches are necessary in order to obtain accurate estimates of transcript abundance. For example, Jiang and Wong (2009) have described a Poisson model

for isoform quantification, showing how to estimate its parameters with a maximum likelihood approach. Other authors have employed more basic (but effective) statistical approaches, for example, Fisher's exact test, to compare levels of AS between treatments, e.g. Wang *et al.*, 2008.

In this paper, we explore the use of RNA-Seq datasets to address the "isoform expression estimation problem" as defined in Jiang and Wong, 2009. It is assumed that the set of splice variants is known; the goal is to estimate the relative expression levels of these isoforms in a mixture. Obtaining precise estimates is necessary because important tissue-specific differences in AS frequently involve a continuum of isoform ratios, rather than all-or-nothing expression (Gupta *et al.*, 2004). Although, currently, the assumption of known isoforms may be limiting in many cases, we will soon be able to construct detailed lists of known isoforms for various organisms and tissue states using high-throughput sequencing (see for example, Salehi-Ashtiani *et al.*, 2008). A key advantage of our method over prior approaches is that our model takes into account the non-uniformity of RNA-Seq reads along the targeted transcripts. In addition, our approach can be easily adapted for use with any high-throughput sequencing technology, including those that employ paired reads. In the following sections we will describe the details of our model, demonstrate its performance on simulated and real data, and outline topics for future research.

Methods

Model Overview

Given a set of *n* unique AS isoforms for a gene, *g*, it is always possible to partition RNA-Seq reads from *g* into 2^n categories according to what subset of these isoforms each read is compatible with. For example, consider two AS isoforms, T_1 and T_2:

```
T1: AAAAAAA UUUUUUUUUU CCCCCCCCCC
T2: AAAAAAA ---------- CCCCCCCCCC
```

In this example, transcript T_1 contains a cassette exon containing only "U" nucleotides. Transcript T_2 skips this exon. Reads aligned to these transcripts can be classified into 4 mutually-exclusive subsets:

- *Subset S_1*: Reads which contain only T's are only compatible with transcript isoform T_1.
- *Subset S_2*: Reads which contain A's followed immediately by C's (e.g. AAACCCCC) are only compatible with T_2.
- *Subset S_3*: Reads which contain only A's or only C's are compatible with both T_1 and T_2.
- *Subset S_4*: Many reads, including any reads containing one or more G's, are not compatible with either T_1or T_2.

In the following we will disregard subset S_n and only consider reads that map to at least one of the known isoforms. Let:

$Pr(S_i)$ denote the probability that one of the gene's reads maps to subset S_i

$Pr(T_i)$ denote the probability that one of the gene's reads maps to transcript T_i

ϕ_i denote the percentage of the transcripts expressed as isoform T_i

Given the subsets introduced above, the following equation describes the probability that an individual read maps to subset S_i:

$$Pr(S_i \mid \vec{\phi}) = \sum_{j=1}^{n} Pr(S_i \mid T_j) Pr(T_j \mid \phi_j) \qquad \forall i \in \{1 \ldots 2^n - 1\} \quad (4.1)$$

In general, we can assume that $Pr(T_j \mid \phi_j)$, the probability an individual read maps to a particular transcript, is dependent on the (unknown) frequency, ϕ_j, of that transcript in the transcript mixture. We will also assume that a given isoform is sampled with probability proportional to its known length. Similarly, $Pr(S_i \mid T_j)$, the probability that an individual read maps to subset S_i, given the read maps to transcript T_j, can be worked out using the known transcript sequence and estimates of the distributions for read length and read start position (for details, see the section "Constructing the Design Matrix"). Let:

Y_i denote the number of reads compatible with subset S_i

R denote the total number of reads for the gene

$p_{ij} = \Pr(S_i \mid T_j)$

$\beta_j = \Pr(T_j \mid \phi_j)$

Assuming that individual reads are iid, we then have $Y_i \sim$ Binomial(R, $\Pr(S_i \mid \vec{\phi})$), and

$$\begin{aligned} E\left(Y_i \mid \vec{\phi}, R\right) &= R \cdot \left(\Pr\left(S_i \mid \vec{\phi}\right)\right) \\ &= R \cdot \left(\sum_{j=1}^{n} \Pr(S_i \mid T_j) \Pr(T_j \mid \phi_j)\right) \quad (4.2) \\ &= R \cdot \left(\sum_{j=1}^{n} p_{ij}\beta_j\right) \end{aligned}$$

For the example shown above, we can express this linear model in matrix form as follows:

$$\begin{bmatrix} Rp_{11} & 0 \\ 0 & Rp_{22} \\ Rp_{31} & Rp_{32} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{bmatrix} \quad (4.3)$$

Because Rp_{21} and Rp_{12} will always be zero, the rank of this matrix is 2, and both β_1 and β_2 are estimable. Although, in general, the number of rows ($2^n - 1$) grows exponentially with the number of possible transcripts, it should be possible to either combine or ignore uninteresting categories.

Distribution of Read Start Position and Read Length

Most methods for estimating isoform abundance assume a uniform sampling distribution for reads along the targeted transcripts (e.g. Mortazavi *et al.,* 2008; Lacroix *et al.,*

2008; Jiang and Wong, 2009). However, it is widely acknowledged that the true distribution for read position deviates substantially from uniformity, and varies with the fragmentation protocol and sequencing technology (Wang *et al.,* 2009). Consequently, accurate methods for isoform quantification must incorporate this critical information.

We believe that these distributions should be consistent properties of the instrument and experimental protocol. With millions of reads often available per experiment, it should be possible to determine these distributions with a high level of accuracy. We used a kernel density approach to estimate read length and read start distributions using the observed empirical distributions observed for well-annotated transcripts (Figure 4-1a,b). The read length was estimated in a similar manner, resulting in an average read length of approximately 30 nucleotides for the Illumina data set, and about 100 nucleotides for the 454 dataset.

To investigate the relationship between experimental protocol and read distribution, we also created a simple simulation that emulates the process of cDNA fragmentation by nebulization. The similarity between Figure 4-1b and Figure 4-1c suggests that our simulation captures the main properties of the nebulization process. We anticipate that more detailed models, which incorporate deep knowledge of the physical processes of fragmentation and sequencing, should be able to accurately describe observed distributions of read length and position.

Constructing the Design Matrix

Let $h(\ k,\ m\ |\ L\ )$ denote the bivariate probability mass function describing the probability that a read has start position k and length m, given that this read aligns to a transcript of length L. We compute $\Pr(S_i \mid T_j)\ \ \forall i \in \{1...2^n\text{-}1\}$ for a particular transcript T_j using the procedure detailed in Figure 4-2.

Estimation of β and ϕ

Given the construction method described above, the design matrix will always be full column rank, so $\hat{\beta}$ will always be fully estimable. Each $Y_i \sim \text{Binomial}(R,\ \Pr(S_i \mid \vec{\phi}))$. For computational simplicity, we use the Normal approximation to the binomial distribution. For the Normal linear model with a known covariance matrix, the maximum likelihood estimate (MLE) obtained using weighted least squares is the best linear unbiased estimator (BLUE). In the system described above, the variances are not known, but can be estimated from the data. In this case, the "feasible weighted least squares" method can be used to approximate the weighted least squares solution. In cases where a resulting $\hat{\beta}$ is not a valid probability, we truncate the estimate at 0 or 1. In addition, we ensure that the total probability is one by dividing each $\hat{\beta}$ by the sum across all i.

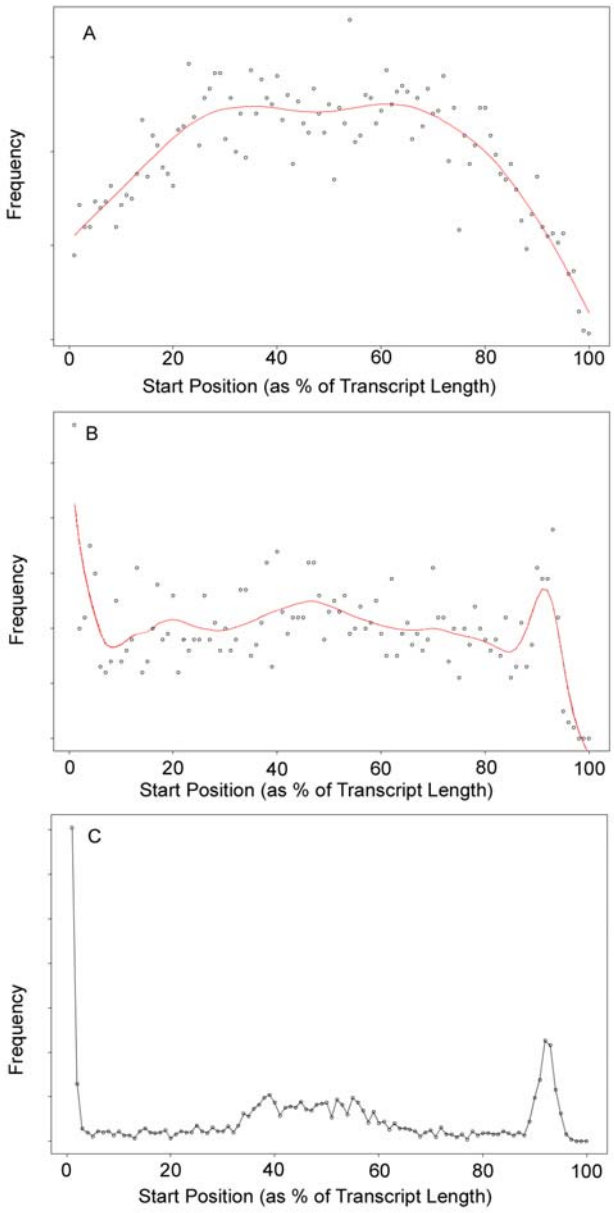


Figure 4-1. Distribution of read start position as percentage of gene length for genes with median gene length of 1200 bp.

The red line is a cubic spline fit. (A) Illumina dataset - RNA fragmentation by sonication (Lister *et al.* 2008), (B) Roche 454 dataset - cDNA fragmentation by nebulization (Weber *et al.*, 2007), and (C) Simulated cDNA fragmentation assuming fragments in size range 500-800 bp.

Initialize $p_{ij} = \Pr(S_i \mid T_j) = 0 \quad \forall i \in \{1...2^n\text{-}1\}$.

For each read start site, *k*, on transcript T_i :

For each possible read length, *m*, for a read starting at *k* :

- Determine the sequence for the read, *r*, aligned to T_j, having start position *k*, and end position *k*+*m*-1: $r = T_i[k, \ldots, k+m\text{-}1]$.
- Determine which category, S_i, the read belongs to, and update its probability:
  $\Pr(S_i \mid T_j) = \Pr(S_i \mid T_j) + h(k, m \mid L_j)$

Figure 4-2. Procedure for computing the design matrix.

Under the commonly employed assumption that a given isoform is sampled with probability proportional to its length, the probability that a given read maps to a transcript isoform can be expressed as follows:

$$\Pr(T_j \mid \phi_j) = \beta_j = \frac{L_j \phi_j}{\sum_{i=1}^{n} L_i \phi_i} \quad (4.4)$$

where L_j is the length of transcript *j* in bases. Given our estimates for the β_j's, the known lengths for each transcript and the fact that $\sum_{i=1}^{n} \phi_i = 1$, the $\hat{\phi}_j$ are uniquely determined and can be obtained by computing the unique solution to a set of linear equations. The sampling distribution of the resulting ϕ_j cannot easily be computed analytically, but confidence intervals can be worked out empirically following a procedure based on (Clopper and Pearson, 1934).

Implementation

The algorithm described was implemented in Java, with matrix computations by the JAMA matrix library (available at http://math.nist.gov/javanumerics/jama/). Additional analyses and simulations were also performed using the R statistical programming language (http://www.r-project.org/).

Results

Simulation

To test our method, we performed a simulated RNA-Seq experiment using the Arabidopsis gene models defined in TAIR 8 (http://arabidopsis.org). A publicly available dataset (NCBI Short Read Archive Accession SRX002554, Lister *et al.*, 2008) was used to estimate the read length and read start position for Illumina reads (Figure 4-1a). The simulation was performed for several of the multi-isoform genes as follows: First, a relative frequency for each of the alternative isoforms was specified as a simulation parameter, along with a predetermined total number of RNA-Seq reads. Each of these reads was then simulated by first selecting an isoform with probability proportional to its length and concentration in the mixture. Then, a read start position k and read length m along the selected isoform were drawn from the distribution, $h(k, m \mid L)$. Using these read coordinates, the nucleotide sequence along the sampled isoform was determined, and this sequence was compared with other isoforms in the mixture to identify which subset of isoforms the read is compatible with. The output of one run of the simulation is a list of subsets and the

corresponding counts of simulated reads assigned to those subsets. For each gene, the entire simulation was repeated 500 times. Given the simulated datasets, we then used the linear model described above to infer the original isoform concentrations from the simulated subset counts. We performed the estimation in two different ways for each gene: first, using a design matrix constructed using the same read position distribution used to generate the simulated reads (Figure 4-1a) and second, using a design matrix constructed from a uniform read position distribution. Note that using the incorrect distribution can introduce a severe bias into the estimates and even change the ordering of the isoform expression levels (e.g. Figure 4-3d).

For each of the 500 replications, we also computed approximate confidence intervals about the estimates. To evaluate the performance of our approximation, we checked each confidence interval to see if it included the true value for the parameter. For the AT1G75410 simulation, 97.4% of the 95% C.I.'s contained the true parameter; 92.2% of the 90% C.I.'s contained the true parameter; and 63.2% of the 65% C.I.'s contained the true parameter.

This entire process was repeated for 100 different genes. Over several replications, the mean estimates were always approximately equal to the simulated isoform concentrations when the correct distribution was used to construct the design matrix. For genes with two isoforms, we found that approximately 750-1000 reads were often needed to obtain a 95% confidence interval with a width of ~20%. However, the number of reads required varies according to mixture composition, number of isoforms, and the read length and start site distributions. (Data not shown).

	Transcript Mixture (ϕ_i)	$\hat{\phi}_i$ - True Read Distribution	$\hat{\phi}_i$ - Uniform Read Distribution
A)	AT1G75410.1=**70%**	**69.8%** (64.2%-75.2%)	**67.5%** (61.5%-73.3%)
	AT1G75410.2=**30%**	**30.2%** (24.8%-35.8%)	**32.5%** (26.7%-38.5%)
B)	AT2G40140.1=**70%**	**70.0%** (55.8%-83.9%)	**55.5%** (43.9%-68.4%)
	AT2G40140.2=**30%**	**30.0%** (16.1%-44.2%)	**44.5%** (31.6%-56.1%)
C)	AT2G01260.1=**20%**	**19.6%** (08.7%-30.5%)	**12.8%** (01.0%-23.9%)
	AT2G01260.2=**70%**	**70.3%** (62.5%-77.8%)	**76.4%** (68.5%-83.8%)
	AT2G01260.3=**10%**	**09.9%** (02.2%-18.8%)	**10.7%** (02.4%-19.9%)
D)	AT1G75380.1=**70%**	**69.5%** (58.9%-78.8%)	**70.9%** (60.8%-80.3%)
	AT1G75380.2=**20%**	**20.7%** (05.9%-33.9%)	**04.4%** (00.0%-21.4%)
	AT1G75380.3=**10%**	**09.4%** (0.00%-22.4%)	**22.9%** (09.3%-35.7%)

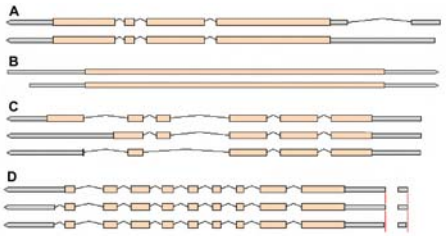


Figure 4-3. Estimates for phi using 2000 simulated reads.

(A) AT1G75410, (B) AT2G40140, (C) AT2G01260 and (D) AT1G75380.

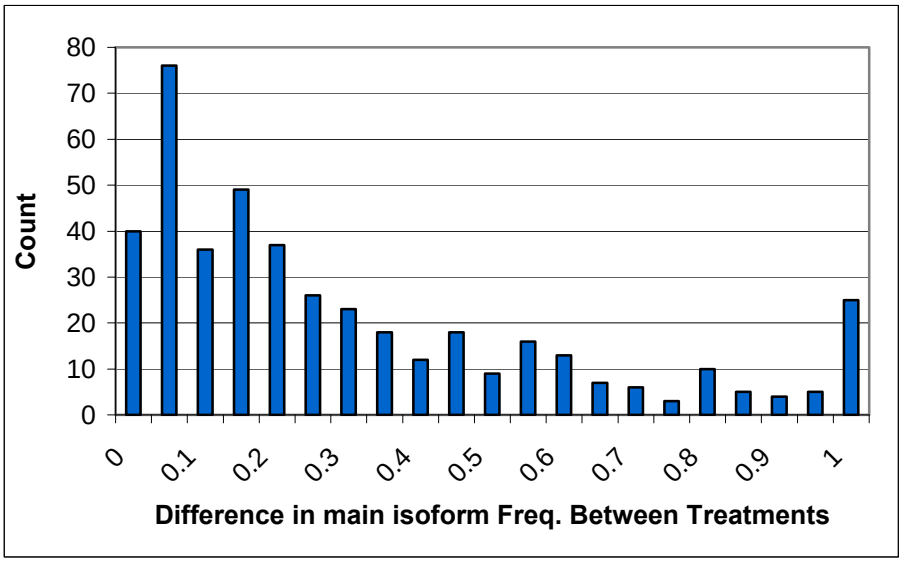


Figure 4-4. Differences in main isoform frequency for 438 AS genes.

Real RNA-Seq Dataset

We next applied our method to the publicly available data sets SRX002554 and SRX002555 (Lister *et al.*, 2008). These two experiments contain approximately 65 million (SRX002554) and 57 million (SRX002555) Illumina reads from floral tissue from two different Arabidopsis strains (SRX002554 = col-0; SRX002555 = ddc). Reads were mapped to the transcriptome using the SOAP v2 alignment program (Li *et al.*, 2009).

TAIR 8 was used to define the tested gene models. Among the 33282 Arabidopsis gene models defined in TAIR 8, 4330 genes had more than 1 isoform. In particular, 3336 genes had 2 isoforms, 739 had 3 isoforms, 186 had 4 isoforms, 48 had 5 isoforms, 14 had 6 isoforms, 4 had 7 isoforms, 2 had 8 isoforms, and 1 had 10 isoforms.

For the SRX002554 dataset, among genes with 2 or more isoforms, 1039 genes had more than 500 mapped reads (31.1%); 481 genes had more than 1000 reads (11.1%); 205 genes had more than 2000 reads (4.7%); and 50 genes had more than 5000 reads (1.5%). Similar coverage was observed in the SRX002555 dataset.

We first used a chi-square test of subset counts to identify genes that were differentially spliced between the two conditions; 438 genes showed significant differences (uncorrected p-value cutoff of .001) between the two samples. We then used the method defined above to infer the isoform ratios in the two different strains. In several cases, genes with highly significant AS levels (according to the Chi-square test), had only very small differences in proportions between the two treatments. These most likely do not represent significant biological differences. On the other hand, we also identified many genes that had

large differences in isoform composition between the two treatments. For example, 286 of the 438 genes had differences of 15% or more in the proportion for the main isoform. Figure 4-4 shows a histogram of the differences in main isoform proportions between the two treatments. We also noted many genes which appeared to exhibit switch-like differential splicing (shown in Figure 4-4 as genes with large differences in isoform expression between treatments). For example, AT4G10610 "RNA-Binding protein 37", which has unknown function, but which has been shown to be highly expressed during development and differentiation of floral tissue (Hecht *et al.*, 1997), was estimated to occur exclusively as isoform AT4G10610.1 in col-0 (SRX002554) and only as AT4G10601.2 in ddc (SRX002555). In addition, many of the genes exhibiting differential splicing had roughly the same number normalized read counts in each treatment, indicating that changes in overall gene expression level and changes in isoform abundance might occur independently (data not shown).

Discussion and Future Work

This paper describes a novel method for inferring AS expression levels. Our method is unique in that it incorporates non-uniform distributions of reads along the targeted transcripts. This is important, since assuming a uniform read distribution might result in serious estimation mistakes. In addition, our method is flexible enough to accommodate a variety of sequencing technologies, including those that incorporate paired reads. We expect that in the near future improved algorithms for isoform quantification will make it possible to investigate the transcriptome at a much higher level of resolution. It is likely that this will

reveal a so far unknown dimension of the transcriptome, including, for example, differentially spliced genes whose overall gene expression level remains unchanged (Blencowe, 2006).

There are several potential avenues for future work. A major limitation of our approach is that it assumes that all transcripts are known, yet the current state of transcriptome annotation is incomplete for most organisms. Because incorrect assumptions regarding potential transcripts in a mixture could lead to erroneous estimates, we are investigating ways to incorporate residual-based diagnostics into our model. These enhancements would serve to identify the presence of unknown "hidden" isoforms in a mixture and would complement isoform quantification with a mechanism for transcript discovery. A second line of research focuses on developing tools for experimental design, in particular, a method for estimating the number of reads required to achieve a given confidence level for any given estimation scenario.

References

Blencowe BJ. **Alternative splicing: new insights from global analyses**. *Cell* 2006, **126(1)**:37-47.

Clopper C, Pearson S. **The use of confidence or fiducial limits illustrated in the case of the binomial**. *Biometrika* 1934, **26**: 404-413.

Gupta S, Zink D, Korn B, Vingron M, Haas SA. **Strengths and weaknesses of EST-based prediction of tissue-specific alternative splicing.** *BMC Genomics* 2004, **5**:72.

Hecht V, Stiefel V, Delseny M, Gallois P. **A new Arabidopsis nucleic-acid-binding protein gene is highly expressed in dividing cells during development.** *Plant Molecular Biology* 1997, **24(1)**:119-24.

Hertel KJ. **Combinatorial control of exon recognition**. *Journal of Biological Chemistry* 2008, **208(3)**:1211-15.

Jiang H, Wong WH. **Statistical inferences for isoform expression in RNA-Seq.** *Bioinformatics* 2009, **25:8**, pp 1026-1032.

Lacroix V, Sammeth M, Guigo R, Bergeron A. **Exact Transcriptome Reconstruction from Short Sequence Reads.** In *Proceedings of WABI*: Sept 15-19, 2008; Karlsrhue, Germany. Edited By Keith A. Crandal and Jens Lagergren:50-63.

Li *et al*. **SOAP2: an improved ultrafast tool for short read alignment.** *Bioinformatics* 2009, **24(15)**:1966-1967.

Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD *et al*. **Highly integrated single-base resolution maps of the epigenome in Arabidopsis.** *Cell* 2008, **133(3)**:523-36

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. **Mapping and quantifying mammalian transcriptomes by RNA-Seq.** *Nature Methods* 2008, **5(7)**:621-8.

Salehi-Ashtiani K, Yan X, Derti A, Tian W, Hao T, Lin C, Makowski K, Shen L, Murray RR, Szeto D, Tusneem N, Smith DR, Cusick ME, Hill DE, Roth FP, Vidal M. **Isoform discovery by targeted cloning deep-well pooling and parallel sequencing.** *Nature Methods* 2008, **5(7)**:597-600.

Wang B-B, Brendel V. **Genomewide comparative analysis of alternative splicing in plants.** *PNAS* 2006, **103(18)**:7175-80.

Wang ET, Sandberg R, Luo S, Hrebtukova I, Zhang Lu, Mayr C, Kingsmore SF, Schroth GP, Burge CB. **Alternative Isoform Regulation in Human Tissue Transcriptomes.** *Nature* 2008, **456(7221)**:470-476.

Wang Z, Gerstein M, Snyder M. **RNA-Seq: a revolutionary tool for transcriptomics.** *Nature Reviews Genetics* 2009, **10(1)**:57-63.

Weber APM, Weber KL, Car K, Wilkerson C, Ohlrogge JB. **Sampling the Arabidopsis transcriptome with massively parallel pyrosequencing.** *Plant Physiology* 2007, Vol **144**, pp 32-42.

Chapter 5

CONCLUSIONS

In chapters 2 and 3 of this dissertation we investigated several aspects of microarray data quality assessment. Chapter 2 introduced a new method for classifying data quality based on the unsupervised assessment of abstract quality indicators. This method is applicable to a wide variety of popular quality metrics and can be regarded as a logical extension of the approach employed *de facto* by many labs. Since our method does not require extensive expert input, it is able to easily accommodate new datasets and chip types as they become available. We suggest that our method could be most useful in scenarios where large datasets need to be automatically checked for quality, perhaps in the context of a screening module embedded in an online public database, or as a component in a meta-analysis pipeline. In addition, this approach would be valuable in scenarios where samples can be re-hybridized when severe quality problems are detected. On the other hand, a weakness of all systems of quality classification based on abstract quality indicators is the issue of what to do when data is flagged as "bad." In practice, it may often be inappropriate to completely discard flawed data. Instead, a system that down-weights low quality data might be more appropriate (Ritchie, *et al.*, 2006). The method outlined in chapter 3 provides a framework for determining whether it is beneficial to actually throw away data under a given experimental design and analysis methodology. Finally, it is important to note that one of the most important factors for obtaining reproducible microarray data is adequate replication (Zhang, *et al.*, 2008). In the early days of microarray experimentation, 3 replicates per treatment was often the norm, leading to a proliferation of under-powered gene expression studies. Now that costs have decreased, and labs have access to shared core analysis pipelines, larger studies are becoming more common and microarrays are

consequently yielding more reliable data. In addition, meta-analysis studies which combine similar, independent, but small, studies have produced promising results (Larsson, *et al.*, 2006).

Meanwhile, an important new technology, high-throughput sequencing of the transcriptome (RNA-Seq), promises to yield an abundance of new data. (Wang, *et al.*, 2009). In fact, most researchers expect RNA-Seq to eventually replace microarrays for most applications (Shendure, 2008). High throughput sequencing-by-synthesis is qualitatively similar to rt-PCR, which remains the "gold standard" for interrogating transcription. The output of a high throughput sequencing analysis consists of actual "counts" of molecules rather than a fluorescence reading having some unknown relationship to the gene expression level, and some have characterized this important difference as "digital" versus "analog" measurement (Shendure, 2008). As such, the dynamic range is expected to be larger for sequencing technologies, conferring a greater ability to detect transcripts occurring at low transcription levels. Furthermore, most researchers expect that sequencing technologies will provide results that are more robust and reproducible than those achieved using microarrays ('t Hoen, *et al.*, 2008). Nevertheless, we are also beginning to appreciate that high-throughput sequencing based expression analysis comes with its own unique set of analytical biases (e.g. Oshlack and Wakefield, 2008). However, since the technology is so new, these biases have yet to be thoroughly explored.

Current research efforts have begun to address quality assessment of high-throughput sequencing data. As with microarrays, quality assessment of RNA-Seq data can occur at various stages in an experiment. Prior to sequencing, the same quality metrics used before

microarray hybridization can be used to verify RNA integrity. As discussed in chapter 1, these include the 28S/18S rRNA ratio (Sambrook, *et al.*, 2001), the RNA integrity number (RIN) (Schroeder, *et al.*, 2006), the RQS Score (Copois, *et al.*, 2007), and the RNA degradation factor (Auer, *et al.*, 2003). After sequencing, the resulting reads often include a quality score for each base, similar in nature to the Phred scores (Ewing, *et al.*, 1998a,b) used in Sanger sequencing (Sanger, *et al.*, 1977). Recent research has shown that judicious use of these quality scores improves alignment accuracy for short read Illumina data (Smith, *et al.* 2008). In chapter 2, we demonstrated how microarray level, post-hybridization quality indicators, including the BioConductor quality indicators, can be used to screen for low quality microarray data. The BioConductor library has recently been extended to include quality assessment metrics for use with high-throughput sequencing data (Morgan, *et al.*, 2009). These quality assessment indicators, which include the number of aligned and unaligned reads, base call frequencies, distribution of read duplicate counts, and alignment quality scores, are aimed at detecting low quality sequencing runs due, for example, to excessive PCR bias or RNA degradation. The method described in chapter 2 could be easily adapted to discriminate between high and low quality RNA-Seq runs based on these indicators. In addition, given a particular experimental design, the simulation methodology described in chapter 3 could likewise be adapted to identify cases in which a sequencing run is so badly flawed that it is better to completely discard the data. These are both potential topics for future research.

In Chapter 4 we introduced a new method for using RNA-Seq data to quantify the alternatively spliced isoforms present in a mixture. One of the key advantages of our method

over similar approaches is that we do not assume uniform sampling of reads over targeted transcripts. We have shown, through simulation, that using the wrong read sampling distribution can lead to incorrect conclusions about the expression of isoforms in a mixture. In an example dataset, we identified 438 genes that exhibit differential splicing between two different Arabidopsis strains. Using our isoform quantification method, we identified several alternatively-spliced genes with switch-like expression properties, as well as a number of other genes that varied more subtly in isoform expression.

The research presented in chapter 4 is ongoing. Currently, we are working on extending the method to accommodate reads that map to more than one gene and which could adversely effect the resulting expression estimates. Similarly, we plan to improve our method to allow for the automatic detection of "hidden" or unannotated spliced isoforms concurrent with the quantification of known transcripts. Improvements to accommodate paired read technologies and to provide visualization of the results are also planned. The ability to reliably compute quantitative isoform expression values will allow us to separate true alternative splicing events from spurious transcripts originating from single mis-spliced transcripts — a major problem in all alternative splicing studies. Given that changes in isoform expression level frequently involve a continuum of isoform ratios, rather than all-or-nothing expression (Blencowe, 2006), and that they are often independent of general gene expression changes (Wang, *et al*., 2008), we anticipate that our research will contribute to revealing a so far uninvestigated layer of the transcriptome. We believe that, in the future, researchers will prioritize genes for functional analysis based not only on observed changes in gene expression levels, but also on changes in alternative splicing.

Currently, the splicing code, or the full set of splice recognition signals, splice enhancers and splice silencers encoded in the genetic sequence, is currently poorly understood. A pre-requisite to deciphering the splicing code is the availability of quantitative isoform expression data from a variety of biological tissues and states. The nucleotide sequences of differentially spliced genes can then be scanned for motifs that occur in association with the observed splicing patterns (e.g. Zhao, *et al.*, 2009; Kim, et al., 2009). There has been some success in generating data of this nature using an exon-junction microarray platform (Pan, *et al.,* 2004), but high-throughput RNA-Seq promises to provide more reliable and complete data at a lower cost. The isoform quantification method we have described in chapter 4 is, therefore, a first step towards unraveling the signals which orchestrate alternative splicing.

An improved understanding of alternative splicing is critical to many important fields of biology. For example, several studies have suggested that between 50-60% of human disease-causing mutations act by disrupting normal splicing patterns (Cartegni, *et al.*, 2002; Pagenstecher, *et al.*, 2006; Lopez-Bigas, *et al.*, 2005). In addition, changes in splicing patterns are known to occur during the normal course of a variety of cancers (Wang and Cooper, 2007). The implications of these observations are potentially enormous, impacting the ways in which we prioritize candidate disease SNPs for functional analysis, design pharmaceutical therapies, and search for biomarkers. Likewise, given the lack of correlation between genome size and phenotypic complexity, a more detailed understanding of alternative splicing will be critical in identifying evolutionarily important mutations. Tissue-specific differences in alternative splicing will also be important in unraveling the intricate

processes of cellular differentiation and growth within complex organisms. Until recently, alternative splicing has often been overlooked in all of these fields, primarily due to an inability to make high-throughput, quantitative measurements of alternatively spliced transcripts. Specialized microarray platforms, and, especially, high-throughput transcriptome sequencing are fundamentally changing the ways in which we investigate these important questions. We hope that the methods detailed in this dissertation can serve to improve the reliability of conclusions derived from data generated by these powerful new technologies.

References

Auer H, Lyianararchchi S, Newsom D, Klisovic MI, Marcucci G, *et al*. **Chipping away at the chip bias: RNA degradation in microarray Analysis**. *Nature Genetics* 2003, **35**:292-293.

Blencowe BJ. **Alternative Splicing: New insights from global analysis.** *Cell* 2006, **126:**37-47.

Cartegni L, Chew SL, Krainer AR. **Listening to silence and understanding nonsense: exonic mutations that affect splicing.** *Nature Reviews Genetics* 2002, **3**:285-298.

Copois V, Bibeau F, Bascoul-Mollevi C, Salvetat N, Chalbos P, *et al*.: **Impact of RNA degradation on gene expression profiles: assessment of different methods to reliably determine RNA quality**. *Journal of Biotechnology* 2007, **127(4)**:549-59.

Ewing B, Hillier L, Wendl MC, Green P. **Base-calling of automated sequencer traces using phred. I. Accuracy assessment**. *Genome Research* 1998, **8**:175-185.

Ewing B, Green P. **Base-calling of automated sequencer traces using phred. II. Error probabilities**. *Genome Research* 1998, **8**:186-194.

Kim J, Zhao S, Howard BE, Heber S. **Mining of cis-regulatory motifs associated with tissue-specific alternative splicing.** In Mandoiu I, Narasimhan G, Zhang Y (Eds.), *Proceedings of the 5th International Symposium on Bioinformatics Research and Applications 2009 (ISBRA 2009)*, Ft. Lauderdale, FL, p 260-271.

Larsson O, Wennmalm K, Sandberg R. **Comparative microarray analysis**. *OMICS: A Journal of Integrative Biology* 2006, **10(3)**: 381-397.

Lopez-Bigas N, Audit B, Ouzounis C, Parra G, Guigo R. **Are splicing mutations the most frequent cause of hereditary disease?** *FEBS Lett.* 2005, **579**:1900-1903.

Morgan M, Anders S, Lawrence M, Aboyoun P, Pages H, Gentleman R. **ShortRead: a Bioconductor package for input, quality assessment, and exploration of high throughput sequence data**. *Bioinformatics* 2009, August 3 [Epub ahead of print].

Oshlack A, Wakefield MJ. **Transcript length bias in RNA-seq data confounds systems biology**. *Biology Direct* 2008, **4:14**.

Pagenstecher C, *et al.* **Aberrant splicing in MLH1 and MSH2 due to exonic and intronic variants**. *Human Genetics* 2006, **119:**9-22.

Pan Q, *et al.* **Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform**. *Mol Cell* 2004, **16**(**6**):929-941.

Ritchie ME, Diyagama D, Neilson J, van Laar R, Dobrovic A, Holloway A, Smyth G: **Empirical array quality weights in the analysis of microarray data.** *BMC Bioinformatics* 2006, **7**:261.

Sambrook J, Russel DW. *Molecular Cloning: A Laboratory Manual, 34d ed*. Cold Spring Harbor Laboratory Press 2001, Cold Spring Harbor, NY.

Sanger F, Nicklen S, Coulson Ar. **DNA sequencing with chain-terminating inhibitors.** *PNAS* 1977, **74**:5463-7.

Schroeder A, Mueller O, Stocker S, Salowsky R, Leiber M, *et al*. **The RIN: an RNA integrity number for assigning integrity values to RNA measurements.** *BMC Molecular Biology* 2006, **7:3**.

Shendure J. **The beginning of the end for microarrays?** *Nature Methods* 2008. **5(7)**:585-87.

Smith AD, Xuan Z, Zhang MQ. **Using quality scores and longer reads improves accuracy of Solexa read mapping**. *BMC Bioinformatics* 2008, **9**:128.

't Hoen PAC, Ariyurek Y, Thygesen HH, Vreugdenhil E, Vossen RHAM, *et al.* **Deep sequencing-based expression analysis shows major advances in robustness,**

resolution and inter-lab portability over five microarray platforms. *Nucleic Acids Research* 2008, **36(21):**e141.

Wang ET, Sandberg R, Luo S, Hrebtukova I, Zhang Lu, Mayr C, Kingsmore SF, Schroth GP, Burge CB. **Alternative Isoform Regulation in Human Tissue Transcriptomes.** *Nature* 2008, **456(7221)**:470-476.

Wang GS, Cooper TA. **Splicing in disease: disruption of the splicing code and the decoding machinery**. *Nature Review Genetics* 2007, **8**:749-761.

Wang Z, Gerstein M, Synder M. **RNA-Seq: a revolutionary tool for transcriptomics**. *Nature Reviews Genetics* 2009, **10(1)**:57-63.

Zhang M, Yao C, Guo Z, Zou J, Zhang L, *et al*. **Apparently low reproducibility of true differential expression discoveries in microarray studies.** *Bioinformatics* 2008, **24(18)**: 2057-2063.

Zhao S, Kim J, Heber S. **Analysis of cis-regulatory motifs in cassette exons by incorporating exon skipping rates.** In Mandoiu I, Narasimhan G, Zhang Y (Eds.), *Proceedings of the 5th International Symposium on Bioinformatics Research and Applications 2009 (ISBRA 2009)*, Ft. Lauderdale, FL, p 260-271.