8-2013

RNA-SEQUENCING APPLICATIONS: GENE EXPRESSION QUANTIFICATION AND METHYLATOR PHENOTYPE IDENTIFICATION

Guoshuai Cai

RNA-SEQUENCING APPLICATIONS: GENE EXPRESSION QUANTIFICATION AND METHYLATOR PHENOTYPE IDENTIFICATION

by
Guoshuai Cai, M.S.

APPROVED:

_____
Supervisory Professor: Shoudan Liang, Ph.D.

_____
Li Zhang, Ph.D.

_____
Qiang Shen, M.D. Ph.D.

_____
Roel Verhaak, Ph.D.

_____
Han Liang, Ph.D.

APPROVED:

_____
Dean, The University of Texas
Graduate School of Biomedical Sciences at Houston

RNA-SEQUENCING APPLICATIONS: GENE EXPRESSION QUANTIFICATION AND METHYLATOR PHENOTYPE IDENTIFICATION

A

DISSERTATION

Presented to the Faculty of

The University of Texas

Health Science Center at Houston

and

The University of Texas

MD Anderson Cancer Center

Graduate School of Biomedical Sciences

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

by

Guoshuai Cai, M.S.

Houston, Texas

August, 2013

To my dear family

ACKNOWLEDGEMENTS

I thank my advisor, Dr. Shoudan Liang.

I thank my advisory and candidacy committee members: Dr. Li Zhang, Dr. Roel Verhaak, Dr. Han Liang, Dr.Qiang Shen, Dr. Christopher Amos, Dr. Keith A. Baggerly, Dr. Peter Müller, Dr. Yuan Ji and Dr. Xueling Huang.

I thank Dr. Victoria Knutson, Ms. Lourdes Perez and other staff members from GSBS.

I thank my family. My parents, my brother and my beloved wife Feifei Xiao, I love you.

I thank all my friends.

ABSTRACT

RNA-SEQUENCING APPLICATIONS: GENE EXPRESSION QUANTIFICATION AND METHYLATOR PHENOTYPE IDENTIFICATION

Publication No._____

Guoshuai Cai, M.S.

Supervisory Professor: Shoudan Liang, Ph.D.

My dissertation focuses on two aspects of RNA sequencing technology. The first is the methodology for modeling the overdispersion inherent in RNA-seq data for differential expression analysis. This aspect is addressed in three sections: (1) Investigation of the relationship between overdispersion and sequencing depth on the gene level and modeling for differential expression analysis. (2) Investigation of the relationship between overdispersion and sequencing depth on the position level and modeling for differential expression analysis. (3) Investigation of the hidden bias on the measurement of spike-in transcripts and modeling to correct this bias. The second aspect of sequencing technology is the application of RNA-seq data to identify the CpG island methylator phenotype (CIMP) by integrating datasets of mRNA expression level and DNA methylation status.

Section 1: The cost of DNA sequencing has reduced dramatically in the past decade. Consequently, genomic research increasingly depends on sequencing technology. To measure gene expression, RNA-seq, sequencing mRNA-converted cDNA, is becoming a widely used method. As it remains elusive how the sequencing capacity influences the accuracy of mRNA expression measurement, an investigation of that relationship is required. First, we empirically calculate the accuracy of the RNA-seq measurement from repeated experiments and identify the source of error to be mainly library preparation procedures. Second, we observe that accuracy improves along with the increasing sequencing depth. However, compared with the accuracy predicted from the binomial distribution, the rate of improvement as a function of sequence reads is globally slower, which indicates that overdispersion exists and is related to sequencing depth. To model the overdispersion,

we therefore use the beta-binomial distribution with a new parameter indicating the dependency between overdispersion and sequencing depth. Our modified beta-binomial model performs better than the binomial or the pure beta-binomial model with a lower false discovery rate.

Section 2: Although a number of methods have been proposed to handle these biases and spurious effects in order to accurately analyze differential RNA expression on the gene level, modeling on the base pair level is required to precisely estimate the mean and variance by taking the non-uniformity of RNA-seq into account. We show in Chapter 1 that the overdispersion rate decreases as the sequencing depth increases on the gene level. Here, we find that the overdispersion rate decreases as the sequencing depth increases on the base pair level, in agreement with what we previously reported for the gene level. Investigating the impact of the sequencing depth and local primer sequence on the overdispersion rate, we observe that the local primer sequence no longer significantly influences the overdispersion rate after stratification by the sequencing depth. Also, we propose four models and compare them with each other and with the DESeq model based on the likelihood value, Akaike information criterion, goodness-of-fit χ^2 test, false discovery rate and the area under the curve. As expected, our beta binomial model with a dynamic overdispersion rate is shown to be superior. Furthermore, this model has many advantages that make it more desirable than DESeq.

Section 3: We investigate biases in RNA-seq by exploring the measurement of the external control, spike-in RNA. This study is based on two datasets with spike-in controls obtained from a recent study. In the ENCODE dataset, 51 replicates of human samples were measured, and in the modENCODE dataset, 6 fly samples from difference scenarios were sequenced. By comparing the patterns of the reads and correlations among samples, we observe an undiscovered bias in the measurement of the spike-in transcripts that arises from the influence of the sample transcripts in RNA-seq. Also, we find that this influence is related to the local sequence of the random hexamer that is used in priming. We suggest a model of the inequality between samples and to correct this type of bias. After corrections, the Pearson correlation coefficient increases by 0.1.

Section 4: The expression of a gene can be turned off when its promoter is highly methylated. Several studies have reported that a clear threshold effect exists in gene silencing that is mediated by DNA methylation. As the transcriptional regulatory system is complicated and has many components, it is reasonable to assume the thresholds are specific for each gene. It is also intriguing to investigate genes that are largely controlled by DNA methylation, as their methylation possibly plays an important role in cancer oncogenesis by significantly inhibiting transcription. These genes are called "L-shaped" genes because they form an "L" shape when plotting mRNA expression level against DNA methylation status. We develop a method to determine the DNA methylation threshold using 997 samples across 7 cancer types from TCGA datasets. Then, from 285 tumor samples and 21 normal samples of breast tissue, we select 128 "L-shaped" genes according to our criteria and identify the CIMP using biclustering and hierarchical clustering. We identify a new CIMP of *BRCA* with 11 markers and observe significant correlation between the CIMP+ subtype and the wild-type *TP53* mutation, ER+/PR+ status, higher age at initial pathologic diagnosis, better treatment response and the possibility of a longer survival time. The 11 CIMP markers are shown to be associated with *TP53* directly or indirectly, and enriched in cancer and other disease networks. Also, we find that 7 epigenetic genes are strongly correlated with both the new CIMP and *TP53* mutation. Based on our findings, we propose a model of the *TP53*-mediated regulatory network with two components: "guidance" genes and "ustainer" genes.

In conclusion, we provide a detailed understanding of the relationship between the overdispersion rate and sequencing depth, which will aid in the analysis of RNA-seq data for detecting and exploring biological problems. Additionally, we demonstrate a novel property of overdispersion in that it improves with sequencing depth. We propose a beta-binomial model with dynamic overdispersion on the position level. We demonstrate that our model provides a better fit for the data. We reveal a new bias in RNA-seq and provide a detailed understanding of the relationship between this new bias and the local sequence, which will aid in understanding RNA-seq technology and in correcting for this bias in the analysis of RNA-seq data. We develop a powerful method to

dichotomize methylation status and consequently we identify a new CIMP of breast cancer with a distinct classification of molecular characteristics and clinical features. Our results suggest that methylation may play an important role in resisting tumor development and that “guidance” genes and genetic modifiers BMI1, IDH1 and TET1 are potential new therapeutic targets.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

1.1 Biases and Spurious Effects in RNA-seq Technology

RNA-seq is becoming a common technique for surveying RNA expression. Because the cost of next-generation sequencing is dropping dramatically, RNA-seq may soon replace microarray analysis in genome-wide surveys of gene expression [1]. Because of the complexity of RNA-seq technology, many inherent biases and unwanted effects exist that make it difficult to develop accurate methods for analyzing RNA-seq data. The various biases and effects that have been identified include a base calling bias, GC content bias, hexamer priming bias, length bias, library effect, and batch effect [2,3,4,5,6,7,8].

Base calling has been reported to be biased. The errors in base calling have been found to increase from the start to the end of the reads. It is known that we can compensate for this error rate by increasing the sequencing depth. Dohm et al. showed that the error rate will shrink to close to 0 when the sequencing depth reaches 20X [9]. Bravo and Irizarry demonstrated that the error rate varies for different nucleotide compositions [7].

The GC content bias arises from the generation of a greater number of reads in regions that are enriched in G-C bases. Thus, sequencing reads have been reported to be differently distributed, over-representing genes with more GC-enriched regions [7,9]. Also hexamer priming introduces bias according to its local sequence by PCR amplification. Li et al. showed non-uniformity in RNA-seq data [4], demonstrating that the number of sequencing reads per nucleotide can vary by 100-fold across the same gene. However, the patterns of reads are highly conserved across tissues. Local sequencing bias has been reported to be a major source of bias in the hybridization step for both microarray and sequencing technologies [3,4,10,11]. On microarrays, probe signals depend on the probe sequence. Li et.al investigated this influence and developed a model of binding interactions that improves the measurement of gene expression. Furthermore, probes may cross-hybridize to off-

target transcripts that have sequences similar to those of the target transcripts. Many studies have been reported on detecting and modeling cross-hybridization events [10,12,13]. The use of a random hexamer primer has been shown to cause bias because of the specific hybridization affinity of its sequence. Both Hansen et al. and Li et al. confirmed that random hexamer priming causes biases in the nucleotide composition at the beginning of the transcriptome sequencing reads [3,4].

Length bias occurs in differential expression analysis of RNA-seq data. For genes with similar expressions, the total sequencing reads on them would correlate with their length such that more reads are observed on longer genes, and more power would be obtained in statistical testing in the differential expression analysis. Oshlack et al. validated that more differential expressions were detected on longer genes using several widely used statistical methods [14].

Furthermore, the sequencing measurement has been found to be influenced systematically by external factors such as the time when the measurement is made, the specific technician performing the measurements, and the specific library preparation procedure. Leek et.al showed a distinct pattern produced by the batch effect on DNA sequencing measurements from the 1000 Genome project [15]. Because of the extra noise introduced by biases or spurious effects, measurements on the same gene often show a level of dispersion that is larger than that given by a Poisson distribution. A comparison of different samples often shows a dispersion that is larger than that given by a binomial distribution. This phenomenon has been termed overdispersion [4].

1.2 Current Methods to Correct Biases and Spurious Effects

These biases and spurious effects have been investigated in several studies and researchers have suggested methods to control or correct them. Several methods have been developed to reduce the error rate of base calling and control the quality of the reads. The use of spike-in transcripts has been suggested as a means of providing quality control and as a standard for normalization. In addition, many statistical methods have been developed to correct for the biases and undesirable effects and to take them into account in the downstream analysis.

1.2.1 Reads Quality Control

Because higher error rates have been found at the end of sequencing reads, some studies have suggested cutting several base pairs from the end of sequencing reads. Ensuring a high sequencing depth is one way to efficiently reduce the error rate. Alternative base calling methods have been developed to decrease the error rate of sequencing reads [16,17,18], and other methods have been developed to enhance the quality of sequencing reads after base calling [19,20,21,22].

1.2.2 Spike-in Transcripts

In order to achieve quality control of measurements, the use of spike-in transcripts was first designed for microarray technology. Spike-in external controls are RNA strands synthesized in vitro that are designed to be fairly different from the genome being studied. Several spike-in sets have been developed, including the GeneChip eukaryotic poly(A) RNA control kit, External RNA Controls Consortium (ERCC) spike-in controls, Agilent Technologies spike-in set and others. The ERCC aims to establish 100 platform-independent controls for evaluating the quality of measurements [23]. Recently, Jiang et al. synthesized ERCC RNAs as a standard for next-generation sequencing as well. The ERCC RNA was synthesized from DNA derived from the deep-sea vent microbe *M. jannaschii* or the *B. subtilis* genome or by in vitro transcription of de novo DNA sequences [24]. The researchers used ERCC RNA to measure the biases produced by the GC content bias, transcript length bias, and measurements correlated with the local sequence of the priming hexamer. The measurement of spike-in controls can be used to measure the performance of data normalization and approaches to differential expression analysis [23,25].

1.2.3 Statistical Methods

Several statistical methods have been suggested to correct for the biases and spurious effects of RNA-seq technology. In order to correct for the uniformity of measurement on the same gene, Li et

al. proposed a statistical method that is based on a linear model with the nucleotide composition at the beginning of the transcriptome sequencing reads as the predicative factor [4]. Li et al. were able to explain more than 50% of the variations and observed better estimations of gene expression on data from both Illumina and Applied Biosystems. Also, Zheng et al. proposed a generalized linear model based on the principal components transformed from dinucleotide compositions and gene length. Their method corrected for the biases on the gene level and could be used for meta-analysis on multiplatform data in terms of gene expression levels [8].

Aiming to take overdispersion into account, the analysis of large datasets produced by RNA-seq requires compatible models and methods. Among those analyses, differential expression (DE) testing is foremost as an essential step [26]. Several methods for DE testing have been proposed, including reads per kilobase of gene length per million mapped reads (RPKM), a 2-stage Poisson model [27], a Bayesian method for calling DE [28], edgeR [29], DESeq [30], and others [26]. Compared with the methods based on Poisson and binomial models, methods based on a quasi-Poisson model and negative binomial model perform better by taking overdispersion into account. Auer and Doerge proposed the 2-stage Poisson model (TSPM) for differential expression analysis. TSPM first tests whether the gene is overdispersed and chooses the Poisson model or quasi-Poisson model automatically for the analysis of genes that are not overdispersed and which are overdispersed, respectively [27]. Further, the generalized linear model could be used to estimate parameters efficiently based on log transformation of the Poisson or quasi-Poisson model. The methods edgeR, DESeq and baySeq were based on a negative binomial model. The method edgeR estimates the variance for each gene specifically by borrowing information from all genes. An empirical Bayes-like approach was used to achieve this aim [29]. Also, DESeq estimates specific dispersions from the local regression from the dispersions and means [30]. Using the negative binomial distribution as the priority distribution, baySeq employs an empirical Bayes method to estimate the specific posterior probability of a DE model for each gene. In addition, methods based on a Bayesian hierarchical

model have been proposed for DE analysis [28]. Oshlack et al. provided a critical review of the methods commonly used for DE analysis [31].

To correct for length bias, Oshlack suggested performing the DE analysis within a fixed-length window on all genes. However, much information would be lost using that approach. Later, Gao et.al proposed two methods to adjust for the length bias in DE analysis based on a Poisson model [2].

The best way to deal with batch effects is to avoid them by using a careful research protocol. A surrogate variable analysis can also be used to correct for batch effects through coefficient estimation in a linear model [32]. Also, methods such as fRMA have been developed to capture batch-to-batch variations for multiarray analysis [33,34].

1.3 Sequencing Analysis is Still Young and Our Research Hypothesis

RNA-seq analysis methods are relatively new, and the biases and spurious effects inherent to RNA-seq technology are still not known clearly. Although many studies have investigated these biases and spurious effects and several statistical methods have been suggested for correcting or modeling them [4,27,29,35], the overdispersion properties of this technology remain elusive. Investigations of the properties of overdispersion will benefit the understanding of sequencing technology and the accuracy of downstream analyses.

1.3.1 The Dependence of Overdispersion and Sequencing Depth

As discussed above, by accounting for overdispersion, the negative binomial and quasi-Poisson models showed much better performance than the former models based on binomial or Poisson distributions. And models with dynamic overdispersion rates were superior to models with constant overdispersion rates [27,29,35]. The constant overdispersion rate is good for describing a genetic difference. However, for genes with no genetic variations, a constant overdispersion rate is inappropriate, based on the intuition that along with increasing sequencing depth the accuracy of the measurement should improve. Therefore, knowing the properties of overdispersion will be crucial

for accurate down streaming analysis, especially for DE analysis. Because the relationship between overdispersion and sequencing depth had not been illustrated, to our knowledge, in this study we first investigated this intuitive notion of dependency.

1.3.2 Modeling on Base Pairs Rather than Genes

Two important and related findings on RNA-seq have been reported. First, Li et al. showed non-uniformity in RNA-seq data [4], including that the number of sequencing reads per nucleotide can vary by 100-fold across the same gene. However, they found that the patterns of reads are highly conserved across tissues. Second, both Hansen et al. and Li et al. confirmed that random hexamer priming causes biases in the nucleotide composition at the beginning of transcriptome sequencing reads [3,4]. These findings indicate that the measurement on each base pair of a gene has a specific mean and variance. Therefore, it is more reasonable to model the measurement based on a base pair rather than a gene. Also, in Chapter 1.3.1, we assumed that the main source of the variance is in the library preparation steps prior to DNA sequencing and showed that the overdispersion rate decreases as the sequencing depth increases on the gene level from the above hypothesis. On the basis of these findings, we developed three corresponding hypotheses. (1) Where there is no difference between two samples, the ratio of the measurements of each base pair on a gene from two samples is a constant across the whole gene. With this assumption, the beta-binomial model is appropriate for comparing samples in two conditions, as modeling on proportion will transform the distribution of reads mapped to base pairs from non-uniformity to uniformity. (2) The overdispersion rate is influenced by random hexamer priming. (3) On the base pair level, as we found on the gene level, the overdispersion rate decreases as the sequencing depth increases. To test these hypotheses, we developed two beta-binomial models. One was a full model based on all three hypotheses, and the other was a reduced model based on hypotheses 1 and 3.

1.3.3 Study of Biases from Spike-in Transcripts

As reviewed above, the local sequence has been reported to influence the measurement from RNA-seq for both microarray and sequencing technologies. Measurement signals in microarrays can be influenced by the affinity of the probe sequence with the target transcript. And cross-hybridization is a well-known cause of bias in microarray technology; therefore, probes could target spike-in transcripts through cross-hybridization [13]. In RNA-seq, the interaction between spike-in and target transcripts remains elusive, although the spike-in transcripts were designed to be quite different from the target transcripts. Also, as a stable quantitative source, spike-in transcripts will offer new insights in understanding biases in RNA-seq.

In this study, aiming to identify potential sources of bias in RNA-seq and develop statistic models to correct for such bias, we investigated the potential biases from the measurement of spike-in transcripts.

1.4 Beta-binomial Distribution and Overdispersion

For RNA-seq analysis, a fundamental question is the relationship between the accuracy of the measurement and the increasing depth of sequencing [36]. The sequencing depth was represented by the total number of mapped reads for each position of a gene, which can be accumulated from multiple lanes for the same sample. Aiming to compare mRNA expression from two samples, we usually assume a binomial distribution. In the binomial model, the uncertainty is $\frac{1}{\sqrt{n}}$, where n is the number of mapped reads on the gene and the uncertainty will shrink to 0 when n is large. It has been shown that the ratio of read counts from two samples follows a binomial distribution [35,37]. However, as discussed above, biological differences and biases introduce extra noise. A distribution with a larger dispersion than that observed from the binomial distribution has been shown by comparing the measurements of different samples [3,4]. A beta-binomial distribution can be used to capture this overdispersion appropriately. Beta-binomial distributions have been used for differential

expression analysis on SAGE data [38], and on peptide counts from label-free tandem mass spectrometry-based proteomics [39]. The probability mass function of a beta-binomial distribution is

$$f(k|n,\alpha,\beta) = \binom{n}{k}\frac{B(k+\alpha,n-k+\beta)}{B(\alpha,\beta)}.$$

This is the probability of obtaining k observations from n sample pools. Parameters α and β were inherited from the beta distribution, followed by the probability of the observation. After reparameterization by $= \frac{\alpha}{\alpha+\beta}$, $\theta = \frac{1}{\alpha+\beta}$, the variance from the beta-binomial distribution can expressed as

$$Var(p) = p(1-p)[\frac{1}{n} + \frac{1-\frac{1}{n}}{\frac{1}{\theta}+1}].$$

Form the equation, the dispersion contains two parts: the first part is from the binomial distribution, which shrinks to 0 when n is large, and when n is large, the second part is a constant. Although using a constant for all genes is appropriate to capture the genetic variant; it may not be correct for genes without genetic variations.

1.5 DNA Methylation and Its Significance in Cancer

It has been shown that DNA methylation is related to cancer through the hypermethylation of cancer suppression genes and hypomethylation of oncogenes. This research on DNA methylation has focused on the methylation status of CpG island promoters. The reason for this focus is that CpG island promoter methylation has been demonstrated to silence genes permanently in mammalian cells. Recent studies have shown that global epigenomic alterations cause silencing in cancer with the altered pathways involved in stem cell growth and differentiation [40].

1.5.1 DNA Methylation and Cancer

It is known that a gene may gain or lose its function through mutation, amplification, or deletion of the genomic neighborhood of the gene. It is also increasingly appreciated that, in cancer, a gene may also lose its function through epigenetic changes—particularly by DNA methylation of its promoter [41,42]. Since DNA methylation as a "silencing" epigenetic change was proposed in 1975, many of its properties have been identified. About 80% of CpG pairs in the human genome are chemically modified by the attachment of a methyl group to the cytosine ring. CpG methylation represses transcription and is thought to be a mechanism to control the transposable elements. The only location where CpG pairs tend not to be methylated is near the transcription start sites, in CpG-rich regions called CpG islands. CpG methylation is epigenetic, i.e., the methylation pattern is preserved mitotically. A genome-wide survey of DNA methylation has shown that the pattern of DNA methylation changes drastically in cancer: there is massive hypomethylation genome-wide, but hypermethylation in CpG island promoters. The hypermethylation of CpG island promoters results in the silencing of a large number of genes. By some estimate, the number of genes silenced by DNA methylation is about ten times the number of mutated genes [43,44], therefore the number of genes silenced by DNA methylation is much larger than the number of mutations.

1.5.2 CpG Island Methylator Phenotype (CIMP)

CIMP was first discovered in colorectal cancer [45] as tumor-specific CpG island hypermethylation of a subset of genes in a subset of tumors. This was confirmed later [46] and has also been found recently in glioma [47]. Previously, several aberrant methylations of genes were reported in breast cancer [48,49,50,51,52,53], and DNA methylation patterns have been claimed to be associated with histological tumor grade [49,54], tumor growth [55,56,57], hormone receptor status, Her2 expression [56,58,59,60], and breast cancer subtypes [61,62,63]. Only a few studies have claimed CIMP with hypermethylated genes for breast cancer [64]. Others have argued that more studies are required in order to confidently state that CIMP exists for breast cancer [51]. However, very recently, TCGA identified methylation clusters that significantly correlate with

mRNA subtypes, and mutations of *TP53*, *PIK3CA*, *MAP3K1* and *MAP2K4* [65]. In order to investigate CIMPs with the new insight of dichotomizing methylation status, the present study focused on breast cancer.

1.5.3 DNA Methylation and mRNA Expression

The expression of a gene can be turned off when its promoter is highly methylated. The exact amount of methylation that will trigger gene silencing depends on the gene as well as the relative position of the CpG site to the transcription start site (TSS). Several studies have reported that a clear threshold effect exists in the gene silencing mediated by DNA methylation, and that significant transcription inhibition occurs only when the methylated CpG islands reach a certain level [66,67]. We also verified this threshold effect genome-wide in this study. The transcription regulatory system is well known as an extremely complicated system with many components, including gene-gene interaction, microRNA regulation, and DNA methylation. Also, the hybridization properties of the probe affect the signal strength for C/T nucleotides and introduce an additional variable from the measured beta value to the gene expression status. Therefore, the threshold of DNA methylation should be gene-specific. To determine whether a gene is turned off by DNA methylation, we therefore must determine a probe-specific threshold in order to dichotomize the methylation status. As we have discussed, the DNA methylation threshold is a stable indicator of whether a gene has lost function. There are several potential benefits, including a better definition of the CIMP. A good way to find the CIMP is by biclustering, in which we search for a sub-cluster of genes and tumors that have the same methylation status. Dichotomizing methylation makes the task of biclustering easier [68,69]. The measurement of DNA methylation can be represented by a β value for each tag, which is the proportion of methylated signals among all signals.

$$\beta_i = \frac{M_i}{U_i + M_i},$$

where for the i-th tag, M_i is the methylation signal and U_i is the unmethylation signal.

1.6 Reduced Representation Bisulfite Sequencing (RRBS)

DNA methylation can be determined genome-wide using reduced representation bisulfite sequencing (RRBS). Bisulfite treatment of DNA converts cytosine to uracil but leaves methylated C unchanged. DNA sequencing of the whole genome and comparing the sequence at CG with the reference therefore allows the methylation status to be determined genome-wide [70,71]. Whole-genome sequencing is expensive. RRBS is cheaper but it is difficult to control the coverage. Most of the genome-wide data on cancer are obtained by reading C/T polymorphism using well-established SNP arrays. Illumina Infinium is an assay specially designed to measure methylation in the promoters of annotated genes. It has on average four probes per gene and therefore measures the methylation of four CpG sites per gene. This is far less expensive and is at present the most cost-effective method of measuring methylation systematically.

1.7 Challenges and Our Hypothesis on DNA Methylation

Many studies have been developed on DNA methylation and many findings have been reported. Several properties of CpG island (CGI) methylation have been classified from studies [72]. We list six such properties here: (1) at a transcription start site (TSS), most CpG islands are hypo methylated; (2) long-term silencing is associated with CGI methylation; (3) sometimes a tissue-specific pattern can be identified in CGIs in gene bodies; (4) compared with that of CGIs, the methylation status of non-CGIs is more tissue-specific and more dynamic; (5) rather than elongation, methylation blocks the start of transcription; and (6) cancer-causing mutations can be the consequence of methylation in gene bodies. However, it is still a big challenge to clarify the mechanisms underlying DNA methylation and its function regarding RNA expression. We sought to investigate genes that are largely regulated by DNA methylation.

Aiming to identify the genes of interest and determine the DNA methylation threshold for the inhibition of mRNA expression as discussed above, in this study we took advantage of computing conditional mutual information scores, which will be robust in determining the methylation threshold

specific to one site on a particular gene. The novelty of this new method is that it relates gene silencing by DNA methylation of promoter CpG islands to gene expression in several tissues, and increases the accuracy of the determination of mRNA methylation status and expression status. Fortunately, some consortia such as TCGA have made efforts to profile a large collection of patients' samples for mRNA expression, miRNA expression, DNA copy number, and methylation status, which makes our study feasible. Based on the genes identified to have large contributions from DNA methylation on transcription regulation, we intended to find a new CIMP of breast cancer in this study.

RNA sequencing technology is complicated and is characterized by many inherent biases and spurious effects. Statistical methods are still limited for accurately estimating overdispersion in DE analysis. Studies have shown that the sequencing reads are not uniformly distributed on genes and are correlated with the nucleotide composition of the hexamer primer local sequence. Also, with the intuition that added increments of sequencing depth will improve the accuracy of the measurement, we investigated the relationship of the overdispersion rate and sequencing depth. We suggested methods based on a beta-binomial model to estimate the overdispersion rate for DE analysis on both the gene level and the position level. Also, inspired by the cross-hybridization issue inherent in microarray technology, we investigated the measurement of spike-in transcripts from RNA-seq, aiming to identify hidden biases. In addition, in order to identify cancer-related genes in terms of methylation status, we developed a method to identify genes for which expression was highly regulated by DNA methylation. From these genes, we identified a new CIMP with correlations between the molecular signature and clinical features.

CHAPTER 2

Accuracy of RNA-seq and Its Dependence on Sequencing Depth

In the past decade, the cost of DNA sequencing has been rapidly and dramatically decreasing. Consequently, sequencing technologies are widely used for genomic research today. Many sequencing platforms have been developed for specific aims, among those, RNA-seq is a key technique to measure gene expression. Sequencing technology is complicated, and many of its properties remain elusive even though numerous studies have been devoted to it. Based on the intuition that increasing the sequencing depth could improve the accuracy of the measurement, we sought to investigate this relationship in order to benefit downstream analyses such as differential expression analysis.

Toward this aim, we empirically evaluated the variance in three RNA-seq datasets. Based on our observation, we concluded that the error of RNA-seq measurements was mainly from the library preparation steps prior to sequencing. And we observed that increasing the sequencing depth indeed improves accuracy. However, in general, with an incremental increase in the depth of the sequencing reads, the overall dispersion decreases more slowly than predicted by the binomial distribution. This indicates that overdispersion exists and decreases along with increments in the sequencing depth. Applying this property, we developed a method based on the beta-binomial distribution with a new parameter to model the relationship between overdispersion and the sequencing depth. We borrowed the information from all genes of replicates to capture this relationship. Then, we estimated the mean and dispersion of each gene specifically according to this relationship. By indicating the specific overdispersion for each gene, our method showed a better performance than the methods based directly on binomial and pure beta-binomial distributions.

We demonstrated a novel property of overdispersion in that it improves with increments in the sequencing depth. Also, we proposed a new form of overdispersion in the beta-binomial model to

borrow the information from all genes to estimate specific parameters for each individual gene. We demonstrated that this new form fits the data better.

2.1 Methods

2.1.1 Peak of Histogram of Proportion Normalization

The normalization procedure using the peak of the histogram of proportion assumes that most genes remain unchanged in the two conditions being compared. In this normalization procedure, we fit the highest peak in the histogram of proportion to a beta function. The maximum of the beta function determines the normalization proportion p_n.

In RPKM normalization, we first count the total number of tags mapped to any gene in the RNA-seq experiment. The number of tags mapped to a particular gene is divided by the total number of tags sequenced (the unit is millions of tags), and then divided by the number of nucleotides in the gene (the unit is thousands). [73]

2.1.2 Datasets Used

The three datasets we used are listed in Table 2.1.

Table 2.1 Three datasets [73]

Data Sets	A	B	
Caltech	Normal Blood	Embryonic Stem Cells	
	Rep1Gm12878CellLongpolyaBow0981x32	PairedRep1H1hescCellPapErng32aR2x75	
	Rep2Gm12878CellLongpolyaBow0981x32	PairedRep2H1hescCellPapErng32aR2x75	
	PairedRep1Gm12878CellLongpolyaBb12x75	PairedRep3H1hescCellPapErng32aR2x75	
	PairedRep2Gm12878CellLongpolyaBb12x75	PairedRep4H1hescCellPapErng32aR2x75	
Chiang	Knock-out of TDP-43	Wild Type	
	GSM546932 A sorted	GSM546935 B sorted	
	GSM546933 D sorted	GSM546936 C sorted	
	GSM546934 E sorted		
Bullard	Brain	UHR library A	UHR library B
	SRR037457	SRR037466	SRR037470
	SRR037458	SRR037467	SRR037471
		SRR037468	SRR037472
		SRR037469	

The *Chiang dataset* consisted of five independent libraries of the deleted TDP-43 gene in the mouse. The data were derived from three independent clones of TDP-43 knockout embryonic stem (ES) cells and two independent clones of control ES cells. Raw reads were mapped to the University of California Santa Cruz mm9 genome library by efficient large-scale alignment of nucleotide databases. One gene deletion is an ideal case for testing normalization procedures with the assumption that most genes do not change.

The *Caltech dataset* consisted of two cells lines: GM12878 (normal blood) and H1hESC (embryonic stem cells), each with four libraries made independently from the same biological sample. The process involved raw Illumina reads on 2x75 datasets (RawData files on the download page, fasta format), which were run through Bowtie, version 0.9.8.1, with up to 2 mismatches. The resulting mappings were stored (RawData2 files, Bowtie format) for up to ten matches per read to the genome, spiked controls and UCSC knownGene splice junctions.

The *Bullard dataset* consisted of human brain reference RNA and human universal reference RNA as two library preparations. We used Bowtie, version 0.12.7, to align the reads to the genome (H. sapiens, NCBI 37.1 assembly). The Bowtie command we used to implement this mapping strategy was ./bowtie -a -v 2 -t -m 1 --best -strata h_sapiens_37_asm. [73]

2.1.3 Maximum-likelihood Estimation (MLE)

Let n_{ip} and m_{ip} be the tags mapped to the i-th gene and p-pair of experiment and control, respectively. The likelihood function according to the beta-binomial distribution is

$$l_{ip} = \binom{n_{ip} + m_{ip}}{n_{ip}} \frac{\prod_{l=0}^{n_{ip}-1}(\alpha_{ip}+l)\prod_{l=0}^{m_{ip}-1}(\beta_{ip}+l)}{\prod_{l=0}^{n_{ip}+m_{ip}-1}(\alpha_{ip}+\beta_{ip}+l)},$$

where α_{ip} and β_{ip} are two parameters of the beta-binomial distribution. This is equivalent to using instead the parameters $p_{ip} = \frac{\alpha_{ip}}{\alpha_{ip}+\beta_{ip}}$and $\theta_{ip} = \frac{1}{\alpha_{ip}+\beta_{ip}}$. It can be shown analytically that the proportion that maximizes the likelihood function is given by $p_{ip} = \frac{n_{ip}}{n_{ip}+m_{ip}}$. We will further assume that θ_{ip} is independent of p; we reparameterize θ_i in terms of parameters D_i and γ: $\theta_{ip} = \frac{D_i}{(n_{ip}+m_{ip})^{\gamma}}$. The parameters were determined by maximizing the likelihood

$$\log L = \sum_{ip} \log l_{ip} \ . \quad (1) \text{ [73]}$$

2.1.4 Likelihood Ratio Test

According to the likelihood ratio test, $2\ln\left(\frac{l(p_i)}{l(p_n)}\right)$ follows a x^2 distribution, where p_i is the proportion for gene i and p_n is the normalized proportion corresponding to no change in gene expression. This is the most convenient way to compute the p-value. [73]

2.1.5 FDR and ROC

To determine the false discovery rate (FDR), we assumed that any gene deemed to be significantly differentially expressed at a given p-value was false when comparing two replicates sequenced from the same biological sample. We computed the FDR by dividing

the number of falsely discovered genes at a given p -value by the number of significantly differentially expressed genes, and comparing the sample to the control at the same p -value.

To determine the receiver operating characteristic (ROC), we first established a gold standard. Approximately 1,000 genes in the Bullard dataset were previously assayed by RT-PCR in four independent experiments [74]. Differentially expressed genes were determined by t-test by Bullard et al. [37]. We used their results to draw an ROC curve when comparing the binomial and beta-binomial distributions for the Bullard dataset. For the Caltech and Chiang datasets, we assumed that the t-test provided a gold standard. In order to reduce errors for small tag counts, we required a gene to have more than 20 mapped tags.

For the Caltech data, the Benjamini & Hochberg adjustment was applied to the p -value calculated by the t-test, using a cutoff of 0.05 [75]. We could not use the FDR p -value adjustment on the Bullard dataset, as much fewer genes had differential expression levels detected from the wild/knockout samples. Therefore, we applied a cutoff of 0.05 to the p -value from the t-test and required a fold change larger than two. [73]

2.1.6 Computing the Fold Change

We related the fold change in the gene expression level FC_i to the optimized ratio p_i and obtained, by definition, $FC_i \propto \frac{p_i}{1-p_i}$. This ratio has to be calibrated against the normalization of the entire experiment. We defined p_n as no change. Therefore, $\log_2(FC_i) = \log_2(\frac{p_i}{p_n}) - \log_2(\frac{1-p_i}{1-p_n})$, where p_n is the normalized ratio as determined over the entire dataset. Infinite values of FC_i can be avoided by adding a pseudo-count to n_i and m_i so that $0 < p_i < 1$. [73]

2.2 Results

2.2.1 Normalization by Proportion

The use of a proportion is a convenient way to compare two samples. Let n_i and m_i be the number of tags mapped to gene i. The proportion is defined as $p_i = \frac{n_i}{n_i+m_i}$. It is convenient to use a proportion because differences in proportion give rise to p-values using established statistics such as binomial and beta-binomial distributions. A proportion is also a convenient component of a normalization procedure.

In order to detect differential expression in two samples, we must determine the ratio of the counts in the two samples that corresponds to the same expression. One method, adapted in calculating the RPKM, assumes that the total number of tags sequenced, and equivalently the total amount of RNA, is a constant. The problem with RPKM normalization is that the number is dominated by a few genes that receive the highest number of sequence reads. These genes may or may not remain constant under the two experimental conditions. One could also use housekeeping genes such as *POLR2A* (polymerase II) or *GAPDH* in a normalization procedure. The problem with relying on a housekeeping gene is that the normalization depends on the choice of genes. Since the number of housekeeping genes is small, this normalization procedure is subject to fluctuation due to relatively small tag counts on these genes. Bullard et al. have shown good results with an upper-quartile normalization method [37].

The most conservative normalization procedure assumes that the maximum number of genes remains unchanged in the two experimental conditions. This corresponds to the maximum in the histogram ratio of tag counts $\frac{n_i}{m_i}$. The tag count proportion p_i is more

convenient to use. The maximum in a histogram of p_i that corresponds to the neutral ratio p_n, where the expression levels are assumed to be equal in the two samples. This maximum can be determined from fitting a Gaussian (or beta function) to the peak of the histogram (Figure 2.1). In this formulation, the RPKM normalization corresponds to choosing $p_n = \frac{N}{N+M}$, where N and M are the total number of tags to genes in the experiment and control.

This peak of histogram normalization is expected to be the most reasonable procedure for the Chiang dataset [76], which consists of the wild-type and knockout versions of the *TDP-43* gene (see Data and Methods for details). For this dataset, we expect the perturbation to the global gene expressions to be smaller than when comparing two different types of cells. Indeed, our peak for the histogram normalization procedure resulted in a median of base-2 logarithm of expression difference ratio between the wild-type and knockout gene of 0.014, which is to be compared to 0.025 for the median under the RPKM normalization procedure. This showed that peak normalization was comparable to and perhaps slightly better than RPKM normalization.

Normalization is performed according to the assumption that most of the genes do not change expression in the two experimental conditions. Although this convenient assumption is probably true in most cases, it has no ironclad biological justification. [73]

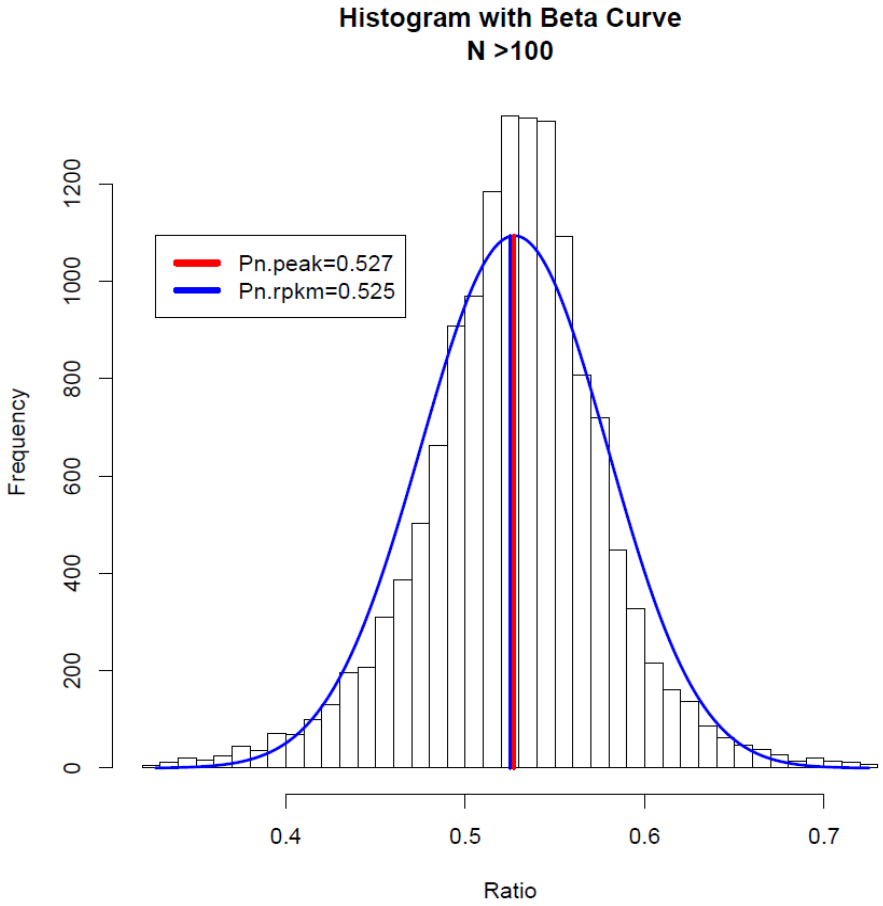


Figure 2.1 Histogram of proportions and peak of histogram of proportion normalization. The peak in the histogram corresponds to the largest density of genes. To determine the peak maximum, the histogram was fitted to a beta function. The blue curve shows the best fit with

the maximum at $p_n = 0.527$. This is to be compared to the proportion corresponding to RPKM normalization, 0.525. [73]

2.2.2 Binomial Distribution Fits the Variance from the Same Library but not from Different Libraries

We empirically studied errors in RNA-seq experiments by examining the variance from replicated measurements. We first examined the fluctuation in reads mapped to a gene from duplicate experiments based on the same biological sample. The p-values of the differences were computed according to a binomial distribution by comparing to a neutral ratio p_n as determined by peak normalization. For the same sample and the same library preparation sequenced in different lanes of the Illumina sequencer, the histogram of the p-values is flat (Figure 2.2 a). This indicates that the errors in different lanes containing samples from the same library are consistent with the binomial distribution. In contrast, the histogram of p-values according to the binomial distribution for two independent library preparations showed clear overabundance of small p-values (Figure 2.2 b). This demonstrated that the binomial distribution does not adequately describe the data---the dispersion of the random fluctuation is stronger than that given by the binomial distribution. We use the term library preparation to refer to an independent extraction of RNA, conversion to DNA and PCR amplification of DNA. Since the experiment and the control must be in separate library preparations, it is important to capture this overdispersion. The overabundance of small p-values for different libraries was also true when we used Fisher's exact test (data not shown). When we used the beta-binomial distribution to compute the p-values for the different libraries, the histogram was flat. This shows that the overdispersion is accounted for by the beta-binomial distribution. A Q-Q plot against either a binomial or beta-binomial distribution (data not shown) also indicated that the beta-binomial distribution better fit the data. [73]

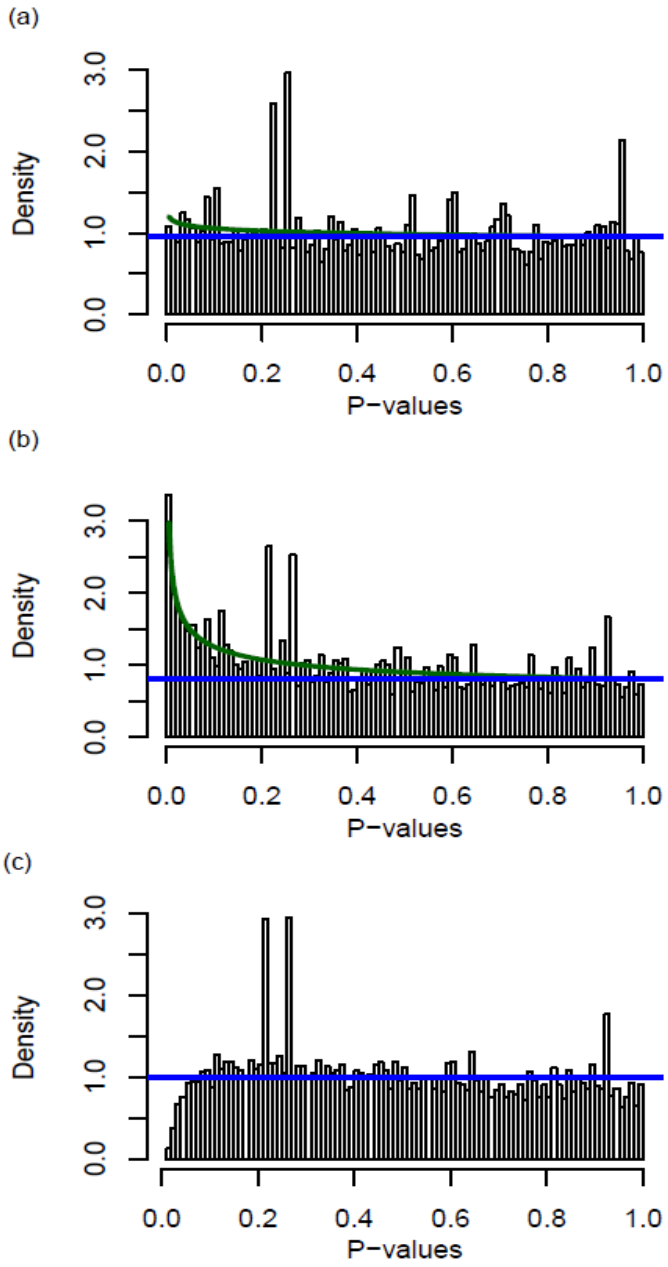


Figure 2.2 Histogram of p-values of gene expression differences from duplicate experiments on the same biological sample. (a) Duplicate experiments were from the same DNA library sequenced in different lanes. The p-values were calculated from the binomial distribution. (Two datasets compared: Bullard SRR037457 vs SRR037458.) (b) When the binomial distribution is applied to the same biological sample prepared in two different libraries, more genes than expected had small probability, which erroneously predicted the existence of significantly differentially expressed genes when there should not be any. (Two datasets compared: Bullard SRR037467 vs SRR037471.) (c) When the same two libraries are compared using the beta-binomial distribution, there is no longer a high density at small p-values. Peak of proportion normalization was used in these calculations. These histograms were drawn using R package Bum-class. [77] [73]

2.2.3 Errors Decreased with Sequencing Depth

We first addressed the uncertainty in the RNA-seq measurement and how uncertainty was related to the sequencing depth empirically from repeated measurements. Specifically, from replicates of the biological sample, we calculated the standard deviation of the proportion. If the proportion satisfied the binomial distribution, we expected $(p_i - p_n)^2 \propto \frac{p_n(1-p_n)}{(n_i+m_i)}$, where n_i and m_i are tags mapped to gene i in two duplicate experiments of the sample (possibly from different libraries), $p_i = \frac{n_i}{n_i+m_i}$ and p_n is the normalization proportion. Figure 2.3 shows a plot of $\frac{(p_i-p_n)^2}{p_n(1-p_n)}$, averaged over pairs of duplicate experiments (Table 2.1), as a function of the mean $n_i + m_i$ for the three sets of experimental data. These figures show that the variance of the proportion continued to decrease at large $n_i + m_i$ and there was no sign of saturation. However, the rates of decrease with the tag counts depended on the dataset and were slower than that given by the binomial distribution. [73]

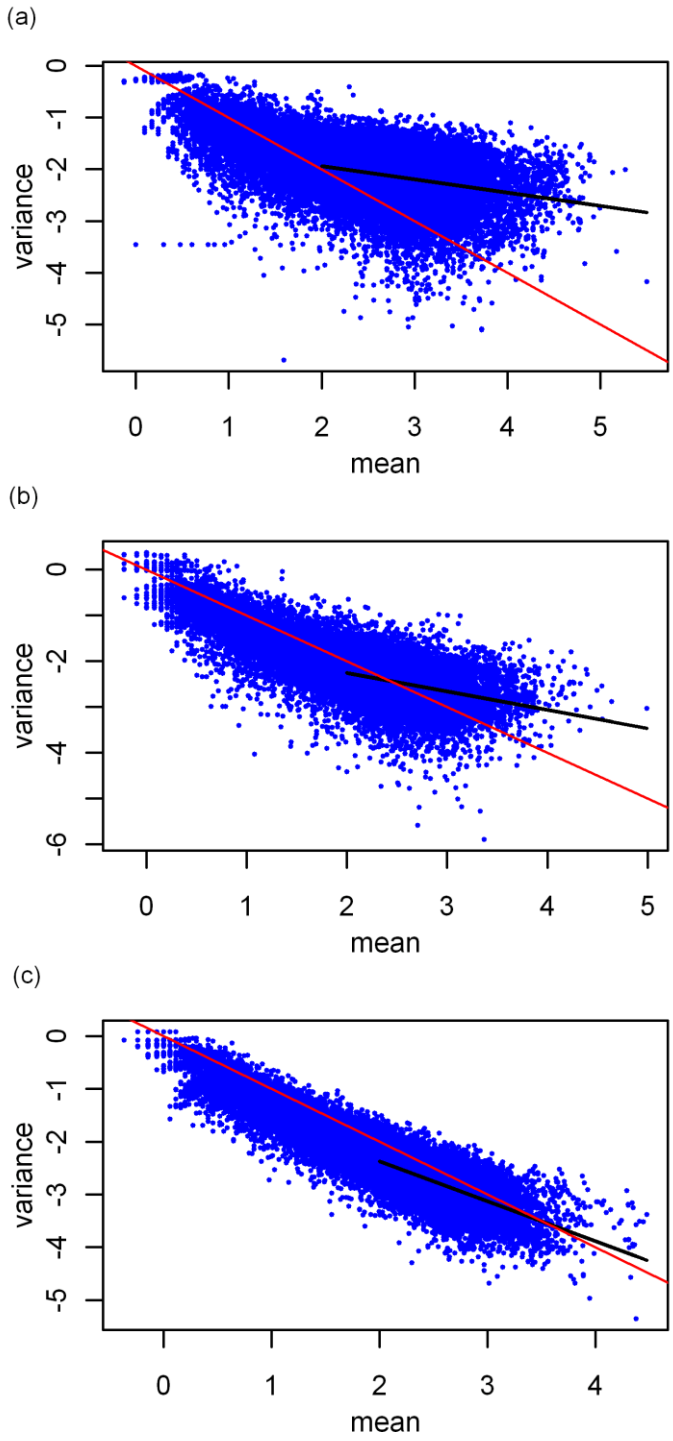


Figure 2.3 The variance of proportion versus the mean tag counts in base-10 log scale. The variances of proportion were computed from replicates of the same biological samples. (a) Caltech dataset; (b) Chiang dataset; (c) Bullard dataset. Each point represents a gene averaged over replicates (see Table 2.1 for the number of replicates for each dataset). The red line has a slope of -1. The black line is fit to the data for a mean (x-axis) larger than 2 (count greater than 100). [73]

2.2.4 Modified Beta Binomial Distribution

We used a beta-binomial distribution to describe the overdispersion in the data, as shown in Figure 2.2 b. However, in the beta-binomial distribution, the standard error approaches a constant as the mean tag counts become very large, whereas empirically, the standard error follows a decreasing trend at large tag counts (Figure 2.3). We therefore made the following assumption about the form of the θ parameter in the beta-binomial distribution (see Method for details). Let n_i and m_i be the number of tags mapped to gene i. We make θ_i depend explicitly on the tag counts.

$$\theta_i = \frac{D_i}{(n_i + m_i)^{\gamma}} \quad (2)$$

Under this assumption, for $0 < \gamma < 1$, the asymptotic form of the variance of the proportion at large tag count $N_i = n_i + m_i$ according to the beta-binomial distribution is $N_i^{-\gamma}$. Therefore the variance of the proportion of the modified beta-binomial distribution does approach zero at large N, but at a slower rate than in the binomial distribution. [73]

2.2.5 Determining the Parameters γ and D_i

Although $\boldsymbol{\gamma}$ can be estimated from the slope and intercept, in the log scale of variance versus the mean tag count (Figure 2.3), it required multiple experiments and had low accuracy due to data scattering. For a better estimation of the parameters \$\gamma\$ and D_i in Eq.2, we used maximum-likelihood estimation (MLE). In this approach, the likelihood was derived from the beta-binomial distribution of tag counts n_i and m_i for gene i, and summed over all the genes and over all the pairs of duplicate experiments. The overdispersion parameters θ_i were given by Eq.2 and the parameter $\boldsymbol{\gamma}$ and parameters $\boldsymbol{D_i}$ for each gene were chosen to maximize the likelihood. The plots in (Figure 2.4) were obtained by performing a full optimization of likelihood Eq.1 (see Methods) with respect to $\boldsymbol{D_i}$ for each $\boldsymbol{\gamma}$, and plotting the optimized likelihood values against $\boldsymbol{\gamma}$. Table 2.2 compares the $\boldsymbol{\gamma}$ from two estimates. The estimated $\boldsymbol{\gamma}$ depended on the data. We computed $\boldsymbol{\gamma}$ for three sets of data. The values ranged from 0.2 to 1.0 (Figure 2.4). These estimates were consistent with those from the standard error (Figure 2.3). [73]

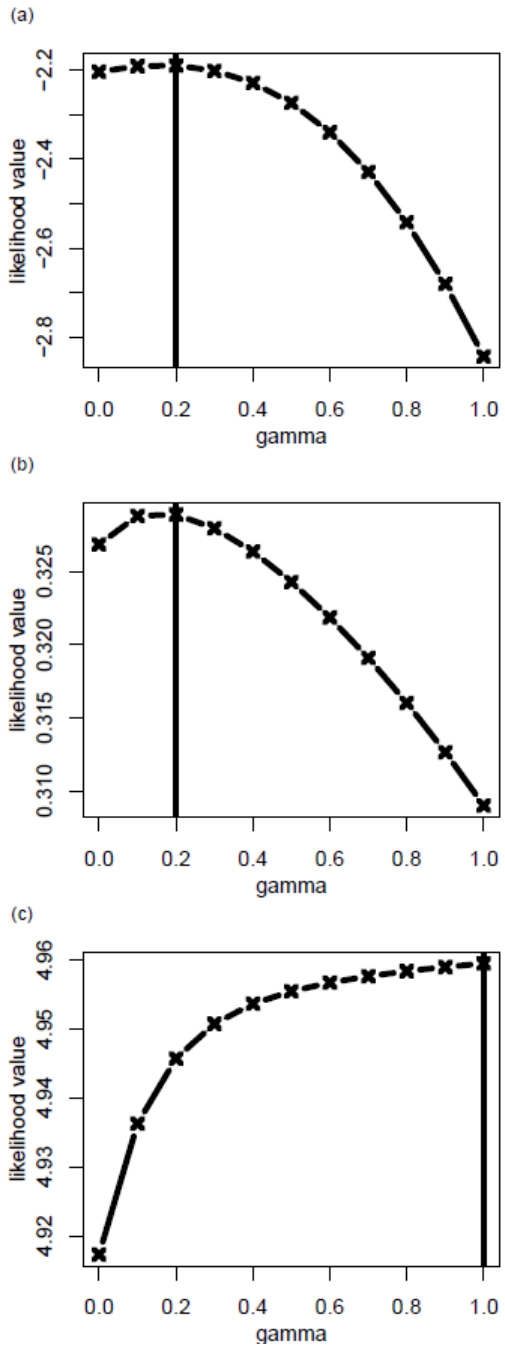


Figure 2.4 Beta-binomial likelihood as a function of the parameter $\boldsymbol{\gamma}$ (a) Caltech dataset; (b) Chiang dataset; (c) Bullard dataset. The vertical lines marked the position of maximum. [73]

Table 2.2 Two estimations of γ from three datasets [73]

Data Set	Pairs of Experiments used in calculation	Standard Error[1]	MLE[2]
Caltech	6[a]	0.26	0.2
Chiang	3[b]	0.40	0.2
Bullard	12[c]	0.76	1.0

1obtained from slope in Figure 2.4

2 from maximizing likelihood Eq.(1)
a from four libraries of same biological sample
b from three knockout replicates and two wild type replicates
c by comparing two di_erent libraries having four and three replicates

2.2.6 Comparison of Beta-binomial and Binomial Distributions

Figure 2.5 shows a comparison of the false discovery rates (FDRs) [75] and receiver operating characteristics (ROCs) [78] for genes deemed to be differentially expressed by the binomial and beta-binomial distributions. For the Bullard dataset, the results were comparable for the two distributions. For the Caltech and Chiang datasets, the beta-binomial distribution was superior (for dataset details, see Methods).

We took the top 300 genes deemed most significantly differentially expressed by a t-test, and by binomial and beta-binomial distributions, and overlaid them in a plot of the fold change versus the average tag counts (see Figure 2.6 and Figure 2.7). We note that the genes identified as significantly differentially expressed by the binomial distribution tended to have large tag counts; whereas many genes identified as significantly differentially expressed from the t-test had small tag counts. Some genes identified as significantly differentially expressed by the binomial distribution (marked by a triangle only) were not identified as significantly differentially expressed by the beta-binomial distribution, even though they had higher fold changes than other genes at similar tag counts. The large fluctuations in the assessment of these genes are evident because they were also not called significantly differentially expressed by the t-test. [73]

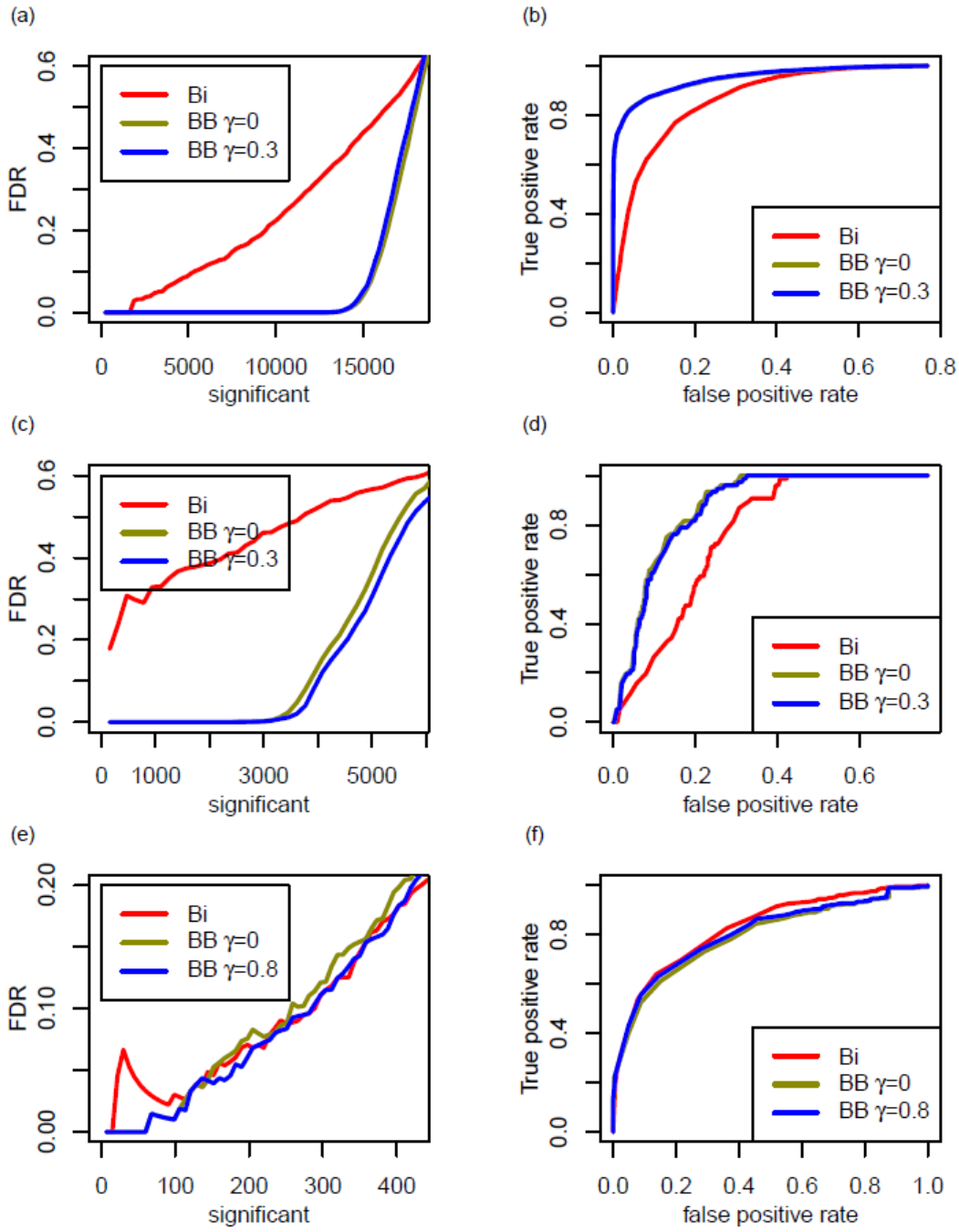


Figure 2.5 False discovery rate (FDR) and receiver operating characteristic (ROC) for three datasets. (a) and (b) Caltech dataset; (c) and (d) Chiang dataset; (e) and (f) Bullard dataset. Three panels on the left indicate the FDR. FDR (on y-axis) is plotted against the number of most significantly differentially expressed genes (on x-axis). Three panels on the right indicate the ROC. Bi denotes binomial distribution; BB denotes beta-binomial distribution. The line for BB $\gamma = 0$ was obtained by setting $\gamma = 0$ and optimizing D_i. It corresponds to the normal beta-binomial distribution. In (b), the line for BB $\gamma = 0$ overlaps with the line for BB $\gamma = 0.3$. [73]

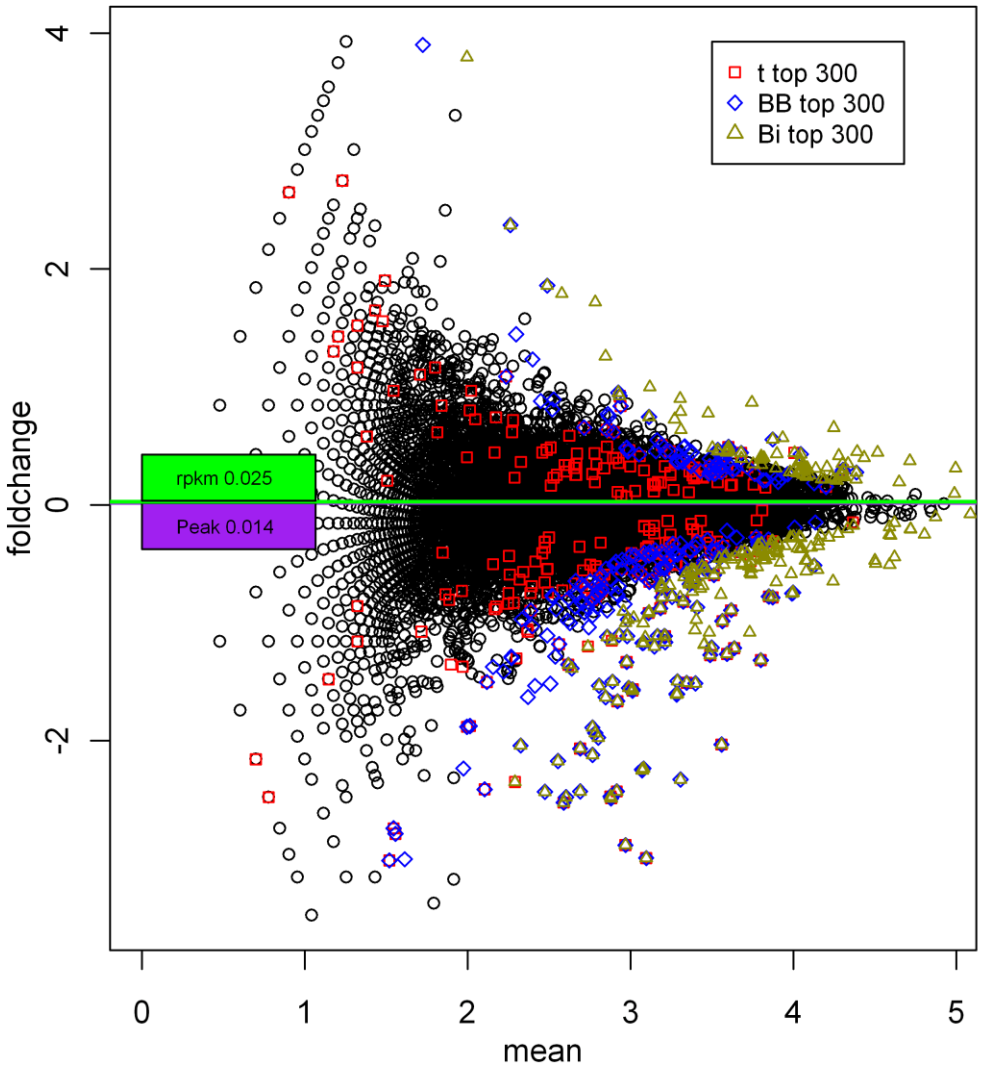


Figure 2.6 Gene expression fold change in the TDP-43 deletion vs wild-type genes (Chiang dataset). Gene expression fold change is plotted against the average tag counts (x-axis in base-10 log; y-axis in base-2 log). The 300 most significantly differentially expressed genes by p-values are depicted by squares (t-test), diamonds (beta-binomial distribution), and triangles (binomial distribution). Black circles represent genes not among the top 300 in any methods. The green and purple boxes and lines indicate the median for RPKM and peak of proportional normalization. The data were from the average of three deletion and two wild-type experiments. [73]

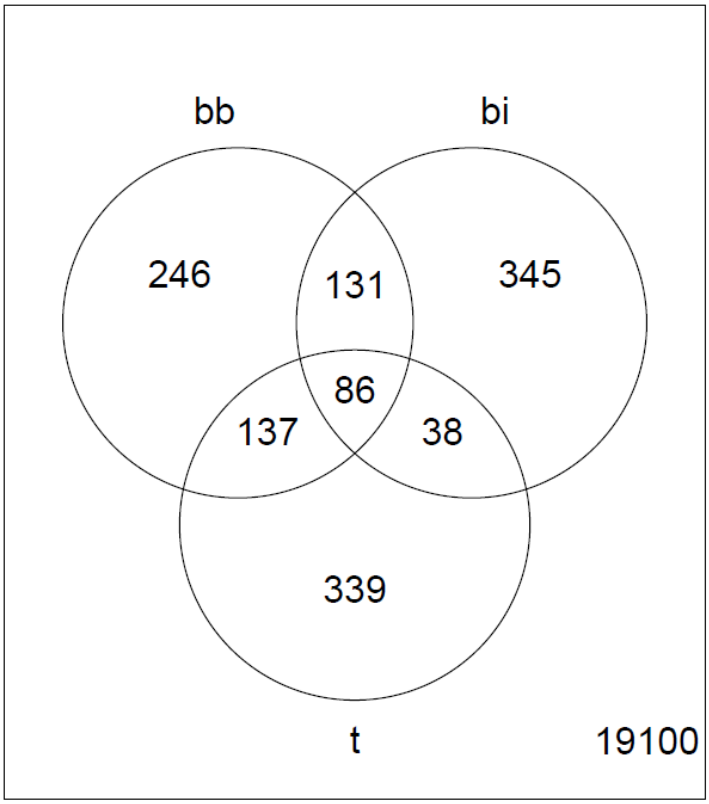


Figure 2.7 Venn diagram comparison. The overlap of top 300 genes identified by beta-binomial (bb) binomial (bi), and the t-test (t) is shown. The number in the lower right of the rectangle indicates the total number of transcripts detected. [73]

2.3 Discussion

In this study, we investigated the error of the gene expression measurement from replicated RNA-seq experiments. We observed that a binomial distribution fit the comparison of the sequencing reads from different sequencing lanes of the same sample, but not from the different preparations of the same biological sample. This observation indicated that a binomial or Poisson distribution fit the nature of the sequencing technology but larger variations could be introduced by the library preparation steps prior to obtaining replicates of the same biological sample. Also we observed that the accuracy of measurement from RNA-seq improved along with increments in sequencing depth. However, compared with the overdispersion predicated by the binomial distribution, the calculated dispersion decreased more slowly along with the increment of the sequencing reads, and many false discoveries were observed from the testing based on the binomial distribution. This indicated that overdispersion exists and is introduced by library preparation, and that it decreases roughly linearly along with the increment in sequencing depth on a log scale. We developed a method based on the beta-binomial distribution with a new parameter to model the relationship between overdispersion and the sequencing depth. We borrowed the information from all genes by introducing the overdispersion parameters in a function of the number of reads to estimate the specific means and dispersion of each gene. We used the maximum likelihood method to determine the parameters. And by comparing the false discovery rate (FDR) and receiver operator

characteristics (ROCs), we showed that our modified beta-binomial model was superior to the binomial model without an overdispersion parameter and to the beta-binomial model with an overdispersion parameter that is constant for all genes.

The advantages of modeling the proportion of measurement counts to detect the differential expression drove us to adapt the beta-binomial distribution rather than the Poisson model. Many studies have reported highly non-uniformly distributed patterns of measurements from RNA-seq on genes [4], and that this non-uniformity is correlated with the local sequence around a random hexamer primer [3,4]. In this situation, the Poisson rate could fluctuate by even a hundred-fold at different positions of the some gene. However, the non-uniform distribution was similar between replicates of the same sample or even among measurements of samples from different tissues [4]. Therefore, we could avoid estimating the highly fluctuating Poisson rate on the same genes by comparing the proportion of measurement of two samples. By modeling the proportion, the non-uniformity was estimated indirectly only through the dispersion. Avoiding estimating the highly fluctuating Poisson rates is therefore advantageous for modeling the proportion of measurements based on the beta-binomial distribution.

In our model, parameter γ in Eq.2 describes the decreasing rate of overdispersion along with the increasing sequencing depth. Our model will reverse back to the pure beta-binomial distribution when the parameter γ goes to 0. Interestingly, we observed that the estimated γ values were dissimilar across the different datasets. This phenomenon is commonly found in measurements from different experiments; for example, the GC count bias in sequencing data has been reported to vary between experiments [9]. Therefore, the experimental protocol might influence parameter γ, which indicates the decreasing rate of overdispersion along with an increasing sequencing depth. Because of the non-uniformity of measurements on genes, it is interesting to investigate the property of overdispersion on each position and it will be more accurate to model measurements on the position level. We carried on the study and proposed another parameterization of overdispersion based on the position, which we describe in the next chapter.

CHAPTER 3

Modeling the Non-uniformity of Measurement from RNA-seq for Differential Expression Analysis

Many biases and spurious effects are inherent in RNA-seq technology. A number of methods have been proposed to handle these biases and undesirable effects in order to accurately analyze differential RNA expression at the gene level. However, modeling at the base pair level is required to precisely estimate the mean and variance of the measurement, because the sequencing reads are non-uniformly distributed on one gene. As a consequence, each position on one gene has a specific mean and variance. It has been reported that priming with a random hexamer contributes to the non-uniformity, which is related to the local sequence around the priming site. In Chapter 2, we showed that the overdispersion rate decreased as the sequencing depth increased on the gene level. On the basis of these findings, we developed three corresponding hypotheses. (1) In comparison of the gene expression of two samples, the null hypothesis is that there is no difference between two samples, even on each position of a gene. Therefore, the proportion of the measurement on each position will be a constant. With this assumption, modeling on the proportion of the measurement based on a beta-binomial distribution will be appropriate, with the advantage that the non-uniformity of the measurement is transformed to a constant mean. (2) On the position level, random hexamer priming influences the overdispersion rate through the local sequence around hexamer primers. (3) Similar with what we observed on the gene level, the overdispersion rate decreases along with the increasing sequencing depth on the position level as well. Based on these hypotheses, we developed two beta-binomial models. One was a full model based on all three hypotheses, and the other was a reduced model based on hypotheses 1 and 3.

First, we investigated the impact of the sequencing depth and local primer sequence on the overdispersion rate. Second, we inspected the impact of different sequencing protocols on the

overdispersion rate. Third, we proposed four models and compared them with each other and DESeq regarding the likelihood value, AIC, goodness-of-fit χ^2 test, and the testing error indicators FDR and AUC.

Similar to our observations on the gene level in Chapter 2, we demonstrated that the overdispersion rate decreased along with the increasing sequencing depth on the position level. Also, the influence of priming with a random hexamer on the overdispersion was validated and relates to the local sequence around the hexamer primer. However, after stratification by sequencing depth, the influence was no longer significant. In addition, our beta-binomial model with a dynamic overdispersion rate on the position level was superior to the other models we proposed in this study. Furthermore, as expected, our proposed model was more desirable than DESeq, which was based on a negative binomial distribution, with many advantages in differential expression analysis for the same biological samples.

The current study provides a thorough understanding of the property of the overdispersion rate on the position level, especially the relationship between the overdispersion rate and sequencing depth. We also clarified that random hexamer priming could influence the overdispersion rate by affecting the sequencing depth of each position. These properties will aid in the quality control and development of statistical methods for downstream analysis. Based on those properties, we suggested a more desirable method to model the non-uniformity measurement. Our method was based on a beta-binomial model with a dynamic overdispersion rate, and a better estimation was obtained from it on each position when compared with the other models, assuming the overdispersion rate is a constant for all points of one gene.

3.1 Methods

3.1.1 Datasets Used

Two datasets were used, the Lichun dataset with spike-in data [24] and the Bullard dataset with the gold standard data [37] (Table 3.1).

Lichun dataset with spike-in data. The Lichun dataset consists of several libraries with different RNA sources (whole cell, cytosol, and nucleolus) and identifications (longPolyA and longNonPolyA). Synthetic spike-in standards from the External RNA Control Consortium (ERCC) were sequenced along with human samples. These libraries were prepared using the dUTP protocol: (1) First-strand synthesis is performed using a random hexamer primer. (2) Second-strand synthesis is performed by RNAse H and DNA polymerase 1. (3) cDNAs are fragmented by sonication, and adapters are ligated to both end of cDNAs. (4) The second strand is eliminated through UNG digestion. (5) Fragments are selected with sizes at 200 base pairs. (6) Paired-end sequencing is performed. The human libraries are mapped to the human genome (hg19) using STAR software, and the ERCC libraries are mapped to the ERCC reference using Bowtie, version 0.11.3 with parameters –v2 –m1. In the present study, we analyzed only libraries from whole cells and longNonPolyA. Because we assumed that the number of reads would influence the overdispersion rate, we selected samples (shown in red in Table 3.1) with approximately the same total counts as the training set to eliminate noise. To avoid the transcription initiation bias in the sequencing [79], we truncated 50 nucleotides on both ends.

Bullard dataset with gold standard data. Two distinct biological samples, brain and UHR, were examined in the Bullard dataset. The UHR samples were from three library preparations, UHR libraries A, B, and C, and the brain samples were from one library preparation. RNA was first fragmented and then converted into cDNA using random hexamer priming. The cDNA was sequenced using the standard Illumina protocol, including adapter ligation, polymerase chain reaction (PCR), size selection, and injection into flow-cells. We used Bowtie, version 0.12.7, to align reads to the genome (H. sapiens, NCBI 37.1 assembly). The Bowtie command for implementing this mapping strategy was ./bowtie –a –v –t –m 1 –best –stratah_sapiens_37_asm. Additionally, about 1000 genes have previously been assayed by real-time PCR; thus, these genes can be applied as a gold standard. All 3 sampled UHR libraries had almost the same yield and thus could be used as the

training set to estimate parameters. Three of them were left for the test set (shown in blue in Table 3.1). Also, 50 nucleotides were truncated on both ends to avoid the transcription initiation bias.

Table 3.1 Two datasets used

Dataset				
Lichun	ERCC	GSM758567 GSM758572 GSM758573 GSM758577 GSM765389 GSM765391 GSM765396 GSM765398 GSM767845 GSM767847 GSM767851 GSM767854 GSM767855 GSM767856		
	Human	GSM767847 GSM758577		
Bullard	Brain	UHR library A	UHR library B	UHR library C
	SRR037455 SRR037456 SRR037457 SRR037458	SRR037466 SRR037467 SRR037468 SRR037469	SRR037470 SRR037471 SRR037472	SRR037473 SRR037474 SRR037475 SRR037476

3.1.2 Normalization

RPKM normalization was applied. RPKM was computed as the number of reads that mapped per kilobase per million mapped reads for each gene, for each sample.

3.1.3 Calculation of Overdispersion Rate θ_{ij} per Base Pair

Let n_{ij} and m_{ij} be the tags mapped to the j-th nucleotide of the i-th gene for the experimental sample and control, respectively. The probability mass function according to the beta-binomial distribution is

$$f(n_{ij}|\alpha_{ij},\beta_{ij}) = \binom{n_{ij}+m_{ij}}{n_{ij}} \frac{B(n_{ij}+\alpha_{ij},m_{ij}+\beta_{ij})}{B(\alpha_{ij}+\beta_{ij})},$$

where α_{ij} and β_{ij} are two parameters of the beta-binomial model. It is equivalent to using the following parameters: $p_i = \frac{\alpha_{ij}}{\alpha_{ij}+\beta_{ij}}$ for all j and $\theta_{ij} = \frac{1}{\alpha_{ij}+\beta_{ij}}$. Analytically, p_i is the expected value of the proportion, which can be estimated in the binomial model as $\frac{\sum_{j=1}^{M} n_{ij}}{\sum_{j=1}^{M} n_{ij}+\sum_{j=1}^{M} m_{ij}}$.

The proportion $\hat{p}_i$ of each gene should equal the proportion of all the reads of all genes, that is,

$$\widehat{p_i} = \widehat{p_n} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M} n_{ij}}{\sum_{i=1}^{N}\sum_{j=1}^{M} n_{ij} + \sum_{i=1}^{N}\sum_{j=1}^{M} m_{ij}} \qquad (1)$$

The variance of the proportion can be obtained as

$$var(p_{ij}) = \frac{p_n(1-p_n)}{n_{ij}+m_{ij}}\left(1 + \frac{n_{ij}+m_{ij}-1}{\frac{1}{\theta_{ij}}+1}\right).$$

Therefore, θ_{ij} can be derived as

$$\theta_{ij} = \frac{\frac{1}{R}\sum_{r}^{R}\left(\frac{var(p_{ij})}{p_n(1-p_n)} - \frac{1}{n_{ij}+m_{ij}}\right)}{1 - \frac{1}{R}\sum_{r}^{R}\frac{var(p_{ij})}{p_n(1-p_n)}}\ , \qquad (2)$$

where r denotes the r-th particular pair among R total pairs of replicates.

We developed a 2-step strategy to calculate θ_{ij}. In the first step, the variance of proportion per base pair was estimated per pair of replicates separately. Then we calculated θ_{ij} according to formula (2).

3.1.4 Base Pair-Based Model

After the reparameterization, the log likelihood of the beta-binomial distribution was derived as

$$L = \sum_{i=1}^{N}\sum_{j=1}^{M}\left[\sum_{k=0}^{n_{ij}-1} log(\,p_i + k\theta_{ij}) + \sum_{k=0}^{m_{ij}-1} log(\,1-p_i + k\theta_{ij}) - \sum_{k=0}^{n_{ij}+m_{ij}-1} log(\,1+k\theta_{ij})\right] \quad (3)$$

Full model. On the basis of all of our assumptions, a full model was suggested, in which θ_{ij} is related to the local sequence around the primer:

$$\theta_{ij} = \frac{De^{\{\sum_{k=1}^{K}\sum_{h\in\{A,T,C\}}\beta_{kh}I(b_{ijk}=h)\}}}{(n_{ij}+m_{ij})^{\gamma}}\ . \qquad (4)$$

In this model, K is the length of the probe around the j-th nucleotide of the i-th gene. We set $K = 80$ as suggested in a previous study [73]. Also, $I(b_{ijk} = h)$ is 1 when the k-th base pair is letter

h, which is A, T, or C exclusively, and 0 otherwise. The parameters we want to estimate are α and β_{kh}, and ε is Gaussian noise.

There are 3 x 80 = 240 parameters on the local sequence in this model, which makes it quite difficult to use the usual ways of estimation, such as maximum-likelihood estimation. We took the log of Eq. 4 and obtained another formula that can facilitate model fitting:

$$log(\theta_{ij}) = log(D) + \sum_{k=1}^{K} \sum_{h\in\{A,T,C\}} \beta_{kh} I(b_{ijk} = h)\} + \gamma\, log(n_{ij} + m_{ij}) \qquad (5)$$

This linear model of θ_{ij} made the estimation fast and robust. The number of parameters, 240, is a really small number compared to the sum of all the positions in all the genes.

Reduced model. On the basis of our third hypothesis, that the overdispersion rate of RNA-seq reads decreases as the sequencing depth increases, we proposed a reduced beta-binomial model for the comparison of two replicates, in which

$$\theta_{ij} = \frac{D}{(n_{ij} + m_{ij})^{\gamma}}\,, \qquad (6)$$

where D is for all nucleotides of any genes and γ represents the decreasing slope of the overdispersion rate plotted against the number of reads.

Counts-excluded model. In order to determine the dependency of the overdispersion rate on the local primer sequence, we excluded the count term from the full model:

$$log(\theta_{ij}) = log(D) + \sum_{k=1}^{K} \sum_{h\in\{A,T,C\}} \beta_{kh} I(b_{ijk} = h)\}\,. \qquad (7)$$

3.1.5 Fitting the Beta-binomial Models

The parameter θ_{ij} was determined by maximizing the log likelihood (Eq. 3) of the reduced model and the full model, respectively. We used the following strategy to fit our models:

1. Initialize $\widehat{p_i}$ as Eq. 1 in the training set.

2. Set $\widehat{p_i}$ as a known parameter and fit the beta-binomial model to obtain D and γ. Apply the least-squares estimation method based on the linear model (Eq. 5).
3. Set θ_{ij} according to Eq. 6 and Eq. 4 as a known parameter in the reduced model and the full model, respectively, to update p_i.
4. Jump to step 2 unless the deviance decreases less than 1%.

The above procedure maximizes the likelihood by iteratively optimizing θ_{ij} in step 2 and p_i in step 3.

3.1.6 Estimating Cross-validation R^2

We used the leave-one-out cross-validation strategy to estimate R^2. The training set was randomly split into five groups of equal size. In each round, we fit our model using four of these five groups, and then calculated R^2 on the remaining subset by the regression sum of squares divided by the total sum of squares. The final cross-validation R^2 was determined as the mean.

3.1.7 Likelihood Ratio Test

According to the likelihood ratio test, $-2\ln \mathrm{L}(p_n) + 2\ln \mathrm{L}(p_i)$ follows the χ^2 distribution, where p_i is the proportion for gene i and p_n is the normalized proportion corresponding to no change in gene expression. In multiple-samples testing, we summed over their pairwise χ^2 scores and obtained *p*-values with a summation of degree of freedom.

3.1.8 Methods for Model Comparison

Goodness of fit was examined for 4 models: the binomial model, the beta-binomial model with a constant overdispersion rate of θ_{ij}, the reduced model with θ_{ij} as in Eq. 6, and the full model with θ_{ij} as in Eq. 4.

Likelihood-value goodness of fit. Proportion p_i was estimated and fixed for all four models. Sequentially, the other parameters were determined by the maximum-likelihood estimation method or the least-squares method, and the likelihood value was calculated by pairwise comparison of the replicate data. The χ^2 test was performed on $D = -2\,ln(L_{null}) + 2\,ln(L_{alternative})$, where L_{null} and $L_{alternative}$ denote the likelihood for the null model and alternative model, respectively.

AIC. Akaike information criterion (AIC) is a measure of the relative goodness of fit of a statistical model. The AIC is calculated by definition as $AIC = 2K - 2\,ln(L)$, where K is the number of parameters and L is the maximum-likelihood value. The final AIC is determined by the mean of all AICs from pairwise replicates.

FDR and AUC. The false discovery rate (FDR) and the area under the receiver operating characteristic curve (AUC) were determined by the method described in our previous study [73]. For the Lichun dataset, which lacked gold standard data, the AUC was not determined.

3.2 Results

3.2.1 Overdispersion Rate on Base Pairs Decreased with Sequencing Depth

In order to test our third hypothesis, we empirically investigated the impact of sequencing depth on the measurement of the overdispersion rate per base pair. Analyzing the spike-in data, we calculated the variance of the proportion of the reads mapped to the j-th base pair of the i-th gene from two replicates of the same sample, then determined the overdispersion rate θ_{ij} (Methods, part C). In Figure 3.1, we plot the estimated θ_{ij} against the number of counts on the corresponding nucleotide position. The results show that the overdispersion rate was strongly inversely correlated with sequencing depth—that is, the overdispersion rate kept decreasing as the sequencing depth increased and without a sign of saturation. This density plot shows that most of the points are concentrated on a line. Examination of the points corresponding to the local sequence, starting with GGGG and AAAA (blue and red points, respectively), also shows that most of the points are on or

close to the line of concentration. However, at positions with large read numbers, the estimated θ_{ij} seemed to depart from the decreasing trend, suggesting that statistical noise may exist and have a greater effect on overdispersion than the library preparation effect at positions with large sequencing depth.

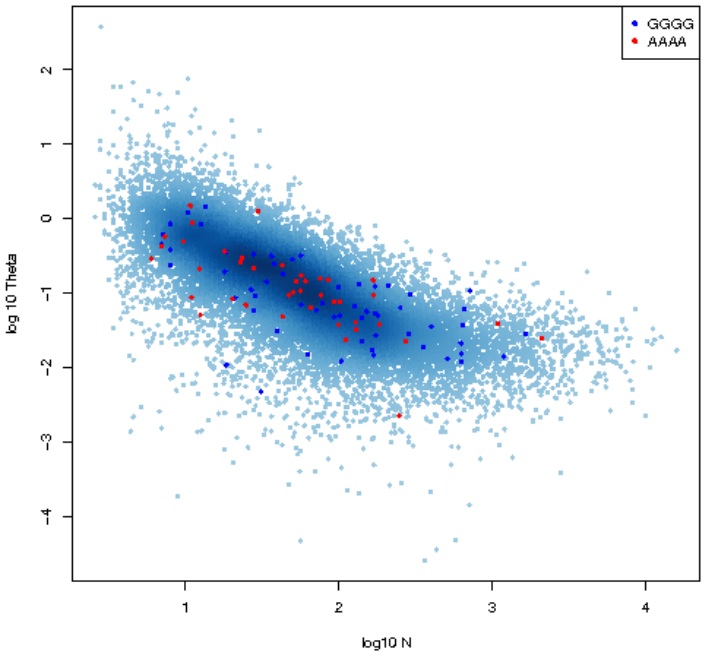


Figure 3.1 The overdispersion of proportion θ_{ij} on base pairs versus the mean tag counts in base 10 log scale. The θ_{ij} values were computed from replicates from the Lichun spike-in training dataset. The blue and red points are for the positions with a local sequence starting with GGGG and AAAA, respectively.

3.2.2 Sequencing Procedure Introduced Extra Noise

Elements of the sequencing procedure (e.g., fragmentation methods, random hexamer priming, etc.) usually introduce bias to RNA-seq measurements. We examined the overdispersion rate estimated from two datasets (Figure 3.2). Interestingly, in the Lichun dataset, the overdispersion rate was significantly larger at the tail of the gene (less than about 200 base pairs). However, no such difference was observed in the Bullard dataset. The same results were obtained in the calculation of the variance (Supplementary Figure S3.1).

To explain these findings, we inspected the Lichun and Bullard protocols. We found three major differences between them. First, Lichun et al. applied the dUTP protocol to obtain strand-specific sequencing, while Bullard et al. used a regular non-strand-specific sequencing protocol. Second, Lichun et al. performed paired-end sequencing, while Bullard et al. tried to obtain single-end sequencing data. Third, fragmentation was carried out before PCR in creating the Lichun dataset; while PCR was performed first in creating the Bullard dataset. The first two differences were ruled out as potential causes for the extra noise on the gene tails in the Lichun dataset, because those differences would not influence the measurement of only part of the gene. However, the third difference can explain the extra noise. In the Lichun dataset, the fact that fragment selection was performed after fragmentation might lead to the loss of many fragments located at the gene tails, thereby introducing an extra error. By contrast, according to the protocol used by Bullard et al., fragmentation was carried out before cDNA PCR and size selection. Thus, it was more like a random process across the whole gene, and thus no difference would be observed.

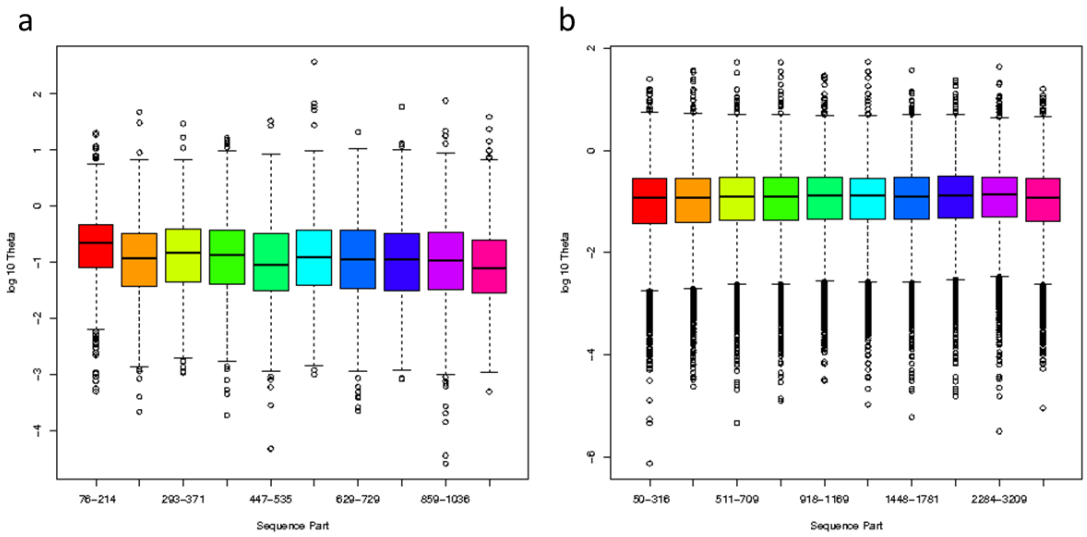


Figure 3.2 The overdispersion rate estimated on any position in 10 equal categories according to the distance of that position from the last nucleotide of the gene. (a) Lichun spike-in dataset. (b) Bullard dataset. Lichun reads start to appear at 76 nt away from the end of the gene because only mate2 on the antisense strand was investigated and the last sequencing reads were mapped to 76nt before the ending. However, Bullard reads start from 50 nt away from the end of the gene because we truncated the genes by 50 nt from the gene head and tail separately.

3.2.3 Models of the Overdispersion Rate

We proposed three models to examine the relationship between the local primer sequence and the overdispersion rate, including the full model, the reduced model, and the counts-excluded model. Using the linear formula transformation (Eq. 5), 240 coefficients of 80 positions around the primer were estimated efficiently. By modeling on the Bullard data, we plotted those coefficients estimated against their corresponding positions (Figure 3.3). We observed a pattern in our counts-excluded model that was similar to the pattern reported in the papers by Hansen et al. and Li et al. [3,4] (Figure 3.3 a,c). However, no such pattern was observed with our full model (Figure 3.3 b,d). Observations were similar for the Lichun spike-in data (Figure 3.4). Both Hansen et al. and Li et al. demonstrated a relationship between hexamer primers and measurement count number. Plus, from Figure 3.1, we conclude that the overdispersion rate on base pairs decreases with increasing sequencing depth. It is reasonable to infer that using a hexamer primer might influence the

overdispersion rate by affecting the count number; thus, upon stratification by counts, the relationship between the use of a hexamer primer and the overdispersion rate would no longer be significant. In addition, we calculated R^2 using the cross-validation method (Methods 3.1.5). R^2 values of 0.481 and 0.488 were obtained for the reduced model and the full model, respectively, with the Bullard data; while values of 0.270 and 0.273, respectively, were obtained with the Lichun spike-in data. About half of the variance was explained for the Bullard dataset, while the relatively lower R^2 was obtained for the Lichun dataset because of its small sample size (only 100 ERCC genes). As expected, compared with the full model, the reduced model achieved a rather similar R^2.

We investigated the influence of primers corresponding to the reads mapped to the antisense and sense strands, respectively. We observed from the Bullard dataset that reads mapped to the antisense and sense strands showed quite similar patterns (Figure 3.3 a,c), which was consistent with the finding of Hansen et al. [3]. However, according to sequencing protocols, the sense strand reads should not have a bias caused by the use of a hexamer primer as the second strand is synthesized by RNAse H niche technology. The explanation of Hansen et al. [3], that the hexamer primer is not completely digested, is quite reasonable. In contrast, different patterns on the sense and antisense strands were observed in the Lichun spike-in dataset (Figure 3.4 a,c). The reason for that is still unknown; one or multiple processes in their strand-specific protocol might impact differently on the paired-end reads. As suggested by the above observations, we estimated coefficients separately for each strand in the present study.

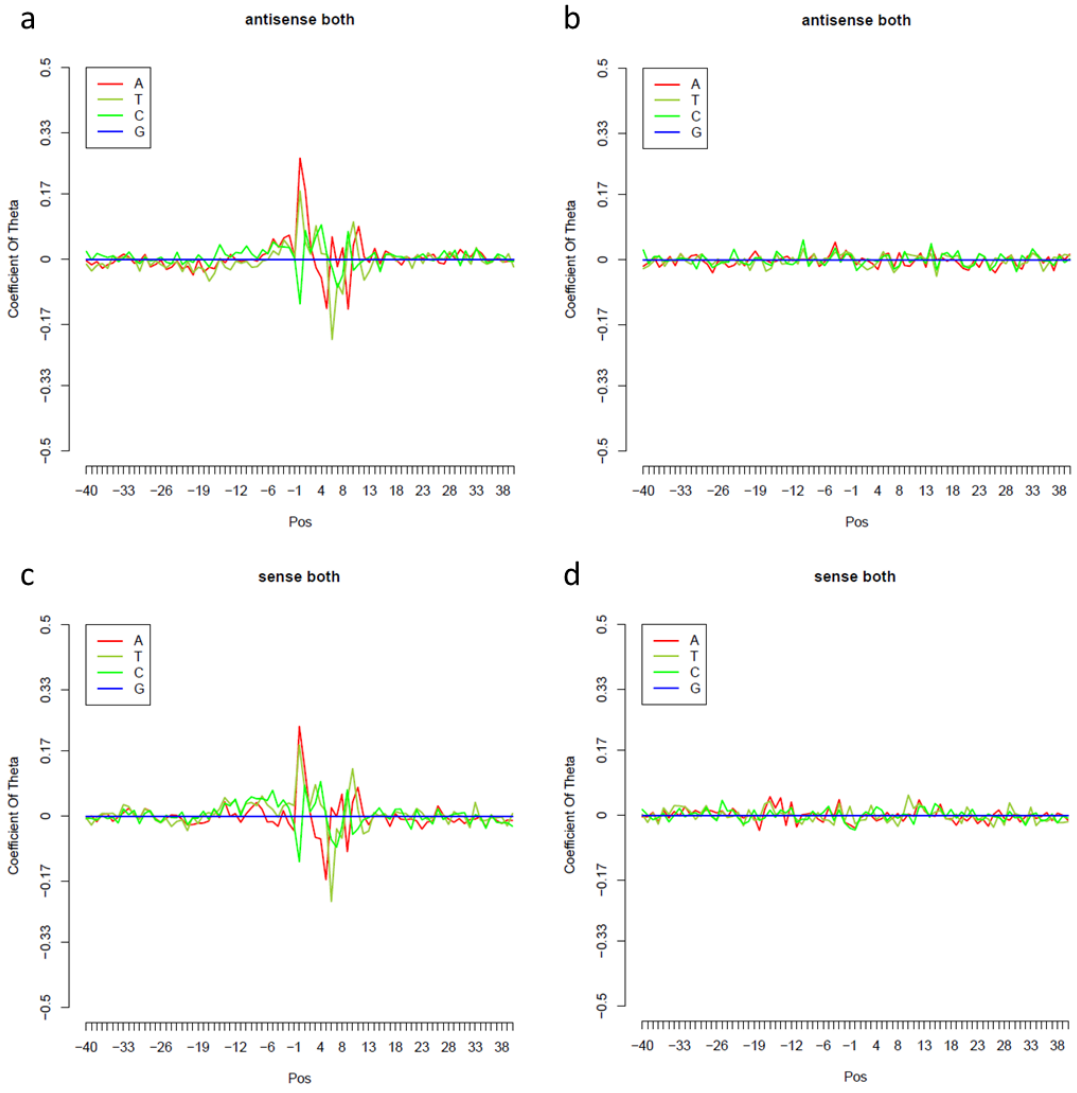


Figure 3.3 Coefficients estimated from linear models from the Bullard dataset. Plotted on the x-axis are the positions around the 5' end of mapped reads, labeled 0. Coefficients were calculated by 2 models. (a) Counts-excluded model on antisense strand. (b) Full model on antisense strand. (c) Counts-excluded model on sense strand. (d) Full model on sense strand.

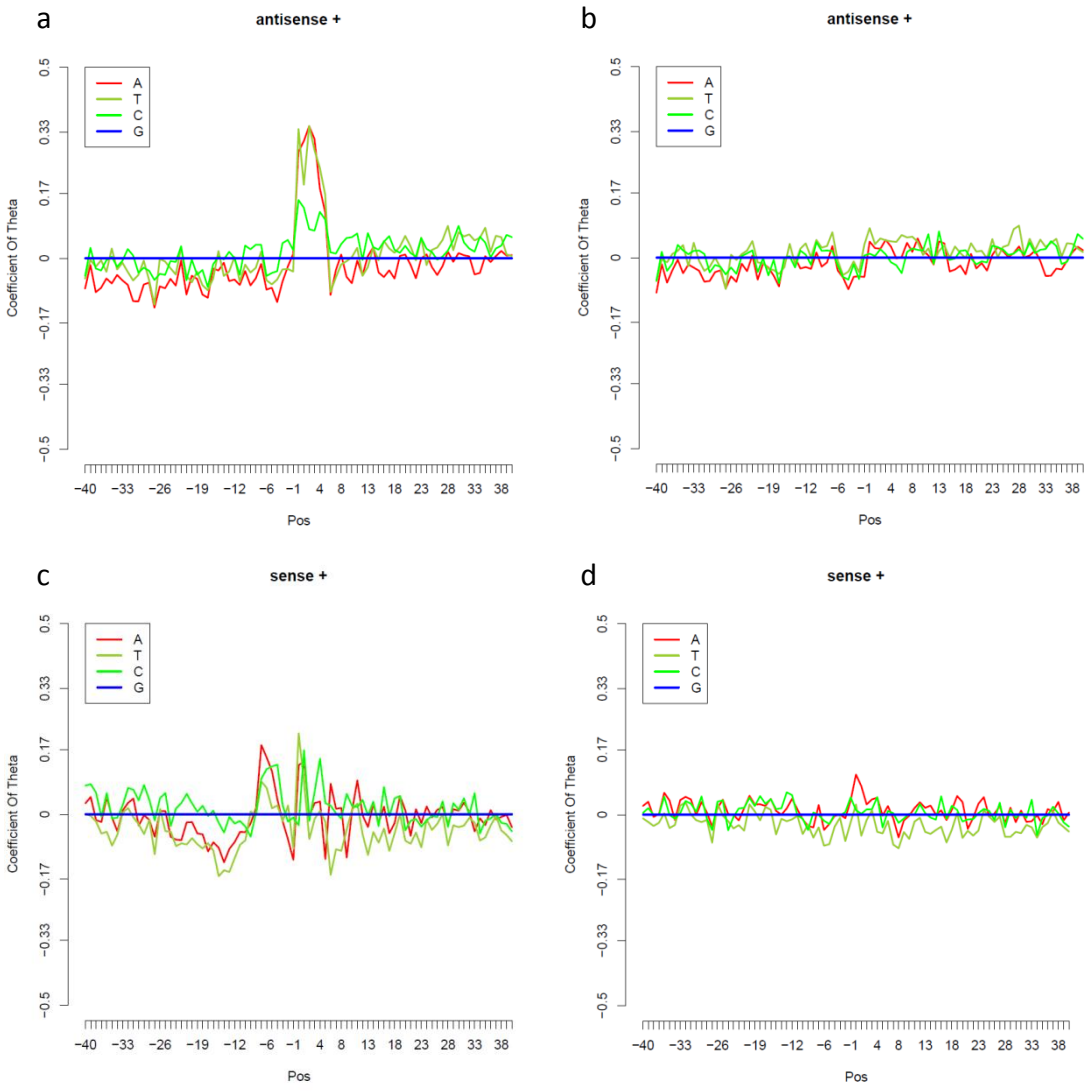


Figure 3.4 Coefficients estimated from linear models from the Lichun dataset. Plotted on the x-axis are the positions around the 5' end of mapped reads, labeled 0. Coefficients were calculated by 2 models. (a) Counts-excluded model on antisense strand. (b) Full model on antisense strand. (c) Counts-excluded model on sense strand. (d) Full model on sense strand.

3.2.4 Comparison of Four Models

Likelihood-value goodness of fit. Comparing maximum likelihood values is a straightforward way to select models. We calculated the likelihood values from four models: the binomial model, beta-binomial model with a constant overdispersion rate of θ_{ij}, reduced beta-binomial model with θ_{ij} as in Eq. 6, and full beta-binomial model with θ_{ij} as in Eq. 4. Then, the percentage of change in

the likelihood value of each model was measured compared with the next neighbor model that preceded it (Figure 3.5 a,b). As expected, models with more parameters had higher maximum likelihood values. The beta-binomial model with a constant overdispersion rate of θ_{ij} made a huge jump from the binomial model (30% to 90%). And the parameter γ in dynamic θ_{ij} (Eq. 6) also improved the fit by roughly 15%. However, the full model we proposed showed almost the same maximum likelihood value as our reduced model. Further, the goodness-of-fit χ^2 test showed that the beta-binomial model with a constant overdispersion rate of θ_{ij} and the reduced beta-binomial model with θ_{ij} as in Eq. 6 (p-value = 0) improved the fit significantly more than the full beta-binomial model with θ_{ij} as in Eq. 4 (p-value = 1). Additionally, AICs measured for the four models showed that the reduced model had the least score and there was an increase for the full model (Figure 3.5 c).

The above results were observed in both the training and the test datasets and suggested that a dynamic overdispersion rate significantly improves the model fit and that our reduced model is a better choice than the other three models. Although we observed the same pattern with the Bullard dataset (Figure 3.5 b,d), no such significant improvement was shown, even for the second model. This is due to the small experimental library effect in the Bullard dataset, which was reported in our previous study [73]. Consistent with our earlier conclusion, we found that overdispersion rates vary dramatically in different studies.

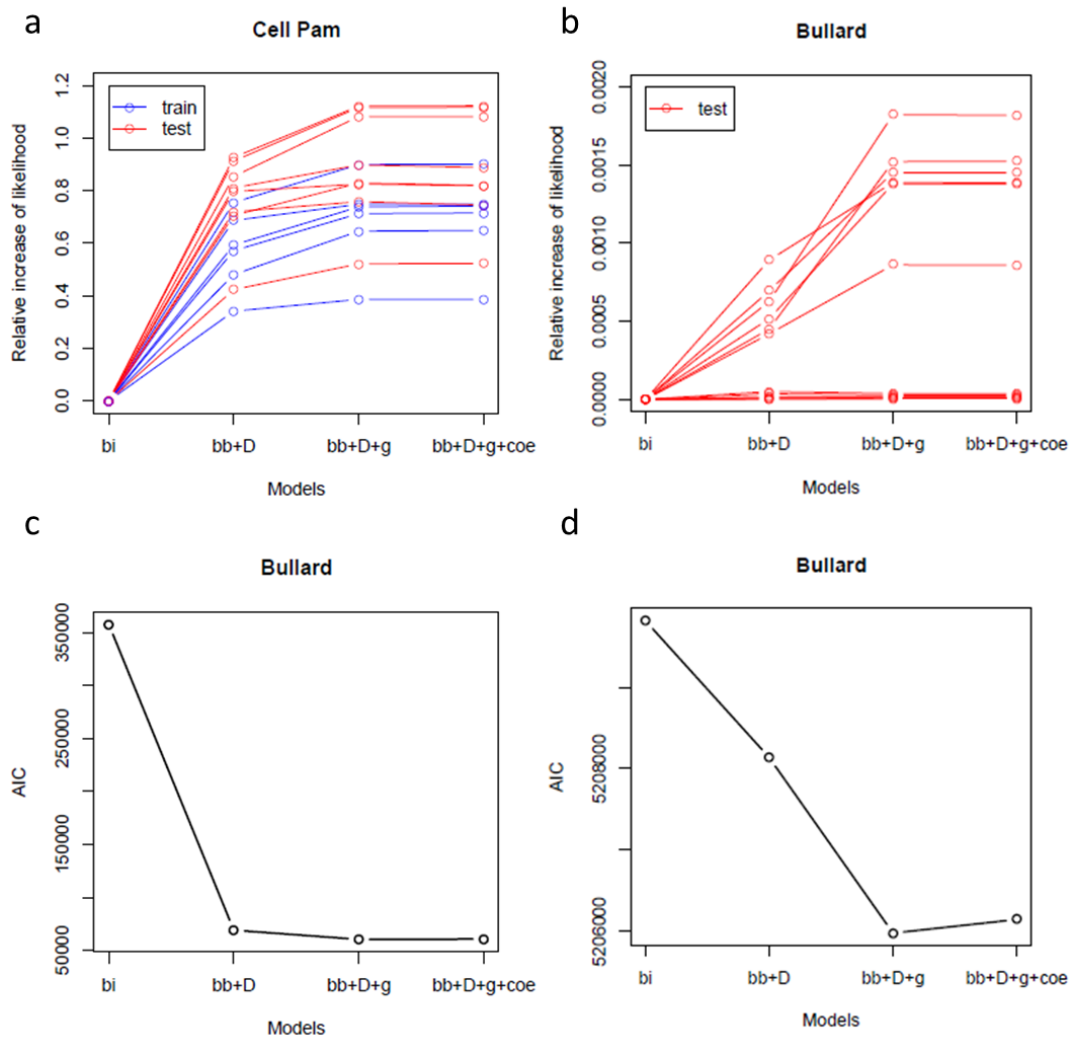


Figure 3.5 Goodness-of-fit examination. (a,c) Lichun dataset. (b,d) Bullard dataset. (a,b) The percentage of change in the likelihood value comparing neighboring models. (c,d) AIC measured for 4 models. bi denotes binomial model, bb+D denotes beta-binomial model with constant overdispersion rate, bb+D+g denotes reduced beta-binomial model, and bb+D+g+coe denotes full beta-binomial model.

FDR and AUC. Further, we compared the FDR and AUC for genes deemed to be differentially expressed by these four models. Bullard UHR sample data showed that our reduced model as well as the full model had the lowest FDRs and largest AUCs (Figure 3.6 c,d). Again, as a result of the small library effect, no big difference was observed between the four models for the Bullard dataset, which agrees with our previous results [73]. Also, our proposed reduced models showed a good

performance for the Lichun spike-in data, though the binomial model with a constant overdispersion rate seemed superior (Figure 3.6 a). Using the same overdispersion rate parameters, our analysis showed a similar result when testing the human samples from the Lichun dataset (Figure 3.6 b). Our FDR and AUC results indicate that among these four models, the binomial model had the worst performance and the beta-binomial models with a dynamic overdispersion rate were preferable, although the beta-binomial model with a constant overdispersion rate had the lowest FDR caused by overfitting. Testing the same sample and different library preparations sequenced by the Illumina sequencer, the beta-binomial model with a constant overdispersion rate shows insufficient small p-values (Figure 3.7 a), and the binomial model has an overabundance of small p-values (Figure 3.7 b). In contrast, the histogram of the p-values is flat for the beta-binomial models with a dynamic overdispersion rate (Figure 3.7 c,d). This indicates that the errors between samples from different library preparations are consistent with the beta-binomial distribution with a dynamic overdispersion rate.

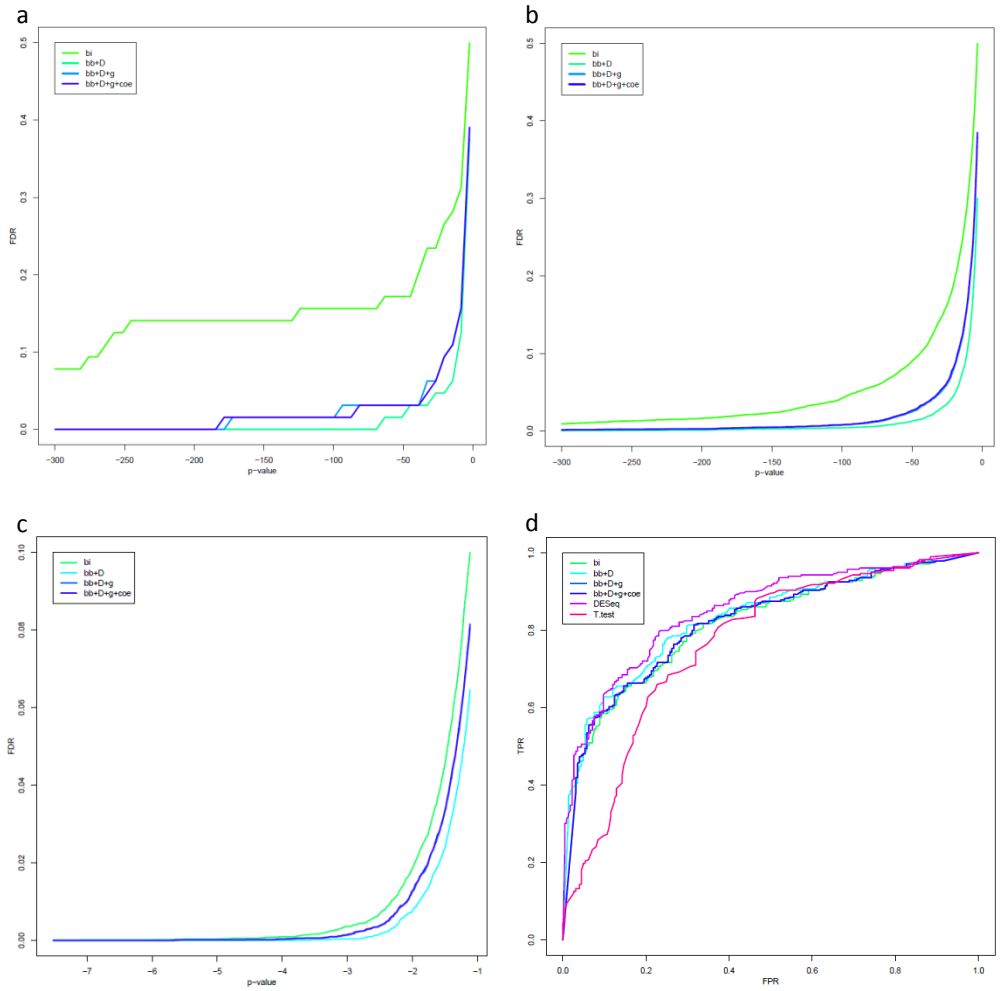


Figure 3.6 FDR and ROC curves for 2 datasets. (a) FDR for Lichun spike-in dataset. (b) FDR for Lichun human dataset. (c) FDR for Bullard dataset. (d) ROC curve for Bullard dataset. About 1000 genes previously assayed by real-time PCR are used as a gold standard to evaluate our test method. In FDR plots, the FDR on the y-axis is plotted against the p-values in log10 scale on the x-axis. bi denotes binomial model, bb+D denotes beta-binomial model with constant overdispersion rate, bb+D+g denotes reduced beta-binomial model, and bb+D+g+coe denotes full beta-binomial model.

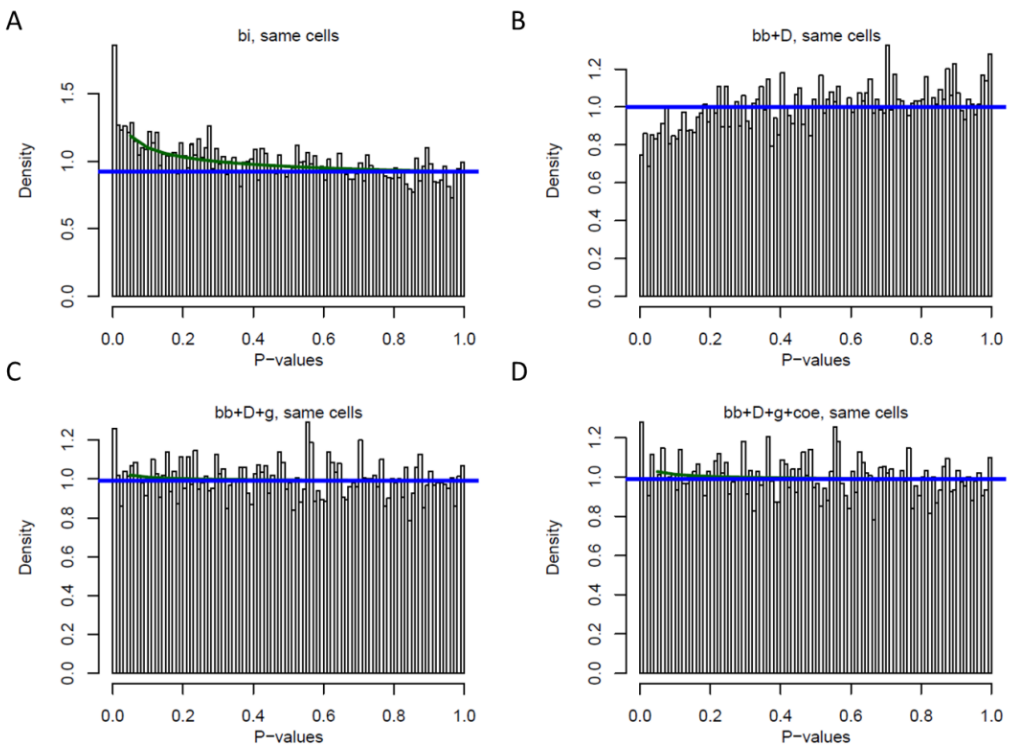


Figure 3.7 Histograms of p-values from replicates of the Bullard dataset. P-values were calculated by (a) binomial model, (b) beta-binomial model with constant θ_{ij}, (c) the reduced beta-binomial model and (d) the full beta-binomial model. In histogram plots, the blue line indicates estimated uniform distributions; green line indicates the mixture distribution of beta distribution and uniform distribution [77].

3.2.5 Comparison of Our Model with DESeq

Compared with our method, DESeq performed slightly better on the AUC (Figure 3.6 d). This is reasonable because DESeq estimated variance by local regression, which is more flexible than our parametric method. Both DEseq and our reduced model were better than the binomial model and significantly superior to the student t-test (Figure 3.6 d). The weak performance of the t-test might be because of the small sample size.

3.3. Discussion

In this study, we found that the overdispersion rate decreases as the sequencing depth increases on the base-pair level, in agreement with what we previously reported on the gene level [73]. Also,

we found that the influence of random hexamer priming on the overdispersion rate is not significant after stratification by sequencing depth. Finally, compared with our other proposed models, we found our beta-binomial model with a dynamic overdispersion rate to be superior. Furthermore, as expected, this model was more desirable than DESeq in a comparison of samples without biological variance.

The property of overdispersion in our model. In this study, we demonstrated that, comparing the reads mapped on specific base pairs, the overdispersion rate decreases as the RNA-seq depth increases. This discovery is consistent with the findings of our previous study at the gene level (described in Chapter 2). First, we observed a strong linear association between the calculated overdispersion rate and the count number. Second, compared with models that ignore the overdispersion rate or which have a constant overdispersion rate, our model that accounts for a dynamic overdispersion rate fit the RNA-seq counts best, demonstrating superior performance based on the likelihood value, AIC, goodness-of-fit χ^2 test, and the testing error indicators FDR and AUC. Rather than considering that the total experimental dataset has a constant variance parameter, other testing methods had been developed as well, including DESeq. DESeq assumes that genes with similar expression levels have the same variance. DESeq performed better than both the models that ignore the overdispersion rate and those with a constant overdispersion parameter [30].

Experimental protocols affect variance. We concluded that experimental protocols have different impacts on the variance of the RNA-seq reads and that even the order of the steps in the protocol matters. We observed extra noise on the tails of genes when fragmentation was performed before PCR. Therefore, we suggest removing the reads on the last 200 base pairs when data from this kind of protocol are analyzed. Because RNA-seq technology involves complex experimental protocols, many biases have been found in library generation, read mapping, and coverage. Systematic errors have been found in differential RNA-seq protocols and platforms [80,81]. Our new finding that conducting fragmentation before PCR introduces extra noise will allow us to develop specific strategies to avoid these biases in our analyses.

Effect of spike-in data on overdispersion rate estimation. Cost issues may prompt consideration of performing experiments without replicates. As measured along with samples that do not have replicates, spike-in samples can be considered as replicates. Our model has only three parameters (Eq. 3 and Eq. 6); therefore, it is quite sufficient to estimate them from roughly 100*1000 counts based on base pairs. Consequently, when testing samples without replicates, one still can borrow the overdispersion rate from the spike-in data. DESeq also tries to handle experiments without replicates by assuming that most genes do not change in expression [30]. However, based on the analysis of samples from a single cell line, our observations indicate that the variance of replicates between two libraries is smaller than the variance between two conditions (Supplementary Figure S3.2). The extra variance caused by gene expression changes would lead to the loss of test power.

Selection of variables. We observed that the overdispersion rate is related to both the local sequence and the sequencing depth. Therefore, we proposed a full model with both variables and a reduced model with only the count number as a variable. The results showed that with the covariate count number, the local sequence has little influence on the overdispersion rate. That is reasonable as the local sequence and count number are dependent, as reported by two groups [3,4]. Consequently, it is preferable to use a model with only the count number as a variable for estimating the overdispersion rate. And we concluded that the reduced model was a better choice for modeling the overdispersion rate and was reasonably economical in terms of time and computing power.

Our model vs DESeq in application. DESeq is widely used in DE testing on RNA-seq. That model performed well when estimating the variance of the counts, including the biological variance. Our model has four main advantages: (1) Modeling based on the proportion in a base-pair unit and modeling the non-uniformity of measurement across the gene. As for the uniformity of measurement, the measurement on each base pair has a specific mean and variance. When considering the total counts as the indicator of the expression level, the estimated expression of one particular gene might be determined by several positions with high counts. Therefore, modeling on the proportion is more desirable because it avoids modeling the highly fluctuating Poisson rate. (2) Tags from strongly

fluctuating positions are down-weighted. Gene expression is no longer the sum over tags from all positions, but is weighted by the overdispersion rate. (3) Our model is based on base pairs with much higher resolution. (4) Spike-in data can be utilized as replicates when estimating the overdispersion rate. (5) With the information from the spike-in data, more accurate estimations can be obtained on samples without replicates, as discussed above. Our model and DESeq are compared in Table 3.2. DESeq testing is performed on the gene expression level, with the hypothesis that the normalized reads on one gene are equal for two samples. In contrast, our model tests on the base pair level with the hypothesis that the normalized reads on each base pair are equal for two samples. Consequently, it depends more on the hypothesis whether there is no difference in the pattern of expression across one specific gene. To reject the null hypothesis that measurements on each base pair are equal, our model is desirable for identifying differential expression on partial genes, but to reject the null hypothesis that measurements on one gene are equal, our model is more suitable for samples with library preparation variance than for samples with biological variance. Also, in this study, we investigated the relationship between overdispersion and the sequencing depth using replicates from the same biological samples. This relationship for different biological samples remains elusive. In other words, our model is desirable for experiments involving samples from a single cell line or the same animal, such as experiments involving one knockout gene from a single cell line. We suggest applying our model to experiments involving samples without biological variance; otherwise, DESeq is more appropriate. In a future study, we will investigate the properties of overdispersion introduced by biological variance.

Table 3.2 Comparison of our model with DESeq

	Our Model	DESeq
Main overdispersion source	Library effect	Library effect + biological variance
Hypothesis	measurement on each base pair are	measurement on each gene are equal

	equal	
Experiments suitable for analysis	All, especially for single cell line	All
Nonuniformity modeling	Yes	No
Borrow information from spike-in data	Can	Can not
Parameter estimation	Maximum likelihood estimation	Local regression (depends on local structure)
Unit for modeling	Base pair	Gene
Expression estimation	Tags were weighted by overdispersion rate	Sum over tags
Replicates required	No, but require hypothesis that most of genes do not change between samples within two conditions, which is inaccurate. Also our method can use spike-in data	No, but require hypothesis that most of genes do not change between samples within two conditions, which is inaccurate.

Many models have been reported to test DE from RNA-seq data. However, it would make much better sense to model the non-uniform measurements of this technology. We modeled the proportion of counts toward this aim, but encountered a limitation of our approach, which is that it may not handle biological variance precisely. Therefore, our model is most appropriate for experiments involving samples from a single cell line or the same animal. The current study provides a detailed understanding of the relationship between the overdispersion rate and sequencing depth, which will aid in the analysis of RNA-seq data for detecting and exploring biological problems. Additionally, we suggest a more desirable beta-binomial model with a dynamic overdispersion rate to cancel the non-uniformity bias and estimate the overdispersion rate more accurately.

CHAPTER 4

A New Type of Bias in RNA-seq

RNA-seq has been widely used in genomic research. However, many studies have reported that inherent biases and spurious effects exist in sequencing technology because of the complexity of the protocol and mechanisms studied. We aim to investigate biases in RNA-seq by exploring the measurement of an external control, spike-in RNA. Tag hybridization has been reported to be the major process through which bias is introduced into microarray analysis. Signals from spike-in transcripts could be influenced by cross-hybridization with tags designed for detecting target transcripts. However, the relationship between spike-in transcripts and sample transcripts has not been fully studied yet. Apart from a concern with cross-hybridization in sequencing technology, it is easy to overlook other possible factors that influence the sequencing measurements. Therefore, this study is important and could aid our understanding of sequencing technology and benefit downstream analysis.

This study is based on two datasets with spike-in controls. The Encode dataset contains measurements from 51 replicates of human samples, and the modENCODE dataset contains sequences from 6 fly samples under difference scenarios. Detailed investigations and correlation analyses were performed among the samples. Also, the alteration of measurements between two samples was modeled with the local sequence as a factor. Furthermore, correction was performed based on the modeling.

We found that an undiscovered bias exists within the measurement of spike-in transcripts, and that it is influenced by the sample transcripts in RNA-seq. Also, we found that this influence is related to the local sequence of the random hexamer used for priming. We suggested modeling the inequality between samples and correcting for this type of bias. After this correction, the Pearson correlation coefficient increased by 0.1. Thus, we revealed a new bias that may be introduced by

resource competition. The current study provides a detailed understanding of the relationship between this new bias and the local sequence, which will aid in our understanding of RNA-seq technology and allow us to correct this bias in the analysis of RNA-seq data.

4.1 Methods

4.1.1 Datasets Used

Two datasets from Lichun et al. [24], ENCODE and modENCODE, containing synthetic spike-in standards from the External RNA Control Consortium (ERCC) were used (Table 4.1).

ENCODE. ENCODE datasets consist of several libraries with different human RNA sources (whole cell, cytosol, and nucleolus) and identifications (longPolyA and longNonPolyA). These data were sequenced along with human samples. The libraries had been prepared using the dUTP protocol for measuring strand-specific transcripts, as described by Lichun et al. [24]. From Illumina GAIIx, 2x76 bp sequencing reads were obtained. The reads were mapped to the human genome (hg19) using STAR software, and the ERCC libraries were mapped using Bowtie, version 0.11.3, with parameters –v2 –m1. To avoid the transcription initiation bias in the sequencing [79], we truncated 50 nucleotides on both ends.

modENCODE . The *D. melanogaster* S2 cell line was used to prepare poly-A+ mRNA. As shown in Table 4.1, modENCODE datasets were obtained from two batches, and each sample was made from a specific library preparation. In batch 1, samples were from four different sample RNA pools, but the sample RNA was from the same pool. The ratio of ERCC and the total RNAs are shown in Table 4.1. Four experiments in batch 1 have ERCC concentrations of 5%, 2.5%, 1% and 100%, which means that only pure ERCC was sequenced. Experiments in batch 2 have the same ERCC concentration of 2.5%. The cDNA was fragmented and the first-strand cDNA was synthesized with random hexamer primers, then the second-strand DNA was synthesized, followed by end repair, poly A addition and adapter ligation. Two methods were used in the preparation. In method A, size selection preceded PCR amplification. This was reversed in method B. The Illumina

GA II platform yielded 36-bp reads. Bowtie version v0.10.0 with parameters -m 1 -v 2 was used to align the reads to the *Drosophila* genome sequence (BDGP release 5, dm3) and ERCC reference sequence. Also, 50 nucleotides were truncated on both ends.

Table 4.1 Two datasets used

	Samples					
ENCODE (51 replicates)	GSM758559 GSM758560 GSM758561 GSM758562 GSM758563 GSM758564 GSM758566 GSM758567 GSM758568 GSM758572 GSM758573 GSM758575 GSM758576 GSM758577 GSM758578 GSM765386 GSM765387 GSM765388 GSM765389 GSM765391 GSM765394 GSM765395 GSM765396 GSM765398 GSM765401 GSM765402 GSM765403 GSM765404 GSM765405 GSM767840 GSM767844 GSM767845 GSM767847 GSM767848 GSM767849 GSM767850 GSM767851 GSM767852 GSM767853 GSM767854 GSM767855 GSM767856 GSM758565 GSM758569 GSM765390 GSM765392 GSM765393 GSM765399 GSM765400 GSM767846 GSM767857					
	Library	Samples	Batch	Sample RNA pool	ERCC %	Method
modENCODE	1	GSM517059	1	1	5	A
	2	GSM517060	1	2	2.5	A
	3	GSM517061	1	3	1	A
	4	GSM517062	1	4	100	A
	5	GSM516588	2	5	2.5	A
	6	GSM516589	2	5	2.5	B

4.1.2 Correlation Analysis of Multiple Samples

To assess the correlation of multiple samples, we calculated the Pearson correlation coefficients on both the gene level and the base level. We sum over the measurement reads on each gene to evaluate the correlation on the gene level and use data on 96 spike-in genes. In comparison, we assess the correlation from the measurement reads on each base pair and use a total of 86,329 reads counts. Hierarchical clustering was performed on the calculated Pearson correlation coefficients to investigate the correlations across multiple samples.

4.1.3 Local Sequence Modeling on Measurement Difference across Samples

In order to model the influence of the local sequence on the measurement difference across samples, we developed a linear model with different sequences on each position as variables. As the outcome variable we used the fold change, which was widely used for measuring change in the expression level of a gene. A log transformation of the fold change would be required in the following linear models.

$$log\, f_i = \alpha + \sum_{k=1}^{K} \sum_{h\in\{A,T,C\}} \beta_{kh} I(b_{ik} = h)\} + \varepsilon \quad (1)$$

In this model, we linked all genes head to tail into one gene. Here, i denotes the i-th position on the imaginary gene; f_i denotes the fold change of reads across two samples on the i-th position by $f_i = \frac{S_{mi}}{S_{ni}}$, where S_{mi} and S_{ni} are the measurements on the i-th position of two samples, m and n. K is the length of the probe around the j-th nucleotide of the imaginary gene. We set $K = 80$ as suggested in a previous study [4]. Also, $I(b_{ik} = h)$ is 1 when the k-th base pair is letter h, which is A, T, or C exclusively, and 0 otherwise. The parameters we want to estimate are α and β_{kh}, and ε is Gaussian noise.

Li et al. used a similar model to predicate measurement reads from RNA-seq [4]. This linear model made the estimation fast and robust. A total of 3 x 80 = 240 parameters on the local sequence were estimated, which is rather a small number compared to the sum of all the positions in all the genes. For the spike-in transcripts, we used reads on all ERCC genes. However, we only used reads on the top 1000 highly expressed genes for the sample transcripts. In order to avoid noise introduced by the low number of reads, we discarded all data points with reads less than 30 mapped to the mRNA transcripts and 5 mapped to ERCC.

4.1.4 Estimating Cross-validation R^2

We used the leave-one-out cross validation strategy to estimate R^2. Data points used in modeling were randomly split into five groups of equal size. In each round, we fit our model using four of these five groups, and then calculated R^2 on the remaining subset by the regression sum of squares divided by the total sum of squares. The final cross-validation R^2 was determined as the mean.

4.1.5 Bias Correction

Once the parameters α and β_{kh} in Eq 1 were estimated, we performed the bias correction based on our model according to Eq 2. We randomly split the data points into five parts and used four of them as the training data and left one out to serve as test data for evaluating the correction.

$$S_{mi}^{*} = S_{ni} * log\, f_i \quad (2)$$

In this equation, S_{mi}^{*} denotes the corrected measurement on the i-th position of the imaginary gene of sample m according to the coefficients on the local sequence of the probe calculated from the fold change of reads across two samples.

4.2 Results

4.2.1 Pattern of Reads on Spike-in Transcripts

In order to investigate the sequencing reads on each position of the spike-in transcripts, we plotted the read counts mapped on the spike-in transcripts. Figure 4.1 shows an example of ERCC-00002, one of the genes with the highest yield of reads. We observed that the pattern of reads of the spike-in transcripts were divergent between samples, although they were consistent between replicates from the ENCODE datasets. And from the modENCODE datasets, we observed that the patterns of the reads differed between batches, but the influence of the different libraries was small. And we noticed a dissimilar pattern in experiments with both ERCC and a sample pool when compared with experiments that used only pure ERCC transcripts. However, no significant

difference was found for the different order of size selection and PCR amplification in sample preparation. We validated our finding statistically by comparing correlations of the sequencing reads from pairwise samples (described in Methods, 4.1.2). From the heatmap of hierarchical clustering based on correlations of counts on each position across samples (Figure 4.2) and pairwise correlation plotting of modERCC datasets (Figure 4.3), we observe that replicates are clustered together for the ENCODE samples and samples in the batch are clustered together, except the samples from sequencing pure ERCC transcripts. Without sample mRNA, the dissimilar of correlation between batches was reduced but increased when other samples were sequenced along with the sample transcripts in the same batch. Also, we concluded that compared with the correlation of the total counts on each gene, the correlation of the counts on each position was more precise (Figure 4.2). The same patterns of correlation across samples were also observed on target transcripts (Figure 4.2 E, F). The top 1000 highly expressed genes were used to calculate the correlation. Distinct patterns were observed on both the base level and the gene level, which might indicate that the discrepancy originated in the sample transcripts.

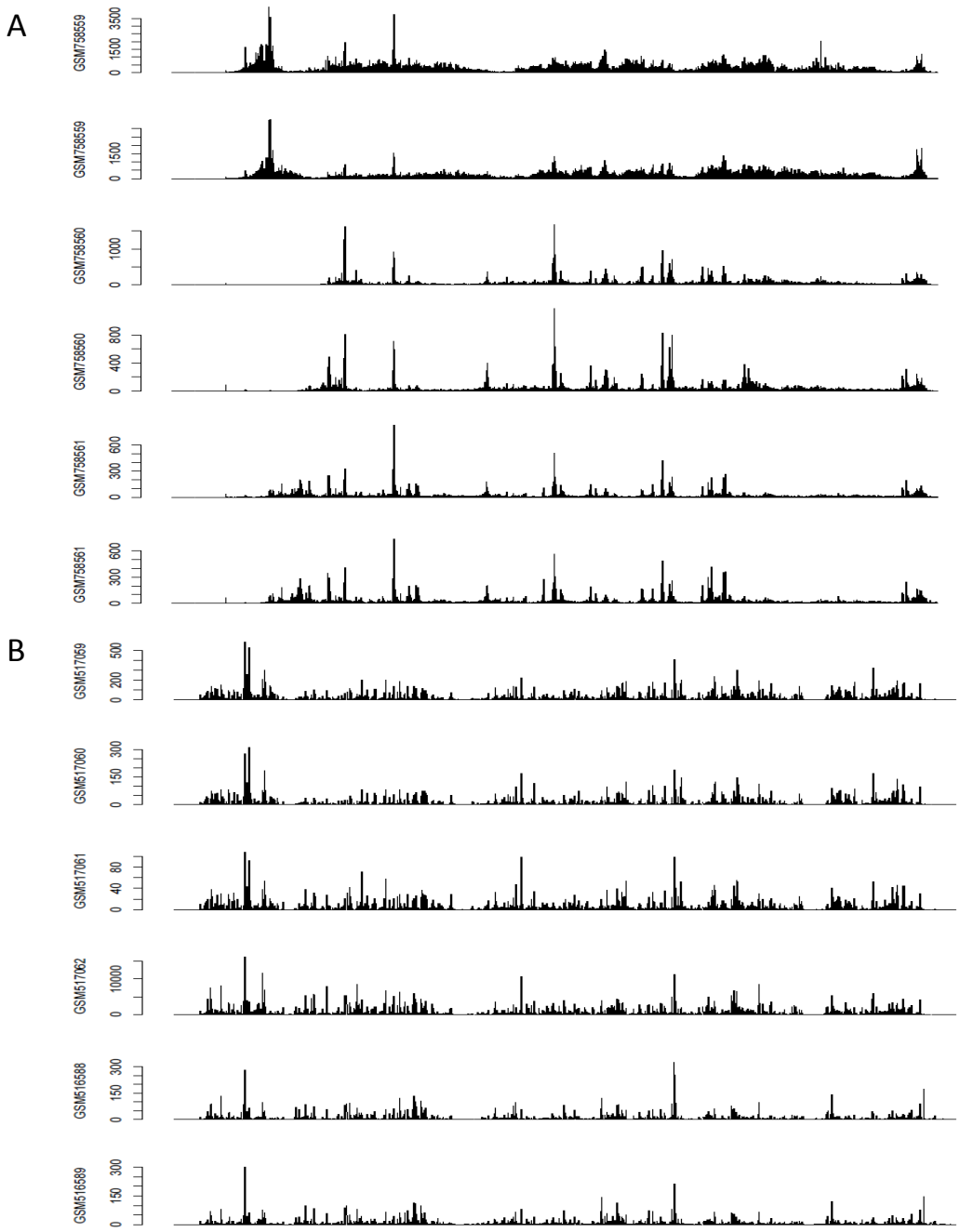


Figure 4.1 Distributions of sequencing reads on ERCC-00002 of different samples. The positions of ERCC-00002 are plotted on the x-axis and the number of sequencing reads are plotted as bars. The distribution of sequencing reads on ERCC-00002 samples from (A) ENCODE datasets and (B) modENCODE. The same label denotes the replicates.

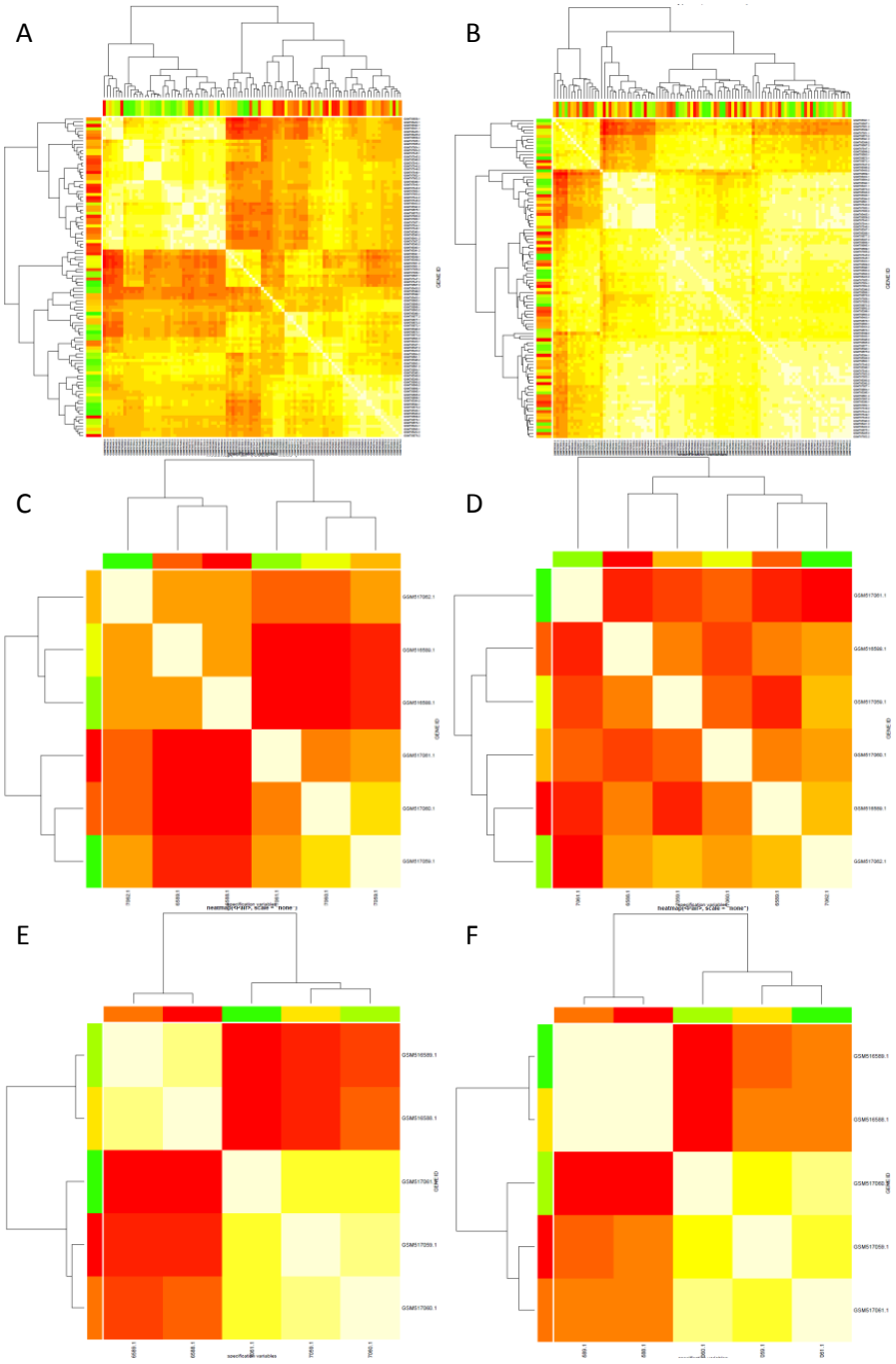


Figure 4.2 Heatmaps of hierarchical clustering on samples. The clustering was based on correlations of (A) counts on each position of spike-in transcripts across samples of ENCODE datasets, (B) counts on each gene of spike-in transcripts across samples of ENCODE datasets, (C) counts on each position of spike-in transcripts across samples of modENCODE datasets, (D) counts on each gene of spike-in transcripts across samples of modENCODE datasets, (E) counts on each position of the top 1000 highly expressed sample transcripts across samples of modENCODE datasets, (F) counts on each gene of the top 1000 highly expressed sample transcripts across samples of modENCODE datasets.

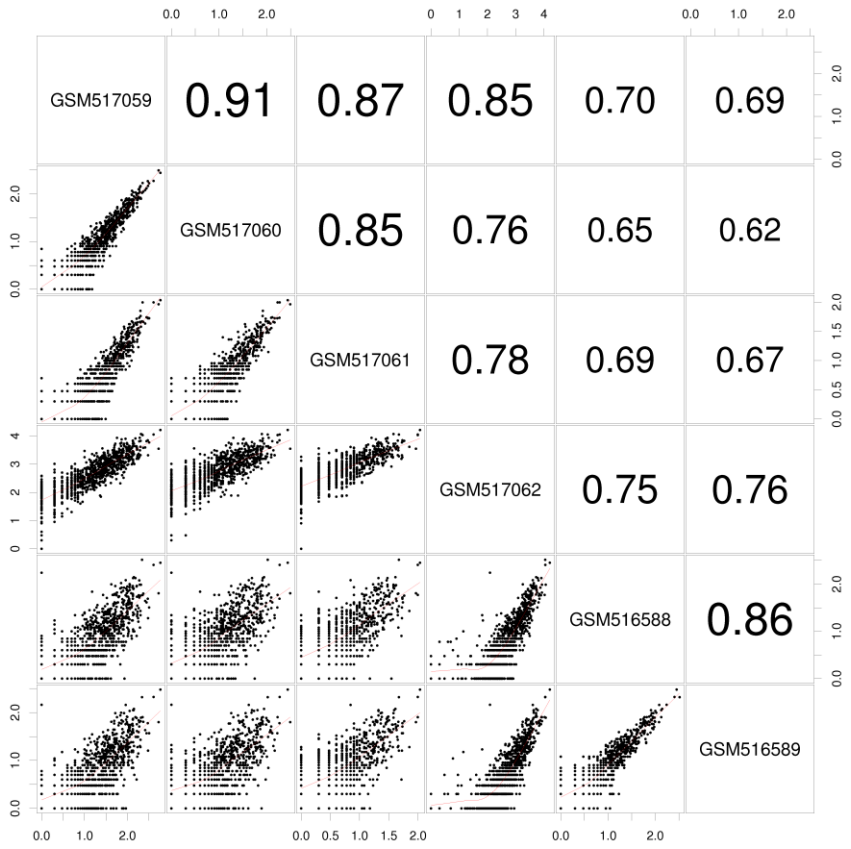


Figure 4.3 Pairwise comparison matrix of the measurement on each base pair of ERCC transcripts from modENCODE dataset. Pairwise plots of sequencing reads mapped on each base pair on the log10 scale are shown in the bottom panel, and calculations of pairwise Pearson coefficients are shown in the top panel.

4.2.2 Modeling on the Local Sequence

We modeled the influence of the local sequence. Using the linear model, we estimated 240 coefficients of 80 positions around the primer from modENCODE mRNA and spike-in reads separately. We plot those estimated coefficients against their corresponding positions in Figure 4.4. We observe that significant coefficients were estimated from the difference between two samples from two batches, as well as separately from the comparison of two samples with and without target transcripts. The significant coefficients expanded to a range from -20 to 15 around the start site of the primer. As expected, no significant coefficients were found by comparing samples with different

ERCC concentrations in the same batch. And interestingly, the order of the size selection and PCR amplification only affected the measurements through the first 2 nt of the hexamer primer. Consistent with our above findings, the patterns of the coefficients were concordant between mRNA and ERCC. Cross-validated r^2 were calculated from the sample mRNA transcripts as shown in Table 4.2, indicating our model could explain around 40% of the differences in two samples.

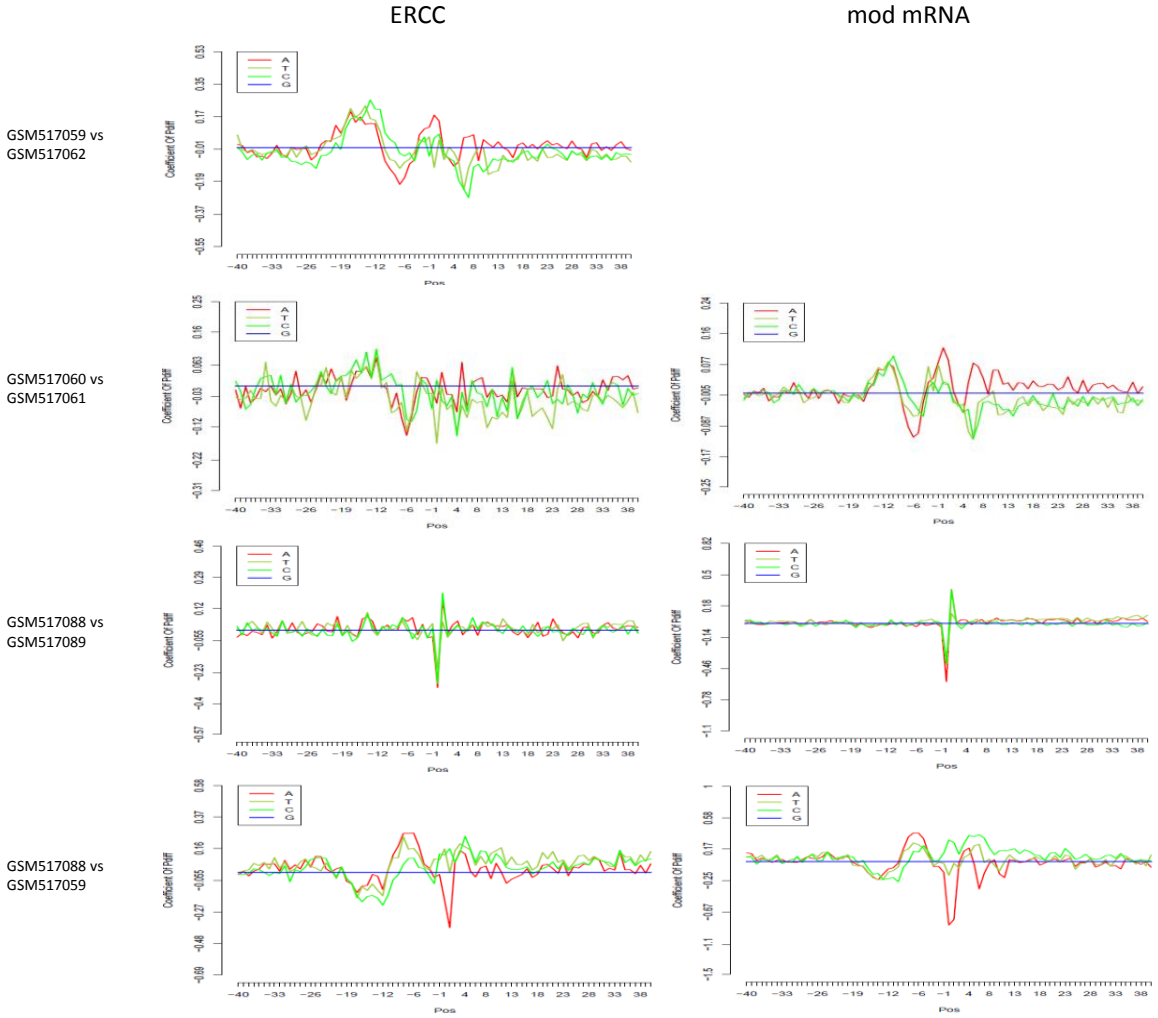


Figure 4.4 Coefficients estimated by the linear model, Eq.1, from the modENCODE dataset. Plotted on the x-axis are the positions around the 5' end of the mapped reads, labeled 0. Coefficients were calculated on ERCC spike-in transcripts (left panel) and mod mRNA transcripts (right panel). Comparing GSM517059 vs GSM517062 captures the discrepancy from ERCC with and without mRNA transcripts. Comparing GSM517060 vs GSM517061 captures the discrepancy from the ERCC ratio (2.5% vs 1%). Comparing GSM517088 vs GSM517089 captures the discrepancy from

the order of the sample preparation (size selection preceding PCR amplification vs the reverse order). Comparing GSM517088 vs GSM517059 captures the discrepancy from different batches.

Table 4.2 Cross-validated r^2 Calculated

	GSM516588	GSM516589	GSM516590	GSM517059	GSM517060	GSM517061
GSM516588	–	0.257	0.274	0.299	0.313	0.260
GSM516589	–	–	0.035	0.382	0.371	0.329
GSM516590	–	–	–	0.408	0.390	0.348
GSM517059	–	–	–	–	0.138	0.172
GSM517060	–	–	–	–	–	0.163
GSM517061	–	–	–	–	–	-

4.2.3 Correction of bias

We estimated the coefficients of the local sequence by comparing the training set randomly selected from two samples. According to our model and estimated coefficients, we performed a correction on the test set. As a result, our correction increased the Pearson correlation from 0.48 to 0.58 (Figure 4.5).

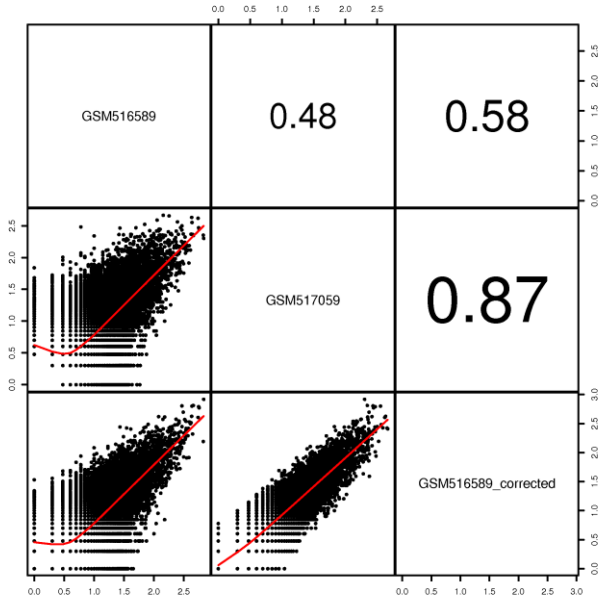


Figure 4.5 Pairwise comparison matrix of the original and corrected measurements on each base pair of mRNA transcripts. Pairwise plots of sequencing reads mapped on each base pair on the log10 scale are shown in the bottom panel; calculations of the pairwise Pearson coefficients are shown in the top panel.

4.3 Discussion

In this study, we found that an unreported bias in the measurement of spike-in transcripts exists, and that it is influenced by sample transcripts in RNA-seq. Also, we found that this influence is related to the local sequence of the random hexamer used for priming. We proposed to model the inequality between samples and suggested a method to correct this bias based on this model.

A new Bias in RNA-seq. In our study, when comparing the reads mapped on specific base pairs, we found similar patterns in replicates but not across different samples. Although many factors could contribute to this discrepancy, we failed to observe a significant difference from the preparation methods and library preparations. This indicates that inconsistency in the library preparation did not contribute to this bias. Interestingly, a rather significant difference was observed on the measurement of the spike-in transcripts sequenced along with different sample transcripts. And diverse sample sources or batches could result in the difference in the sample transcripts. Cross-hybridization has been reported as an inherently problematic issue in microarrays. Unspecific tags could hybridize with other sequences besides the target and introduce a bias in measurements [10,11]. The measurement of spike-in transcripts could be influenced by incorrect interactions with probes designed for target transcripts [13]. Similar to cross-hybridization in microarrays, a mechanism may exist that affects the measurement of spike-in transcripts by sample transcripts. This mechanism may be from the competition of sequencing resources, such as dNTP and the hexamer primer. The source of this bias needs to be studied further for clarification..

In this study, we proposed a statistical method to model the influence of the local sequence on this new bias. And we observed significant coefficients ranging from -20 to +15 around the beginning of the hexamer primer. This range of coefficients has been reported in the study of non-uniformity by Li et al. [4] and in our study of overdispersion in Chapter 3. This finding may indicate that the bias from hexamer priming affects more than one property. Besides affinity abilities, resource competition involving the hexamer primer may play a role in introducing bias. Our model

can be used to correct this type of bias, and, indeed, the Pearson correlation coefficient increased by 0.1 after we made this correction.

Novelty and strength of our study. First, we discovered a new type of bias that may be generated from resource competition and which is related to the local sequence, as described above. We suggested a statistical model to model the bias and perform the correction. Our findings will contribute to understanding RNA-seq technology, exploring inherent biases and estimating true measurements for downstream analysis.

Second, for the first time, to our knowledge, we found that the order of size selection and PCR amplification could influence the measurement through the first 2 nt of the hexamer primer in RNA-seq. One possible explanation is that PRC amplification is influenced by the first 2nt of the reads and causes the bias, together with the size selection by the imprecise isolation of agarose gel electrophoresis.

Third, we demonstrated that a comparison of measurements on each position is more precise than a comparison based on the gene. In this study, we observed much more precise correlations of measurements between replicates with higher resolution compared with correlations on genes. Therefore, we suggest utilizing all information from all base pairs in analyzing sequencing data. However, more exhaustive research on sequencing bias and sophisticated methods were required in this analysis. We have suggested a powerful method to estimate the overdispersion rate based on base pairs (Chapter 2).

Fourth, rather than modeling based on measurements of one sample, we modeled the fold change between two samples. Benefiting from this effort, we discovered the new bias and were able to offer a method of correction.

Many biases have been reported in RNA-seq data and several methods have been proposed for bias correction. However, the research on RNA-seq technology is still in its infancy. Here, we revealed a new bias that may be introduced by resource competition. Our study provides a detailed

understanding of the relationship between this new bias and the local sequence, which will aid in understanding RNA-seq technology and in correcting for this bias in the analysis of RNA-seq data.

CHAPTER 5

New CpG Island Methylator Phenotype (CIMP) and Biomarker Identification by Integrating Methylation and mRNA Expression

The expression of a gene can be turned off when its promoter is highly methylated. Several studies have reported that a clear threshold effect exists in the gene silencing that is mediated by DNA methylation. It is reasonable to assume that a specific DNA methylation threshold exists for each gene because of the complicated transcription regulatory system. Therefore, we must determine that threshold in order to predicate whether the gene was inhibited by DNA methylation. According to the estimated thresholds, DNA methylation status could be dichotomized and makes the task of biclustering easier, which is a good way to identify CIMP. We aimed to develop a method to determine the DNA methylation threshold and investigate whether CIMP exists in breast cancer. Only limited research has claimed the identification of CIMP with hypermethylated genes in breast cancer.

We developed a method to determine the DNA methylation threshold from 997 samples across 7 cancer types from TCGA datasets obtained from Illumina Infinium Hman DNA Methylation27 arrays and Illumina GA II and HiSeq platforms. Then, from 285 tumor samples and 21 normal samples of breast tissue, we selected 128 "L-shaped" genes according to our criteria and identified CIMP by biclustering and hierarchical clustering. Gene-set enrichment analysis and correlation analysis on expression, mutation and clinical features were performed.

We suggested a method based on mutual information calculation to determine the threshold of DNA methylation and distinguish the genes for which the expression levels were significantly regulated by DNA methylation. Based on the dichotomized methylation status predicated on 128 "L-shaped" genes, we identified a new CIMP of *BRCA* with 11 markers. We observed significant correlations of CIMP+ with wild-type *TP53* mutation, ER+/PR+ positive status, higher age at initial

pathologic diagnosis, better treatment response and perhaps a longer survival time. The 11 CIMP markers were shown to be associated with *TP53* directly or indirectly, and were enriched in cancer and other disease networks. Also, we found that 7 epigenetic genes were correlated strongly with both the new CIMP and *TP53* mutation. Based on our findings, we proposed a model of a *TP53*-mediated regulatory network with two components: "Guidance" and "Sustainer."

We developed a powerful method to dichotomize the methylation status and identify a CIMP of breast cancer with a distinct classification of molecular characteristics and clinical features. Our results suggest that methylation may play an important role in resisting tumor development. The regulatory component of "Guidance" which we defined, and genetic modifiers BMI1, IDH1 and TET1 might be potential targets for new treatments.

5.1 Methods

5.1.1 Datasets Used

Methylation datasets. We obtained methylation datasets from TCGA generated by the Illumina Infinium Human DNA Methylation27 array for 3382 samples across 12 cancer types (Table 5.1). For breast cancer, 318 tumor and 29 normal samples were measured. Level 3 preprocessed data were available for beta values, which is the ratio of the methylated probes among all probes for each detected site.

mRNA expression datasets. Also from TCGA, we downloaded mRNA expression datasets generated by the Illumina GA II and HiSeq platforms. Data for 2271 samples across 9 cancer types were available as of February 25, 2012 (Table 5.1). For breast cancer, 775 tumor and 102 normal samples were measured. In each sample, RPKM for 20532 genes were calculated as level 3 data. In order to avoid the 0 value, we replaced them with the minimum non-zero value of the same gene among all samples. And we took log2 scale of the RPKM value.

Overlapping datasets used in mutual information (MI) calculation. A total of 997 samples across 7 cancer types had both methylation status and mRNA expression data (Table 5.1). Among these 997 samples, 285 tumor and 21 normal samples were measured for breast cancer.

Table 5.1 Datasets used

		BRCA	COAD	GBM	KIRC	KIRP	READ	LAML	LUSC	LUAD	OV	STAD	UCEC	LIHC	HNSC
DNA Methylation	Tumor	318	168	296	438	16	70	384	134	128	576	82	117		
(3382)	Normal	29	45	6	410	6	11	0	32	27	25	61	3		
mRNA Expression	Tumor	775	192		468	16	71		221				306	17	18
(2271)	Normal	102	0		68	0	0		17				3	8	0
Overlapped	Tumor	285	161		207	16	67		126				113		
(997)	Normal	21	0		0	0	0		0				0		

5.1.2 Determining Methylation Threshold

Mutual information computation. Mutual information computation of the density of the distribution can be improved by taking into account the natural measurement of uncertainty. At low ranges of gene expression, especially for the log-transformed RNA-seq data, the large difference is not real, but is due to the randomness of the measurement. This can introduce noise into the mutual information calculation, especially when the number of samples is small. To solve this problem, a pair of methylation-expression measurements is not assigned to its bin, instead, it is represented by a smeared density function centered at the methylation-expression values. The uncertainties in both directions are taken from the estimated measurement errors.

Measurement values were assumed to be distributed as normal distributions. Their uncertainties were calculated from 6 replicates. Subsequently, we calculated the probability of expression and the methylation value for each patient by summing up the probabilities from all patients. We can write the joint and marginal probabilities as

$$p_{x_i} = \frac{1}{N}\sum_{j=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_{x_j}} e^{-\frac{(x_i - x_j)^2}{2\sigma_{x_j}^2}}$$

$$\widetilde{p_{x_i}} = \frac{1}{\Delta x \sum_{i=1}^{S_{,1}} p_{x_i}} \, p_{x_i} \qquad i \in [1,2,\dots S_{,1}] \qquad (1)$$

$$p_{\beta_i} = \frac{1}{N} \sum_{j=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_{\beta_j}} e^{-\frac{(\beta_i - \beta_j)^2}{2\sigma_{\beta_j}^2}}$$

$$\widetilde{p_{\beta_i}} = \frac{1}{\Delta\beta \sum_{i=1}^{S_{1,}} p_{\beta_i}} p_{\beta_i} \qquad i \in [1,2,\dots S_{1,}] \qquad (2)$$

$$p_{x_i\beta_i} = \frac{1}{N} \sum_{j=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_{x_j}\sigma_{\beta_j}} e^{[-\frac{(x_i - x_j)^2}{2\sigma_{x_j}^2} - \frac{(\beta_i - \beta_j)^2}{2\sigma_{\beta_j}^2}]}$$

$$\widetilde{p_{x_i\beta_i}} = \frac{1}{\Delta x \Delta\beta \sum_{i=1}^{S} p_{x_i\beta_i}} p_{x_i\beta_i} \quad i \in [1,2,\dots S] \quad (3)$$

$$where\ \Delta x = \frac{x_{max} - x_{min}}{n_x}\ and\ \Delta\beta = \frac{1}{n_\beta}$$

where p_{x_i} is the marginal probability of the i-th pseudo patient's expression level of one particular gene. Similarly, p_{β_i} is the marginal probability of the methylation β value. In Eq.(3), $p_{x_i\beta_i}$ is the joint probability of the mRNA expression level and methylation β value. N denotes the total number of patients and i pseudo and j indicate the i-th and j-th patients, respectively. The expression is denoted by x and the methylation values are denoted by β. In addition, $\sigma_{x_j}^2$ and $\sigma_{\beta_j}^2$ are the uncertainties in both directions for the j-th patient.

In order to determine the methylation threshold, we slide the cutoff point from the minimum to the maximum methylation values. Mutual information values are calculated for two parts besides the cutoff, and the sum is taken as Eq.(4). In the same way, we can calculate expression mutual information. Using Eq.(5), we calculated the "2-way" mutual information integrating mRNA expression and DNA methylation.

$$MI_\beta = w_S \int_{x_{min}}^{x_{max}} \int_{0}^{\beta_l} \widetilde{p_{x_i\beta_i}} \, log\left(\frac{\widetilde{p_{x_i\beta_i}}}{\widetilde{p_{x_i}}\widetilde{p_{\beta_i}}}\right) d_{x_i} d_{\beta_i} + w_{N-S} \int_{x_{min}}^{x_{max}} \int_{\beta_l}^{1} \widetilde{p_{x_i\beta_i}} \, log\left(\frac{\widetilde{p_{x_i\beta_i}}}{\widetilde{p_{x_i}}\widetilde{p_{\beta_i}}}\right) d_{x_i} d_{\beta_i}$$

$$MI_x = w_S \int_{x_{min}}^{x_k} \int_0^1 \widetilde{p_{x_\iota\beta_\iota}} \log\left(\frac{\widetilde{p_{x_\iota\beta_\iota}}}{\widetilde{p_{x_\iota}}\widetilde{p_{\beta_\iota}}}\right) d_{x_i} d_{\beta_i} + w_{N-S} \int_{x_k}^{x_{max}} \int_0^1 \widetilde{p_{x_\iota\beta_\iota}} \log\left(\frac{\widetilde{p_{x_\iota\beta_\iota}}}{\widetilde{p_{x_\iota}}\widetilde{p_{\beta_\iota}}}\right) d_{x_i} d_{\beta_i}$$

$$where\ x_k = x_{min} + k\Delta x \quad k \in (0,1,2 \dots n_x]\ and\ \beta_l = l\Delta\beta \quad l \in (0,1,2 \dots n_\beta]$$

$$MI_\beta and\ MI_x\ can\ be\ calculated\ as$$

$$MI_1 = w_S\, \Delta x\Delta\beta \sum_{i=1}^{S} \widetilde{p_{x_\iota\beta_\iota}} \log\left(\frac{\widetilde{p_{x_\iota\beta_\iota}}}{\widetilde{p_{x_\iota}}\widetilde{p_{\beta_\iota}}}\right) + w_{N-S}\Delta x\Delta\beta \sum_{i=S+1}^{N-S} \widetilde{p_{x_\iota\beta_\iota}} \log\left(\frac{\widetilde{p_{x_\iota\beta_\iota}}}{\widetilde{p_{x_\iota}}\widetilde{p_{\beta_\iota}}}\right)$$

$$= \Delta x\Delta\beta \sum_{i=1}^{S} p_{x_i\beta_i} \log\left(\frac{\widetilde{p_{x_\iota\beta_\iota}}}{\widetilde{p_{x_\iota}}\widetilde{p_{\beta_\iota}}}\right) + \Delta x\Delta\beta \sum_{i=S+1}^{N-S} p_{x_i\beta_i} \log\left(\frac{\widetilde{p_{x_\iota\beta_\iota}}}{\widetilde{p_{x_\iota}}\widetilde{p_{\beta_\iota}}}\right) \quad (4)$$

$$where\ w_S = \sum_{i=1}^{S} p_{x_i\beta_i}\ and\ w_{N-S} = \sum_{i=1}^{N-S} p_{x_i\beta_i}$$

$$MI_2 = w_{S1} \int_{x_{min}}^{x_k} \int_0^{\beta_l} \widetilde{p_{x_\iota\beta_\iota}} \log\left(\frac{\widetilde{p_{x_\iota\beta_\iota}}}{\widetilde{p_{x_\iota}}\widetilde{p_{\beta_\iota}}}\right) d_{x_i} d_{\beta_i} + w_{S2} \int_{x_k}^{x_{max}} \int_0^{\beta_l} \widetilde{p_{x_\iota\beta_\iota}} \log\left(\frac{\widetilde{p_{x_\iota\beta_\iota}}}{\widetilde{p_{x_\iota}}\widetilde{p_{\beta_\iota}}}\right) d_{x_i} d_{\beta_i}$$

$$+ w_{S3} \int_{x_{min}}^{x_k} \int_{\beta_l}^{1} \widetilde{p_{x_\iota\beta_\iota}} \log\left(\frac{\widetilde{p_{x_\iota\beta_\iota}}}{\widetilde{p_{x_\iota}}\widetilde{p_{\beta_\iota}}}\right) d_{x_i} d_{\beta_i} + w_{S4} \int_{x_k}^{x_{max}} \int_{\beta_l}^{1} \widetilde{p_{x_\iota\beta_\iota}} \log\left(\frac{\widetilde{p_{x_\iota\beta_\iota}}}{\widetilde{p_{x_\iota}}\widetilde{p_{\beta_\iota}}}\right) d_{x_i} d_{\beta_i}$$

$$= \Delta x\Delta\beta \sum_{i=1}^{S1} p_{x_i\beta_i} \log\left(\frac{\widetilde{p_{x_\iota\beta_\iota}}}{\widetilde{p_{x_\iota}}\widetilde{p_{\beta_\iota}}}\right) + \Delta x\Delta\beta \sum_{i=1}^{S2} p_{x_i\beta_i} \log\left(\frac{\widetilde{p_{x_\iota\beta_\iota}}}{\widetilde{p_{x_\iota}}\widetilde{p_{\beta_\iota}}}\right)$$

$$+ \Delta x\Delta\beta \sum_{i=1}^{S3} p_{x_i\beta_i} \log\left(\frac{\widetilde{p_{x_\iota\beta_\iota}}}{\widetilde{p_{x_\iota}}\widetilde{p_{\beta_\iota}}}\right) + \Delta x\Delta\beta \sum_{i=1}^{S4} p_{x_i\beta_i} \log\left(\frac{\widetilde{p_{x_\iota\beta_\iota}}}{\widetilde{p_{x_\iota}}\widetilde{p_{\beta_\iota}}}\right) \quad (5)$$

$$where\ x_k = x_{min} + k\Delta x \quad k \in (0,1,2 \dots n_x]\ and\ \beta_l = l\Delta\beta \quad l \in (0,1,2 \dots n_\beta]$$

$$and\ S1 + S2 + S3 + S4 = N.$$

Criteria. Aiming to identify "L-shaped" genes, we applied the following three criteria: (1) the range of mutual information is no less than 0.3, (2) samples are split into 4 quadrants by the thresholds of mRNA expression and DNA methylation, with at least 200 samples located in each of the first and fourth quadrants, (3) no fewer than 600 samples are in the first and the fourth quadrants.

Binary coding. Once we obtained the threshold of the DNA methylation beta value, each gene was assigned to 1 or 0, where 1 means the methylation status of a gene changed compared to that of normal samples.

5.1.3 Gene Set Enrichment Analysis

MsigDB. MsigDB (http://www.broadinstitute.org/gsea/msigdb/index.jsp) has more than 6000 integrated datasets and a comprehensive analysis platform. It covers positional gene sets, curated gene sets, motif gene sets, computational gene sets and GO gene sets.

BioProfiling.de . BioProfiling.de (http://www.bioprofiling.de/index.html) is another online tool we used for comprehensive analysis of gene sets. It provides a handful of types of analysis, including GO gene function, IntAct protein interaction and KEGG pathway relationships, cancer relationships and miRNA regulatory predication.

Ingenuity Pathway Analysis (IPA). We used IPA software to obtain gene sets of enriched signal pathways and found potential upstream regulators.

5.1.4 CIMP Identification

Biclustering. Finding contiguous blocks with changed methylation status is of interest and we used the biclustering algorithm BicBin to handle our binary and sparse data matrices.

Hierarchical clustering. For the block found by Bicbin, we applied hierarchical clustering to find biomarker genes of the CIMP. We used R package hclust and OOMPA, applying the "ward" agglomeration method and the "binary" distance measure.

5.1.5 Clinical Correlation Analysis

We downloaded from TCGA a clinical dataset for breast cancer that contained 919 patient samples, 316 of which overlapped with both the methylation and mRNA expression datasets. We applied a generalized linear model to investigate the correlation of the methylation status with

clinical features, including "pretreatment_history," "ajcc_cancer_metastasis_stage_code," "prior_diagnosis," "ajcc_neoplasm_disease_stage," "ajcc_tumor_stage_code," "gender," "age_at_initial_pathologic_diagnosis," "days_to_death," "breast_carcinoma_progesterone_receptor_status," "breast_carcinoma_estrogen_receptor_status," and "lab_proc_her2_neu_immunohistochemistry_receptor_status."

5.1.6 Survival Analysis

From the "days_to_death" values in the above clinical dataset, we applied the Cox model for right-censored survival analysis using the R library "survival." Also, we examined survival analysis from BioProfiling.de based on texting mining information.

5.1.7 Mutation Analysis

We downloaded a genetic mutation pre-processed dataset from TCGA that contained 507 samples, 301 of which overlapped with both the methylation and mRNA expression datasets. Again, we used a generalized linear model to inspect the correlation with methylation status and mRNA expression.

5.1.8 Identifying *TP53*-Mediated "Rescue" Genes

"Guardian." In this study, we identified the "guardian" genes of the *TP53* system according to the following criteria: (1) a significant different value exists for CIMP+ compared with normal samples and CIMP-, with the same trend; and (2) no significant difference exists between CIMP- and normal samples.

"Sustainer." We identified the "sustainer" genes of the *TP53* system according to the following criteria: a significant different value exists for CIMP- compared with normal samples and CIMP+, with the same trend.

5.2 Results

5.2.1 Determining Methylation Threshold

For each gene, we calculated the mutual information score (MI score) on both the methylation and mRNA expression dimensions according to Eq (5). As shown in Figure 5.1, along with the increments in the measurement value on each dimension, the MI score decreased and then increased. The point with the lowest MI score indicated the optimal threshold. A plot of mRNA expression against DNA methylation would exhibit an "L" shape for a gene that could be turned off by DNA methylation. As an example, the ESR1 gene is a typical "L-shaped" gene that encodes an estrogen receptor that is an important breast cancer biomarker. From our analysis, 271 among 285 breast cancer samples and all 21 paired normal samples showed ESR1 hypomethylation status and high mRNA expression level (Figure 5.1 A). Compared with ESR1, HOXA9 showed a reverse "L" shape, as the threshold of methylation was much higher and it was hypomethylated in breast cancer samples but was hypermethylated in normal samples (Figure 5.1 B). The 3-D MI score of HOXA9 is shown in Figure 5.2. Also, the results show that mRNA expression levels are dissimilar across cancer types (Figure 5.1).

In Figure 5.3, for all genes, histograms of the thresholds identified show that they were enriched in small methylation values and large mRNA expression levels. However, for genes with MI differences larger than 0.3, the thresholds estimated were highly gene-specific, and they were enriched in small mRNA expression levels, which is expected as the character of "L-shaped" genes. Also, a small peak that is seen on the right tail of the histogram for mRNA expression might have been formed by a reverse "L-shaped" gene, such as HOXA9 (Figure 5.3 D).

As expected, the MI scores of the "L-shaped" genes show a deep "U" shape (Figure 5.1). And the depth of this kind of "U" shape shows the difference in the maturity of the information before and after being split apart by the threshold cutoff. Therefore, the depth of this "U" shape is a good indicator of "L-shaped" genes. The histogram of the depth, displayed in Figure 5.4, shows that 449 genes reached our criteria depth of 0.3. We performed a differential analysis between breast cancer

and normal samples for both DNA methylation and mRNA expression. After adjusting for the FDR, we plotted the p-values of differential methylation against that of mRNA expression. Using the normal samples as the baseline, a minus score means the measurement of tumor samples is lower than that of the normal samples. We observe in Figure 5.4 that genes with a deep "U" shape are enriched in the lower-right and upper-left quadrants, which indicates that they tend to have correlated methylation status and mRNA expression, in the way that high expression correlates with low methylation status or low expression correlates with high methylation status. However, we also observed reverse correlations of methylation status and mRNA expression for many genes. And, based on depths larger than 0.3 as one of our criteria (see Methods), most genes were filtered out because of moderate changes in MI scores.

In order to validate the hypothesis that beta value thresholds are tailored to each microarray probe, we examined the MI scores of 29 genes as epigenetic prognostic signatures from the colon cancer study by Yi et al [82]. As the dichotomized methylation status will more accurately reflect the on-off state of a gene due to DNA methylation, we found that 16 of 25 overlapping genes have a depth of MI score larger than 0.1 (compared with 2,363 genes among 12,783; chi squared test, p-value 9.397e-05) (Table S5.1) and exhibit "L" shapes when the methylation value is plotted against the mRNA expression value (data not shown).

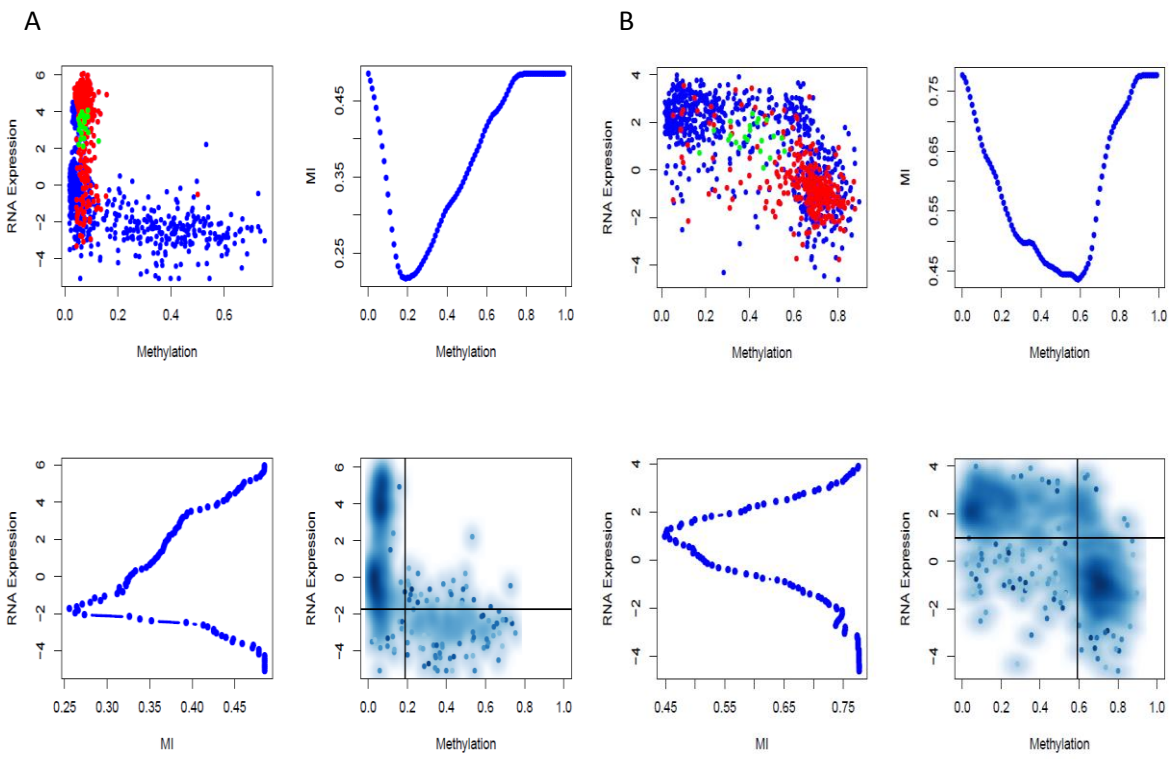


Figure 5.1 Threshold determination from integrating mRNA expression and DNA methylation datasets. A) *ESR1* gene; B) *HOXA9* gene. Threshold determination of ESR1 genes and pattern of mRNA expression against DNA methylation status is investigated for 997 patient samples across 7 cancer types (top left); red points are from *BRCA* samples; green points are from normal samples; and blue points are from other cancer types. Mutual information score is calculated for DNA methylation (top right) and mRNA expression (bottom left) by sliding the cutoff point. Thresholds are determined with the minimal mutual information score (bottom right). B exhibits results in the same way for the *HOXA9* gene.

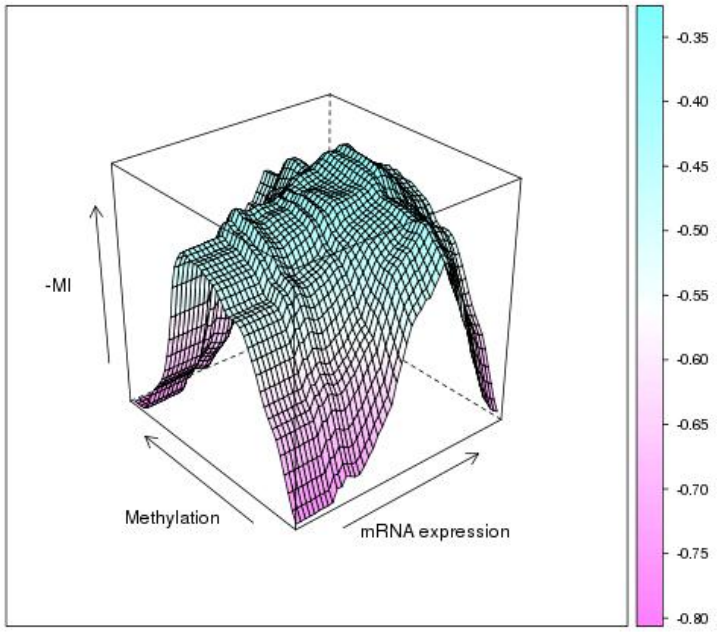


Figure 5.2 3-D plotting of mutual information score calculated for *HOXA9* genes.

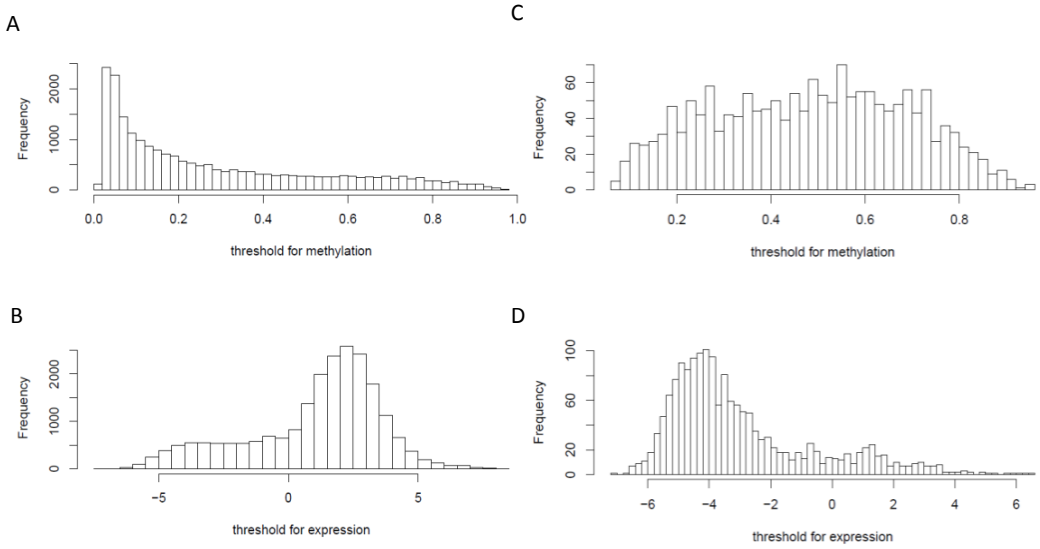


Figure 5.3 Histogram of thresholds estimated. A, B are for all genes; C, D are for differences in MI larger than 0.3; A, C are for DNA methylation status; and C, D are for mRNA expression level.

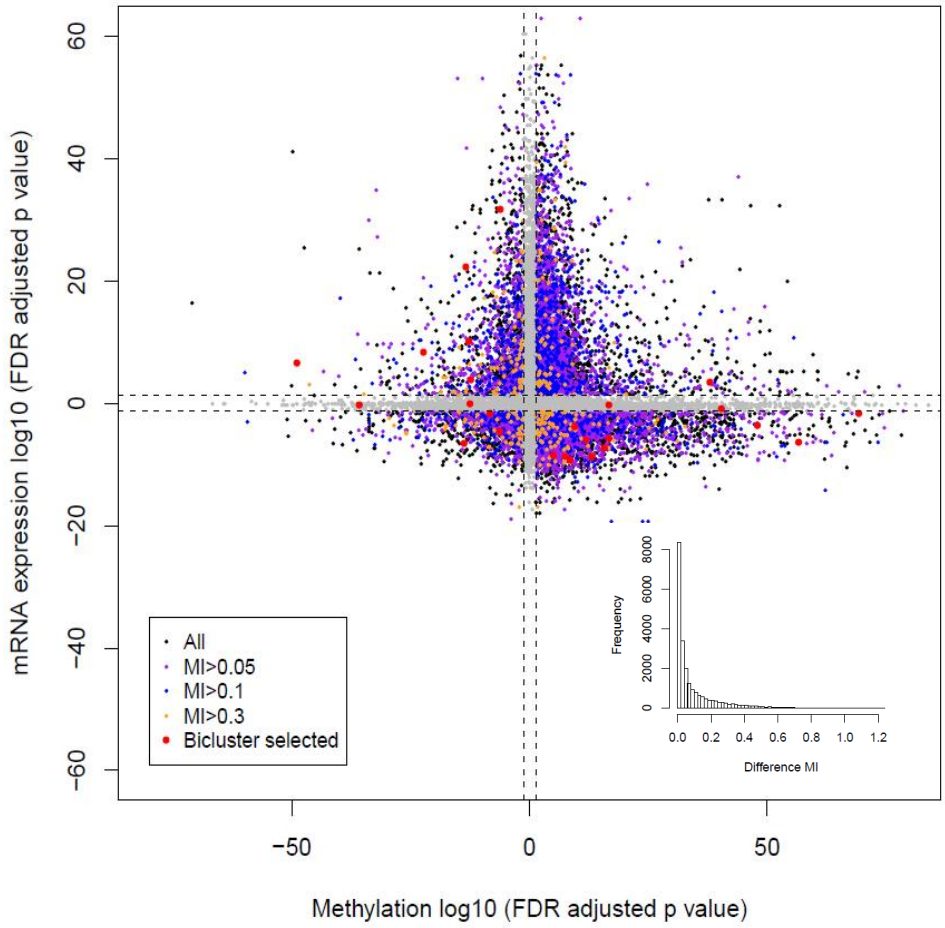


Figure 5.4 Comparison of transcriptome versus epigenetic differences between *BRCA* and normal samples. Starburst plot is shown for comparison of DNA methylation and mRNA expression data for 12,783 unique genes. Log10 (FDR-adjusted p-value) is plotted for DNA methylation (x-axis) and gene expression (y-axis) for each gene. If a mean DNA methylation β-value or mean gene expression value is higher (greater than zero) in the *BRCA* samples, −1 is multiplied to log10 (FDR-adjusted p-value), providing positive values. The dashed black lines indicate FDR-adjusted p-value at 0.05. Data points in grey indicate no significance in the comparison. Data points in purple indicate mutual information scores larger than 0.05; points larger than 0.1 are denoted by blue, and points larger than 0.3 are denoted by yellow. Points in red indicate genes identified to be differentially methylated when comparing *BRCA* with normal samples by biclustering. A histogram of the marginal mutual information score before and after data splitting is shown (bottom right).

5.2.2 Identification of "L-shaped" Genes

Aiming to investigate genes for which DNA methylation plays a large regulatory role, we identified "L-shaped" genes according to the following three criteria on MI scores and a plotting

pattern: (1) the range of mutual information scores is not less than 0.3; (2) samples are split into 4 quadrants by the thresholds of mRNA expression and DNA methylation, and at least 200 samples are located in each of the first and fourth quadrants; and (3) no fewer than 600 samples are in the first and the fourth quadrants. As shown in Figure 5.1, both o*ESR1* and *HOXA9* are selected as "L-shaped" genes, which satisfy our criteria. With our criteria, a total of 128 "L-shaped" genes out of 12,783 genes were selected and expected to be selected with high specificity. Table S5.2 shows the details of the genes sets, including the gene names, thresholds, MI scores and depths. Among these 128 "L-shaped" genes, 17 are transcription factors (chi squared test, p-value=0. 27), including 7 homeobox genes (chi squared test, p-value=2.02e-06) *CDX1, HNF1A, HNF1B, HOXA9, PAX8, POU3F3* and *POU4F1*. Table S5.3 shows the top curated gene sets, GO gene sets and oncogene signature enriched sets, which is from MsigDB gene enrichment analysis. From the results, we found that these 128 enriched genes were associated with multiple cancer types. And 10 genes were reported to be hypermethylated in lung cancer samples (p-value=6.53e-05). The top results of IPA associated network functions and biofunctions are shown in Table S5.4, which indicates that these 128 "L-shaped" genes are tightly associated with cancer, cellular disorder and disease development.

5.2.3 A New CIMP

5.2.3.1 CIMP Identification

In this study, we focus on CIMP identification for *BRCA*. The values of DNA methylation status were binary coded based on the estimated methylation threshold. Aiming to identify the most differentially methylated genes in the tumor samples compared with normal samples, we performed biclustering and identified 25 out of 128 genes by discriminating a block of 1 in breast cancer samples and 0 in normal samples (Figure 5.5; Table 5.2). Most of the selected genes had significant differences in DNA methylation status and mRNA expression levels between cancer and normal samples (Figure 5.4). These included *CFI, HOXA9, HSPB2, COL17A1, AQP1, POU3F3, PLD5, IL1A, POU4F1, CRYAB, LAMB3, TRIM29, SLC10A4, SCTR, MEP1A, IL20RA, SLC44A4, TFF1,*

C1orf64, *C10orf81*, *ZG16B*, *SPDEF*, *RERG*, *PTK6*, and *BNIPL*. Among these 25 "L-shaped" genes, 4 were transcription factors (chi squared test, p-value=0. 53), including 3 homeodomain proteins (chi squared test, p-value=6.916e-06), HOXA9, POU3F3 and POU4F1. Table S5.5 shows the top curated gene sets, GO gene sets and oncogene signature enriched sets, which indicate that these 25 genes are associated intensively with breast cancer. The top results of IPA of associated network functions and biofunctions are shown in Table S5.6, which indicates that these 25 genes are tightly associated with cancer, cellular disorder and disease development, as expected, and are also associated with the cell cycle, cellular movement and cell death.

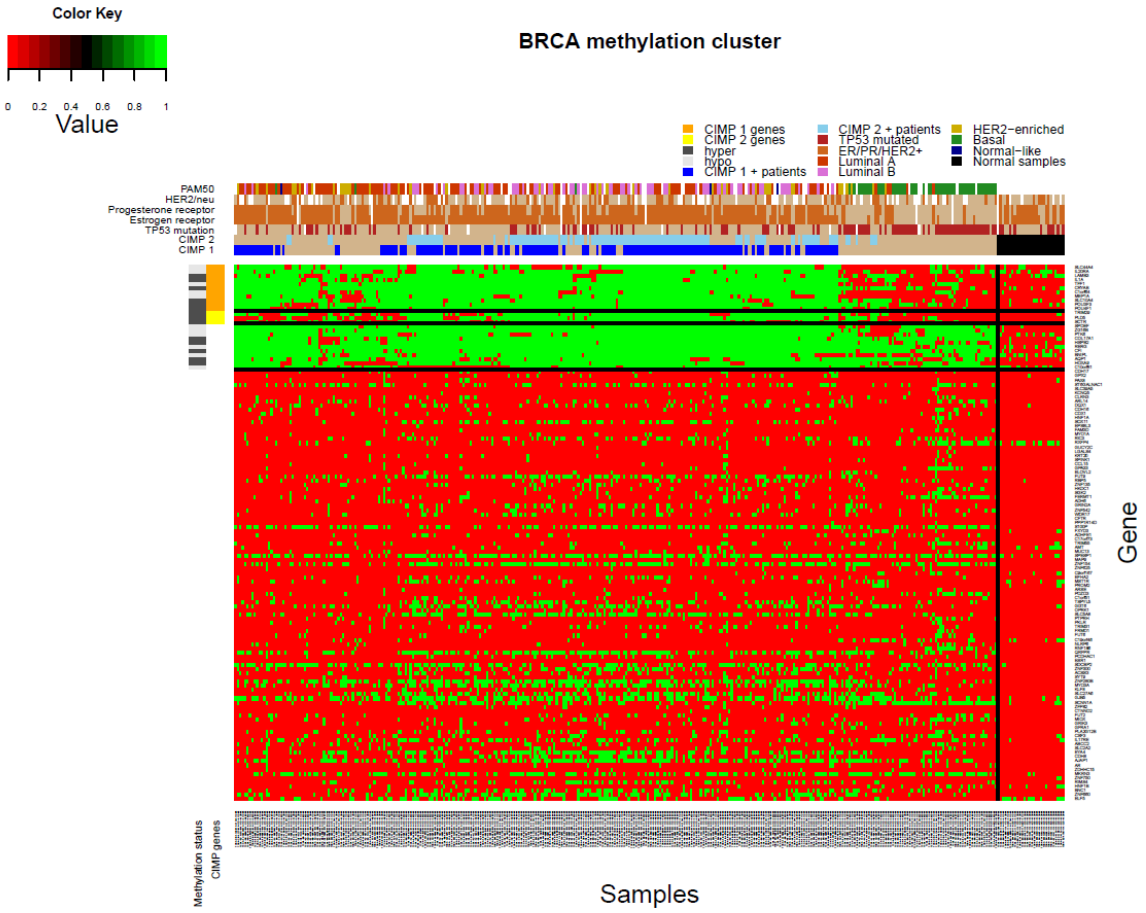


Figure 5.5 Coordinated analysis of breast cancer CIMP defined from dichotomized methylation status. CIMPs are identified by biclustering and a sequential supervised hierarchical clustering on the 128 "L-shaped" genes. The green and red heat map displays the sample and gene consensus. For each breast cancer sample, genes with unchanged methylation status compared with that of the normal samples are denoted in green; and genes with changed relative methylation status are denoted in red. The vertical and bottom horizontal black lines indicate the boundary of the bi-clusters. The other two black lines indicate empirically identified CIMP markers. The methylation status of CIMP

biomarkers is shown in the left panel; associations with molecular and clinical features are shown in the top panel.

Table 5.2 Demographics of CIMP subtypes in this study.

		CIMP1 +	CIMP1 -	Normal	P-value
TP53	Mutated	31	81	10	2.38e-11
	Wild	130	61	17	
ER	+	164	76	21	2.20e-16
	-	6	65	6	
PR	+	137	66	19	1.00e-09
	-	33	75	8	
HER2	+	37	34	7	0.660
	-	95	74	16	
PAM50 subtypes	Luminal A	68	43	0	0.209
	Luminal B	57	11	0	9.068e-06
	HER2-enriched	13	28	0	0.012
	Basal	1	49	0	2.586e-11
	Normal-like	2	1	0	1

A supervised hierarchical clustering analysis of the binary DNA methylation data was performed on the 25 genes selected from biclustering, and 3 clusters were identified (Figure 5.5). Cluster 1 contains 11 genes (*SLC44A4*, *IL20RA*, *LAMB3*, *IL1A*, *TFF1*, *CRYAB*, *C1orf64*, *MEP1A*, *SLC10A4*, *POU3F3*, and *POU4F1*) while cluster 2 contains 3 genes (*TRIM29*, *PLD5*, and *SCTR*) and cluster 3 contains the remaining 11 genes. Based on clusters 1 and 2, we defined two CIMPs. We allowed for 10% measurement error tolerance, which means that for each patient in the CIMP+ group, at least 10 out of 11 genes as biomarkers had methylation status that changed compared with that of the normal group. Among all 286 *BRCA* tumor samples, we defined 149 CIMP 1+ and 136 CIMP 2+ samples. As there were more CIMP1 markers than CIMP2, we focused on CIMP 1 in this study. Most of the identified CIMP 1 markers have the most significant difference in DNA methylation and mRNA expression levels between CIMP + and CIMP – samples and are significantly down- or up-regulated and hyper- or hypomethylated (Figure 5.6, lower-right and upper-left quadrants). A histogram of the methylation frequency distribution for the set of CIMP1 biomarkers shown in Figure 5.6 indicates that the distribution of the CIMP1 markers creates a good

bimodal distribution, with two methylation frequency peaks at 0.2 and 0.9. Compared with the normal samples, 5 (MEP1A, IL20RA, SLC44A4, TFF1, and C1orf64) out of 11 were hypomethylated and 6 (POU3F3, IL1A, POU4F1, CRYAB, LAMB3, and SLC10A4) were hypermethylated (Figure 5.5).

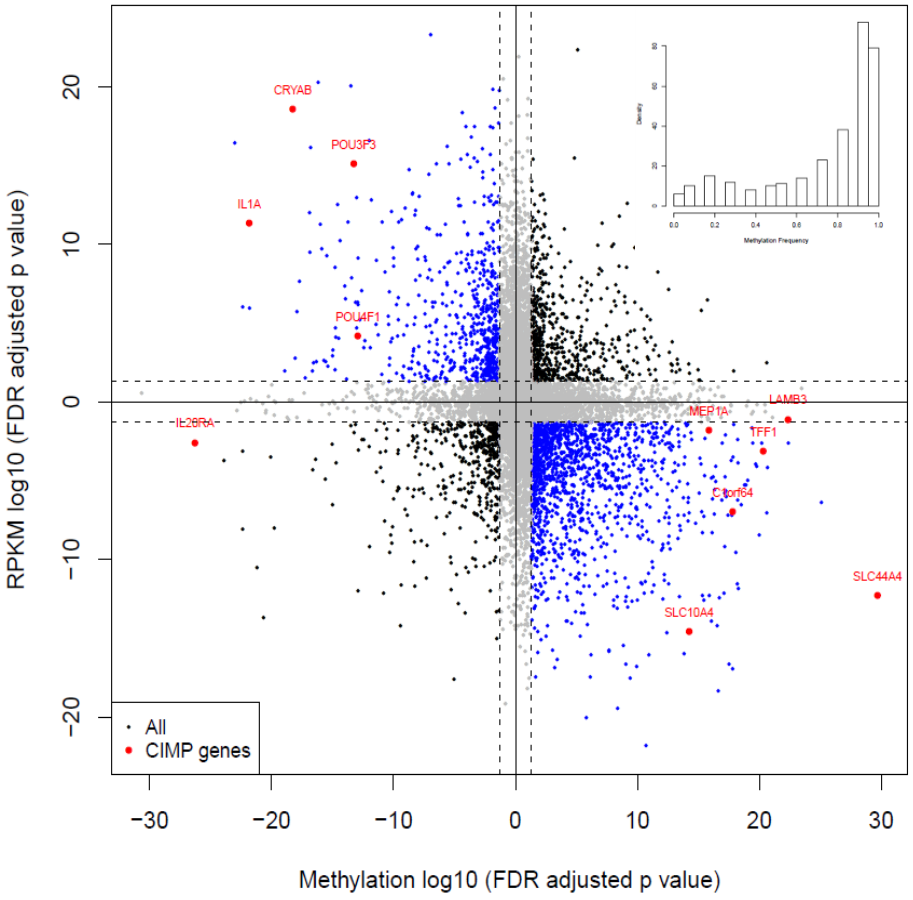


Figure 5.6 Comparison of transcriptome versus epigenetic differences between *BRCA* CIMP+ and CIMP- samples. Starburst plot is shown for comparison of DNA methylation and mRNA expression data for 12,783 unique genes. The x-axis and y-axis are defined in the same way as in Figure 5.4, as well as the black lines and grey data points. Points in red indicate CIMP markers identified; points in blue indicate significant up- and down-regulation in the gene expression levels and significant hyper- or hypomethylation in *BRCA* CIMP+ tumors compared to CIMP- tumors. A histogram of the methylation frequency of CIMP markers is shown (top right).

5.2.3.2 CIMP Markers Enrichment Analysis

Among 11 CIMP markers, POU3F3 and POU4F1 are transcription factors (chi squared test, p-value=0. 93) and both are homeodomain proteins (chi squared test, p-value=0. 01). From GSEA of MsigDB, all significant enriched GO terms and oncogenic signatures are displayed in Table S5.7. Our 11 CIMP markers are enriched in the gene sets with the following characteristics: (1) differentially expressed between carcinoma and normal cells and between luminal-like breast cancer cell lines and the basal-like cell lines (*LAMB3, TRIM29, IL20RA, CRYAB* and *IL1A*); (2) part of the validated nuclear estrogen receptor alpha network (*TFF1* and *POU4F1*); (3) discriminate between ESR1+ and ESR1 tumors (*TRIM29, TFF1* and *SLC44A4*); (4) respond to bystander irradiation (*LAMB3* and *IL1A*); (5) targets of polycomb gene EED, SUZ12 and BMI1 (*IL20RA, SLC44A4, SCTR, SLC10A4, IL1A* and *LAMB3*); (6) down-regulated in metastases from malignant melanoma compared to the primary tumors (*LAMB3* and *TRIM29*); and (7) related to anti-apoptosis and negative regulation of development (*CRYAB* and *IL1A*). Only two significantly associated network functions were found by IPA. Interestingly, all 11 markers were shown to be associated with *TP53* directly or indirectly. Seven of them are involved in the network "cancer, hematological disease, immunological disease" and the other three are involved in "amino acid metabolism, cellular compromise, cellular movement" (Figure 5.7).

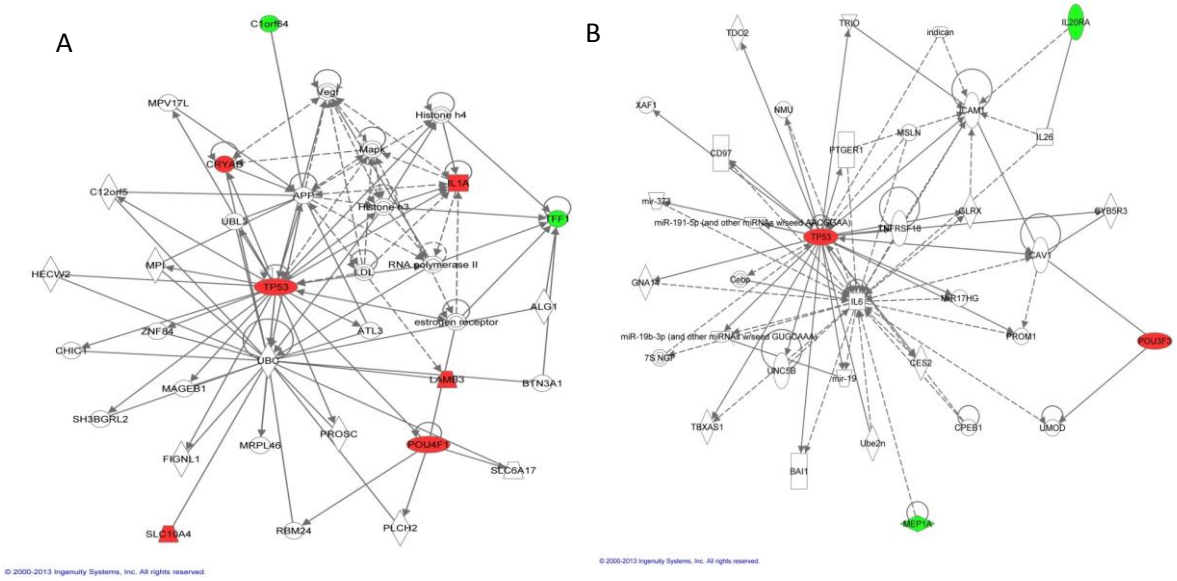


Figure 5.7 CIMP markers involved in networks. A. Network "cancer, hematological disease, immunological disease." B. Network "amino acid metabolism, cellular compromise, cellular movement." Genes in color (except *TP53*) indicate CIMP markers; red indicates hypermethylation in CIMP+ samples and green indicates hypomethylation. *TP53* in red indicates the wild-type *TP53* gene and overexpression of *TP53* as well.

5.2.3.3 Correlation with *TP53* Mutation and Subtypes

We performed logistic regression analysis to investigate the correlation of our defined CIMP subtype with the classic breast cancer markers estrogen receptor, progesterone receptor, HER2 receptor, status of *TP53* mutation and TCGA defined CIMP PAM50 subtypes (Figure 5.5) [65]. Consistent with the strong correlation of CIMP and *BRCA* basal subtypes, our defined CIMP was found to strongly correlate with *TP53* mutation, estrogen receptor status, and progesterone receptor status, but not with HER2 receptor status. No association was observed between CIMP and *BRCA*1 or *BRCA*2 mutations (data not shown). We further investigated the correlation to *BRCA* subtypes and found a very strong correlation between CIMP1 and basal and luminal B subtypes. We observed a mild correlation with the HER2-enriched subtype. Next, we performed genome-wide correlation analysis, and found that *TP53* was the only one gene that has significant correlation with the CIMP in terms of its mutation status (Figure 5.8). We also checked the correlation of CIMP+ with mRNA

expression of *TP53* and other *TP53*-related genes, including two *TP53* binding proteins (TP53BP1 and TP53BP2), four TP53-induced proteins (TP53I11, TP53I3, TP53INP1, TP53INP2), two *TP53* target genes (*TP53TG1* and *TP53TG5*) and one *TP53* regulating kinase (TP53RK). Interestingly, moderately significant correlations were found between CIMP and mRNA expression of *TP53, TP53BP1, TP53I3* and *TP53TG5*, and strong correlations existed with *TP53BP2, TP3I11, TP53INP1* and *TP53TG1* (Table 5.3). Also, we found significantly negative associations between *TP53* mutation and expressions of *TP53BP1* (p-value= 1.29e-07), *TP53I11* (p-value= 1.28e-05), *TP53INP1* (p-value= 2.07e-12), *TP53TG1* (p-value= 2. 93e-09) and *TP53TG5* (p-value= 6.10e-04) .We found the expression of *TP53BP2* to be significantly and positively correlated with *TP53* mutation (p-value= 1.64e-06).

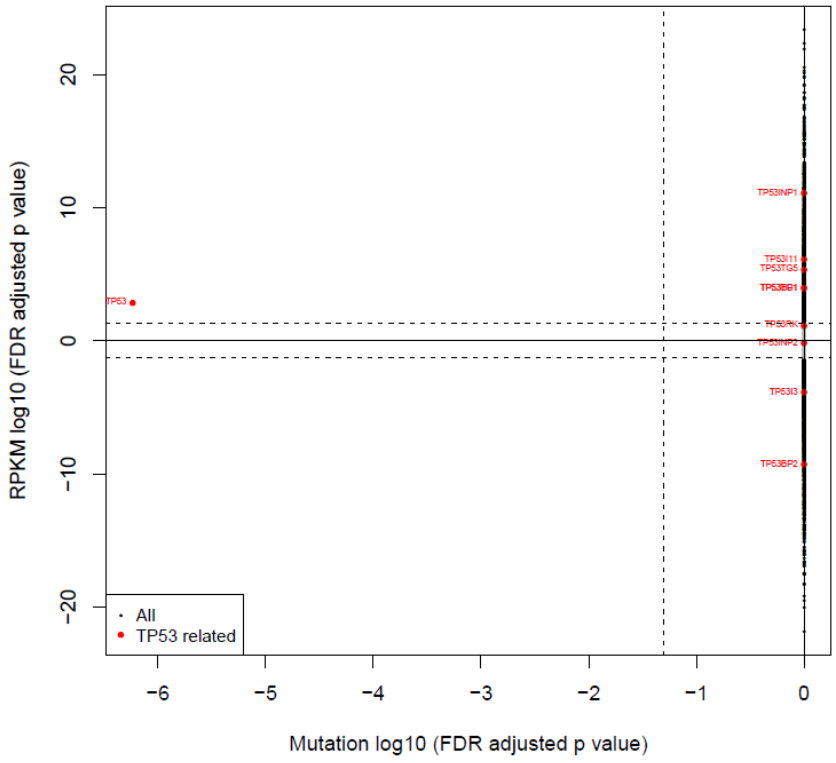


Figure 5.8 Comparison of transcriptome versus mutation differences between *BRCA* CIMP+ and CIMP- samples. Starburst plot is shown for comparison of mutation status and mRNA expression data for 12,783 unique genes. The x-axis and y-axis are defined in the same way as in Figure 5.4, as well as the black lines and grey data points. Data points in red indicate *TP53* genes and several *TP53*-related genes.

Table 5.3 Association study of CIMP subtype with molecular and clinical features.

Object	Term	Odd ratio	Lower boundary	Higher Boundary	P-value
estrogen receptor	positive	23.39	9.67	56.60	2.73e-12
progesterone receptor	positive	4.78	2.87	7.97	1.78e-09
HER2 receptor	positive	0.80	0.45	1.41	0.44
TP53	mutation	0.18	0.11	0.30	5.87e-11
TP53	mRNA expression	1.32	1.13	1.54	4.48e-04
TP53BP1	mRNA expression	1.25	1.12	1.39	2.91e-05
TP53BP2	mRNA expression	0.62	0.54	0.71	2.46e-11
TP53I11	mRNA expression	1.40	1.24	1.59	1.05e-07
TP53I3	mRNA expression	0.75	0.66	0.86	2.78e-05
TP53INP1	mRNA expression	1.87	1.59	2.19	2.38-13
TP53INP2	mRNA expression	0.96	0.82	1.11	0.55
TP53RK	mRNA expression	1.08	1.00	1.18	0.04
TP53TG1	mRNA expression	1.38	1.19	1.61	2.68e-05
TP53TG5	mRNA expression	1.51	1.30	1.77	6.81e-07
pretreatment history	Yes	Inf	1.23	Inf	0.017
prior diagnosis	Yes	13.59	1.77	104.66	0.012
ajcc cancer metastasis stage	Higher stage	3.02	0.61	14.98	0.18
ajcc neoplasm disease stage	Higher stage	1.29	0.82	2.037	0.27
ajcc tumor stage	Higher stage	1.19	0.77	1.86	0.43
gender	Male	Inf	0.35	Inf	0.25
age at initial diagnosis	Older	55.62	2.96	1044.40	0.0076

5.2.3.4 Clinical Correlation Analysis

Clinical correlation analysis was performed on 7 clinical features, including pretreatment history, prior diagnosis, American Joint Committee on Cancer (AJCC) cancer metastasis stage, AJCC neoplasm disease stage, AJCC tumor stage, gender and age at initial pathologic diagnosis (Table 5.3). The contingency table is shown in Table 5.4. Significant correlation was found between CIMP and age at initial pathologic diagnosis (p-value 0.0076), the mean age for patients with CIMP+ tumors was 62 years, while it was 55 for patients with CIMP- tumors (Figure 5.9). Stratified by age at initial pathologic diagnosis, prior diagnosis showed significant correlation with CIMP (p-value 0.012). Because of the small sample sizes of patients in particular categories, we applied Fisher's exact test on the pretreatment history and gender, and found significant correlations between CIMP and pretreatment (p-value 0.017).

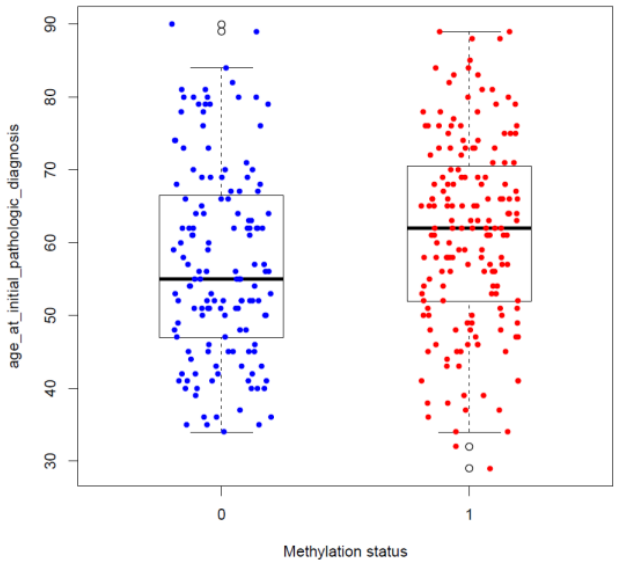


Figure 5.9 Association study between age and CIMP subtypes. Category 0 indicates CIMP- subtype and 1 indicates CIMP+ subtype.

Table 5.4 Demographics of CIMP subtype according to clinical features

		Methylation status	
		0	1
Pretreatment history	NO	143	164
	Yes	0	7
prior diagnosis	No	142	154
	Yes	1	17
ajcc cancer metastasis stage	M0	140	151
	M1	2	7
ajcc neoplasm disease stage	0	0	1
	I	27	30
	II	90	94
	III	19	32
	IV	2	7
ajcc tumor stage	T1	37	46
	T2	93	93
	T3	11	17
	T4	2	14
gender	Male	0	3
	Female	143	168

5.2.3.5 Survival Analysis

Stratified by age at initial pathologic diagnosis, relative risk of methylation status was estimated using a Cox model on survival data. Compared to patients with CIMP- tumors, those with CIMP+ tumors tend to have longer survival times (odds ratio=1.29); although this difference is not significant (p-value 0.457). We also investigated the survival analyses of each CIMP marker based on curated datasets using GENESURV tools in bioprofiling.de. We found that 5 (SLC44A4,

IL20RA, TFF1, C1orf64, and POU4F1) of 6 markers that show significant survival difference for patients with breast cancer agreed with our finding that patients with tumors in which the methylation status change have longer survival time. The last gene, which was the only gene that did not agree with our finding, showed the least significant correlation, with a p-value of 0.0257 (Table S5.8).

Clustering of correlations. We performed hierarchical clustering analysis on p-values from the correlation analysis of each biomarker (Fig 5.10). All CIMP markers are clustered in one block with 9 features, including *TP53* mutation status, *TP53BP2* mRNA expression, age at initial diagnosis, *TP53BP1* mRNA expression, *TP53I11* mRNA expression, *TP53TG5* mRNA expression, *TP53TG5* mRNA expression, *TP53INP1* mRNA expression, estrogen receptor status, and progesterone receptor status. This result coincided with the above correlation analysis, indicating a strong relationship between our identified CIMP and several factors, including *TP53* mutation, estrogen receptor status, progesterone receptor status, age at initial pathologic diagnosis, and the expression of *TP53*-related genes such as *TP53BP1, TP53BP2, TP53I11, TP53TG1, TP53TG5, TP53INP1* and *TP53RK*.

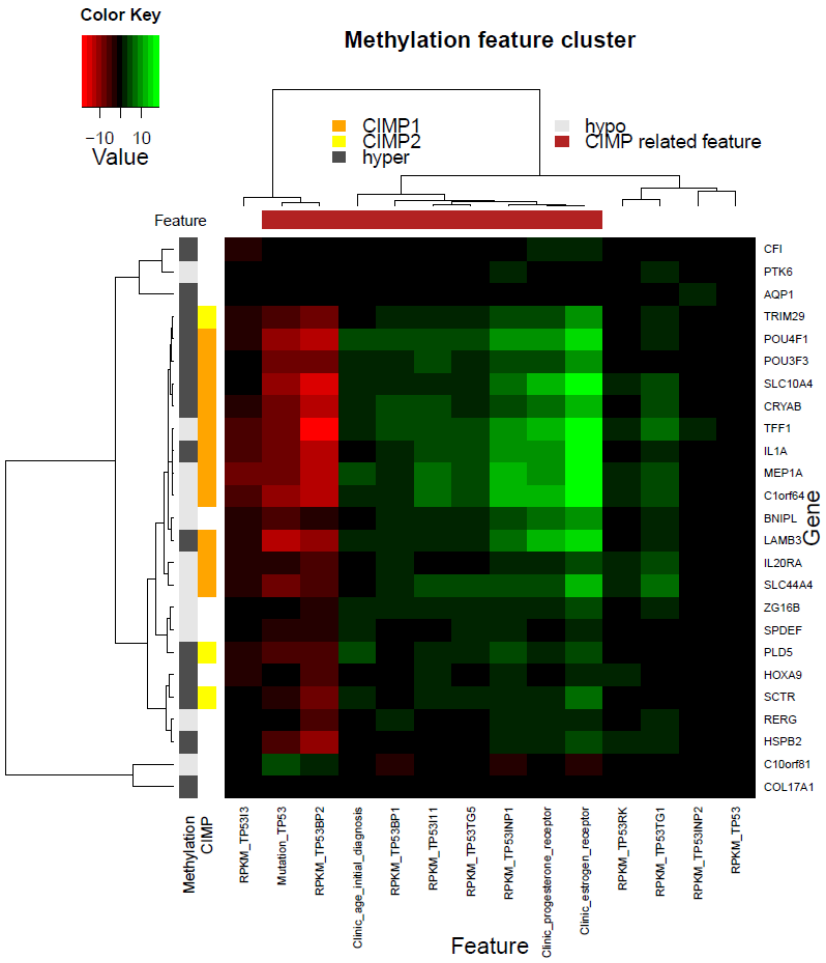


Figure 5.10 Consensus clustering analysis of association study for each marker. Unsupervised hierarchical clustering was performed on 25 genes identified by biclustering. The green and red heat map displays association significance and gene consensus. Red indicates a positive correlation between methylation status and molecular or clinical features and blue indicates a negative correlation. The methylation status of CIMP biomarkers is shown in the left panel, and CIMP identified in Figure 5.5 is shown in the top panel.

5.2.3.6 Epigenetic Modifiers

GSEA has shown that 11 CIMP+ markers were enriched in polycomb target gene sets. We investigated the mRNA expression level of 12 epigenetic modifiers from ploycomb repressive complex 2 (PRC2) (*BMI1*), PRC1 (*EED*, *SUZ12* and *EZH2*), DNA methyltransferases (DNMT1, DNMT3A and DNMT3B), H3K4 histone methyltransferase (MLL), isocitrate dehydrogenases (IDH1 and IDH2) and tet methylcytosine dioxygenases (TET1 and TET2). The mRNA expression levels were compared among CIMP+ samples, CIMP- samples and normal samples pairwise, and

their correlations with *TP53* mutations were calculated as well (Table 5.5). Our results show that (1) BMI1, IDH1 and TET1 have significant differential expression between CIMP+ samples and both of the other two groups; (2) DNMTA, DNMT3B, EZH1 and IDH2 have significant differential expression between CIMP- samples and both of the other two groups; and (3) all 7 genes are significantly correlated with *TP53* mutation (Table 5.5; Figure 5.11). Unfortunately, the mRNA expression value of SUZ12 was missing. Unexpectedly, EED does not shown significant differential expression between the CIMP+ subtype and the other two groups because of large variance (Figure 5.10 D). These findings indicate that (1) many epigenetic modifiers are tightly associated with *TP53*; (2) eleven CIMP+ markers may be regulated by *BMI1*, *IDH1* and *TET1*; (3) the expression of a large proportion of these 12 epigenetic modifiers belongs to two distinct patterns; and (4) at least 2 types of epigenetic modifiers that are functional in the *TP53* system exist.

Table 5.5 Association study of epigenetic modifiers with CIMP subtype and TP53 mutation

		CIMP+ vs CIMP-		CIMP+ vs Normal		CIMP- vs Normal		TP53 mutation	
		OR	P-value	OR	P-value	OR	P-value	OR	P-value
PRC1	BMI1	1.28	2.19E-04	1.36	1.40E-02	1.06	6.30E-01	0.75	7.00E-05
PRC2	EED	0.85	1.41E-03	0.91	3.28E-01	1.07	4.53E-01	1.21	3.84E-04
	SUZ12	NA	NA	NA	NA	NA	NA	NA	NA
	EZH2	0.76	4.49E-04	3.11	1.61E-16	4.09	1.85E-06	1.82	2.88E-14
DNA methyltransferases	DNMT1	0.91	8.65E-02	1.64	1.06E-07	1.80	7.62E-07	1.28	8.88E-06
	DNMT3A	0.82	1.36E-04	1.47	2.49E-05	1.79	5.95E-08	1.27	1.12E-05
	DNMT3B	0.69	8.19E-05	1.66	1.44E-03	2.41	9.40E-06	2.16	7.14E-15
histone	MLL	0.98	7.42E-01	0.46	1.71E-14	0.46	2.02E-12	0.93	1.77E-01
isocitrate dehydrogenase	IDH1	0.87	4.32E-02	0.70	4.78E-03	0.80	9.59E-02	1.21	8.22E-03
	IDH2	0.73	2.55E-05	2.07	3.15E-07	2.85	4.21E-13	1.62	1.36E-10
Tet methylcytosine dioxygenase	TET1	0.71	4.03E-05	0.57	1.54E-05	0.80	2.05E-02	1.57	1.40E-07
	TET2	1.01	8.79E-01	0.70	5.00E-04	0.70	1.12E-02	0.95	3.57E-01

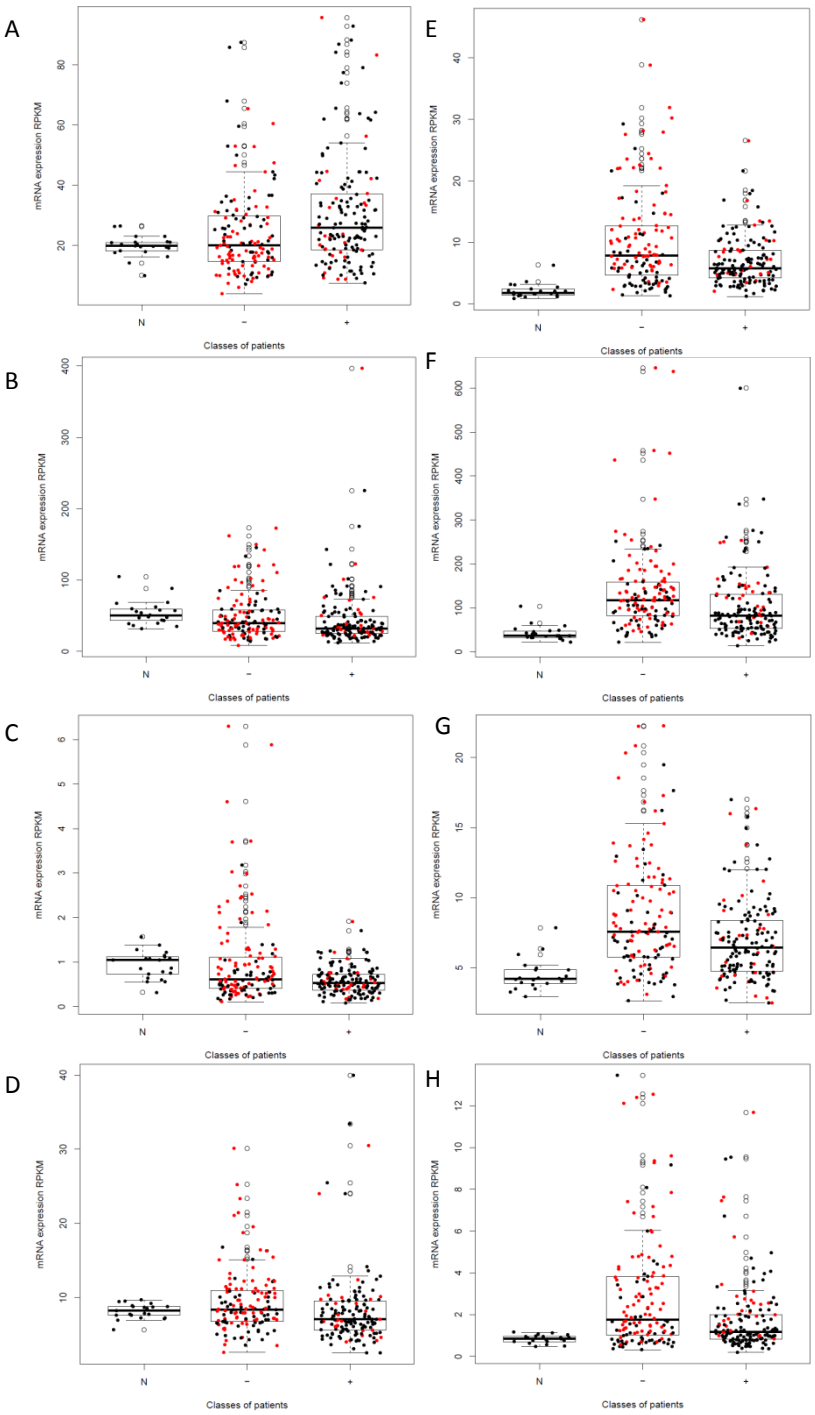


Figure 5.11 Boxplot of expression of epigenetic modifiers in *BRCA* CIMP subtype and normal samples. (A-D) BMI1, IDH1, TET1 and EED. (E-H) EZH2, IDH2, DNMT3A and DNMT3B. Category N indicates normal samples; - indicates CIMP- subtype; + indicates CIMP+ subtype. Data points in red indicate samples with *TP53* mutation.

5.3 Discussion

In this study, we proposed a method to discriminate genes for which the mRNA expression levels are affected by the DNA methylation level, and to determine the threshold of DNA methylation as an indicator of whether one gene is expressed or depressed. Based on the estimated dichotomized methylation status, we identified a new CIMP of *BRCA* with 11 markers. Interestingly, the new CIMP+ samples we identified were correlated significantly with negative *TP53* mutation status, estrogen receptor-positive status, progesterone receptor-positive status, higher age at initial pathologic diagnosis, pretreatment history and a possibly longer survival time. The 11 CIMP markers were shown to be associated with *TP53* directly or indirectly, and enriched in cancer and diseases networks. Also, we found that 7 among 12 epigenetic genes were correlated strongly with both the new CIMP and *TP53* mutation. Based on our findings, we proposed a model in which there are at least two groups of members in the *TP53* regulatory network, which are named "guidance" and "sustainer."

Method for determining methylation threshold. In this paper, we estimated the methylation threshold based on the calculation of the conditional mutual information score. DNA methylation is considered to suppress gene expression and is often described as a binary measurement (hypermethylation or hypomethylation) [83]. Intuitively, the genes with expression regulated significantly by methylation will exhibit an "L" or reverse "L" shape when DNA methylation values are plotted against mRNA expression values (and the data split by the threshold cutoffs should be independent). *ESR1* and *HOXA9* are two examples of "L-shaped" genes, which also exhibit dramatic changes in methylation patterns in different cancers. This finding is consistent with that of Qiu et al. [83]. *ESR1* encodes an estrogen receptor that has well-known involvement in pathological processes of breast cancer and has been shown to have hypermethylation status [84]. Also, *HOXA9* genes have been characterized under epigenetic silencing in tumors [85]. Our method was validated by the observation that most of the 29-gene set from the colon cancer study has a large difference in the values of the mutual information score measured before and after determining the methylation

threshold [82]. Compared with the method that naively determines the methylation threshold by searching for the turning point in the mRNA expression [86], our method is more precise and flexible and has two advantages. First, we determined the threshold through measuring the dependence between DNA methylation and mRNA expression data; therefore, our method used more information than the naïve method that was based on only the change in mRNA expression. Second, we estimated the smeared densities of both DNA methylation and mRNA expression for each patient, which allows for a much smoother and more precise calculation.

For all genes, the estimated thresholds tend to be small for methylation but large for mRNA expression. This observation is consistent with previous findings that most genes in the genome are hypomethylated [87,88]. However, we found that for genes with MI score differences larger than 0.3, the thresholds of methylation corresponding to gene expression regulation are gene-specific, which is concordant with our assumption. Also, it is reasonable as we found small thresholds for mRNA expression levels for the gene sets enriched with "L-shaped"genes.

Identification of "L-shaped" genes. "L-shaped" genes were selected according to three criteria (see Methods) from two aspects. First, "L-shaped" genes should have large differences in the mutual information score as methylation value and expression value would be much more independent after being split by threshold cutoffs. Second, for each "L-shaped" gene, both samples with hypomethylation and high expression levels and samples with hypermethylation and low expression levels should occupy considerable portions.

A total of 128 "L-shaped" genes were identified. These genes are tightly associated with cancer and disease and are enriched in curated gene sets associated with multiple cancer types. These results are reasonable because these "L-shaped" genes were selected from TCGA data of multiple cancer types. As expected, 10 hyper-methylated genes are significantly related to lung cancer, which supports our intention to investigate the relationship between DNA methylation, mRNA expression and cancer. Interestingly, 7 of the 128 "L-shaped" genes were homeobox genes (p-value=2.02e-06),

which is consistent with reports that homeobox genes are associated with tumorigenesis and are epigenetically regulated by the polycomb complex [89,90].

Identification of a new CIMP. We identified 25 markers differentially methylated in tumor samples compared with normal samples. In this study, we focused on one CIMP with 11 markers, including SLC44A4, IL20RA, LAMB3, IL1A, TFF1, CRYAB, C1orf64, MEP1A, SLC10A4, POU3F3, and POU4F1. By integrating gene expression and DNA methylation data, we found that selected markers had rather significant DNA hypermethylation and down-regulated gene expression or hypomethylation and up-regulated gene expression when comparing either tumor to normal samples or CIMP+ to CIMP- samples. Also, using the strategy of integrating gene expression and DNA methylation data, Noushmehr et al. identified 300 genes with significant hypermethylation and gene expression changes for a CIMP of glioma [47]. However, those genes may not be "L-shaped."

Most of our new CIMP markers have been reported to have significant associations between their methylation status and various cancers. IL20RA has been reported to be hypermethylated in lung cancer cell lines [91]. TFF1 has been validated as hypomethylated and overexpressed in breast carcinoma [48,92]. Mahapatra et al. identified hypermethylated POU3F3 as a biomarker for systemic progression of prostate cancer [93]. POU4F1 has been demonstrated to have lower methylation frequency in leukemia cell lines compared to primary acute lymphoblastic leukemia samples, but higher methylation levels in low-grade breast cancer compared to normal samples [49,94]. Methylation of the CRYAB gene promoter was reported to occur in distinct anaplastic thyroid carcinomas [95]. The frequent up-regulation of LAMB3 by promoter demethylation has been reported in breast cancer, gastric cancer and bladder cancer, but the opposite has been reported in prostate cancer [53,96,97,98].

Among the 11 CIMP biomarkers we identified, we found 5 (MEP1A, IL20RA, SLC44A4, TFF1, and C1orf64) to be hypermethylated and 6 (POU3F3, IL1A, POU4F1, CRYAB, LAMB3, and SLC10A4) to be hypomethylated. Although most CIMP samples that have been identified have been accompanied by hypermethylation markers, which were named CIMP-high samples (widespread

promoter methylation) in some studies, the identification of CIMP-low samples (with less widespread promoter methylation) is ongoing [99,100,101,102]. Also, a previous breast cancer study revealed that hypomethylation at many CpG islands was significantly associated with epithelia to mesenchymal transition [103]. In this study, we proposed a new CIMP — "CIMP-M" ("CIMP-mixture") because of the mixture of epigenetic markers with hypo- and hyper methylation status. Although only a few studies have claimed CIMP with hypermethylated genes for breast cancer [63,64,65], consistent with a study by Fang et al. and TCGA group, this study suggests clearly that CIMP exists with a bimodal distribution and is significantly associated with certain molecular markers and clinical features.

Consistent with the conclusion of Tommasi et al. that the methylation of homeobox genes frequently occurs in breast cancer [104], the new CIMP biomarkers are enriched in homeodomain proteins. Also, they were found to be enriched in gene sets corresponding to metastasis, anti-apoptosis and negative regulation of development. And they were enriched in the *TP53* and estrogen receptor alpha signal pathway networks. In addition, they were significantly involved in networks associated with cancer and disease, amino acid metabolism, cellular compromise and cellular movement. These finding indicate that the new CIMP biomarkers play important roles in cancer.

In this study, we demonstrated that the new CIMP identified was significantly correlated with *TP53* mutation, estrogen receptor status and progesterone receptor status. Excitingly, we found that CIMP+ was significantly associated with the luminal B breast cancer subtype, and CIMP- was significantly associated with the basal breast cancer subtype. These findings were also exhibited by GSEA of 11 CIMP biomarkers that were enriched in the *TP53* network, ER1 network, gene sets for anti-apoptosis and genes discriminating between the luminal-like and basal-like breast cancer cell lines. Ronneberg et al. reported 3 clusters associated with luminal tumors and basal-like tumors and found them to be significantly different in association with estrogen receptor status, and *TP53* mutation [62]. However, only 1 gene, *TFF1*, among our 11 CIMP biomarkers overlapped with the

findings from the study by Ronneberg et al. A possible reason for this discrepancy is poor correlations between methylation status and mRNA expression levels for many genes because of the complexity of the regulatory system. For this newly identified CIMP, we observed only a mild association with the HER2-enriched subtype and no significant associations with *BRCA1* or *BRCA2* mutations. Associations with the HER2-enriched subtype and *BRCA1* and *BRCA2* mutations and methylation clusters were found by Holm et al. [61].

TP53 was the only gene for which mutation was found to be associated with CIMP. And consistent with previous studies [62,105], we found a strong correlation between *TP53* mutation and ER negativity. Several studies have been developed to investigate the relationship of *TP53* and ER expression. Estrogen receptor α has been reported to bind to *TP53* and inhibit its transcription, resulting in the inhibition of p53-mediated cell death. Therefore, benefits from a reactivated *TP53* pathway, good prognosis and treatment response from anti-estrogens such as tamoxifen have been reported [106,107,108]. In our study, we observed an older age at initial pathologic diagnosis, prior diagnosis, pretreatment history and longer survival time in patients with the CIMP+ subtype with wild-type *TP53*, which indicates that patients with the CIMP+ subtype are at risk later in life and are more susceptible to treatment. Shirley et al. stated that p53 regulates ER expression through transcriptional control of the ER promoter [105], which may be the reason most *TP53* mutations were found in patients with ER- tumors. Taken together, we observe that a feedback loop exists between *TP53* and ER expression; however, this complicated regulatory system remains elusive. We also investigated 9 *TP53*-related genes and found 8 of them (*TP53, TP53BP1, TP53BP2, TP53I3, TP3I11, TP53INP1, TP53TG1* and *TP53TG5*) to be significantly associated with CIMP subtypes and 5 of them (*TP53BP1, TP53I11, TP53INP1, TP53TG1* and *TP53TG5*) to be significantly negatively associated with *TP53* mutation. We found a positive association with *TP53* mutation for one gene, *TP53BP2*. It is known that TP53BP1 and TP53BP2 bind to *TP53* and enhance *TP53*-mediated transcriptional activation. The association we found between their expression and *TP53* indicated that feedback relationships might exist between *TP53* and its binding proteins. And the opposite

associations of TP53BP1 and TP53BP2 with *TP53* mutation indicate that different feedback mechanisms are involved. In summary, significant associations were found between both *TP53* mutation and CIMP subtypes and *TP53* binding proteins, induced proteins and target proteins. Whether these *TP53*-related genes are involved in maintaining the specific methylation pattern of CIMP will be of interest in future research.

Regulation through epigenetic modifiers. In this study, we investigated the expression of 12 epigenetic modifiers in samples of CIMP+ and CIMP- subtypes and normal samples. We found that aberrant expressions of *BMI1, IDH1* and *TET1* exist when comparing the CIMP+ subtype to the other two groups, and *DNMT3A, DNMT3B, EZH2* and *IDH2* were expressed differently in the CIMP- subtype compared to the other two groups. All 7 genes were significantly correlated with *TP53* mutation. These findings are in accordance with the findings of Pietersen et al., that *EZH2* and *BMI1* are inversely correlated with *TP53* mutation and prognosis in breast cancer. They claimed that tumors with high expression levels of *BMI1* are associated with a good prognosis [109]. In contrast, Guo et al. claimed that *BMI1* promotes invasion and metastasis [110]. In our study, we observed that higher expression of *BMI1* correlated well with ER+/PR+ cancer subtypes, CIMP+ subtype and wild-type *TP53* and with older age at diagnosis, good response to treatment and longer survival time. *TET1* has been reported to suppress the invasion ability of breast tumors through demethylation [111]. *EZH2* has been found commonly overexpressed in breast cancer and has been reported to repress DNA repair, leading to tumor progression [76,112]. A significant association has been found for *DNMT3B* with breast cancer subtypes discriminated by methylation profiling [63]. All these findings suggest that the 11 CIMP biomarkers we identified might be regulated by *BMI1, IDH1* and *TET1*, and that there is another group of epigenetic modifiers functional in the CIMP- subtype specifically that includes *DNMT3A, DNMT3B, EZH2* and *IDH2*. Other studies have reported that EZH2 is functional in stem cell maintenance [113] and basal-like tumors are more like stem cells compared with other subtypes [114,115]. As expected, 6 of 11 CIMP biomarkers are curated targets of polycomb genes *EED, SUZ12* and *BMI1*. Squazzo et al. found that a common set of promoters

occupied by *SUZ12* exist in *MCF3* tumor cells and embryonic tumors [116]. However, we cannot determine whether that gene is differently expressed between the CIMP+ and CIMP- groups because of missing data.

Model of TP53-mediated regulatory system. In this study, biomarkers exhibited a different methylation status in the CIMP+ group compared with the CIMP- and normal groups. We observed that patients with the CIMP+ subtype were older at the time of the initial pathologic diagnosis, and had a longer survival time compared with patients in the CIMP- subtype. It seems a paradox that the methylation status of the CIMP- subtype was similar to that of the normal samples (Figure 5.5). Patients with the CIMP+ subtype were older at the clinical onset of the disease, had a better treatment response, and had the potential to survive longer. Also, the CIMP- subtype was found to be tightly related to *TP53* mutation. As *TP53* is well known to be associated with response to genotoxic and non-genotoxic stresses [117,118], we infer that aberrant CIMP+ methylation status results from the response of *TP53* to cellular stresses from a tumor.

Here, we infer that there is an in emergency rescue system in which a group of genes play an important role along with *TP53*. We call this group of genes "guidance" genes because they react when the cell is at risk and contain the 11 CIMP+ markers we found. We call another group of genes that functional in the *TP53* system the "skeleton" genes. "Sustainer" genes will not function when *TP53* mutation occurs. We suggest a model of the *TP53* regulatory network that contains two components, the "guidance" and "sustainer" genes, as shown in Figure 5.12. "Guidance" genes play an emergency rescue role in the response to stresses, while "sustainer" genes act as the maintenance department, taking responsibility for the operation of the *TP53*-mediated regulatory system. Interestingly, as the "guidance" genes, the 11 CIMP biomarkers were found to be enriched in response to bystander irradiation. For the CIMP+ group, enrichment of the estrogen and progesterone receptors might be the stress signal from breast cancer that induces the response of the wild-type *TP53* gene acting in a "guidance" capacity. For the CIMP- group, the *TP53*-mediated regulatory

system was silenced by the mutation of *TP53*, and the "sustainer" genes lost their function. We predicated 1594 "guidance" genes and 2016 "sustainer" genes in terms of the mRNA expression levels and 1591 "guidance" genes and 557 "sustainer" genes in term of the DNA methylation status. Both "guidance" and "sustainer" genes were enriched in important processes of disease resistance and metabolism. However, the "guidance" genes were enriched in the immune system process for rescue action, while the "sustainer" genes were found to be enriched in gene sets responding to changes in the cytoskeleton. We predicated "guidance" and "sustainer" genes according to our criteria (see Method I). We identified 1594 "guidance" and 2106 "sustainer" genes in terms of the mRNA expression level and 1591 "guidance" and 557 "sustainers" in terms of the DNA methylation status. The 11 CIMP markers were significantly selected as "guidance" genes (Figure 5.13). IPA showed that "guidance" genes were enriched in cell signaling, molecular transport, nucleic acid metabolism, and development disorders, while "sustainer" genes were enriched in the cell cycle, cell assembly and organization, cellular growth and proliferation and tumor morphology. MsigDB GSEA showed that "guidance" and "sustainer" genes were very significantly enriched in chemical reactions, response to stress, signal transduction, transport, transcription and cellular metabolic processes. As expected, GSEA showed that "guidance" genes were enriched in the immune system process for rescue action. And "sustainer" genes were found to be enriched in gene sets responding to changes in the cytoskeleton, which is consistent with the findings from IPA that "sustainer" genes play a role in cell assembly and organization and are responsible for tumor morphology. In addition, "guidance" genes are found to be enriched as the targets of *EED* and *SUZ12*. Interestingly, the *LET-7* family was found to regulate "guidance" genes and the *MIR-506*, *MIR-30*, and *MIR-17* families and *MIR-124A* were found to regulate both "guidance" and "sustainer" genes.

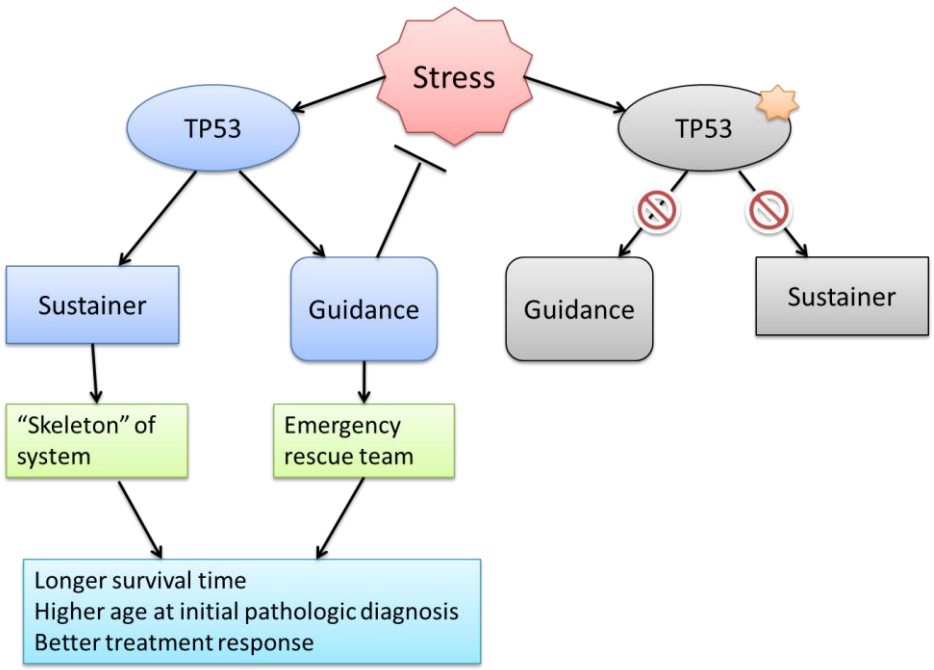


Figure 5.12 Predicated model of *TP53*-mediated regulatory system. We proposed a model of the *TP53*-mediated regulatory system with two components: "guidance" and "sustainer" genes. "Sustainer" genes serve as the "skeleton" of the system and "guidance" genes serve as an emergency rescue team. When cells are under stress, "guidance" genes will respond to resist the stress. However, when *TP53* is mutated, "sustainer" genes will lose their function and "guidance" genes will not work under stress.

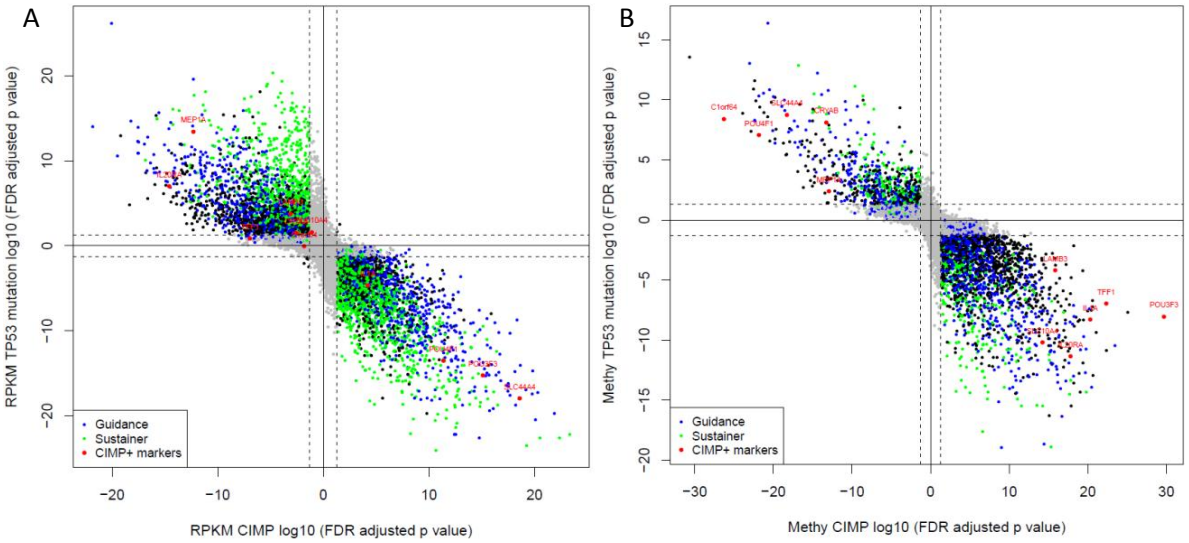


Figure 5.13 Comparison of transcriptome, epigenetic and mutation differences between *BRCA* CIMP subtypes. A. Comparison of transcriptome and mutation differences. B. Comparison of epigenetic and mutation differences. For both A and B, the x-axis and y-axis are defined in the same way as in Figure 5.4, as are the black lines and grey data points. Data points in red indicate the identified CIMP markers. Data points in blue denote predicated "guidance" genes and green denote "sustainer" genes.

We also observed two distinct patterns of the expression of epigenetic modifiers, which were in concordance. Therefore, we inferred that *TP53* acts through at least two epigenetic mediation systems, one for "guidance" genes, containing *EED, SUZ12* and *BMI1*, and another for "sustainer" genes, containing *DNMT3A, DNMT3B, EZH2* and *IDH2*. Therefore, for patients with the CIMP- subtype, *DNMT3A, DNMT3B, EZH2* and *IDH2* might be potential targets of treatment for restoring the *TP53* regulatory system, and *EED, SUZ12* and *BMI1* might be targets for resisting tumor development. The *EZH2* inhibitor DZNep has been reported to induce robust apoptosis in breast cancer cells [119]. Nevertheless, more research is needed to validate this *TP53*-mediated regulatory model and predicate the "guidance" and "sustainer" genes.

In summary, we have presented a method based on the calculation of mutual information scores to determine the threshold of DNA methylation and distinguish the genes for which the levels of expression are significantly regulated by DNA methylation. We have identified a CpG island phenotype (CIMP) of breast cancer with strong correlation with the wild-type *TP53* mutation, ER+/PR+ subtypes, higher age at the time of disease diagnosis, better treatment response and possibly a longer survival time. Both hypermethylation and hypomethylation status were shown on biomarkers of the CIMP+ group. In addition, we have suggested a model of the *TP53*-mediated regulatory system in which *TP53* might regulate "guidance" and "sustainer" genes through two epigenetic mediation systems.

CONCLUSIONS

To investigate the properties of overdispersion in RNA-seq data, we empirically calculated the variance from replicated experiments. We observed a dependency relationship between overdispersion and sequencing depth on both the gene and position levels, which is consistent with the intuition that increments in sequencing depth will improve the sequencing accuracy. Based on this property, we developed a function for estimating the overdispersion rate by borrowing information from all genes. The Poisson distribution is usually applied to model discrete counts, but it is not appropriate for modeling the RNA-seq data as the Poisson rate on each position fluctuates in a large range because of the uniformity of mapped reads. Therefore, compared with the Poisson model, a model based on the proportion of the sequencing counts between two samples has the advantage of avoiding the estimation of the fluctuating measurement by converting non-uniform measurements into a constant ratio for one gene. We adopted the beta-binomial distribution to model the ratio of two measurements and the overdispersion. In the first study, we developed a statistical model on the gene level for differential expression analysis by utilizing the property of overdispersion. Our model obtained a better performance than models that did not incorporate the overdispersion rate or which used a constant overdispersion rate. Next, aiming to more accurately model the measurement, we modeled on the position level with a specific dispersion rate for each position. We also investigated the influence of random hexamer priming on overdispersion and found that the use of a random hexamer primer influenced the overdispersion mainly by affecting the sequencing depth. Therefore, it is desirable to estimate overdispersion utilizing its dependency upon sequencing depth. And, consistent with our model on the gene level, our model on the position level was superior to the models that did not incorporate the overdispersion rate or which used a constant overdispersion rate. Compared with DESeq which is a widely used method based on a negative binomial model on the gene level, our model is technically more desirable because it avoids modeling the highly fluctuating Poisson rate.

Inspired by the cross-hybridization issue inherent in microarrays, we investigated the measurement of spike-in transcripts sequenced along with different sample transcripts. Interestingly, we observed that the measurement of spike-in transcripts was influenced by the sample transcripts. Although the precise mechanism is still unknown, we developed an efficient statistical method to correct for this bias introduced by sample transcripts, and obtained an increase of 0.1 in the Pearson correlation coefficient after correction. The new type of bias observed in this study will aid the understanding of sequencing technology and contribute to better accuracy in downstream analysis.

Aiming to identify "L-shaped" genes that were largely regulated by DNA methylation on transcription, we developed a statistical approach for determining the DNA methylation threshold on the inhibition of mRNA expression. We used a mutual information technique to determine the threshold from 997 samples across 7 cancer types in TCGA datasets, and identified a total of 128 "L-shaped" genes. We performed biclustering and hierarchical clustering on *BRCA* samples and identified a new CIMP with 11 biomarkers. We found a strong correlation between this new CIMP+ subtype in breast cancer and TP53 mutation, ER+ status, PR+ status, basal subtype, higher age at initial diagnosis, prior treatment history and the possibility of a longer survival time. We also found a strong correlation between 7 epigenetic genes and both of the new CIMP subtypes and *TP53* mutation. Based on our findings in this study, we predicated a model of the *TP53*-mediated regulatory network with two components: "guidance" genes that serve in an emergency capacity and "sustainer" genes that serve in a supporting capacity.

We developed new methods for RNA-seq data analysis for differential expression analysis, bias correction and integration of mRNA expression levels and DNA methylation status. We believe that the new methods will be useful in furthering the accuracy of differential expression analyses, in understanding the biases and spurious effects inherent in sequencing technology, and in CIMP identification and biomarker discovery related to DNA methylation.

Next, we will extend our method to handle biological samples in differential expression analysis. For the new bias identified in RNA-seq technology, which is introduced by using different

sample transcripts, we will explore the underlying mechanism and develop a more powerful method for correcting this bias. For the method to identify “L-shaped” genes, we plan to apply our method to other cancer types. Also, we will work on integrating more datasets, with the aim of obtaining a comprehensive analysis.

APPENDIX

Appendix 1: Supplementary Figures

Figure S3.1 The variance estimated on any position in 10 equal categories according to the distance of that position from the last nucleotide of the gene. (a) Lichun spike-in dataset. (b) Bullard dataset. Lichun reads start to appear at 76 nt away from the end of the gene because only mate2 on the antisense strand were investigated and the last sequencing reads were mapped to 76 nt before the ending. However, the Bullard reads start from 50 nt away from the end of the gene because we truncated the genes by 50 nt from the gene head and tail separately.

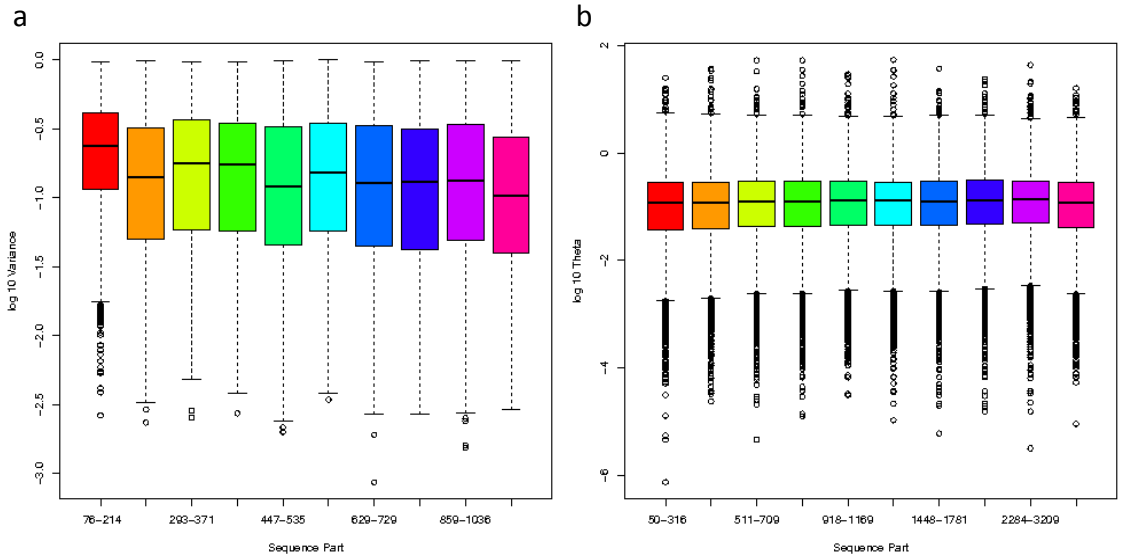
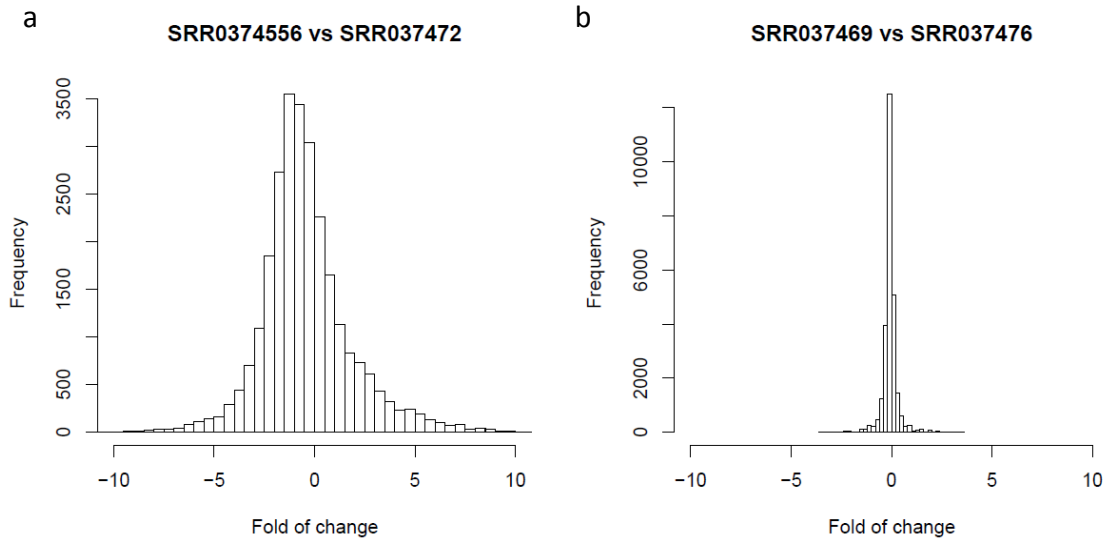


Figure S3.2 Histogram comparing the sequencing reads from two samples. The fold change for each gene was calculated as the log scale of the ratio of the reads counts of two samples. (a) SRR037456 is a brain sample and SRR037472 is a UHR sample. (b) SRR03469 and SRR037476 were two UHR samples with different library preparations. Apparently, the variance between two different samples is much larger than that between library preparations.



Appendix 2: Supplementary Tables

Table S5.1 The mutual information scores calculated on 25 genes selected from the study by Yi et al [82]. The column headings are defined as follows: "Gene" is the gene symbol, "Probe" is the probe id from Illumina Infinium Human DNA Methylation27 array, MI_max is the maximum mutual information score, MI_min_methy is the minimum mutual information score on data split by the threshold on methylation, MI_diff_methy is the difference between the maximum and minimum mutual information scores on methylation.Threshold_methy_1way is the threshold predicated methylation by the 1-way mutual information method. MI_min_exp is the minimum mutual information score on data split by the threshold on expression, MI_diff_exp is the difference between the maximum and minimum mutual information scores on expression.Threshold_exp_1way is the threshold predicated on expression by the 1-way mutual information method, MI_min_12 is the minimum mutual information score calculated by the 2-way mutual information method, and MI_diff_12 is the difference between the maximum and minimum mutual information calculated by the 2-way mutual information method.

Gene	Probe	MI_max	MI_min_methy	MI_diff_methy	Threshold_methy_1way	MI_min_exp	MI_diff_exp	Threshold_exp_1way	MI_min_12	MI_diff_12
APC2	cg18133957	0.918801	0.714133	0.204668	0.64	0.841674	0.077128	-3.11018	0.718529	0.200272
CD109	cg23442323	0.317559	0.170615	0.146944	0.15	0.185321	0.132239	0.903185	0.29325	0.024309
CHD5	cg00282347	0.672821	0.43322	0.239601	0.15	0.604483	0.068338	-2.39189	0.532149	0.140672
EVL	cg17813891	0.242147	0.200957	0.04119	0.77	0.212001	0.030146	3.887943	0.207849	0.034298
EYA4	cg01805282	0.698954	0.364727	0.334227	0.56	0.648932	0.050022	-1.73657	0.356336	0.342618
EYA4	cg07327468	0.499959	0.258202	0.241757	0.3	0.407461	0.092498	-1.64256	0.327041	0.172918
FBN2	cg25084878	0.203767	0.109825	0.093943	0.44	0.161119	0.042649	0.983988	0.112375	0.091393
FLNC	cg25664034	0.180049	0.121759	0.058291	0.22	0.155523	0.024526	-1.1728	0.11843	0.061619
GPNMB	cg17274742	0.239358	0.209117	0.030241	0.27	0.222848	0.01651	4.917683	0.215993	0.023365
GUCY1A2	cg23984434	0.433081	0.304618	0.128464	0.52	0.364053	0.069029	-1.45504	0.285112	0.14797
HAPLN1	cg09893305	0.485798	0.361431	0.124367	0.67	0.451865	0.033933	-2.10028	0.363029	0.122769
ICAM5	cg11373429	0.330354	0.229248	0.101107	0.29	0.277045	0.05331	-2.49657	0.22655	0.103805
IGFBP3	cg22083798	0.270092	0.114852	0.15524	0.19	0.129001	0.141091	4.244468	0.162303	0.107789
LAMA1	cg07846220	0.344199	0.14664	0.197558	0.48	0.169132	0.175067	0.448955	0.116187	0.228012
MMP2	cg12317456	0.23921	0.195411	0.0438	0.18	0.202031	0.037179	3.006276	0.201968	0.037242
NRCAM	cg17885062	0.467722	0.31834	0.149382	0.25	0.366087	0.101635	-0.87651	0.309591	0.158131
NTNG1	cg02361557	0.370772	0.289216	0.081555	0.1	0.319383	0.051389	-2.71286	0.279319	0.091453
PPM1E	cg19141563	0.771768	0.454593	0.317175	0.11	0.571661	0.200108	-3.55218	0.445726	0.326043
PRKD1	cg21794225	0.354138	0.239484	0.114654	0.34	0.269377	0.084761	0.505021	0.269189	0.084949
RET	cg05621401	0.355159	0.222573	0.132586	0.39	0.307011	0.048148	-1.59387	0.203522	0.151637
SH3TC1	cg07816074	0.314486	0.136964	0.177522	0.54	0.167381	0.147105	1.117468	0.114844	0.199642
STARD8	cg20832009	0.214374	0.149339	0.065034	0.09	0.166904	0.04747	1.090363	0.192478	0.021896
SYNE1	cg26620959	0.384867	0.245108	0.139759	0.37	0.263686	0.121181	1.276008	0.212871	0.171995
TCERG1L	cg10175795	0.78407	0.516304	0.267766	0.29	0.64614	0.13793	-3.2857	0.491611	0.292459
ZNF569	cg03884783	0.378155	0.200037	0.178118	0.14	0.228017	0.150139	-1.15158	0.198254	0.179902

Table S5.2 The mutual information scores calculated on128 selected "L-shaped" genes. The column headings are defined as follows: "Gene" is the gene symbol, "Probe" is the probe id from Illumina Infinium Human DNA Methylation27 array, MI_max is the maximum mutual information score, MI_min_methy is the minimum mutual information score on data split by the threshold on methylation, MI_diff_methy is the difference between the maximum and minimum mutual information scores on methylation.Threshold_methy_1way is the threshold predicated methylation by the 1-way mutual information method. MI_min_exp is the minimum mutual information score on data split by the threshold on expression, MI_diff_exp is the difference between the maximum and minimum mutual information scores on expression.Threshold_exp_1way is the threshold predicated on expression by the 1-way mutual information method, MI_min_12 is the minimum mutual information score calculated by the 2-way mutual information method, and MI_diff_12 is the difference between the maximum and minimum mutual information calculated by the 2-way mutual information method.

gene	probe	MI_max	MI_min_methy	MI_diff_methy	threshold_methy_1way	MI_min_exp	MI_diff_exp	threshold_exp_1way	MI_min_12	MI_diff_12
ELF5	cg01473816	0.735515	0.365517	0.369998	0.65	0.403495	0.332021	-0.56323	0.433501	0.302014
ZNF660	cg22598028	0.810154	0.529791	0.280363	0.56	0.645475	0.164679	-3.13433	0.507875	0.302279
BNC1	cg10398682	0.7974	0.530258	0.267142	0.34	0.694086	0.103314	-3.43322	0.494801	0.3026
AQP1	cg04551925	0.488725	0.216458	0.272267	0.65	0.226185	0.26254	4.347634	0.186021	0.302704
HNF1B	cg12788467	0.706421	0.242299	0.464122	0.21	0.283626	0.422795	1.065114	0.401953	0.304468
SCTR	cg15250797	0.899921	0.592828	0.307093	0.34	0.766125	0.133797	-1.54074	0.594811	0.305111
RIMS4	cg19332710	0.967803	0.675715	0.292088	0.09	0.831937	0.135866	-4.23222	0.662516	0.305287
ZNF750	cg27285599	1.081205	0.554931	0.526274	0.71	0.584408	0.496797	-1.27621	0.769256	0.311949
MKRN3	cg23234999	0.569254	0.233338	0.335915	0.69	0.225356	0.343898	-0.60685	0.256656	0.312598
ZDHHC15	cg11272332	0.458103	0.162324	0.295779	0.61	0.175179	0.282924	-1.63846	0.144598	0.313506
AR	cg07780118	1.410697	1.139401	0.271296	0.55	1.185463	0.225234	-2.3349	1.096967	0.31373
AJAP1	cg17525406	0.749099	0.441601	0.307498	0.24	0.573167	0.175931	-1.48321	0.434429	0.31467
COL17A1	cg13448625	0.688208	0.377329	0.310878	0.73	0.408947	0.279261	2.316326	0.373186	0.315021
CDH8	cg27444994	0.785918	0.512933	0.272985	0.21	0.616816	0.169102	-2.44332	0.468805	0.317113
EYA4	cg01401376	0.628242	0.320893	0.307349	0.19	0.532458	0.095784	-4.83861	0.310776	0.317466
SLC2A2	cg17142134	1.391408	0.929798	0.46161	0.77	0.961738	0.42967	-2.60575	1.073395	0.318013
BNIPL	cg11584936	0.377669	0.112681	0.264988	0.48	0.090151	0.287518	1.390571	0.059333	0.318337
ABCC2	cg17044311	0.615384	0.29788	0.317505	0.61	0.353671	0.261714	-0.77229	0.29691	0.318475
IL17RE	cg15095327	0.47712	0.19049	0.28663	0.41	0.212958	0.264162	0.697661	0.158556	0.318564
CSF3	cg21432842	0.921367	0.583928	0.337439	0.44	0.675398	0.245969	-1.53257	0.602621	0.318746
PLA2G12B	cg02044879	0.700094	0.391152	0.308942	0.67	0.461068	0.239025	-0.61664	0.381121	0.318972
GFRA1	cg23898073	0.812224	0.507971	0.304252	0.38	0.604805	0.207418	-1.24464	0.491249	0.320975
PTK6	cg21484834	0.495062	0.200641	0.294421	0.59	0.204655	0.290407	1.509966	0.172645	0.322417
GRIK3	cg06722633	0.800611	0.447245	0.353366	0.38	0.600492	0.200119	-0.159	0.477079	0.323532
SLC10A4	cg08209133	0.975948	0.7216	0.254348	0.24	0.829085	0.146864	-3.69717	0.651721	0.324227
MIOX	cg24867501	0.596044	0.269926	0.326118	0.53	0.305587	0.290457	-1.49346	0.27115	0.324894
FUT2	cg19025034	0.689034	0.421987	0.267047	0.64	0.412492	0.276541	1.776775	0.36339	0.325644

C1orf64	cg08887581	0.753526	0.45742	0.296106	0.55	0.510592	0.242934	-1.15974	0.427017	0.326509
CTNND2	cg25302419	0.946794	0.591773	0.355021	0.15	0.74347	0.203324	-2.34925	0.619508	0.327286
ZFP82	cg25886284	0.556009	0.189931	0.366078	0.44	0.230526	0.325483	-0.45555	0.226114	0.329895
TFF1	cg02643667	0.765564	0.390236	0.375328	0.7	0.412194	0.35337	1.537345	0.435639	0.329925
SCNN1A	cg18738906	0.812605	0.399356	0.413249	0.84	0.422281	0.390324	2.530125	0.480903	0.331702
RERG	cg19205533	0.535646	0.258539	0.277107	0.52	0.266642	0.269004	1.029177	0.203494	0.332152
GJB5	cg01333788	0.684365	0.366705	0.317661	0.59	0.438701	0.245664	-0.46991	0.351345	0.33302
SLC27A6	cg07103493	0.921344	0.626766	0.294578	0.19	0.720875	0.200468	-2.0485	0.587686	0.333657
KLF8	cg06655100	0.414627	0.128344	0.286283	0.47	0.13434	0.280287	-0.44184	0.079397	0.33523
HSPB2	cg12598198	0.528822	0.181982	0.34684	0.73	0.187582	0.34124	1.587704	0.192111	0.336712
SPDEF	cg07705908	0.4796	0.188741	0.29086	0.48	0.215167	0.264433	0.329094	0.142267	0.337334
MYO3A	cg23771603	0.744835	0.396946	0.347889	0.3	0.523339	0.221497	-2.12864	0.406577	0.338258
ZNF280B	cg16184943	0.740464	0.410234	0.33023	0.32	0.541364	0.1991	-2.0221	0.402119	0.338345
SYT9	cg08185661	0.771767	0.496032	0.275735	0.36	0.57061	0.201157	-2.83131	0.433329	0.338438
ACSS3	cg01283289	0.512383	0.196294	0.316089	0.32	0.242489	0.269895	-0.74502	0.171965	0.340418
ZNF300	cg19014419	0.674167	0.3365	0.337667	0.45	0.343022	0.331146	-1.06142	0.333666	0.340502
SDCBP2	cg16173067	0.793389	0.496307	0.297082	0.54	0.507857	0.285532	1.690591	0.451909	0.34148
ESR1	cg11251858	0.556531	0.237457	0.319074	0.23	0.276133	0.280398	-1.95442	0.214283	0.342248
PCDHAC1	cg12629325	1.245624	0.917178	0.328446	0.7	1.004481	0.241143	-2.72548	0.900754	0.34487
QRFPR	cg00015770	0.775014	0.455846	0.319168	0.26	0.546744	0.22827	-2.1042	0.429645	0.345368
RNF186	cg09195271	0.488165	0.175055	0.31311	0.49	0.226956	0.261209	0.434018	0.139986	0.34818
NLRP6	cg09205751	1.032516	0.739321	0.293196	0.58	0.837747	0.194769	-1.31783	0.683545	0.348971
TRIM29	cg13625403	0.698953	0.38551	0.313444	0.72	0.410654	0.288299	3.132584	0.349665	0.349289
C19orf46	cg18542098	0.561077	0.229331	0.331746	0.58	0.21359	0.347488	1.251329	0.209131	0.351946
FUT6	cg00579402	0.48988	0.19361	0.29627	0.55	0.200781	0.289098	1.244701	0.135357	0.354523
FRMD1	cg00350478	0.915153	0.57194	0.343213	0.46	0.764102	0.15105	-1.17138	0.559785	0.355368
TRIM31	cg00679556	1.196369	0.776016	0.420353	0.32	0.78661	0.409759	1.542483	0.838973	0.357396
PKLR	cg02280309	0.566715	0.229336	0.337379	0.54	0.272666	0.294049	-0.56814	0.203941	0.362775
LAMB3	cg01580568	0.673686	0.347255	0.326431	0.64	0.379195	0.294491	2.501904	0.309883	0.363803
PTPRH	cg11261264	0.71989	0.383755	0.336135	0.49	0.438244	0.281647	0.992225	0.355374	0.364516
SLC5A8	cg10141715	0.832332	0.497111	0.335221	0.45	0.602577	0.229755	-3.29278	0.466493	0.365839
OPRK1	cg25990647	0.957108	0.526838	0.43027	0.5	0.618263	0.338845	-2.99389	0.590702	0.366406
CRYAB	cg15227610	0.502165	0.162421	0.339744	0.49	0.152555	0.34961	4.876306	0.135562	0.366604
GGT6	cg04511534	0.628586	0.312934	0.315651	0.62	0.30334	0.325246	1.38697	0.261784	0.366801
TSPYL5	cg15747595	0.607299	0.260296	0.347003	0.62	0.287044	0.320255	0.134197	0.240189	0.36711
C1orf51	cg09563216	0.479246	0.12577	0.353476	0.42	0.150387	0.328859	-0.03753	0.107255	0.371991
PDZD3	cg09799714	0.99307	0.633584	0.359486	0.61	0.737601	0.255468	-1.66304	0.620425	0.372645
POU4F1	cg08097882	0.570379	0.201465	0.368913	0.28	0.435486	0.134893	-4.37757	0.195556	0.374823
SLC44A4	cg07363637	0.531739	0.200281	0.331459	0.52	0.212227	0.319512	3.354669	0.156722	0.375017
IL1A	cg00839584	1.219214	0.862595	0.356618	0.45	1.002302	0.216912	-2.47846	0.842711	0.376503
ARSE	cg11964613	0.867826	0.469613	0.398213	0.43	0.493215	0.374611	1.468561	0.48941	0.378415
PROM2	cg20775254	0.646546	0.227767	0.418779	0.6	0.251476	0.39507	1.43697	0.267849	0.378697
MST1R	cg08687163	0.799316	0.34714	0.452176	0.74	0.358414	0.440902	1.225448	0.419	0.380316

HOXA9	cg27009703	0.776843	0.435609	0.341234	0.59	0.448497	0.328346	0.966311	0.395496	0.381347
EFHA2	cg26831415	0.895025	0.546818	0.348207	0.41	0.618256	0.276769	-1.42114	0.51238	0.382645
C9orf167	cg07717632	0.464618	0.09617	0.368449	0.24	0.102045	0.362573	1.68074	0.081905	0.382714
ZNF625	cg17892556	0.51436	0.173167	0.341193	0.5	0.192065	0.322295	-1.0213	0.128129	0.38623
ZNF154	cg21790626	0.844424	0.502406	0.342019	0.55	0.537558	0.306867	-0.68034	0.45538	0.389044
IL20RA	cg22487322	0.527202	0.173394	0.353807	0.67	0.204838	0.322364	0.408141	0.136437	0.390765
ZG16B	cg26259865	0.699068	0.335499	0.363568	0.58	0.342143	0.356925	0.91889	0.305258	0.39381
MAP9	cg03616357	0.569406	0.215317	0.354089	0.5	0.228202	0.341203	0.202835	0.174758	0.394648
SPESP1	cg09886641	0.700947	0.30479	0.396157	0.7	0.373778	0.327169	0.327288	0.301781	0.399167
MUC13	cg09081544	0.60281	0.203562	0.399248	0.53	0.23206	0.37075	3.883348	0.199972	0.402839
C10orf81	cg10368842	0.56222	0.217218	0.345001	0.46	0.221419	0.340801	0.027331	0.159156	0.403064
AMT	cg20191453	0.529286	0.183579	0.345707	0.66	0.177	0.352287	1.208175	0.125002	0.404284
TRIM55	cg23322523	0.786631	0.400507	0.386124	0.53	0.472923	0.313708	-1.9437	0.380077	0.406554
C17orf73	cg03016571	0.538502	0.15689	0.381612	0.5	0.148801	0.389701	0.568318	0.128526	0.409976
ADHFE1	cg08090772	0.567394	0.196004	0.371391	0.46	0.22613	0.341264	-0.59164	0.155751	0.411644
FXYD3	cg02633817	0.525605	0.140403	0.385202	0.62	0.15792	0.367685	2.933435	0.109581	0.416024
S100P	cg22266967	0.555592	0.176513	0.379079	0.51	0.199093	0.356499	3.061998	0.139353	0.41624
PPP1R14D	cg04968426	0.562222	0.178981	0.383241	0.58	0.186838	0.375385	2.089626	0.144708	0.417514
CFTR	cg25509184	1.091757	0.682676	0.409081	0.3	0.753909	0.337847	0.894448	0.671309	0.420448
WDR17	cg18443378	0.917938	0.545883	0.372055	0.21	0.613637	0.304301	-2.11554	0.496885	0.421053
ZNF542	cg26309134	0.564676	0.200377	0.364298	0.44	0.22219	0.342485	0.185751	0.143361	0.421315
PLD5	cg12613383	1.105161	0.714951	0.39021	0.33	0.95884	0.146322	-3.91296	0.683786	0.421375
GRIN2A	cg01722994	0.865231	0.521814	0.343417	0.28	0.69796	0.167271	-3.80076	0.442957	0.422274
ADH6	cg06518271	0.763355	0.295122	0.468233	0.64	0.34885	0.414505	-1.51351	0.338687	0.424668
FERMT1	cg09539538	0.80119	0.39755	0.40364	0.42	0.393227	0.407963	1.802201	0.374811	0.42638
CFI	cg12243271	0.556621	0.192373	0.364248	0.59	0.190559	0.366061	2.738497	0.125797	0.430824
SGK2	cg17463527	0.683206	0.285248	0.397958	0.57	0.309921	0.373285	0.40685	0.24935	0.433856
POU3F3	cg20291049	0.83464	0.422907	0.411733	0.23	0.446055	0.388585	0.173539	0.39827	0.43637
HKDC1	cg11762346	0.792341	0.397796	0.394545	0.5	0.476037	0.316304	-0.82915	0.354975	0.437366
ZNF135	cg16638540	0.656561	0.266618	0.389943	0.61	0.282597	0.373965	-0.8111	0.214471	0.442091
RBP5	cg24441911	0.609581	0.193144	0.416438	0.67	0.204222	0.40536	2.562149	0.162078	0.447504
FUT9	cg01837719	1.514984	1.10021	0.414775	0.23	1.290737	0.224247	-4.87768	1.064773	0.450211
ELOVL2	cg13562911	1.006106	0.576278	0.429828	0.3	0.691245	0.314861	-1.94761	0.55532	0.450786
MEP1A	cg20980592	0.691537	0.309015	0.382523	0.65	0.509764	0.181773	0.561969	0.239937	0.4516
GPA33	cg06665322	0.683287	0.276509	0.406778	0.47	0.290014	0.393273	2.688244	0.229175	0.454112
CCL15	cg23743114	0.572053	0.135469	0.436584	0.46	0.170003	0.402049	-0.16002	0.11699	0.455063
SPINK1	cg04577715	0.571031	0.259503	0.311528	0.65	0.175335	0.395696	2.964529	0.112849	0.458182
KRT20	cg25124433	0.597884	0.172336	0.425548	0.73	0.179377	0.418507	1.989879	0.13845	0.459434
LGALS4	cg06394229	0.71857	0.301996	0.416574	0.45	0.339338	0.379232	3.4325	0.258952	0.459618
GUCY2C	cg18754342	0.639586	0.17	0.469586	0.5	0.193055	0.446531	1.373467	0.162648	0.476939
RXFP4	cg08403419	0.978926	0.572869	0.406057	0.32	0.657796	0.32113	-2.15265	0.494636	0.48429
RIC3	cg08383315	1.306606	0.949637	0.356969	0.34	1.082187	0.22442	-2.97625	0.82194	0.484666
MYO1A	cg09541248	1.331314	0.910194	0.421119	0.38	0.975095	0.356219	0.692855	0.842628	0.488686

FAM3D	cg02194211	0.703317	0.230127	0.47319	0.67	0.286637	0.41668	3.942016	0.202327	0.50099
EPS8L3	cg00491404	1.112369	0.662543	0.449826	0.51	0.692608	0.419762	2.764433	0.603092	0.509278
SOX11	cg08432727	1.557355	1.177898	0.379457	0.46	1.333619	0.223737	-2.77406	1.041207	0.516148
HNF1A	cg16175725	0.692391	0.152521	0.53987	0.63	0.172907	0.519485	0.840162	0.169552	0.522839
CDX1	cg15452204	0.852634	0.411413	0.441222	0.51	0.41054	0.442094	1.512887	0.328988	0.523647
CDH16	cg14221831	1.104623	0.658576	0.446046	0.62	0.684173	0.42045	1.071852	0.578087	0.526536
DQX1	cg02034222	0.784156	0.318513	0.465643	0.66	0.330438	0.453718	-0.31324	0.256062	0.528094
ARL14	cg24147596	0.999617	0.536725	0.462892	0.51	0.55707	0.442547	-0.89091	0.463837	0.535779
CLRN3	cg23817637	0.79768	0.346046	0.451633	0.47	0.366644	0.431036	1.132192	0.244632	0.553048
KCNQ5	cg15717808	1.032089	0.486842	0.545247	0.31	0.583496	0.448593	-2.60956	0.463127	0.568962
SLC39A5	cg00668685	0.961775	0.476375	0.4854	0.61	0.505763	0.456012	1.003672	0.391803	0.569973
ST6GALNAC1	cg13015534	1.16578	0.660012	0.505768	0.39	0.6851	0.48068	-1.12498	0.582409	0.583371
PAX8	cg07403255	0.80501	0.221437	0.583574	0.54	0.222126	0.582885	1.642063	0.193171	0.611839
GPX2	cg09643186	1.091144	0.529561	0.561583	0.64	0.561502	0.529642	3.109897	0.472359	0.618785
CDH17	cg12038710	0.958454	0.432057	0.526397	0.59	0.43411	0.524344	2.838026	0.314921	0.643533

Table S5.3 Gene set enrichment analysis (GSEA) on 128 identified "L-shaped" genes from MsigDB. Top 50 significant gene set enrichments, with p-value < 0.05, are shown in three categories: curated gene sets, GO gene set and oncogene signatures.

Curated gene sets					
Gene Set Name	**# Genes in Gene Set (K)**	**Description**	**# Genes in Overlap (k)**	**k/K**	**p value**
DOANE_BREAST_CANCER_ESR1_UP	112	Genes up-regulated in breast cancer samples positive for ESR1 [GeneID=2099] compared to the ESR1 negative tumors.	8	0.0714	2.11E-07
DODD_NASOPHARYNGEAL_CARCINOMA_UP	1821	Genes up-regulated in nasopharyngeal carcinoma (NPC) compared to the normal tissue.	30	0.0154	5.55E-07
SERVITJA_ISLET_HNF1A_TARGETS_DN	109	Genes down-regulated in pancreatic islets upon knockout of HNF1A [GeneID=6927].	7	0.0642	2.56E-06
TURASHVILI_BREAST_LOBULAR_CARCINOMA_VS_DUCTAL_NORMAL_DN	91	Genes down-regulated in lobular carcinoma vs normal ductal breast cells.	6	0.0659	1.18E-05
HUPER_BREAST_BASAL_VS_LUMINAL_UP	54	Genes up-regulated in basal mammary epithelial cells compared to the luminal ones.	5	0.0926	1.24E-05
JAEGER_METASTASIS_DN	258	Genes down-regulated in metastases from malignant melanoma compared to the primary tumors.	9	0.0349	1.42E-05
TURASHVILI_BREAST_DUCTAL_CARCINOMA_VS_DUCTAL_NORMAL_DN	198	Genes down-regulated in ductal carcinoma vs normal ductal breast cells.	9	0.0404	1.51E-05
ONDER_CDH1_TARGETS_2_DN	464	Genes down-regulated in HMLE cells (immortalized nontransformed mammary epithelium) after E-cadhedrin (CDH1) [GeneID=999] knockdown by RNAi.	11	0.0237	5.65E-05
CERVERA_SDHB_TARGETS_1_DN	38	Genes turned off in Hep3B cells (hepatocellular carcinoma, HCC) upon knockdown of SDHB [GeneID=6390] by RNAi.	4	0.1053	5.79E-05
HATADA_METHYLATED_IN_LUNG_CANCER_UP	390	Genes with hypermethylated DNA in lung cancer samples.	10	0.0256	6.53E-05
LEE_LIVER_CANCER_SURVIVAL_UP	185	Genes highly expressed in hepatocellular carcinoma with good survival.	7	0.0378	7.95E-05
GOZGIT_ESR1_TARGETS_DN	781	Genes down-regulated in TMX2-28 cells (breast cancer) which do not express ESR1 [GeneID=2099]) compared to the parental MCF7 cells which do.	15	0.0179	1.12E-04
SMID_BREAST_CANCER_RELAPSE_IN_BRAIN_DN	85	Genes down-regulated in brain relapse of breast cancer.	5	0.0588	1.12E-04
VECCHI_GASTRIC_CANCER_ADVANCED_VS_EARLY_DN	138	Down-regulated genes distinguishing between two subtypes of gastric cancer: advanced (AGC) and early (EGC).	6	0.0435	1.23E-04
PID_A6B1_A6B4_INTEGRIN_PATHWAY	46	a6b1 and a6b4 Integrin signaling	4	0.087	1.24E-04
MIKKELSEN_ES_LCP_WITH_H3K4ME3	142	Genes with low-CpG-density promoters (LCP) bearing histone H3 trimethylation mark at K4 (H3K4me3) in embryonic stem cells (ES).	7	0.0423	1.44E-04
SENGUPTA_NASOPHARYNGEAL_CARCINOMA_DN	349	Genes down-regulated in nsopharyngeal carcinoma relative to the normal tissue.	9	0.0258	1.47E-04
SMID_BREAST_CANCER_RELAPSE_IN_BONE_UP	97	Genes up-regulated in bone relapse of breast cancer.	5	0.0515	2.09E-04
YOSHIMURA_MAPK8_TARGETS_UP	1305	Genes up-regulated in vascular smooth muscle cells (VSMC) by MAPK8 (JNK1) [GeneID=5599].	18	0.0138	3.10E-04

WAGNER_APO2_SENSITIVITY	25	Genes whose expression most significantly correlated with cancer cell line sensitivity to the proapoptotic ligand APO2 [GeneID=8797].	3	0.12	3.52E-04
KEGG_GLYCOSPHINGOLIPID_BIOSYNTHESIS_LACTO_AND_NEOLACTO_SERIES	26	Glycosphingolipid biosynthesis - lacto and neolacto series	3	0.1154	3.97E-04
FEVR_CTNNB1_TARGETS_UP	682	Genes up-regulated in intestinal crypt cells upon deletion of CTNNB1 [GeneID=1499].	12	0.0176	4.19E-04
MIKKELSEN_IPS_LCP_WITH_H3K4ME3	174	Table 2S. Genes in MEF, MCV6, MCV8.1 and ES cells by epigenetic mark of their promoter	7	0.0345	4.29E-04
LIEN_BREAST_CARCINOMA_METAPLASTIC_VS_DUCTAL_DN	114	Genes down-regulated between two breast carcinoma subtypes: metaplastic (MCB) and ductal (DCB).	5	0.0439	4.42E-04
HOEGERKORP_CD44_TARGETS_DIRECT_UP	27	Genes directly up-regulated by CD44 [GeneID=960] stimulation of B lymphocytes.	3	0.1111	4.44E-04
LIM_MAMMARY_LUMINAL_MATURE_UP	116	Genes consistently up-regulated in mature mammary luminal cells both in mouse and human species.	5	0.0431	4.78E-04
FARMER_BREAST_CANCER_BASAL_VS_LULMINAL	330	Genes which best discriminated between two groups of breast cancer according to the status of ESR1 and AR [GeneID=2099;367]: basal (ESR1- AR-) and luminal (ESR1+ AR+).	8	0.0242	5.21E-04
CHARAFE_BREAST_CANCER_BASAL_VS_MESENCHYMAL_UP	121	Genes up-regulated in basal-like breast cancer cell lines as compared to the mesenchymal-like ones.	5	0.0413	5.80E-04
LI_PROSTATE_CANCER_EPIGENETIC	30	Genes affected by epigenetic aberrations in prostate cancer.	3	0.1	6.09E-04
CREIGHTON_ENDOCRINE_THERAPY_RESISTANCE_3	720	The 'group 3 set' of genes associated with acquired endocrine therapy resistance in breast tumors expressing ESR1 and ERBB2 [GeneID=2099;2064].	13	0.0167	6.76E-04
KIM_RESPONSE_TO_TSA_AND_DECITABINE_UP	129	Genes up-regulated in glioma cell lines treated with both decitabine [PubChem=451668] and TSA [PubChem=5562].	5	0.0388	7.75E-04
POOLA_INVASIVE_BREAST_CANCER_DN	134	Genes down-regulated in atypical ductal hyperplastic tissues from patients with (ADHC) breast cancer vs those without the cancer (ADH).	5	0.0373	9.20E-04
CHARAFE_BREAST_CANCER_LUMINAL_VS_MESENCHYMAL_UP	450	Genes up-regulated in luminal-like breast cancer cell lines compared to the mesenchymal-like ones.	10	0.02	9.29E-04
SMID_BREAST_CANCER_BASAL_UP	648	Genes up-regulated in basal subtype of breast cancer samles.	11	0.017	9.81E-04
LIU_CDX2_TARGETS_UP	36	Genes up-regulated in HET1A cells (esophagus epithelium) engineered to stably express CDX2 [GeneID=1045].	3	0.0833	1.05E-03
YANG_BREAST_CANCER_ESR1_UP	36	Genes up-regulated in early primary breast tumors expressing ESR1 [GeneID=2099] vs the ESR1 negative ones.	3	0.0833	1.05E-03
WATANABE_COLON_CANCER_MSI_VS_MSS_DN	81	Down-regulated genes discriminating between MSI (microsatellite instability) and MSS (microsatellite stability) colon cancers.	4	0.0494	1.09E-03
SMID_BREAST_CANCER_LUMINAL_B_DN	564	Genes down-regulated in the luminal B subtype of breast cancer.	10	0.0177	1.21E-03
BENPORATH_ES_WITH_H3K27ME3	1118	Set 'H3K27 bound': genes posessing the trimethylated H3K27 (H3K27me3) mark in their promoters in human embryonic stem cells, as identified by ChIP on chip.	16	0.0134	1.35E-03
SCHAEFFER_PROSTATE_DEVELOPMENT_48HR_U	487	Genes up-regulated in the urogenital sinus (UGS) of day E16 females exposed to the androgen	9	0.0185	1.61E-03

P		dihydrotestosterone [PubChem=10635] for 48 h.			
ACEVEDO_FGFR1_TARGETS_IN_PROSTATE_CANCER_MODEL_DN	308	Genes down-regulated during prostate cancer progression in the JOCK1 model due to inducible activation of FGFR1 [GeneID=2260] gene in prostate.	7	0.0227	1.69E-03
MIKKELSEN_MEF_HCP_WITH_H3K27ME3	590	Genes with high-CpG-density promoters (HCP) bearing histone H3 trimethylation mark at K27 (H3K27me3) in MEF cells (embryonic fibroblast).	10	0.0169	1.69E-03
SMID_BREAST_CANCER_BASAL_DN	701	Genes down-regulated in basal subtype of breast cancer samles.	11	0.0157	1.83E-03
PID_HNF3APATHWAY	44	FOXA1 transcription factor network	3	0.0682	1.88E-03
TURASHVILI_BREAST_LOBULAR_CARCINOMA_VS_LOBULAR_NORMAL_UP	94	Genes up-regulated in lobular carcinoma vs normal lobular breast cells.	4	0.0426	1.88E-03
JI_CARCINOGENESIS_BY_KRAS_AND_STK11_UP	12	Cluster A: genes up-regulated in primary lung tumors driven by KRAS [GeneID=3845] activation and loss of STK11 [GeneID=6794]; also up-regulated in human squamous cell carcinoma (SCC) vs adenocarcinoma subtype of NSCLC (non-small cell lung cancer).	2	0.1667	1.95E-03
MIKKELSEN_MCV6_LCP_WITH_H3K4ME3	162	Genes with low-CpG-density promoters (LCP) bearing the tri-methylation mark at H3K4 (H3K4me3) in MCV6 cells (embryonic fibroblasts trapped in a differentiated state).	6	0.0309	2.13E-03
MADAN_DPPA4_TARGETS	46	Genes differentially expressed in ES cells with DPPA4 [GeneID=55211] knockout.	3	0.0652	2.14E-03
REACTOME_TRANSMEMBRANE_TRANSPORT_OF_SMALL_MOLECULES	413	Genes involved in Transmembrane transport of small molecules	8	0.0194	2.18E-03
RAY_ALZHEIMERS_DISEASE	13	A biomarker of plasma signaling proteins that predicts clinical Alzheimer's diagnosis.	2	0.1538	2.30E-03
GO gene sets					
FUCOSYLTRANSFERASE_ACTIVITY	10	Genes annotated by the GO term GO:0008417. Catalysis of the transfer of a fucosyl group to an acceptor molecule, typically another carbohydrate or a lipid.	3	0.3	3.07E-04
IONOTROPIC_GLUTAMATE_RECEPTOR_ACTIVITY	10	Genes annotated by the GO term GO:0004970. Combining with glutamate to initiate a change in cell activity through the regulation of ion channels.	2	0.2	8.28E-03
EXCRETION	36	Genes annotated by the GO term GO:0007588. The elimination by an organism of the waste products that arise as a result of metabolic activity. These products include water, carbon dioxide (CO2), and nitrogenous compounds.	3	0.0833	1.40E-02
EPIDERMIS_DEVELOPMENT	71	Genes annotated by the GO term GO:0008544. The process whose specific outcome is the progression of the epidermis over time, from its formation to the mature structure. The epidermis is the outer epithelial layer of a plant or animal, it may be a single layer that produces an extracellular material (e.g. the cuticle of arthropods) or a complex stratified squamous epithelium, as in the case of many vertebrate species.	4	0.0563	1.78E-02
APICAL_PART_OF_CELL	17	Genes annotated by the GO term GO:0045177. The apical region of a cell.	2	0.1176	2.35E-02
GLUTAMATE_SIGNALING_PATHWAY	17	Genes annotated by the GO term GO:0007215. The series of molecular signals generated as a consequence of glutamate binding to a cell surface receptor.	2	0.1176	2.35E-02
ACTIN_FILAMENT	18	Genes annotated by the GO term GO:0005884. A filamentous structure formed of a two-stranded helical	2	0.1111	2.61E-02

		polymer of the protein actin and associated proteins. Actin filaments are a major component of the contractile apparatus of skeletal muscle and the microfilaments of the cytoskeleton of eukaryotic cells. The filaments, comprising polymerized globular actin molecules, appear as flexible structures with a diameter of 5-9 nm. They are organized into a variety of linear bundles, two-dimensional networks, and three dimensional gels. In the cytoskeleton they are most highly concentrated in the cortex of the cell just beneath the plasma membrane.			
ECTODERM_DEVELOPMENT	80	Genes annotated by the GO term GO:0007398. The process whose specific outcome is the progression of the ectoderm over time, from its formation to the mature structure. In animal embryos, the ectoderm is the outer germ layer of the embryo, formed during gastrulation.	4	0.05	2.63E-02
CHLORIDE_CHANNEL_ACTIVITY	19	Genes annotated by the GO term GO:0005254. Catalysis of facilitated diffusion of an chloride (by an energy-independent process) involving passage through a transmembrane aqueous pore or channel without evidence for a carrier-mediated mechanism.	2	0.1053	2.90E-02
ANION_CHANNEL_ACTIVITY	20	Genes annotated by the GO term GO:0005253. Catalysis of the energy-independent passage of anions across a lipid bilayer down a concentration gradient.	2	0.1	3.19E-02
GLUTAMATE_RECEPTOR_ACTIVITY	20	Genes annotated by the GO term GO:0008066. Combining with glutamate to initiate a change in cell activity.	2	0.1	3.19E-02
NEGATIVE_REGULATION_OF_TRANSPORT	20	Genes annotated by the GO term GO:0051051. Any process that stops, prevents or reduces the frequency, rate or extent of the directed movement of substances (such as macromolecules, small molecules, ions) into, out of, within or between cells.	2	0.1	3.19E-02
CHANNEL_REGULATOR_ACTIVITY	24	Genes annotated by the GO term GO:0016247.	2	0.0833	4.47E-02
Oncogene signatures					
LEF1_UP.V1_DN	190	Genes down-regulated in DLD1 cells (colon carcinoma) over-expressing LEF1 [Gene ID=51176].	6	0.0316	1.43E-02
RELA_DN.V1_DN	141	Genes down-regulated in HEK293 cells (kidney fibroblasts) upon knockdown of RELA [Gene ID=5970] gene by RNAi.	5	0.0355	1.58E-02
ESC_J1_UP_LATE.V1_UP	191	Genes up-regulated during late stages of differentiation of embryoid bodies from J1 embryonic stem cells.	5	0.0262	4.90E-02
P53_DN.V1_DN	192	Genes down-regulated in NCI-60 panel of cell lines with mutated TP53 [Gene ID=7157].	5	0.026	4.99E-02

Table S5.4 Associated network functions and biofunctions analysis on 128 identified "L-shaped" genes by IPA.

Top Networks		
ID	**Associated Network Functions**	**Score**
1	Gastrointestinal Disease, Hepatic System Disease, Liver Cholestasis	49
2	Endocrine System Disorders, Gastrointestinal Disease, Hereditary Disorder	38
3	Cellular Compromise, Neurological Disease, Organismal Injury and Abnormalities	36
4	Cell-To-Cell Signaling and Interaction, Hair and Skin Development and Function, Tissue Development	27
5	Hereditary Disorder, Metabolic Disease, Renal and Urological Disease	24

Diseases and Disorders		
Name	**p-value**	**#Molecules**
Cancer	5.55E-08 - 1.36E-02	54
Gastrointestinal Disease	3.94E-05 - 1.36E-02	37
Inflammatory Response	4.91E-05 - 1.36E-02	10
Organismal Injury and Abnormalities	1.00E-04 - 1.36E-02	21
Developmental Disorder	2.75E-04 - 1.36E-02	23

Molecular and Cellular Functions		
Name	**p-value**	**#Molecules**
Lipid Metabolism 4.62E-05	4.62E-05 - 1.15E-02	12
Molecular Transport 4.62E-05	4.62E-05 - 8.10E-03	13
Small Molecule Biochemistry	4.62E-05 -1.26E-02	20
Cell Death and Survival	4.91E-05 - 1.36E-02	15
Gene Expression	1.02E-04 - 1.06E-02	6

Table S5.5 Gene set enrichment analysis (GSEA) on 25 genes from MsigDB. These 25 genes were identified to have the most differential methylation when comparing tumor samples with normal samples. Top 50 significant gene set enrichments, with p-value < 0.05, are shown in three categories: curated gene sets, GO gene set and oncogene signatures.

Curated gene sets					
Gene Set Name		**Description**	**# Genes in Overlap (k)**	**k/K**	**p value**
TURASHVILI_BREAST_DUCTAL_CARCINOMA_VS_DUCTAL_NORMAL_DN	198	Genes down-regulated in ductal carcinoma vs normal ductal breast cells.	7	0.0303	7.51E-08
SMID_BREAST_CANCER_LUMINAL_B_DN	564	Genes down-regulated in the luminal B subtype of breast cancer.	7	0.0124	2.23E-06
HOEGERKORP_CD44_TARGETS_DIRECT_UP	27	Genes directly up-regulated by CD44 [GeneID=960] stimulation of B lymphocytes.	3	0.1111	3.74E-06
SMID_BREAST_CANCER_NORMAL_LIKE_UP	476	Genes up-regulated in the normal-like subtype of breast cancer.	6	0.0126	1.23E-05
PID_A6B1_A6B4_INTEGRIN_PATHWAY	46	a6b1 and a6b4 Integrin signaling	3	0.0652	1.92E-05
BIOCARTA_HSP27_PATHWAY	15	Stress Induction of HSP Regulation	2	0.1333	1.30E-04
CHARAFE_BREAST_CANCER_LUMINAL_VS_BASAL_DN	455	Genes down-regulated in luminal-like breast cancer cell lines compared to the basal-like ones.	5	0.011	1.40E-04
TURASHVILI_BREAST_LOBULAR_CARCINOMA_VS_DUCTAL_NORMAL_DN	91	Genes down-regulated in lobular carcinoma vs normal ductal breast cells.	3	0.033	1.48E-04
TURASHVILI_BREAST_LOBULAR_CARCINOMA_VS_LOBULAR_NORMAL_UP	94	Genes up-regulated in lobular carcinoma vs normal lobular breast cells.	3	0.0319	1.63E-04
SMID_BREAST_CANCER_RELAPSE_IN_BONE_UP	97	Genes up-regulated in bone relapse of breast cancer.	3	0.0309	1.79E-04
LIM_MAMMARY_STEM_CELL_UP	489	Genes consistently up-regulated in mammary stem cells both in mouse and human species.	5	0.0102	1.96E-04
DOANE_BREAST_CANCER_ESR1_UP	112	Genes up-regulated in breast cancer samples positive for ESR1 [GeneID=2099] compared to the ESR1 negative tumors.	3	0.0268	2.74E-04
FARMER_BREAST_CANCER_BASAL_VS_LULMINAL	330	Genes which best discriminated between two groups of breast cancer according to the status of ESR1 and AR [GeneID=2099;367]: basal (ESR1- AR-) and luminal (ESR1+ AR+).	4	0.0121	4.92E-04
SMID_BREAST_CANCER_BASAL_UP	648	Genes up-regulated in basal subtype of breast cancer samles.	5	0.0077	7.12E-04
SMID_BREAST_CANCER_RELAPSE_IN_BRAIN_UP	39	Genes up-regulated in brain relapse of breast cancer.	2	0.0513	9.00E-04
HAN_SATB1_TARGETS_UP	395	Genes up-regulated in MDA-MB-231 cells (breast cancer) after knockdown of SATB1 [GeneID=6304] by RNAi.	4	0.0101	9.64E-04
SMID_BREAST_CANCER_BASAL_DN	701	Genes down-regulated in basal subtype of breast cancer samles.	5	0.0071	1.01E-03
COWLING_MYCN_TARGETS	43	Genes down-regulated by MYCN [GeneID=4613] but not by its transactivation-defficient, trunkated form N-Myc-delta-73.	2	0.0465	1.09E-03
BECKER_TAMOXIFEN_RE	50	Genes up-regulated in a breast cancer cell line resistant	2	0.04	1.48E-03

SISTANCE_UP		to tamoxifen [PubChem=5376] compared to the parental line sensitive to the drug.			
HALMOS_CEBPA_TARGETS_UP	52	Genes up-regulated in H358 cells (lung cancer) by inducible expression of CEBPA [GeneID=1050] off plasmid vector.	2	0.0385	1.60E-03
EBAUER_TARGETS_OF_PAX3_FOXO1_FUSION_UP	207	Genes up-regulated in Rh4 cells (alveolar rhabdomyosarcoma, ARMS) after knockdown of the PAX3-FOXO1 [GeneiD=5077;2308] fusion protein by RNAi for 72 hr.	3	0.0145	1.63E-03
HUPER_BREAST_BASAL_VS_LUMINAL_UP	54	Genes up-regulated in basal mammary epithelial cells compared to the luminal ones.	2	0.037	1.72E-03
SU_PANCREAS	54	Genes up-regulated specifically in human pancreas.	2	0.037	1.72E-03
ONDER_CDH1_TARGETS_2_DN	464	Genes down-regulated in HMLE cells (immortalized nontransformed mammary epithelium) after E-cadhedrin (CDH1) [GeneID=999] knockdown by RNAi.	4	0.0086	1.75E-03
ISSAEVA_MLL2_TARGETS	62	Genes down-regulated in HeLa cells upon knockdown of MLL2 [GeneID=8085] by RNAi.	2	0.0323	2.26E-03
PID_ERA_GENOMIC_PATHWAY	65	Validated nuclear estrogen receptor alpha network	2	0.0308	2.48E-03
JAEGER_METASTASIS_DN	258	Genes down-regulated in metastases from malignant melanoma compared to the primary tumors.	3	0.0116	3.05E-03
REACTOME_CELL_JUNCTION_ORGANIZATION	78	Genes involved in Cell junction organization	2	0.0256	3.55E-03
CHIANG_LIVER_CANCER_SUBCLASS_POLYSOMY7_UP	79	Marker genes up-regulated in the 'chromosome 7 polysomy' subclass of hepatocellular carcinoma (HCC); characterized by polysomy of chromosome 7 and by a lack of gains of chromosome 8q.	2	0.0253	3.64E-03
CROMER_METASTASIS_DN	81	Metastatic propensity markers of head and neck squamous cell carcinoma (HNSCC): down-regulated in metastatic vs non-metastatic tumors.	2	0.0247	3.82E-03
SMID_BREAST_CANCER_RELAPSE_IN_BRAIN_DN	85	Genes down-regulated in brain relapse of breast cancer.	2	0.0235	4.20E-03
GHANDHI_BYSTANDER_IRRADIATION_UP	86	Genes significantly (FDR < 10%) up-regulated in IMR-90 cells (fibroblast) in response to bystander irradiation.	2	0.0233	4.30E-03
BILANGES_SERUM_SENSITIVE_GENES	90	Genes translationally regulated in MEF cells (embryonic fibroblasts) in response to serum starvation but not by rapamycin (sirolimus) [PubChemID=6610346].	2	0.0222	4.70E-03
CADWELL_ATG16L1_TARGETS_UP	93	Genes up-regulated in Paneth cell (part of intestiinal epithelium) of mice with hypomorphic (reduced function) form of ATG16L1 [GeneID=55054].	2	0.0215	5.00E-03
SMID_BREAST_CANCER_RELAPSE_IN_BONE_DN	315	Genes down-regulated in bone relapse of breast cancer.	3	0.0095	5.32E-03
LEI_MYB_TARGETS	318	Myb-regulated genes in MCF7 (breast cancer) and lung epithelial cell lines overexpressing MYBL2, MYBL1 or MYB [GeneID=4605;4603;4602].	3	0.0094	5.47E-03
BENPORATH_EED_TARGETS	1062	Set 'Eed targets': genes identified by ChIP on chip as targets of the Polycomb protein EED [GeneID=8726] in human embryonic stem cells.	5	0.0047	6.17E-03
GHANDHI_DIRECT_IRRADIATION_UP	110	Genes significantly (FDR < 10%) up-regulated in IMR-90 cells (fibroblast) in response to direct irradiation.	2	0.0182	6.93E-03
SENGUPTA_NASOPHARYNGEAL_CARCINOMA_DN	349	Genes down-regulated in nsopharyngeal carcinoma relative to the normal tissue.	3	0.0086	7.07E-03
LIEN_BREAST_CARCINOMA_METAPLASTIC_VS_DUCTAL_DN	114	Genes down-regulated between two breast carcinoma subtypes: metaplastic (MCB) and ductal (DCB).	2	0.0175	7.43E-03
LIM_MAMMARY_LUMINAL_MATURE_UP	116	Genes consistently up-regulated in mature mammary luminal cells both in mouse and human species.	2	0.0172	7.68E-03

CERVERA_SDHB_TARGETS_1_UP	118	Genes turned on in Hep3B cells (hepatocellular carcinoma, HCC) upon knockdown of SDHB [GeneID=6390] by RNAi.	2	0.0169	7.94E-03
REACTOME_CELL_CELL_COMMUNICATION	120	Genes involved in Cell-Cell communication	2	0.0167	8.20E-03
CHARAFE_BREAST_CANCER_LUMINAL_VS_BASAL_UP	380	Genes up-regulated in luminal-like breast cancer cell lines compared to the basal-like ones.	3	0.0079	8.92E-03
YEGNASUBRAMANIAN_PROSTATE_CANCER	128	Genes expressed in at least one prostate cancer cell line but not in normal prostate epithelial cells or stromal cells	2	0.0156	9.28E-03
SENESE_HDAC2_TARGETS_DN	133	Genes down-regulated in U2OS cells (osteosarcoma) upon knockdown of HDAC2 [GeneID=3066] by RNAi.	2	0.015	9.99E-03
LIM_MAMMARY_STEM_CELL_DN	428	Genes consistently down-regulated in mammary stem cells both in mouse and human species.	3	0.007	1.23E-02
RIGGI_EWING_SARCOMA_PROGENITOR_UP	430	Genes up-regulated in mesenchymal stem cells (MSC) engineered to express EWS-FLI1 [GeneID=2130;2321] fusion protein.	3	0.007	1.25E-02
CHARAFE_BREAST_CANCER_LUMINAL_VS_MESENCHYMAL_UP	450	Genes up-regulated in luminal-like breast cancer cell lines compared to the mesenchymal-like ones.	3	0.0067	1.41E-02
DODD_NASOPHARYNGEAL_CARCINOMA_UP	1821	Genes up-regulated in nasopharyngeal carcinoma (NPC) compared to the normal tissue.	7	0.0033	1.44E-02
GO gene sets					
EXCRETION	36	Genes annotated by the GO term GO:0007588. The elimination by an organism of the waste products that arise as a result of metabolic activity. These products include water, carbon dioxide (CO2), and nitrogenous compounds.	2	0.0556	4.78E-03
DIGESTION	44	Genes annotated by the GO term GO:0007586. The whole of the physical, chemical, and biochemical processes carried out by multicellular organisms to break down ingested nutrients into components that may be easily absorbed and directed into metabolism.	2	0.0455	7.08E-03
NEGATIVE_REGULATION_OF_CELL_PROLIFERATION	156	Genes annotated by the GO term GO:0008285. Any process that stops, prevents or reduces the rate or extent of cell proliferation.	3	0.0192	9.94E-03
REGULATION_OF_GROWTH	58	Genes annotated by the GO term GO:0040008. Any process that modulates the frequency, rate or extent of the growth of all or part of an organism so that it occurs at its proper speed, either globally or in a specific part of the organism's development.	2	0.0345	1.21E-02
EPIDERMIS_DEVELOPMENT	71	Genes annotated by the GO term GO:0008544. The process whose specific outcome is the progression of the epidermis over time, from its formation to the mature structure. The epidermis is the outer epithelial layer of a plant or animal, it may be a single layer that produces an extracellular material (e.g. the cuticle of arthropods) or a complex stratified squamous epithelium, as in the case of many vertebrate species.	2	0.0282	1.77E-02
GROWTH	77	Genes annotated by the GO term GO:0040007. The increase in size or mass of an entire organism, a part of an organism or a cell.	2	0.026	2.07E-02
CYTOSOL	205	Genes annotated by the GO term GO:0005829. That part of the cytoplasm that does not contain membranous or particulate subcellular components.	3	0.0146	2.07E-02
ECTODERM_DEVELOPME	80	Genes annotated by the GO term GO:0007398. The	2	0.025	2.22E-02

NT		process whose specific outcome is the progression of the ectoderm over time, from its formation to the mature structure. In animal embryos, the ectoderm is the outer germ layer of the embryo, formed during gastrulation.			
ENDOPEPTIDASE_ACTIVITY	117	Genes annotated by the GO term GO:0004175. Catalysis of the hydrolysis of nonterminal peptide linkages in oligopeptides or polypeptides, and comprising any enzyme of sub-subclasses EC:3.4.21-99. They are classfied according to the presence of essential catalytic residues or ions at their active sites.	2	0.0171	4.47E-02
ANTI_APOPTOSIS	118	Genes annotated by the GO term GO:0006916. A process which directly inhibits any of the steps required for cell death by apoptosis.	2	0.0169	4.53E-02
Oncogene signatures					
LEF1_UP.V1_DN	190	Genes down-regulated in DLD1 cells (colon carcinoma) over-expressing LEF1 [Gene ID=51176].	3	0.0158	7.39E-03
BMI1_DN_MEL18_DN.V1_UP	145	Genes up-regulated in DAOY cells (medulloblastoma) upon knockdown of BMI1 and PCGF2 [Gene ID=648, 7703] genes by RNAi.	2	0.0138	3.78E-02

Table S5.6 Associated network functions and biofunctions analysis on 25 genes from IPA. These 25 genes were identified to have the most differential methylation when comparing tumor samples with normal samples.

Top Networks		
ID	**Associated Network Functions**	**Score**
1	Molecular Transport, Nucleic Acid Metabolism, Small Molecule Biochemistry	29
2	Cellular Movement, Reproductive System Development and Function, Cell Morphology	23
3	Cell-To-Cell Signaling and Interaction, Cellular Assembly and Organization, Tissue Development	3

Diseases and Disorders		
Name	**p-value**	**#Molecules**
Cancer	6.29E-06 - 7.39E-03	14
Connective Tissue Disorders	2.09E-05 - 6.49E-03	9
Dermatological Diseases and Conditions	2.09E-05 - 5.32E-03	4
Developmental Disorder	2.09E-05 - 6.49E-03	7
Hereditary Disorder	2.09E-05 - 5.91E-03	5

Molecular and Cellular Functions		
Name	**p-value**	**#Molecules**
Cell Cycle	6.29E-06 - 7.39E-03	3
Cellular Movement	1.09E-05 - 5.91E-03	12
Cell Death and Survival	6.78E-05 - 7.39E-03	16
Cellular Development	9.38E-05 - 7.39E-03	7
Cellular Growth and Proliferation	1.28E-04 - 7.39E-03	13

Table S5.7 Gene set enrichment analysis (GSEA) on 128 identified "L-shaped" genes from MsigDB. Significant gene set enrichments, with p-value < 0.05, are shown in three categories: curated gene sets, GO gene set and oncogene signatures.

Curated gene sets					
Gene Set Name	**# Genes in Gene Set (K)**	**Description**	**# Genes in Overlap (k)**	**k/K**	**p value**
TURASHVILI_BREAST_LOBULAR_CARCINOMA_VS_DUCTAL_NORMAL_DN	91	Genes down-regulated in lobular carcinoma vs normal ductal breast cells.	3	0.033	2.75E-05
TURASHVILI_BREAST_LOBULAR_CARCINOMA_VS_LOBULAR_NORMAL_UP	94	Genes up-regulated in lobular carcinoma vs normal lobular breast cells.	3	0.0319	3.03E-05
CHARAFE_BREAST_CANCER_LUMINAL_VS_BASAL_DN	455	Genes down-regulated in luminal-like breast cancer cell lines compared to the basal-like ones.	4	0.0088	1.82E-04
TURASHVILI_BREAST_DUCTAL_CARCINOMA_VS_DUCTAL_NORMAL_DN	198	Genes down-regulated in ductal carcinoma vs normal ductal breast cells.	3	0.0152	2.77E-04
PID_A6B1_A6B4_INTEGRIN_PATHWAY	46	a6b1 and a6b4 Integrin signaling	2	0.0435	4.18E-04
HUPER_BREAST_BASAL_VS_LUMINAL_UP	54	Genes up-regulated in basal mammary epithelial cells compared to the luminal ones.	2	0.037	5.77E-04
ISSAEVA_MLL2_TARGETS	62	Genes down-regulated in HeLa cells upon knockdown of MLL2 [GeneID=8085] by RNAi.	2	0.0323	7.60E-04
PID_ERA_GENOMIC_PATHWAY	65	Validated nuclear estrogen receptor alpha network	2	0.0308	8.34E-04
FARMER_BREAST_CANCER_BASAL_VS_LULMINAL	330	Genes which best discriminated between two groups of breast cancer according to the status of ESR1 and AR [GeneID=2099;367]: basal (ESR1- AR-) and luminal (ESR1+ AR+).	3	0.0091	1.22E-03
GHANDHI_BYSTANDER_IRRADIATION_UP	86	Genes significantly (FDR < 10%) up-regulated in IMR-90 cells (fibroblast) in response to bystander irradiation.	2	0.0233	1.45E-03
SMID_BREAST_CANCER_RELAPSE_IN_BONE_UP	97	Genes up-regulated in bone relapse of breast cancer.	2	0.0206	1.85E-03
GHANDHI_DIRECT_IRRADIATION_UP	110	Genes significantly (FDR < 10%) up-regulated in IMR-90 cells (fibroblast) in response to direct irradiation.	2	0.0182	2.36E-03
DOANE_BREAST_CANCER_ESR1_UP	112	Genes up-regulated in breast cancer samples positive for ESR1 [GeneID=2099] compared to the ESR1 negative tumors.	2	0.0179	2.45E-03
ONDER_CDH1_TARGETS_2_DN	464	Genes down-regulated in HMLE cells (immortalized nontransformed mammary epithelium) after E-cadhedrin (CDH1) [GeneID=999] knockdown by RNAi.	3	0.0065	3.23E-03
LIU_PROSTATE_CANCER_DN	481	Genes down-regulated in prostate cancer samples.	3	0.0062	3.58E-03
LIM_MAMMARY_STEM_CELL_UP	489	Genes consistently up-regulated in mammary stem cells both in mouse and human species.	3	0.0061	3.75E-03
BENPORATH_EED_TARGETS	1062	Set 'Eed targets': genes identified by ChIP on chip as targets of the Polycomb protein EED [GeneID=8726] in human embryonic stem cells.	4	0.0038	4.30E-03
SMID_BREAST_CANCER_LUMINAL_B_DN	564	Genes down-regulated in the luminal B subtype of breast cancer.	3	0.0053	5.59E-03
MIKKELSEN_MEF_HCP_WITH_H3K27ME3	590	Genes with high-CpG-density promoters (HCP) bearing histone H3 trimethylation mark at K27 (H3K27me3) in MEF cells (embryonic fibroblast).	3	0.0051	6.33E-03

WU_CELL_MIGRATION	184	Genes associated with migration rate of 40 human bladder cancer cells.	2	0.0109	6.46E-03
SMID_BREAST_CANCER_BASAL_UP	648	Genes up-regulated in basal subtype of breast cancer samles.	3	0.0046	8.21E-03
BENPORATH_PRC2_TARGETS	652	Set 'PRC2 targets': Polycomb Repression Complex 2 (PRC) targets; identified by ChIP on chip on human embryonic stem cells as genes that: posess the trimethylated H3K27 mark in their promoters and are bound by SUZ12 [GeneID=23512] and EED [GeneID=8726] Polycomb proteins.	3	0.0046	8.35E-03
ZHANG_RESPONSE_TO_IKK_INHIBITOR_AND_TNF_UP	223	Genes up-regulated in BxPC3 cells (pancreatic cancer) after treatment with TNF [GeneID=7124] or IKI-1, an inhibitor of IkappaB kinase (IKK).	2	0.009	9.35E-03
OSWALD_HEMATOPOIETIC_STEM_CELL_IN_COLLAGEN_GEL_UP	233	Genes up-regulated in hematopoietic stem cells (HSC, CD34+ [GeneID=947]) cultured in a three-dimentional collagen gel compared to the cells grown in suspension.	2	0.0086	1.02E-02
SMID_BREAST_CANCER_BASAL_DN	701	Genes down-regulated in basal subtype of breast cancer samles.	3	0.0043	1.02E-02
JAEGER_METASTASIS_DN	258	Genes down-regulated in metastases from malignant melanoma compared to the primary tumors.	2	0.0078	1.24E-02
KEGG_CYTOKINE_CYTOKINE_RECEPTOR_INTERACTION	267	Cytokine-cytokine receptor interaction	2	0.0075	1.32E-02
HADDAD_B_LYMPHOCYTE_PROGENITOR	293	Genes up-regulated in hematopoietic progenitor cells (HPC) of B lymphocyte lineage CD34+CD45RA+CD10+ [GeneID=947;5788;4311].	2	0.0068	1.57E-02
ZHANG_TLX_TARGETS_60HR_UP	293	Genes up-regulated in neural stem cells (NSC) at 60 h after cre-lox knockout of TLX (NR2E1) [GeneID=7101].	2	0.0068	1.57E-02
MARTENS_TRETINOIN_RESPONSE_UP	857	Genes up-regulated in NB4 cells (acute promyelocytic leukemia, APL) in response to tretinoin [PubChem=5538]; based on Chip-seq data.	3	0.0035	1.75E-02
SMID_BREAST_CANCER_RELAPSE_IN_BONE_DN	315	Genes down-regulated in bone relapse of breast cancer.	2	0.0063	1.80E-02
LEI_MYB_TARGETS	318	Myb-regulated genes in MCF7 (breast cancer) and lung epithelial cell lines overexpressing MYBL2, MYBL1 or MYB [GeneID=4605;4603;4602].	2	0.0063	1.84E-02
SENGUPTA_NASOPHARYNGEAL_CARCINOMA_DN	349	Genes down-regulated in nsopharyngeal carcinoma relative to the normal tissue.	2	0.0057	2.19E-02
GRUETZMANN_PANCREATIC_CANCER_UP	358	Genes up-regulated in pancreatic ductal adenocarcinoma (PDAC) identified in a meta analysis across four independent studies.	2	0.0056	2.29E-02
CHARAFE_BREAST_CANCER_LUMINAL_VS_BASAL_UP	380	Genes up-regulated in luminal-like breast cancer cell lines compared to the basal-like ones.	2	0.0053	2.56E-02
CHEMNITZ_RESPONSE_TO_PROSTAGLANDIN_E2_DN	391	Genes down-regulated in CD4+ [GeneID=920] T lymphocytes after stimulation with prostaglandin E2 [PubChem=5280360].	2	0.0051	2.70E-02
HAN_SATB1_TARGETS_UP	395	Genes up-regulated in MDA-MB-231 cells (breast cancer) after knockdown of SATB1 [GeneID=6304] by RNAi.	2	0.0051	2.75E-02
BENPORATH_SUZ12_TARGETS	1038	Set 'Suz12 targets': genes identified by ChIP on chip as targets of the Polycomb protein SUZ12 [GeneID=23512] in human embryonic stem cells.	3	0.0029	2.90E-02
RIGGI_EWING_SARCOMA_PROGENITOR_UP	430	Genes up-regulated in mesenchymal stem cells (MSC) engineered to express EWS-FLI1 [GeneID=2130;2321] fusion protein.	2	0.0047	3.22E-02
MIKKELSEN_MCV6_HCP_	435	Genes with high-CpG-density promoters (HCP) bearing	2	0.0046	3.29E-02

WITH_H3K27ME3		the tri-methylation mark at H3K27 (H3K27me3) in MCV6 cells (embryonic fibroblasts trapped in a differentiated state).			
DELYS_THYROID_CANCER_UP	443	Genes up-regulated in papillary thyroid carcinoma (PTC) compared to normal tissue.	2	0.0045	3.40E-02
BENPORATH_ES_WITH_H3K27ME3	1118	Set 'H3K27 bound': genes posessing the trimethylated H3K27 (H3K27me3) mark in their promoters in human embryonic stem cells, as identified by ChIP on chip.	3	0.0027	3.51E-02
SENESE_HDAC1_TARGETS_UP	457	Genes up-regulated in U2OS cells (osteosarcoma) upon knockdown of HDAC1 [GeneID=3065] by RNAi.	2	0.0044	3.60E-02
ZHOU_INFLAMMATORY_RESPONSE_LIVE_UP	485	Genes up-regulated in macrophage by live P.gingivalis.	2	0.0041	4.02E-02
ENK_UV_RESPONSE_KERATINOCYTE_UP	530	Genes up-regulated in NHEK cells (normal epidermal keratinocytes) after UVB irradiation.	2	0.0038	4.72E-02
GO gene sets					
DIGESTION	44	Genes annotated by the GO term GO:0007586. The whole of the physical, chemical, and biochemical processes carried out by multicellular organisms to break down ingested nutrients into components that may be easily absorbed and directed into metabolism.	2	0.0455	2.41E-03
ANTI_APOPTOSIS	118	Genes annotated by the GO term GO:0006916. A process which directly inhibits any of the steps required for cell death by apoptosis.	2	0.0169	1.64E-02
TRANSCRIPTION_FACTOR_ACTIVITY	354	Genes annotated by the GO term GO:0003700. The function of binding to a specific DNA sequence in order to modulate transcription. The transcription factor may or may not also interact selectively with a protein or macromolecular complex.	3	0.0085	1.99E-02
NEGATIVE_REGULATION_OF_APOPTOSIS	150	Genes annotated by the GO term GO:0043066. Any process that stops, prevents or reduces the frequency, rate or extent of cell death by apoptosis.	2	0.0133	2.57E-02
NEGATIVE_REGULATION_OF_PROGRAMMED_CELL_DEATH	151	Genes annotated by the GO term GO:0043069. Any process that stops, prevents or reduces the frequency, rate or extent of programmed cell death, cell death resulting from activation of endogenous cellular processes.	2	0.0132	2.61E-02
NEGATIVE_REGULATION_OF_DEVELOPMENTAL_PROCESS	197	Genes annotated by the GO term GO:0051093. Any process that stops, prevents or reduces the rate or extent of development, the biological process whose specific outcome is the progression of an organism over time from an initial condition (e.g. a zygote, or a young adult) to a later condition (e.g. a multicellular animal or an aged adult).	2	0.0102	4.25E-02
Oncogene signatures					
LEF1_UP.V1_DN	190	Genes down-regulated in DLD1 cells (colon carcinoma) over-expressing LEF1 [Gene ID=51176].	3	0.0158	1.50E-03
BMI1_DN_MEL18_DN.V1_UP	145	Genes up-regulated in DAOY cells (medulloblastoma) upon knockdown of BMI1 and PCGF2 [Gene ID=648, 7703] genes by RNAi.	2	0.0138	1.36E-02

Table S5.8 Significance of biomarkers in term of methylation status in survival analysis by GENESURV tools in bioprofiling.de. The methylation status determined by our method is shown in columns under the headings of Methylation_CIMP+ and Methylation_CIMP-.

gene	Methylation_CIMP+	methylation_CIMP-	GENESURV_survival longer_expression	GENESURV_p-value	GENESURV_link
SLC44A4	hypo	hyper	high	9.77E-05	http://www.bioprofiling.de/cgi-bin/GEO/GENESURV/display_survival_details.GENE.pl?ID=GSE30682&affy=ILMN_1730977&ncbi=80736&geneA=SLC44A4&tmp_dir=dir_10084_1355434594
IL20RA	hypo	hyper	high	7.83E-05	http://www.bioprofiling.de/cgi-bin/GEO/GENESURV/display_survival_details.GENE.pl?ID=GSE22220&affy=3390504&ncbi=53832&geneA=IL20RA&tmp_dir=dir_10084_1355434594
TFF1	hypo	hyper	high	0.00154	http://www.bioprofiling.de/cgi-bin/GEO/GENESURV/display_survival_details.GENE.pl?ID=GSE30682&affy=ILMN_1722489&ncbi=7031&geneA=TFF1&tmp_dir=dir_10084_1355434594
C1orf64	hypo	hyper	high	0.00637	http://www.bioprofiling.de/cgi-bin/GEO/GENESURV/display_survival_details.GENE.pl?ID=GSE22226&affy=39434&ncbi=149563&geneA=C1ORF64&tmp_dir=dir_10084_1355434594
MEP1A	hypo	hyper	low	0.0257	http://www.bioprofiling.de/cgi-bin/GEO/GENESURV/display_survival_details.GENE.pl?ID=GSE30682&affy=ILMN_1659984&ncbi=4224&geneA=MEP1A&tmp_dir=dir_10084_1355434594
POU4F1	hyper	hypo	low	1.99E-05	http://www.bioprofiling.de/cgi-bin/GEO/GENESURV/display_survival_details.GENE.pl?ID=GSE30682&affy=ILMN_1738691&ncbi=5457&geneA=POU4F1&tmp_dir=dir_10084_1355434594

BIBLIOGRAPHY

[1] P.A. t Hoen, Y. Ariyurek, H.H. Thygesen, E. Vreugdenhil, R.H. Vossen, R.X. de Menezes, J.M. Boer, G.J. van Ommen, J.T. den Dunnen, Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms, Nucleic Acids Res 36 (2008) e141.

[2] L. Gao, Z. Fang, K. Zhang, D. Zhi, X. Cui, Length bias correction for RNA-seq data in gene set analyses, Bioinformatics 27 (2011) 662-669.

[3] K.D. Hansen, S.E. Brenner, S. Dudoit, Biases in Illumina transcriptome sequencing caused by random hexamer priming, Nucleic Acids Res 38 (2010) e131.

[4] J. Li, H. Jiang, W.H. Wong, Modeling non-uniformity in short-read rates in RNA-Seq data, Genome Biol 11 (2010) R50.

[5] A. Roberts, C. Trapnell, J. Donaghey, J.L. Rinn, L. Pachter, Improving RNA-Seq expression estimates by correcting for fragment bias, Genome Biol 12 (2011) R22.

[6] S. Schwartz, R. Oren, G. Ast, Detection and removal of biases in the analysis of next-generation sequencing reads, PLoS One 6 (2011) e16685.

[7] M.A. Taub, H. Corrada Bravo, R.A. Irizarry, Overcoming bias and systematic errors in next generation sequencing data, Genome Med 2 (2010) 87.

[8] W. Zheng, L.M. Chung, H. Zhao, Bias detection and correction in RNA-Sequencing data, BMC Bioinformatics 12 (2011) 290.

[9] J.C. Dohm, C. Lottaz, T. Borodina, H. Himmelbauer, Substantial biases in ultra-short read data sets from high-throughput DNA sequencing, Nucleic Acids Res 36 (2008) e105.

[10] K. Kapur, H. Jiang, Y. Xing, W.H. Wong, Cross-hybridization modeling on Affymetrix exon arrays, Bioinformatics 24 (2008) 2887-2893.

[11] L. Zhang, M.F. Miles, K.D. Aldape, A model of molecular interactions on short oligonucleotide microarrays, Nat Biotechnol 21 (2003) 818-821.

[12] P. Uva, E. de Rinaldis, CrossHybDetector: detection of cross-hybridization events in DNA microarray experiments, BMC Bioinformatics 9 (2008) 485.

[13] C. Wu, R. Carta, L. Zhang, Sequence dependence of cross-hybridization on short oligo microarrays, Nucleic Acids Res 33 (2005) e84.

[14] A. Oshlack, M.J. Wakefield, Transcript length bias in RNA-seq data confounds systems biology, Biol Direct 4 (2009) 14.

[15] J.T. Leek, R.B. Scharpf, H.C. Bravo, D. Simcha, B. Langmead, W.E. Johnson, D. Geman, K. Baggerly, R.A. Irizarry, Tackling the widespread and critical impact of batch effects in high-throughput data, Nat Rev Genet 11 (2010) 733-739.

[16] H.C. Bravo, R.A. Irizarry, Model-based quality assessment and base-calling for second-generation sequencing data, Biometrics 66 (2010) 665-674.

[17] W.C. Kao, Y.S. Song, naiveBayesCall: an efficient model-based base-calling algorithm for high-throughput sequencing, J Comput Biol 18 (2011) 365-377.

[18] W.C. Kao, K. Stevens, Y.S. Song, BayesCall: A model-based base-calling algorithm for high-throughput short-read sequencing, Genome Res 19 (2009) 1884-1895.

[19] D.R. Kelley, M.C. Schatz, S.L. Salzberg, Quake: quality-aware detection and correction of sequencing errors, Genome Biol 11 (2010) R116.

[20] J. Schroder, H. Schroder, S.J. Puglisi, R. Sinha, B. Schmidt, SHREC: a short-read error correction method, Bioinformatics 25 (2009) 2157-2163.

[21] X. Yang, K.S. Dorman, S. Aluru, Reptile: representative tiling for short read error correction, Bioinformatics 26 (2010) 2526-2533.

[22] X. Zhao, L.E. Palmer, R. Bolanos, C. Mircean, D. Fasulo, G.M. Wittenberg, EDAR: an efficient error detection and removal algorithm for next generation sequencing data, J Comput Biol 17 (2010) 1549-1560.

[23] T. Wilkes, H. Laux, C.A. Foy, Microarray data quality - review of current developments, OMICS 11 (2007) 1-13.

[24] L. Jiang, F. Schlesinger, C.A. Davis, Y. Zhang, R. Li, M. Salit, T.R. Gingeras, B. Oliver, Synthetic spike-in standards for RNA-seq experiments, Genome Res 21 (2011) 1543-1551.

[25] I.V. Yang, Use of external controls in microarray experiments, Methods Enzymol 411 (2006) 50-63.

[26] V.M. Kvam, P. Liu, Y. Si, A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data, Am J Bot 99 (2012) 248-256.

[27] P.L. Auer, R.W. Doerge, A Two-Stage Poisson Model for Testing RNA-Seq Data Statistical Applications in Genetics and Molecular Biology, Vol. 10, No. 1. (2011) Key: citeulike:9608729 10 (2011) 1-26.

[28] J. Lee, Y. Ji, S. Liang, G. Cai, P. Muller, On differential gene expression using RNA-Seq data, Cancer Inform 10 (2011) 205-215.

[29] M.D. Robinson, D.J. McCarthy, G.K. Smyth, edgeR: a Bioconductor package for differential expression analysis of digital gene expression data, Bioinformatics 26 (2010) 139-140.

[30] S. Anders, W. Huber, Differential expression analysis for sequence count data, Genome Biol 11 (2010) R106.

[31] A. Oshlack, M.D. Robinson, M.D. Young, From RNA-seq reads to differential expression results, Genome Biol 11 (2010) 220.

[32] J.T. Leek, J.D. Storey, Capturing heterogeneity in gene expression studies by surrogate variable analysis, PLoS Genet 3 (2007) 1724-1735.

[33] M.N. McCall, B.M. Bolstad, R.A. Irizarry, Frozen robust multiarray analysis (fRMA), Biostatistics 11 (2010) 242-253.

[34] M.N. McCall, H.A. Jaffee, R.A. Irizarry, fRMA ST: frozen robust multiarray analysis for Affymetrix Exon and Gene ST arrays, Bioinformatics 28 (2012) 3153-3154.

[35] L. Wang, Z. Feng, X. Wang, X. Zhang, DEGseq: an R package for identifying differentially expressed genes from RNA-seq data, Bioinformatics 26 (2010) 136-138.

[36] J.M. Toung, M. Morley, M. Li, V.G. Cheung, RNA-sequence analysis of human B-cells, Genome Res 21 (2011) 991-998.

[37] J.H. Bullard, E. Purdom, K.D. Hansen, S. Dudoit, Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments, BMC Bioinformatics 11 (2010) 94.

[38] K.A. Baggerly, L. Deng, J.S. Morris, C.M. Aldaz, Differential expression in SAGE: accounting for normal between-library variation, Bioinformatics 19 (2003) 1477-1483.

[39] T.V. Pham, S.R. Piersma, M. Warmoes, C.R. Jimenez, On the beta-binomial model for analysis of spectral count data in label-free tandem mass spectrometry-based proteomics, Bioinformatics 26 (2010) 363-369.

[40] R.S. Jones, Epigenetics: reversing the 'irreversible', Nature 450 (2007) 357-359.

[41] P.A. Jones, S.B. Baylin, The fundamental role of epigenetic events in cancer, Nat Rev Genet 3 (2002) 415-428.

[42] P.A. Jones, S.B. Baylin, The epigenomics of cancer, Cell 128 (2007) 683-692.

[43] M. Greaves, C.C. Maley, Clonal evolution in cancer, Nature 481 (2012) 306-313.

[44] K.D. Siegmund, P. Marjoram, Y.J. Woo, S. Tavare, D. Shibata, Inferring clonal expansion and cancer stem cell dynamics from DNA methylation patterns in colorectal cancers, Proc Natl Acad Sci U S A 106 (2009) 4828-4833.

[45] M. Toyota, N. Ahuja, M. Ohe-Toyota, J.G. Herman, S.B. Baylin, J.P. Issa, CpG island methylator phenotype in colorectal cancer, Proc Natl Acad Sci U S A 96 (1999) 8681-8686.

[46] D.J. Weisenberger, K.D. Siegmund, M. Campan, J. Young, T.I. Long, M.A. Faasse, G.H. Kang, M. Widschwendter, D. Weener, D. Buchanan, H. Koh, L. Simms, M. Barker, B. Leggett, J. Levine, M. Kim, A.J. French, S.N. Thibodeau, J. Jass, R. Haile, P.W. Laird, CpG island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with BRAF mutation in colorectal cancer, Nat Genet 38 (2006) 787-793.

[47] H. Noushmehr, D.J. Weisenberger, K. Diefes, H.S. Phillips, K. Pujara, B.P. Berman, F. Pan, C.E. Pelloski, E.P. Sulman, K.P. Bhat, R.G. Verhaak, K.A. Hoadley, D.N. Hayes, C.M. Perou, H.K. Schmidt, L. Ding, R.K. Wilson, D. Van Den Berg, H. Shen, H. Bengtsson, P. Neuvial, L.M. Cope, J. Buckley, J.G. Herman, S.B. Baylin, P.W. Laird, K. Aldape, Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma, Cancer Cell 17 (2010) 510-522.

[48] B. Davidson, Z. Zhang, L. Kleinberg, M. Li, V.A. Florenes, T.L. Wang, M. Shih Ie, Gene expression signatures differentiate ovarian/peritoneal serous carcinoma from diffuse malignant peritoneal mesothelioma, Clin Cancer Res 12 (2006) 5944-5950.

[49] M. Faryna, C. Konermann, S. Aulmann, J.L. Bermejo, M. Brugger, S. Diederichs, J. Rom, D. Weichenhan, R. Claus, M. Rehli, P. Schirmacher, H.P. Sinn, C. Plass, C. Gerhauser, Genome-wide methylation screen in low-grade breast cancer identifies novel epigenetically altered genes as potential biomarkers for tumor diagnosis, FASEB J 26 (2012) 4937-4950.

[50] Y. Huang, Y. Wang, M. Wang, B. Sun, Y. Li, Y. Bao, K. Tian, H. Xu, Differential methylation of TSP50 and mTSP50 genes in different types of human tissues and mouse spermatic cells, Biochem Biophys Res Commun 374 (2008) 658-661.

[51] P.K. Lo, S. Sukumar, Epigenomics and breast cancer, Pharmacogenomics 9 (2008) 1879-1902.

[52] P. Novak, T. Jensen, M.M. Oshiro, G.S. Watts, C.J. Kim, B.W. Futscher, Agglomerative epigenetic aberrations are a common event in human breast cancer, Cancer Res 68 (2008) 8616-8625.

[53] U.G. Sathyanarayana, A. Padar, C.X. Huang, M. Suzuki, H. Shigematsu, B.N. Bekele, A.F. Gazdar, Aberrant promoter methylation and silencing of laminin-5-encoding genes in breast carcinoma, Clin Cancer Res 9 (2003) 6389-6394.

[54] Y.K. Bae, A. Brown, E. Garrett, D. Bornman, M.J. Fackler, S. Sukumar, J.G. Herman, E. Gabrielson, Hypermethylation in histologically distinct classes of breast cancer, Clin Cancer Res 10 (2004) 5998-6005.

[55] J.M. Garcia, J. Silva, C. Pena, V. Garcia, R. Rodriguez, M.A. Cruz, B. Cantos, M. Provencio, P. Espana, F. Bonilla, Promoter methylation of the PTEN gene is a common molecular change in breast cancer, Genes Chromosomes Cancer 41 (2004) 117-124.

[56] S. Li, M. Rong, B. Iacopetta, DNA hypermethylation in breast cancer and its association with clinicopathological features, Cancer Lett 237 (2006) 272-280.

[57] P.S. Yan, M.R. Perry, D.E. Laux, A.L. Asare, C.W. Caldwell, T.H. Huang, CpG island arrays: an application toward deciphering epigenetic signatures of breast cancer, Clin Cancer Res 6 (2000) 1432-1438.

[58] W. Feng, L. Shen, S. Wen, D.G. Rosen, J. Jelinek, X. Hu, S. Huan, M. Huang, J. Liu, A.A. Sahin, K.K. Hunt, R.C. Bast, Jr., Y. Shen, J.P. Issa, Y. Yu, Correlation between CpG methylation profiles and hormone receptor status in breast cancers, Breast Cancer Res 9 (2007) R57.

[59] E. Sunami, M. Shinozaki, M.S. Sim, S.L. Nguyen, A.T. Vu, A.E. Giuliano, D.S. Hoon, Estrogen receptor and HER2/neu status affect epigenetic differences of tumor-related genes in primary breast tumors, Breast Cancer Res 10 (2008) R46.

[60] M. Widschwendter, K.D. Siegmund, H.M. Muller, H. Fiegl, C. Marth, E. Muller-Holzner, P.A. Jones, P.W. Laird, Association of breast cancer DNA methylation profiles with hormone receptor status and response to tamoxifen, Cancer Res 64 (2004) 3807-3813.

[61] K. Holm, C. Hegardt, J. Staaf, J. Vallon-Christersson, G. Jonsson, H. Olsson, A. Borg, M. Ringner, Molecular subtypes of breast cancer are associated with characteristic DNA methylation patterns, Breast Cancer Res 12 (2010) R36.

[62] J.A. Ronneberg, T. Fleischer, H.K. Solvang, S.H. Nordgard, H. Edvardsen, I. Potapenko, D. Nebdal, C. Daviaud, I. Gut, I. Bukholm, B. Naume, A.L. Borresen-Dale, J. Tost, V. Kristensen, Methylation profiling with a panel of cancer related genes: association with estrogen receptor, TP53 mutation status and expression subtypes in sporadic breast cancer, Mol Oncol 5 (2011) 61-76.

[63] I. Van der Auwera, W. Yu, L. Suo, L. Van Neste, P. van Dam, E.A. Van Marck, P. Pauwels, P.B. Vermeulen, L.Y. Dirix, S.J. Van Laere, Array-based DNA methylation profiling for breast cancer subtype discrimination, PLoS One 5 (2010) e12616.

[64] F. Fang, S. Turcan, A. Rimner, A. Kaufman, D. Giri, L.G. Morris, R. Shen, V. Seshan, Q. Mo, A. Heguy, S.B. Baylin, N. Ahuja, A. Viale, J. Massague, L. Norton, L.T. Vahdat, M.E. Moynahan, T.A. Chan, Breast cancer methylomes establish an epigenomic foundation for metastasis, Sci Transl Med 3 (2011) 75ra25.

[65] TCGA, Comprehensive molecular portraits of human breast tumours, Nature 490 (2012) 61-70.

[66] M. Curradi, A. Izzo, G. Badaracco, N. Landsberger, Molecular mechanisms of gene silencing mediated by DNA methylation, Mol Cell Biol 22 (2002) 3157-3173.

[67] P. Qiu, L. Zhang, Identification of markers associated with global changes in DNA methylation regulation in cancers, BMC Bioinformatics 13 Suppl 13 (2012) S7.

[68] J. Ihmels, S. Bergmann, N. Barkai, Defining transcription modules using large-scale gene expression data, Bioinformatics 20 (2004) 1993-2003.

[69] M. van Uitert, W. Meuleman, L. Wessels, Biclustering sparse binary genomic data, J Comput Biol 15 (2008) 1329-1345.

[70] R. Lister, J.R. Ecker, Finding the fifth base: genome-wide sequencing of cytosine methylation, Genome Res 19 (2009) 959-966.

[71] W.A. Pastor, U.J. Pape, Y. Huang, H.R. Henderson, R. Lister, M. Ko, E.M. McLoughlin, Y. Brudno, S. Mahapatra, P. Kapranov, M. Tahiliani, G.Q. Daley, X.S. Liu, J.R. Ecker, P.M. Milos, S. Agarwal, A. Rao, Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells, Nature 473 (2011) 394-397.

[72] P.A. Jones, Functions of DNA methylation: islands, start sites, gene bodies and beyond, Nat Rev Genet 13 (2012) 484-492.

[73] G. Cai, H. Li, Y. Lu, X. Huang, J. Lee, P. Muller, Y. Ji, S. Liang, Accuracy of RNA-Seq and its dependence on sequencing depth, BMC Bioinformatics 13 Suppl 13 (2012) S5.

[74] R.D. Canales, Y. Luo, J.C. Willey, B. Austermiller, C.C. Barbacioru, C. Boysen, K. Hunkapiller, R.V. Jensen, C.R. Knight, K.Y. Lee, Y. Ma, B. Maqsodi, A. Papallo, E.H. Peters, K. Poulter, P.L. Ruppel, R.R. Samaha, L. Shi, W. Yang, L. Zhang, F.M. Goodsaid, Evaluation of DNA microarray results with quantitative gene expression platforms, Nat Biotechnol 24 (2006) 1115-1122.

[75] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, J Roy Statist Soc B 57 (1995) 289-300.

[76] P.M. Chiang, J. Ling, Y.H. Jeong, D.L. Price, S.M. Aja, P.C. Wong, Deletion of TDP-43 down-regulates Tbc1d1, a gene linked to obesity, and alters body fat metabolism, Proc Natl Acad Sci U S A 107 (2010) 16320-16324.

[77] K.R. Coombes, OOMPA package.

[78] T. Fawcett, An introduction to ROC analysis, Pattern Recognition Letters 27 (2006) 861-874.

[79] L.J. Core, J.J. Waterfall, J.T. Lis, Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters, Science 322 (2008) 1845-1848.

[80] O. Harismendy, P.C. Ng, R.L. Strausberg, X. Wang, T.B. Stockwell, K.Y. Beeson, N.J. Schork, S.S. Murray, E.J. Topol, S. Levy, K.A. Frazer, Evaluation of next generation sequencing platforms for population targeted sequencing studies, Genome Biol 10 (2009) R32.

[81] T. Raz, P. Kapranov, D. Lipson, S. Letovsky, P.M. Milos, J.F. Thompson, Protocol dependence of sequencing-based gene expression measurements, PLoS One 6 (2011) e19287.

[82] J.M. Yi, M. Dhir, L. Van Neste, S.R. Downing, J. Jeschke, S.C. Glockner, M. de Freitas Calmon, C.M. Hooker, J.M. Funes, C. Boshoff, K.M. Smits, M. van Engeland, M.P. Weijenberg, C.A. Iacobuzio-Donahue, J.G. Herman, K.E. Schuebel, S.B. Baylin, N. Ahuja, Genomic and epigenomic integration identifies a prognostic signature in colon cancer, Clin Cancer Res 17 (2011) 1535-1545.

[83] P. Qiu, A.J. Gentles, S.K. Plevritis, Reducing the computational complexity of information theoretic approaches for reconstructing gene regulatory networks, J Comput Biol 17 (2010) 169-176.

[84] M.M. Gaudet, M. Campan, J.D. Figueroa, X.R. Yang, J. Lissowska, B. Peplonska, L.A. Brinton, D.L. Rimm, P.W. Laird, M. Garcia-Closas, M.E. Sherman, DNA hypermethylation of ESR1 and PGR in breast cancer: pathologic and epidemiologic associations, Cancer Epidemiol Biomarkers Prev 18 (2009) 3036-3043.

[85] L. Gao, M.A. Smit, J.J. van den Oord, J.J. Goeman, E.M. Verdegaal, S.H. van der Burg, M. Stas, S. Beck, N.A. Gruis, C.P. Tensen, R. Willemze, D.S. Peeper, R. van Doorn, Genome-wide promoter methylation analysis identifies epigenetic silencing of MAPK13 in primary cutaneous melanoma, Pigment Cell Melanoma Res 26 (2013) 542-554.

[86] A. Etcheverry, M. Aubry, M. de Tayrac, E. Vauleon, R. Boniface, F. Guenot, S. Saikali, A. Hamlat, L. Riffaud, P. Menei, V. Quillien, J. Mosser, DNA methylation in glioblastoma: impact on gene expression and clinical outcome, BMC Genomics 11 (2010) 701.

[87] M. Szyf, DNA demethylation and cancer metastasis: therapeutic implications, Expert Opin Drug Discov 3 (2008) 519-531.

[88] M. Szyf, P. Pakneshan, S.A. Rabbani, DNA demethylation and cancer: therapeutic implications, Cancer Lett 211 (2004) 133-143.

[89] C. Abate-Shen, Deregulated homeobox gene expression in cancer: cause or consequence?, Nat Rev Cancer 2 (2002) 777-785.

[90] N. Shah, S. Sukumar, The Hox genes and their roles in oncogenesis, Nat Rev Cancer 10 (2010) 361-371.

[91] M. Tessema, R. Willink, K. Do, Y.Y. Yu, W. Yu, E.O. Machida, M. Brock, L. Van Neste, C.A. Stidley, S.B. Baylin, S.A. Belinsky, Promoter methylation of genes in and around the candidate lung cancer susceptibility locus 6q23-25, Cancer Res 68 (2008) 1707-1714.

[92] D. Dietrich, R. Lesche, R. Tetzner, M. Krispin, J. Dietrich, W. Haedicke, M. Schuster, G. Kristiansen, Analysis of DNA methylation of multiple genes in microdissected cells from formalin-fixed and paraffin-embedded tissues, J Histochem Cytochem 57 (2009) 477-489.

[93] S. Mahapatra, E.W. Klee, C.Y. Young, Z. Sun, R.E. Jimenez, G.G. Klee, D.J. Tindall, K.V. Donkena, Global methylation profiling for risk prediction of prostate cancer, Clin Cancer Res 18 (2012) 2882-2895.

[94] T. Dunwell, L. Hesson, T.A. Rauch, L. Wang, R.E. Clark, A. Dallol, D. Gentle, D. Catchpoole, E.R. Maher, G.P. Pfeifer, F. Latif, A genome-wide screen identifies frequently methylated genes in haematological and epithelial cancers, Mol Cancer 9 (2010) 44.

[95] I. Mineva, W. Gartner, P. Hauser, A. Kainz, M. Loffler, G. Wolf, R. Oberbauer, M. Weissel, L. Wagner, Differential expression of alphaB-crystallin and Hsp27-1 in anaplastic thyroid carcinomas because of tumor-specific alphaB-crystallin gene (CRYAB) silencing, Cell Stress Chaperones 10 (2005) 171-184.

[96] O.H. Kwon, J.L. Park, M. Kim, J.H. Kim, H.C. Lee, H.J. Kim, S.M. Noh, K.S. Song, H.S. Yoo, S.G. Paik, S.Y. Kim, Y.S. Kim, Aberrant up-regulation of LAMB3 and LAMC2 by promoter demethylation in gastric cancer, Biochem Biophys Res Commun 406 (2011) 539-545.

[97] U.G. Sathyanarayana, R. Maruyama, A. Padar, M. Suzuki, J. Bondaruk, A. Sagalowsky, J.D. Minna, E.P. Frenkel, H.B. Grossman, B. Czerniak, A.F. Gazdar, Molecular detection of noninvasive and invasive bladder tumor tissues and exfoliated cells by aberrant promoter methylation of laminin-5 encoding genes, Cancer Res 64 (2004) 1425-1430.

[98] U.G. Sathyanarayana, A. Padar, M. Suzuki, R. Maruyama, H. Shigematsu, J.T. Hsieh, E.P. Frenkel, A.F. Gazdar, Aberrant promoter methylation of laminin-5-encoding genes in prostate cancers and its relationship to clinicopathological features, Clin Cancer Res 9 (2003) 6395-6400.

[99] J.R. Jass, Classification of colorectal cancer based on correlation of clinical, morphological and molecular features, Histopathology 50 (2007) 113-130.

[100] T. Kawasaki, M. Ohnishi, K. Nosho, Y. Suemoto, G.J. Kirkner, J.A. Meyerhardt, C.S. Fuchs, S. Ogino, CpG island methylator phenotype-low (CIMP-low) colorectal cancer shows not only few methylated CIMP-high-specific CpG islands, but also low-level methylation at individual loci, Mod Pathol 21 (2008) 245-255.

[101] S. Ogino, T. Kawasaki, G.J. Kirkner, M. Loda, C.S. Fuchs, CpG island methylator phenotype-low (CIMP-low) in colorectal cancer: possible associations with male sex and KRAS mutations, J Mol Diagn 8 (2006) 582-588.

[102] J.M. Teodoridis, C. Hardie, R. Brown, CpG island methylator phenotype (CIMP) in cancer: causes and implications, Cancer Lett 268 (2008) 177-186.

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

[104] S. Tommasi, D.L. Karm, X. Wu, Y. Yen, G.P. Pfeifer, Methylation of homeobox genes is a frequent and early epigenetic event in breast cancer, Breast Cancer Res 11 (2009) R14.

[105] S.H. Shirley, J.E. Rundhaug, J. Tian, N. Cullinan-Ammann, I. Lambertz, C.J. Conti, R. Fuchs-Young, Transcriptional regulation of estrogen receptor-alpha by p53 in human breast cancer cells, Cancer Res 69 (2009) 3405-3414.

[106] S.T. Bailey, H. Shin, T. Westerling, X.S. Liu, M. Brown, Estrogen receptor prevents p53-dependent apoptosis in breast cancer, Proc Natl Acad Sci U S A 109 (2012) 18060-18065.

[107] C.E. Berger, Y. Qian, G. Liu, H. Chen, X. Chen, p53, a target of estrogen receptor (ER) alpha, modulates DNA damage-induced growth suppression in ER-positive breast cancer cells, J Biol Chem 287 (2012) 30117-30127.

[108] S.D. Konduri, R. Medisetty, W. Liu, B.A. Kaipparettu, P. Srivastava, H. Brauch, P. Fritz, W.M. Swetzig, A.E. Gardner, S.A. Khan, G.M. Das, Mechanisms of estrogen receptor antagonism

toward p53 and its implications in breast cancer therapeutic response and stem cell regulation, Proc Natl Acad Sci U S A 107 (2010) 15081-15086.

[109] A.M. Pietersen, H.M. Horlings, M. Hauptmann, A. Langerod, A. Ajouaou, P. Cornelissen-Steijger, L.F. Wessels, J. Jonkers, M.J. van de Vijver, M. van Lohuizen, EZH2 and BMI1 inversely correlate with prognosis and TP53 mutation in breast cancer, Breast Cancer Res 10 (2008) R109.

[110] B.H. Guo, Y. Feng, R. Zhang, L.H. Xu, M.Z. Li, H.F. Kung, L.B. Song, M.S. Zeng, Bmi-1 promotes invasion and metastasis, and its elevated expression is correlated with an advanced stage of breast cancer, Mol Cancer 10 (2011) 10.

[111] C.H. Hsu, K.L. Peng, M.L. Kang, Y.R. Chen, Y.C. Yang, C.H. Tsai, C.S. Chu, Y.M. Jeng, Y.T. Chen, F.M. Lin, H.D. Huang, Y.Y. Lu, Y.C. Teng, S.T. Lin, R.K. Lin, F.M. Tang, S.B. Lee, H.M. Hsu, J.C. Yu, P.W. Hsiao, L.J. Juan, TET1 suppresses cancer invasion by activating the tissue inhibitors of metalloproteinases, Cell Rep 2 (2012) 568-579.

[112] B. Shi, J. Liang, X. Yang, Y. Wang, Y. Zhao, H. Wu, L. Sun, Y. Zhang, Y. Chen, R. Li, M. Hong, Y. Shang, Integration of estrogen and Wnt signaling circuits by the polycomb group protein EZH2 in breast cancer cells, Mol Cell Biol 27 (2007) 5105-5119.

[113] M.E. Valk-Lingbeek, S.W. Bruggeman, M. van Lohuizen, Stem cells and cancer; the polycomb connection, Cell 118 (2004) 409-418.

[114] G. Honeth, P.O. Bendahl, M. Ringner, L.H. Saal, S.K. Gruvberger-Saal, K. Lovgren, D. Grabau, M. Ferno, A. Borg, C. Hegardt, The CD44+/CD24- phenotype is enriched in basal-like breast tumors, Breast Cancer Res 10 (2008) R53.

[115] E. Lim, F. Vaillant, D. Wu, N.C. Forrest, B. Pal, A.H. Hart, M.L. Asselin-Labat, D.E. Gyorki, T. Ward, A. Partanen, F. Feleppa, L.I. Huschtscha, H.J. Thorne, S.B. Fox, M. Yan, J.D. French, M.A. Brown, G.K. Smyth, J.E. Visvader, G.J. Lindeman, Aberrant luminal progenitors as the candidate target population for basal tumor development in BRCA1 mutation carriers, Nat Med 15 (2009) 907-913.

[116] S.L. Squazzo, H. O'Geen, V.M. Komashko, S.R. Krig, V.X. Jin, S.W. Jang, R. Margueron, D. Reinberg, R. Green, P.J. Farnham, Suz12 binds to silenced regions of the genome in a cell-type-specific manner, Genome Res 16 (2006) 890-900.

[117] E. Appella, C.W. Anderson, Post-translational modifications and activation of p53 by genotoxic stresses, Eur J Biochem 268 (2001) 2764-2772.

[118] S.P. Hussain, C.C. Harris, p53 biological network: at the crossroads of the cellular-stress response pathway and molecular carcinogenesis, J Nippon Med Sch 73 (2006) 54-64.

[119] J. Tan, X. Yang, L. Zhuang, X. Jiang, W. Chen, P.L. Lee, R.K. Karuturi, P.B. Tan, E.T. Liu, Q. Yu, Pharmacologic disruption of Polycomb-repressive complex 2-mediated gene repression selectively induces apoptosis in cancer cells, Genes Dev 21 (2007) 1050-1063.

VITA

Guoshuai Cai, the son of Yiming Cai and Zhonghua Yang, was born on January 9, 1984 in Arong Qi, Nei Mongol, the People's Republic of China. He graduated from Zhalantun No.1 High School in 2002. Then he attended Wuhan University in Wuhan, Hubei, China and obtained a Bachelor's of Science degree in biotechnology in 2006. From the same school, he earned a Master's of Science degree in microbiology in 2009. In August 2009, he entered the Graduate School of Biomedical Sciences at The University of Texas Health Science Center at Houston and The University of Texas MD Anderson Cancer Center. He expects to earn a Doctor of Philosophy degree in bioinformatics and biostatistics in August 2013.