2010

Inferential Methods for High-Throughput Methylation Data

Maria Capparuccini
Virginia Commonwealth University

Follow this and additional works at: https://scholarscompass.vcu.edu/etd

Part of the Biostatistics Commons

© The Author

Downloaded from https://scholarscompass.vcu.edu/etd/156

This Dissertation is brought to you for free and open access by the Graduate School at VCU Scholars Compass. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of VCU Scholars Compass. For more information, please contact libcompass@vcu.edu.
INFERENTIAL METHODS FOR HIGH-THROUGHPUT METHYLATION DATA

A Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at Virginia Commonwealth University.

by

MARIA I. CAPPARUCCINI
Bachelor of Arts in Mathematics, University of Virginia, 2005

Director: KELLIE J. ARCHER
ASSOCIATE PROFESSOR
DEPARTMENT OF BIOSTATISTICS

Virginia Commonwealth University
Richmond, Virginia
December 2010
Acknowledgement

I am so thankful to both Dr. Archer and Dr. Gennings for the roles that they have had throughout my graduate studies. To Dr. Archer, I am so appreciative for the patient and constant guidance she gave me both in my classroom work and throughout my research that has indeed made my graduate experience rewarding. To Dr. Gennings, I am so honored that she invited me as a first year transfer student to join her training grant.

In all things I do, I give thanks and glory to God. I am so thankful for my greatest gifts, my husband David, our daughter Irene, and our baby on the way. Your unwavering love, encouragement and joy bring me so much happiness. A sincere thank you, as well, for my parents, David’s parents and our siblings for all their words of encouragement and prayers. I especially thank my mom who provided many hours of babysitting. While I know there are few things she would rather do than spend time with her granddaughter, I thank her for selflessly rearranging her schedule so many times to accommodate mine.

Last, but not least, thank you to each of my committee members. I am so excited and grateful to you both for agreeing to be on my committee and for your guidance.
# Table of Contents

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Overview of DNA Methylation and the Illumina Technology</td>
</tr>
<tr>
<td></td>
<td>1.1 Genetic Packaging and Epigenetic Modifications</td>
</tr>
<tr>
<td></td>
<td>1.2 DNA Methylation: An Overview</td>
</tr>
<tr>
<td></td>
<td>1.3 Aberrant DNA Methylation</td>
</tr>
<tr>
<td></td>
<td>1.4 The Role of Environmental toxins in aberrant DNA methylation</td>
</tr>
<tr>
<td></td>
<td>1.5 The Illumina Technology</td>
</tr>
<tr>
<td>2</td>
<td>Current Methods for Analyzing DNA Methylation Data</td>
</tr>
<tr>
<td></td>
<td>2.1 Hepatitis C Virus (HCV) Data Set</td>
</tr>
<tr>
<td></td>
<td>2.2 Analysis of Methylation Data</td>
</tr>
<tr>
<td></td>
<td>2.3 Hypothesis testing using the two sample t-test</td>
</tr>
<tr>
<td></td>
<td>2.4 Filtered Two Sample T-Test</td>
</tr>
<tr>
<td></td>
<td>2.5 Limitations of current analysis methods</td>
</tr>
<tr>
<td>3</td>
<td>Determination of Minimum Level of Discernable Differential Methylation</td>
</tr>
</tbody>
</table>

Acknowledgements: ii  
List of Tables: vii  
List of Figures: ix
3.1 Motivation and Methods: Hierarchical model building to Bibikova et al. dilution data .................................................................20

3.1(a) Nonlinear Mixed Model Formulation: Logistic Regression Model ........................................................................................24

3.1(b) Model Selection .................................................................................................................................................................26

3.1(c) Establishment of a significant relationship between Proportion of Methylation and Mixing Ratio .................................................30

3.2 Establishing a Minimum Level of Detectable Differentiation ..........31

3.3 Verification of fixed asymptote assumption .....................................35

3.3(a) Nonlinear Mixed Model Formulation: Asymptotic Regression Model ..................................................................................36

3.3(b) Nonlinear Mixed Model Formulation: Logistic Regression Model ..................................................................................................37

3.3(c) Model Selection .................................................................................................................................................................38

3.3(d) Predicted CpG site specific response asymptotes ........................................46

3.3 Concluding Remarks .........................................................................................................................................................48

4 Proposed Two Group Inferential Methods for Testing Proportion

Methylated ........................................................................................................50

4.1 Alternative Analysis Exploration ...............................................................50

4.2 Kolmogorov-Smirnov Test ........................................................................53
4.3 Testing for Differentially Methylated CpG sites:

Beta Distribution Motivated Two-Sample Test ............................................55

4.3(a) Test Based on a Beta Distributed Random Variable ..........................55

4.3(b) Likelihood Ratio Test ..................................................................57

4.3(c) Nelder Mead Simplex .................................................................60

4.4 Testing for Differentially Methylated CpG sites:

Bivariate Normal Distribution Motivated Test ..........................................61

5 Simulation Study......................................................................................65

5.1 Testing Type I error rates ...................................................................66

5.1(a) Type 1 error data simulation methods and results:

Beta distribution .........................................................................................67

5.1(b) Type 1 error data simulation methods and results:

Bivariate Normal distribution .................................................................69

5.2 Testing Type II Error Rates ..............................................................71

5.2(a) Type II error data simulation methods and results:

Beta distribution .........................................................................................73

5.2(b) Type II error data simulation methods and results:

Bivariate Normal distribution .................................................................80

5.3 Application of the Beta Test to Hepatitis C Virus data set ...............85
5.4 Conclusions: Determination of a well-defined statistical testing measure ................................................................. 86

6 Conclusions and Recommendations ................................................................. 90

Literature Cited ........................................................................................................ 95

Appendix A Table of CpG sites found significant using test based on the Beta Distribution ................................................................. 99

Appendix B Source Code for the construction of CpG site specific nonlinear mixed models in Chapter 3 ................................................................. 109

Appendix C Source code for the establishment of a minimum level of detectable differential level of methylation ......................................... 112

Appendix D Source code for Chapter 3.3 CpG site specific hierarchical models ........ 115

Appendix E Source code for Chapter 5 data simulations for the computation of Type I error under the assumption of Beta distributed data .......... 119

Appendix F Source code for Chapter 5 data simulations for the computation of Type I error under the assumption of Bivariate Normal distributed data ........................................ 130

Appendix G Source code for Chapter 5 data simulations for the computation of Type II error under the assumption of Beta distributed data .......... 142
Appendix H  Source code for Chapter 5 data simulations for the computation of Type II error under the assumption of Bivariate Normal distributed data ................................................................. 156

VITA ........................................................................................................................................ 172
# List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Expected approximate methylation level for X linked genes given mixing ratio</td>
<td>19</td>
</tr>
<tr>
<td>3.2</td>
<td>Correlation among 17 CpG sites</td>
<td>24</td>
</tr>
<tr>
<td>3.3</td>
<td>Pairwise comparisons across adjacent mixing ratios</td>
<td>34</td>
</tr>
<tr>
<td>3.4</td>
<td>AIC comparison by CpG site between Asymptotic and Logistic Regression Models</td>
<td>39</td>
</tr>
<tr>
<td>3.5</td>
<td>Results of Likelihood Ratio test to determine the significance of a random effect for $\phi_{zmj}$ parameter in Asymptotic Regression Model</td>
<td>41</td>
</tr>
<tr>
<td>3.6</td>
<td>Point estimates and 95% confidence intervals of predicted proportion methylated for mixtures demonstrating 100% Female DNA</td>
<td>47</td>
</tr>
<tr>
<td>5.1</td>
<td>Type I Error rate comparison where data has been computed under the assumption of a Beta distribution</td>
<td>68</td>
</tr>
<tr>
<td>5.2</td>
<td>Type I Error rate comparison across varying sample sizes where data have been computed under the assumption of a Beta distribution</td>
<td>69</td>
</tr>
<tr>
<td>5.3</td>
<td>Type I Error rate comparison where data has been computed under the assumption of a Bivariate Normal distribution</td>
<td>70</td>
</tr>
<tr>
<td>5.4</td>
<td>Type I Error rate comparison across varying sample sizes where data have been computed under the assumption of a Bivariate Normal distribution</td>
<td>71</td>
</tr>
<tr>
<td>5.5</td>
<td>Type II Error rate comparison where data has been computed under the assumption of a Beta distribution</td>
<td>75</td>
</tr>
<tr>
<td>5.6</td>
<td>Type II Error rate comparison across varying levels of differential proportion methylated where data have been generated with 20 samples in outcome group 1 and 16 samples in outcome group 2 according to a Beta distribution</td>
<td>78</td>
</tr>
</tbody>
</table>
Table 5.7: Type II Error rate comparison across varying levels of differential proportion methylated where data have been generated with 40 samples each in outcome groups 1 and 2 according to a Beta distribution ........................................78

Table 5.8: Type II Error rate comparison across varying levels of differential proportion methylated where data have been generated with 80 samples each in outcome groups 1 and 2 according to a Beta distribution ........................................80

Table 5.9: Type II Error rate comparison across varying levels of differential proportion methylated where data have been generated with 20 samples in outcome group 1 and 16 samples in outcome group 2 according to a Bivariate Normal distribution ................................................................................................................81

Table 5.10: Type II Error rate comparison across varying levels of differential proportion methylated where data have been generated with 20 samples in outcome group 1 and 16 samples in outcome group 2 according to a Bivariate Normal distribution ................................................................................................................83

Table 5.11: Type II Error rate comparison across varying levels of differential proportion methylated where data have been generated with 40 samples each in outcome groups 1 and 2 according to a Bivariate Normal distribution ........................................84

Table 5.12: Type II Error rate comparison across varying levels of differential proportion methylated where data have been generated with 80 samples each in outcome groups 1 and 2 according to a Beta distribution ........................................84
List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Boxplot of proportion methylation by percent of female genomic DNA in the hybridized sample</td>
<td>21</td>
</tr>
<tr>
<td>3.2</td>
<td>Mean proportion methylated level across CpG site by percent of female genomic DNA in sample for the first dilution order (top panel) and the second dilution order (bottom panel)</td>
<td>22</td>
</tr>
<tr>
<td>3.3</td>
<td>Proportion methylated level by Percent of female genomic DNA for each CpG site (1-17) within dilution order (1-2)</td>
<td>22</td>
</tr>
<tr>
<td>3.4</td>
<td>Standardized residuals plotted as a function of fitted values for the original Logistic regression model</td>
<td>27</td>
</tr>
<tr>
<td>3.5</td>
<td>Standardized residuals plotted as a function of fitted values using an exponential variance structure</td>
<td>28</td>
</tr>
<tr>
<td>3.6</td>
<td>Mean fitted proportion methylated curve overlaid upon observed proportion methylated values</td>
<td>29</td>
</tr>
<tr>
<td>3.7</td>
<td>Standardized residuals versus fitted values in final mixed ANOVA model</td>
<td>33</td>
</tr>
<tr>
<td>3.8</td>
<td>Standardized residuals plotted as a function of fitted values for the 17 CpG site-specific models</td>
<td>43</td>
</tr>
<tr>
<td>3.9</td>
<td>Mean fitted proportion methylated curve overlaid upon observed proportion methylated values for each mixing ratio level by CpG site</td>
<td>45</td>
</tr>
<tr>
<td>4.1</td>
<td>Histogram of Green (top panel) and Red (bottom panel) intensities for a representative beadarray</td>
<td>51</td>
</tr>
<tr>
<td>4.2</td>
<td>Histogram of Green uncorrected (top panel), Green corrected, and Red (bottom panel) channel intensities for a representative beadarray. Green corrected intensities were post-constant-scale normalized</td>
<td>52</td>
</tr>
</tbody>
</table>
The role of abnormal DNA methylation in the progression of disease is a growing area of research that relies upon the establishment of sound statistical methods. The common method for declaring there is differential methylation between two groups at a given CpG site, as summarized by the difference between proportions methylated \( \Delta \beta = \beta_1 - \beta_2 \), has been through use of a Filtered Two Sample t-test, using the recommended filter of 0.17 (Bibikova et al., 2006b). In this dissertation, we performed a re-analysis of the data used in recommending the threshold by fitting a mixed-effects ANOVA model. It was determined that the 0.17 filter is not accurate and conjectured that application of a Filtered
Two Sample t-test likely leads to loss of power. Further, the Two Sample t-test assumes that data arise from an underlying distribution encompassing the entire real number line, whereas $\beta_1$ and $\beta_2$ are constrained on the interval $[0,1]$. Additionally, the imposition of a filter at a level signifying the minimum level of detectable difference to a Two Sample t-test likely reduces power for smaller but truly differentially methylated CpG sites. Therefore, we compared the Two Sample t-test and the Filtered Two Sample t-test, which are widely used but largely untested with respect to their performance, to three proposed methods. These three proposed methods are a Beta distribution test, a Likelihood ratio test, and a Bootstrap test, where each was designed to address distributional concerns present in the current testing methods. It was ultimately shown through simulations comparing Type I and Type II error rates that the (unfiltered) Two Sample t-test and the Beta distribution test performed comparatively well.
CHAPTER 1 Overview of DNA Methylation and the Illumina Technology

Cancer is a class of diseases marked by a continuing increase in aberrant gene function. This aberrant gene function likely arises from either a genetic mutation or an epigenetic change. An epigenetic change is a modification to genomic DNA that has only implications in gene expression capacity but leaves the primary sequence of the DNA unchanged. Research has long recognized the role of genetic mutations, whether inherited through the germ line or arising in somatic tissues later in life, in contributing to cancer. However, more recent studies evidencing their role in the progression of tumors have brought epigenetic changes to the forefront of cancer research (Jones, 2002).

1.1 Genetic Packaging and Epigenetic Modifications

To understand the role of epigenetic changes in contribution to cancer, one must first take a deeper look at the mechanisms of genetic packaging and expression. The genetic information of eukaryotic cells is stored as DNA in their nucleus. The enormous amount of DNA is compactly stored in the nucleus of each cell by means of a complex combination with proteins to form chromatin. The fundamental repeating storage unit of this chromatin is called a nucleosome; each nucleosome contains approximately 146 base pairs of DNA wrapped around it. In addition to providing a compact storage mechanism
of DNA, chromatin also preserves the appropriate access to it through the interaction of the nucleosome with specific protein complexes that associate with, manipulate and epigenetically modify the nucleosome (Rountree, 2001). These epigenetic modifications, while not affecting the primary structure of the genetic code, can have implications in the secondary interactions that are critical to the regulation of gene expression. The interaction of specific interest here is transcription status.

Transcription is the process of forming an equivalent RNA copy of a sequence of DNA (MedicineNet.com, 2010). The regulation of transcription by means of an epigenetic modification could have the ability to alter a wide range of gene function from high-level expression to complete silencing (Rountree, 2001). Two epigenetic modifications, in particular, have garnered much interest in their inverse roles in transcriptional regulation and hence in their role in the progression of cancer. The first epigenetic modification, histone acetylation involves the acetylation of the amino-terminal tails of histones H3 and H4 by histone acetyltransferases (HATs) and is implicated in transcriptional activation. The second epigenetic modification, which is of primary concern for study here, is DNA methylation.

1.2 DNA Methylation: An Overview

DNA methylation is the addition of a methyl group to the 5′-carbon of a cytosine. The only methylated base in human DNA is a 5-methylcytosine located in a cytosine-guanine (CG) dinucleotide, also referred to as a CpG site. DNA methylation occurs by
means of an S-adenosyl-methionine acting as a methyl donor by means of adding a methyl
group to the cytosine ring to form methyl cytosine (Eng, 2000). This reaction is catalyzed
by DNA methyltransferases (DNMTs), of which there are three types: Dnmt1, Dnmt3a,
and Dnmt3b (Rountree, 2001). Through the interplay of these three enzymes and their
associated factors which target and regulate their enzymatic activity, genomic methylation
patterns are established during embryogenesis.

CpG sites, and consequently DNA methylation sites, are not distributed evenly
throughout the genome. Rather, CpG sites are relatively scarce throughout the genome as
a whole, and they are clustered in regions of the DNA strand called “CpG islands,”
broadly defined as CG rich regions of DNA (Herman, 2003). More specifically, CpG
islands are characterized as sequences longer than 200 base pairs in length with a CG
dinucleotide content of greater than 50% and an observed over expected ratio of CpG
dinucleotides of at least 0.6 (Gardiner-Garden and Frommer, 1987, Takai and Jones, 2002).
CpG islands are primarily found in the 5′-regions of housekeeping genes, spanning from
the promoter region to the first exon. The rationale behind the existence of CpG islands is
as follows: it has been shown that most (~80%) of CpG dinucleotides that are located
outside the bounds of a CpG island are methylated (Herman, 2003). Over time, this
methylation has reduced the overall frequency of CpGs in the genome as a whole due to
the fact that methylated cytosines have an increased vulnerability to deamination (Bird,
1980). On the other hand, most of the CpG dinucleotides within the bounds of CpG
islands are unmethylated, rendering them less vulnerable to deamination, hence preserving
the expected frequency of CpG dinucleotides in these regions (Rountree, 2001). However,
while most CpG islands are normally unmethylated, areas of dense DNA methylation do naturally occur in these regions and are primarily associated with germ line differentiation through the inactivation of genes on the X-chromosome in somatic cells and with the silencing of alleles of imprinted genes.

1.3 Aberrant DNA methylation

The condition of study here for its role in the progression of cancer was aberrant DNA methylation and its effect on the transcription of RNA. Abnormal DNA methylation in gene promoter regions is shown to interfere in the transcription of DNA to RNA, ultimately suppressing gene expression (Michalowsky, 1989). When gene expression is suppressed in a gene responsible for inhibiting the progression to cancer, like the tumor suppressor gene, the chance for the development of a cancerous lesion increases. The interaction of methylcytosine binding proteins with other structural components of the chromatin renders the DNA inaccessible to transcription factors through histone deacetylation and chromatin structure changes, thereby accomplishing this gene silencing effect.

1.4 The role of environmental toxins in aberrant DNA methylation

While its exact cause is not always known, DNA methylation can be either heritable or as a novel response to toxins (Dolinoy et al., 2007; Illumina, 2006; Liu et al.,
The incidence of aberrant DNA methylation as a response to environmental toxins is well documented, and three recent and important studies are here highlighted.

In Dolinoy et al. (2007), the effects of maternal exposure to the estrogenic xenobiotic chemical bisphenol A (BPA) on the fetal epigenome was explored by studying the Agouti gene in viable yellow agouti mice. The Agouti gene controls hair color and was studied because it is a metastable epiallele, an allele that, owing to differential methylation patterns, can be expressed variably in genetically identical individuals. Female mice were fed food supplemented with BPA beginning two weeks prior to mating through pregnancy and lactation. Dolinoy et al. examined the coat color of the litter and measured DNA methylation at nine sites in the Agouti gene. Compared to a control litter, the litter of mice exposed maternally to BPA exhibited a different distribution of coat color that was skewed towards a lighter coat. Furthermore, they showed that BPA exposure reduced the percentage of cells with methylation at the nine measured sites in the Agouti gene. It was also found that greater variability in methylation patterns existed between animals than within a range of tissue types of a single mouse which suggested that the methylation patterns were likely to have been established early in the embryonic development.

As a second example of the role of an environmental toxin in aberrant DNA methylation we turn to Weihrauch et al. (2001). Here, the mechanism of the effect of Vinyl chloride (VC) exposure on hepatocellular carcinomas (HCC) was examined. VC is a colorless toxic gas widely employed as a refrigerant and intermediate in organic synthesis. It is well-documented both that in VC-associated liver angiosarcomas (LAS), mutations of the K-ras-2 gene exist and that K-ras-2 mutations induce p16 methylation.
accompanied by the inactivation of the \( p16 \) gene, i.e., the tumor suppressor gene. Weihrauch et al. hypothesized that there was a similar relation between VC-associated HCC and mutations of the K-\( ras \)-2 gene and subsequent methylation of the \( p16 \) gene. In comparing a sample of patients with VC associated HCCs against a control group of patients with HCC due to other demonstrated factors, it was determined that there was indeed an increased proportion of K-\( ras \)-2 gene mutation and abnormal hypermethylation of the \( p16 \) gene in the test sample. This suggested that exposure to the toxin VC does, in fact, play an important role in the pathogenesis of VC-associated HCC by means of K-\( ras \)-2 mutations and resulting \( p16 \) gene methylation.

Finally, in Liu et al. (2008), the effect of combined exposure to the inhaled diesel exhaust particles toxin and a fungus allergen Aspergillus fumigatus on methylation levels was observed. Resulting from a 3-week course of inhaled diesel exhaust particle exposure while undergoing intranasal sensitization to A. fumigates, hypermethylation was induced at CpG\(^{-45}\), CpG\(^{-53}\), and CpG\(^{-205}\) sites of the cytokine interferon (IFN)-\( \gamma \) promoter and hypomethylation at CpG\(^{-408}\) of the T helper (Th) 2 interleukin (IL)-4 cytokine promoter. The induced methylation changes of the promoters in both genes were significantly correlated with changes in immunoglobulin (Ig) E levels. While much research has demonstrated both the role of Th 2 IL-4 cytokines in promoting allergic sensitization and asthma and the role of Th1 cytokine IFN-\( \gamma \) in protecting against allergic sensitization and asthma, this study was the first to show that inhaled environmental exposures can affect asthma pathogenesis by altering methylation patterns of Th genes \textit{in vivo}. 
Therefore, as demonstrated by the aforementioned articles, aberrant DNA methylation is a research topic of current and specific interest. DNA methylation has been implicated in several cancers ranging from breast cancer to Acute Lymphoblastic Leukemia and colon cancer by means of silencing the tumor suppressor gene (Yan et al., 2000; Roman-Gomez et al., 2002; Suzuki et al., 2002). In this thesis, aberrant DNA methylation was examined as it relates the hepatocellular carcinoma due to hepatitis C infection.

1.5 The Illumina Technology

Of particular interest here was the detection of methylation sites for the quantitative analysis of differential methylation patterns. Thus, the need for a technology to accurately and completely assess the methylation status of the large number of existing CpG sites arises. The GoldenGate Methylation BeadArray Cancer Panel I platform is technology developed by Illumina for cancer-focused methylation analysis. It enables the simultaneous analysis of up to 1,505 CpG loci selected from 807 genes, and it runs two or more CpG assays for over 74% of these genes (Illumina, 2006). This GoldenGate BeadArray begins with a bisulfite treatment to DNA that converts unmethylated cytosines to uracils while leaving methylated cytosines unchanged. This bisulfite treated genomic DNA is then immobilized on paramagnetic beads. Then, for each CpG site, an allele-specific oligonucleotide and a locus specific oligonucleotide are assembled for each the methylated and unmethylated state, where separate labels (red and green, respectively) are
used to design each of the two states. Thus, two pairs of probes are designed for each site. For all loci simultaneously, the pooled oligonucleotides are assayed by annealing to the target sequence and then washed to remove excess or mishybridized oligonucleotides. Extension of the allele specific oligonucleotides to the locus specific oligonucleotides and subsequent ligation creates PCR templates, which are amplified with fluorescently labeled universal primers. The resulting products are ultimately hybridized to a beadarray bearing the complementary address sequences. The fluorescent expression for each the methylated and unmethylated states are quantified and reported as two channel array data. Therefore, each array consists of expression values for the methylated target sequence (Red) as well as expression values for the unmethylated target sequence (Green).

The Illumina assay represents each CpG site as a specific beadtype, and it incorporates approximately 30 beads (e.g. oligonucleotides) per beadtype to ensure the assay’s reproducibility. As some examples, the Illumina GoldenGate BeadArray platform has been used in studies to show that human embryonic stem cells are distinguished from all other cell types by their methylation profile (Bibikova et al., 2006a), to differentiate methylation patterns between lung adenocarcinomas and normal lung tissues (Bibikova et al., 2006b), to identify genes that show aberrant DNA methylation in hematologic neoplasms (Martin-Subero et al., 2009), and to identify CpG loci exhibiting de novo DNA methylation in two molecular subtypes of diffuse large B-cell lymphoma (Wang et al., 2010).
CHAPTER 2 Current Methods for Analyzing DNA Methylation Data

In this chapter, we introduce the application dataset that motivated this research, identifying differentially methylated CpG sites in HCV-induced hepatocellular carcinoma. We then describe the two current inferential methods commonly applied when comparing two groups with respect to DNA methylation, the Two Sample t-test and the Filtered Two Sample t-test. At the conclusion of this chapter, we also highlight their perceived deficiencies.

2.1 Hepatitis C Virus (HCV) Data Set

Hepatitis C is an infectious disease caused by the Hepatitis C virus (HCV). The Hepatitis C virus is an RNA virus belonging to the family of flaviviruses, which primarily targets the hepatocytes, the main tissue of the liver (Lauer, 2001). HCV is spread by blood to blood contact, most often through injection drug use, blood transfusion (although increasingly unlikely), and in rare cases, sexual contact. Acute Hepatitis C, referring to the first six months after infection with HCV, is relatively asymptomatic, with only vague symptoms including jaundice, malaise, and nausea. This silent onset of HCV makes the assessment of this infection difficult. In the majority of cases, the Hepatitis C virus produces a chronic infection lasting longer than six months. Most chronic infections
result in some degree of fibrosis of the liver, while 15 to 20 percent of those infected develop cirrhosis.

The time frame in which the various stages of liver disease develop, as a result of HCV, varies widely from patient to patient. In approximately one third of patients, serious liver disease develops 20 years or less after infection, while in another third, serious liver disease takes 30 years or longer to develop. Several factors have been proposed to accelerate the clinical progression of HCV including alcohol intake, coinfection with HIV-1 or HBV, male gender, or an older age at infection. However, once cirrhosis is established, the risk of hepatocellular carcinoma (HCC) is approximately 1 to 4 percent per year. Broadly speaking, treatment for HCV infection consists of a combination therapy of interferon and ribavirin, although liver transplantation remains the only viable treatment option for patients with either decompensated HCV-related cirrhosis or early stage HCC. Current studies, such as the methylation data set used here, focus on identifying molecular events that may be useful in the early diagnosis of the progression from HCV cirrhosis to HCC.

This research was motivated by a subset of a published methylation data set (Archer et al., 2010) which is publicly available at Gene Expression Omnibus, accession number GSE 18081 (Edgar et al., 2002; Barrett et al., 2009). The subset considered in this thesis consists of methylation data for 36 patients with either HCV cirrhosis without concomitant Hepatocellular Carcinoma (HCC) (N=16) or HCV cirrhosis with concomitant HCC (N=20). Each patient’s DNA sample was prepared according to the
Illumina technology with primers for the Methylated (Red) and Unmethylated (Green) channels.

2.2 Analysis of Methylation Data

In methylation studies such as the one studied here, data are presented as expression levels for both the methylated (Red) and unmethylated (Green) channels across multiple samples for a given collection of CpG sites. The data are further divided between samples originating from two groups of patients, one group of which presents a certain disease. Then, it is of clinical interest to analyze differences in methylation patterns between the two groups. Here, the first group was comprised of the samples which demonstrated HCV cirrhosis without concomitant HCC versus the second group that was comprised of samples which demonstrated HCV cirrhosis with concomitant HCC. These CpG sites, once identified, would be marked for future study as locations which could factor in the progression of HCV cirrhosis to the development of HCC.

All of the raw data were read into the R programming environment. Neither background correction nor normalization steps were implemented on this initial read-in. Bead summary data were created according to the default method used by Illumina as follows:

First, beads which had intensities greater than 3 median absolute deviations (MADs) from the bead median intensity on the original (un-logged) scale were removed for each bead type on each array. Once outliers were removed, the mean intensity of the
remaining beads was calculated for each the red and green channels using the unlogged scale. More specifically, for each array \((i)\) and beadtype (i.e. CpG site) \((j)\), beadtype expression for the red (methylated) and green (unmethylated) channels was estimated by averaging the intensities over the beads within the beadtype,

\[
R_{ij} = \frac{1}{k} \sum_{k} r_{ijk}, \quad G_{ij} = \frac{1}{k} \sum_{k} g_{ijk}
\]

where \(i=1,2,\ldots,36\) indexes the samples, \(j=1,2,\ldots,1624\) indexes the beadtypes, \(k\) indexes the number of beads per beadtype \((k \approx 30)\), and \(r_{ijk}\) and \(g_{ijk}\) are the red and green bead intensities, respectively, for the \(i\)th sample, \(j\)th beadtype, and \(k\)th bead.

Illumina BeadArrays contain approximately 77 blank beadtypes which do not have associated gene annotation data. Therefore only those 1,547 beadtypes with associated gene information files were retained for future analyses. Note that as previously stated, the Illumina platform enables the simultaneous analysis of up to 1,505 CpG loci, not including control beadtypes. Thus these 1,547 retained beadtypes indeed include control beadtypes.

2.3 Hypothesis testing using the Two Sample t-test

Upon the collection and preprocessing of methylation data, it was of interest to conduct hypothesis testing to deduce differences in methylation patterns between groups. CpG sites with differential methylation between the two groups mark sites for future studies as sites which possibly contributed to the progression of disease. Hypothesis tests
were performed on a beadtype level summary statistic measuring “proportion methylated” which is often computed as

$$\beta_{jl} = \frac{\max(R_{jl},0)}{\max(R_{jl},0) + \max(G_{jl},0)} \quad (2.2)$$

Again $i$ indexes the array ($i=1,\ldots,36$), $j$ indexes the beadtype ($j=1,\ldots,1547$), and here, $l$ indexes outcome group ($l=1,2$). Note that for the purposes of this paper, no background subtraction was performed in this portion of the analysis. Hypothesis tests for identifying differentially expressed CpG sites were performed using the random variable, ‘proportion methylated.’ Currently, the two-sample t-test is a widely accepted test in this scenario. So, for each beadtype, a two-sample t-test was applied to the sample proportion methylated to test the null hypothesis of the equality of methylation levels in two groups of samples. More specifically, letting $\beta_{j1}$ and $\beta_{j2}$ represent the mean population proportion methylated across arrays in each group for the $j$th CpG site, the null hypothesis is

$$H_0 : \beta_{j1} = \beta_{j2} \quad (2.3)$$

which is tested against the alternative hypothesis,

$$H_1 : \beta_{j1} \neq \beta_{j2} \quad (2.4)$$

The test statistic for the two-sample t-test is

$$t_j = \frac{\bar{\beta}_{j1} - \bar{\beta}_{j2}}{SE(\bar{\beta}_{j1} - \bar{\beta}_{j2})} = \frac{\bar{\beta}_{j1} - \bar{\beta}_{j2}}{\sqrt{\frac{\hat{\sigma}_{j1}^2}{n_1} + \frac{\hat{\sigma}_{j2}^2}{n_2}}} \quad (2.5)$$
where $\bar{\beta}_{j1}$, $\sigma^2_{j1}$, and $n_1$ are the sample mean proportion methylated, the sample variance, and the sample size, respectively, of the group 1 samples, and $\bar{\beta}_{j2}$, $\sigma^2_{j2}$, and $n_2$ are those corresponding to the group 2 samples for the $j$th CpG site summarized over the arrays in the group. Next, the degrees of freedom (df) assuming unequal variances in the two groups are calculated as

$$
df_j = \frac{\left(\frac{\hat{\sigma}_{j1}^2}{n_1} + \frac{\hat{\sigma}_{j2}^2}{n_2}\right)^2}{\frac{\hat{\sigma}_{j1}^4}{n_1} + \frac{\hat{\sigma}_{j2}^4}{n_2}}. \tag{2.6}
$$

The two-sided p-value is calculated,

$$
p_j = 2(1 - F_{df_j} \left| t_j \right|) \tag{2.7}
$$

where $F$ is the cumulative distribution function of a $t$ distribution.

Results of a t-test are commonly expressed in terms of a p-value, or the probability that a test statistic computed under the condition that the null hypothesis is true is at least as extreme as the value of the test statistic that was actually obtained. A large p-value (close to 1) indicates a high probability that the obtained test statistic is aligned with one that would typically be computed if the null hypothesis were true. So this is considered as evidence in favor of not rejecting the null hypothesis. A small p-value, on the other hand, supports rejection of the null hypothesis and the conclusion that the two means are not equivalent.
2.4 Filtered Two Sample T-test

In the landmark paper describing Illumina’s GoldenGate Methylation BeadArray assay, Bibikova et al. (2006b) introduced the Filtered t-test, and this testing framework quickly gained in popularity. Under this testing framework, a CpG site would be identified as differentially methylated between two groups of arrays if both the p-value of the Two Sample t-test was less than some pre-determined threshold and the difference between the mean proportion methylated values, $\Delta \beta_j = \bar{\beta}_{j1} - \bar{\beta}_{j2}$, was larger than an additionally specified threshold. Bibikova, et al. (2006b) estimated that the Illumina technology can discriminate between levels of differentiation that differ by 0.17 or greater, so they used $\Delta \beta = 0.17$ as the threshold for the filtered t-test. However, since this threshold recommendation, more stringent thresholds have been used. For example, in Richter et al. (2009), a two sample t-test was conducted at the $\alpha = 0.001$ level with an additional filter of $\Delta \beta > 0.30$. Additionally, responding to a concern of too little statistical power of a performed Mann-Whitney U test, Martin-Subero et al. (2009) filtered $\Delta \beta$ at 0.50.

2.5 Limitations of current analysis methods

Despite a lack of study into the performance quality of these two tests on this data type, the Two Sample t-test and the Filtered Two Sample t-test have attained widespread use. However, the Two Sample t-test assumes that the sample data follow a normal
distribution ranging the entire real number line. Although, considering the scale of the proportion methylated variable (i.e., $\beta_{ijl} \in [0,1] \forall i, j,$ and $l$), one could argue that a two-sample t-test is not statistically rigorous. Furthermore, the application of a Filtered Two Sample t-test is only as good as the accuracy of the filter. That is, applying a filter to the two sample t-test may reduce power for many CpG sites with smaller but truly different proportions, where power is defined as the probability that a test will reject a false null hypothesis. Thus, in this thesis we performed a detailed analysis into the accuracy of the 0.17 estimated level of minimum discrimination between methylation levels. We also proposed alternative inferential methods for comparing two groups with respect to proportion methylated that utilize the scale of the proportion methylated variable (i.e., $\beta_{ijl} \in [0,1] \forall i, j,$ and $l$), and to thus fully consider all known facets of the distribution. In so doing, we both propose a new computed minimum level of discrimination, and we further solidify recommendations for future testing methodology.
CHAPTER 3 Determination of Minimum Level of Discernable Differential Methylation
According to the Illumina Technology

The application of a Filtered Two Sample t-test hinges on the task of identifying an appropriate filter to apply to the delta beta, $\Delta \beta$, values in conjunction with the Two Sample t-test. Bibikova, et al. (2006b) estimated that the Illumina technology can discriminate between levels of differentiation that differ by 0.17 or greater, so often $\Delta \beta = 0.17$ has been used as a threshold for the filtered t-test. The determination of a 0.17 threshold was based on a dilution study performed by diluting female genomic DNA to male genomic DNA at ratios of 100:0, 50:50, 20:80, 10:90, 5:95, and 0:100 prior to bisulfite conversion. This dataset consists of 47 samples from an Illumina Universal Array Matrix wherein the GoldenGate Methylation Cancer Panel I was used as the methylation assay. Specifically, the 47 samples were derived by mixing male and female genomic DNA in the previously specified proportions. Two unique female and 2 unique male samples were used in creating the mixtures that were subsequently hybridized: Female 1 and Male 1 dilutions were each technically replicated 4 times (with exception of the 95:5 which was replicated 3 times) and Female 2 and Male 2 dilutions were each replicated 4 times. 17 CpG sites associated with X-linked genes were examined on the X chromosome to assess methylation levels. For each of 17 genes on the X chromosome, Bibikova et al. computed the maximum standard deviation over the replicate runs and found that the standard deviation of the $\beta$- value obtained across the four replicates was <0.06 in 99% of
cases. They concluded that the Illumina technology could thus discriminate levels of methylation that differ by at least $1.96 \times \sqrt{2} \times 0.06 \equiv 0.17$.

Moreover, since this threshold recommendation, more stringent thresholds have been used. For example, in Richter et al., a two sample t-test was conducted at the $\alpha = 0.001$ level with an additional filter of $\Delta \beta > 0.30$ (2009). Additionally, responding to a concern of too little statistical power of a performed Mann-Whitney U test, Martin-Subero et al. (2009) filtered $\Delta \beta$ at 0.50.

However, we conjecture that a Filtered Two-Sample t-test may severely reduce power for CpG sites with small but truly differential $\Delta \beta$’s. Nevertheless, if a filter is to be applied, the level of the filter should be computed as a function of the analyzed data since various factors such as tissue type and handling can affect the amount of variability in CpG site methylation across technical replicates. Therefore, the application of a standard filtering level such as that developed by Bibikova et al. or a more “conservative” filter level as seen in other research studies could in actuality be missing very real differential methylation patterns present in the data.

To investigate the proposed filter, herein we perform a re-analysis of the dilution data (Bibikova et al., 2006). This data set provided what was needed to compute a minimum discernable level of differential methylation because it is essentially a clinical data set in which the true outcome is known. As alluded to in Chapter 1, DNA methylation is the means by which one of the two X chromosomes in a female is inactivated for germ line differentiation in somatic cells. Because males only possess one X chromosome, methylation for this purpose does not occur. Therefore, for genes linked to the X
chromosome, one could expect approximately 50% methylation of these genes in a sample arising from a female and 0% methylation of these genes in a sample arising from a male. The dilution data set of Bibikova et al. consisted of 17 X linked genes measured in mixtures of female and male genomic DNA samples where the mixtures were according to specific proportions. It can then be expected that a sample with 100% Female: 0% Male would exhibit approximately 50% methylation, and so on, as summarized in Table 3.1 below.

<table>
<thead>
<tr>
<th>Mixing Ratio</th>
<th>Expected approximate methylation level</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. 100% Female: 0% Male</td>
<td>50%</td>
</tr>
<tr>
<td>2. 50% Female: 50% Male</td>
<td>25%</td>
</tr>
<tr>
<td>3. 20% Female: 80% Male</td>
<td>10%</td>
</tr>
<tr>
<td>4. 10% Female: 90% Male</td>
<td>5%</td>
</tr>
<tr>
<td>5. 5% Female: 95% Male</td>
<td>2.5%</td>
</tr>
<tr>
<td>6. 0% Female: 100% Male</td>
<td>0%</td>
</tr>
</tbody>
</table>

Table 3.1 Expected approximate methylation level for X linked genes given mixing ratio.

To conclude a minimum level of discernable differential methylation, we sought to make the following five comparisons to determine if differential methylation could be detected:

Mixing Ratio A versus Mixing Ratio B

100% Female : 0% Male versus 50% Female : 50% Male
50% Female : 50% Male versus 20% Female : 80% Male
20% Female : 80% Male versus 10% Female : 90% Male
10% Female : 90% Male versus 5% Female : 95% Male
5% Female : 95% Male versus 0% Female : 100% Male

Briefly, across the 17 X-chromosome linked CpG sites, it was sought to model methylation levels as a function of the mixing ratio of female to male DNA. By means of this resulting
model, we both obtained an accurate description of and established the significance of the relationship between the proportion methylation level and the mixing ratio. Additionally, a mixed ANOVA model was employed to perform five pairwise comparisons listed above testing for differences in the mean methylation levels between two mixing ratios. In so doing, we can determine that differential methylation can be identified at values as small as the minimum mixing ratio comparison showing a significant difference in methylation. This chapter concludes with a summary of the results of these hypothesis tests and the conclusions drawn.

3.1 Motivation and Methods: Hierarchical model building to Bibikova et al. dilution data

We sought a model to appropriately define the relationship of the proportion of methylation across CpG sites as a function of the mixing ratio. As evidenced in Figure 3.1, the ratio of Female to Male DNA in a sample possesses a nonlinear relation with methylation level.
Figure 3.1. Boxplot of proportion methylation by percent of female genomic DNA in the hybridized sample

Moreover, it is also graphically observed that this nonlinear relation appears to depend both on dilution order (Female 1: Male 1 vs. Female 2: Male 2; see Figure 3.2) and on the particular CpG site within this dilution order that the methylation is measured (see Figure 3.3).
Figure 3.2 Mean proportion methylated level across CpG site by percent of female genomic DNA in sample for the first dilution order (top panel) and the second dilution order (bottom panel).

Figure 3.3 Proportion methylated level by Percent of female genomic DNA for each CpG site (1-17) within dilution order (1-2).
By design, while neither the CpG site nor the dilution order should affect the relationship between mixing ratio and methylation level, Figures 3.2 and 3.3 demonstrate evidence of random differences within these grouping levels. Therefore, owing to these differing nonlinear shapes at both the dilution order level and the CpG site level, it was concluded that a nonlinear mixed effects model was the most appropriate model building procedure with which to progress. Additionally, owing to the fact that four technical replicates were produced for each mixing ratio (with exception of the 95:5 which was replicated 3 times), we hypothesized that random differences among the order of the replicates could potentially exist. Therefore, we chose to fit a nonlinear mixed effects model summarizing the mean proportion of methylation at each mixing ratio while employing random effects for the dilution order (1 or 2), the technical replicate order (1, 2, 3, or 4) nested within each dilution order, and the particular CpG site (1,…,17) within each replicate order that was nested within each dilution order.

Given that the nature of the data was CpG site methylation from each of 17 sites within each of two females and two males, one would expect the proportion of methylation across each of the 17 CpG sites to be correlated. Indeed, this is the case, as illustrated in Table 3.2, and the presence of such correlations further motivates the construction of a single nonlinear model to summarize the proportion of methylation across the 17 CpG sites.
Table 3.2 Correlation among 17 CpG sites.

The model building procedure was performed using the `nlme` library in R version 2.11.1 (R Development Core Team, 2010). Owing to the shape of the methylation curves in Figure 3.3, a Logistic Regression model was chosen for our model building procedure. Model formulation for the mixed Logistic Regression model is as follows.

3.1(a) Nonlinear Mixed Model Formulation: Logistic Regression Model

An initial model (Equation 3.1) was fit including independent random effects for the $j$th CpG site within the $m$th dilution order level and the $n$th technical replicate order.

Generally, the nonlinear mixed model employing three nested random effects is represented
where \( y \) summarizes the mean proportion of methylation as a function of \( x \), mixing ratio (Pinheiro and Bates, 2000). Here, \( \beta \) is a vector of fixed effects, with design matrix \( A_{mnj} \).

The first level random effects, \( b_m \), those corresponding to the \( m \)th dilution order, are independently distributed vectors with variance-covariance matrix \( \Psi_1 \) and a corresponding design matrix, \( B_{mnj} \). The second level random effects, \( b_{mn} \), those corresponding to the \( n \)th technical replicate within the \( m \)th dilution order, are independently distributed vectors with variance-covariance matrix \( \Psi_2 \) and a corresponding design matrix, \( B_{mn,j} \). And, the third level random effects, \( b_{mnj} \), those corresponding to the \( j \)th CpG site nested within the \( m \)th dilution order and the \( n \)th technical replicate, are independently distributed vectors with variance-covariance matrix \( \Psi_3 \) and a corresponding design matrix, \( B_{mnj} \).

Using the framework of the Logistic regression model, this model is expressed

\[
\begin{eqnarray*}
  f(\varphi_{mnj}, x) &=& \frac{0.5}{1 + \exp[(\varphi_{2mnj} - x) / \varphi_{3mnj}]} \\
  \varphi_{mnj} &=& [\varphi_{2mnj}, \varphi_{3mnj}]^T \\
  \varphi_{2mnj} &=& \beta_2 + b_2 + b_{2m} + b_{2mn} + b_{2mnj} \\
  \varphi_{3mnj} &=& \beta_3 + b_3 + b_{2m} + b_{2mn} + b_{2mnj}
\end{eqnarray*}
\]

(Eq. 3.2)
Note that in Equation 3.2, the asymptote as the percent of Female genomic DNA approaches 100 is fixed at 0.5 because, biologically, this is the expected proportion of methylation for this mixing ratio level. Furthermore, \( \phi_2 \) represents the value of percent of Female genomic DNA present in the mixture at the inflection point of the curve and \( \phi_3 \) is a scale parameter on the x-axis quantifying the relationship between mixing ratio and proportion of methylation.

3.1(b) Model Selection

Now that the methodology underlying the candidate model has been outlaid, the model was estimated and the adequacy of the model fit was evaluated by means of a residual plot. A plot of the standardized residuals plotted as a function of the fitted values revealed systematic departures from zero as demonstrated in Figure 3.4.
In response to these systematic departures, an alternative Logistic regression model was fit employing an alternative variance structure of the within group errors. This variance model was in the form in which the variance at each CpG site increases exponentially with the percent of Female DNA present in the mixture. This property can be graphically observed in Figure 3.3 by noting that the curves describing the expected proportion of methylation become much more highly variable as the percent of Female DNA increases. Indeed the application of this structure yielded a more appropriate plot of the standardized residuals plotted as a function of the fitted values:

**Figure 3.4** Standardized residuals plotted as a function of the fitted values for the original Logistic regression model
As a final verification of model adequacy, a plot of the fitted methylation values overlaid upon the observed methylation values in the mixture data set for the updated model using the exponential variance structure is included in Figure 3.6 below. Given that the fitted mean curve adheres well to observed data, we concluded that the model provides an appropriate fit.
Our model building procedure concluded with significance testing for the inclusion of the three random effects in the model. We began with significance testing for the random component associated with the $j$th CpG site, which was nested within the $m$th dilution order and the $n$th technical replicate. This was assessed by fitting a new Logistic regression model where the random effects associated with CpG site were dropped from the model. The resulting reduced model was compared to the original full model by means of a Likelihood Ratio Test, where the test statistic is computed

$$D = -2 \ln \left( \frac{L(H_0)}{L(H_1)} \right)$$

(3.4)

Here, $L(H_0)$ is the likelihood of the of the Logistic Regression model under the null hypothesis of a full model containing a random term for the nested CpG site effect in the
model. And, $L(H_A)$ is the likelihood under the alternative hypothesis of a reduced model where either $b_{2mnj} = 0$ and $b_{3mnj} = 0$. The likelihood ratio test statistic was compared to a chi-squared distribution with $n_1 - n_2$ degrees of freedom, where $n_1$ and $n_2$ are the degrees of freedom of the full and reduced models, respectively. The p-value of the performed Likelihood Ratio test was 0.06 which was considered only marginally significant. Therefore, it was concluded that evidence for dropping the random components for the nested CpG site effect did not exist.

3.1(c) Establishment of a significant relationship between Proportion of Methylation and Mixing Ratio

By means of this resulting model, we obtained an accurate description of the relationship between the proportion methylation level and the mixing ratio. We used the fitted mixed Logistic regression model to establish the significance of this dose-response relationship. We examined the $\phi_3$ parameter from the fitted model because this parameter defines the slope of our nonlinear model, and thus, it describes the nature of the relation between our independent and dependent variables. In the estimated model, $\hat{\phi}_3 = 8.40$, with a 95% confidence interval ($\hat{\phi}_3 \pm 2 \times s.e.$) of (5.61, 11.89). Because this confidence interval does not include zero, we conclude there is a significant relationship between percent of female genomic (dose) and proportion methylation level (response). Therefore
we proceeded in performing pairwise comparisons of adjacent levels for better discerning
the minimum level of detectable differential methylation,

3.2 Establishing a Minimum Level of Detectable Differentiation

Upon establishing both the nature and the significance of the relationship between
mixing ratio and the mean level of proportion of methylation at each CpG site, we wished
to establish a minimum level of detectable differentiation for the Illumina technology. In
this manner, we evaluated whether the threshold of 0.17 established in Bibikova et al.
(2006b) that was claimed to be the minimum level of differentiation that the Illumina
technology could detect was appropriate.

In establishing the minimum level of discrimination, we sought to perform the five
comparisons previously outlined, and listed again here:

Mixing Ratio A versus Mixing Ratio B
100% Female : 0% Male versus 50% Female : 50% Male
50% Female : 50% Male versus 20% Female : 80% Male
20% Female : 80% Male versus 10% Female : 90% Male
10% Female : 90% Male versus 5% Female : 95% Male
5% Female : 95% Male versus 0% Female : 100% Male

A mixed ANOVA model was employed to perform these five pairwise comparisons testing
for differences in the mean methylation levels between two mixing ratios. Because, by
biological principle, we could anticipate the methylation levels present at each mixing ratio
(see Table 3.1), we could also anticipate the differential level of methylation present for
each of the five pairwise comparisons. For example, for the comparison of 100% Female:
0% Male versus 50% Female: 50% Male, we could have expected a difference in mean methylation of approximately 0.50-0.25 = 0.25, and so on, down to the last pairwise comparison designed to query a difference in mean methylation of 0.025. The minimum level of discernable discrimination was then concluded to be the minimum queried difference in mean methylation for which the pairwise comparison determined significantly differentially methylated between the two mixing ratios.

The form of the mixed ANOVA model was as follows:

\[
y_{mnj} = \mu + \alpha_r + \beta_m + \gamma_{n(m)} + \delta_{j(mn)} + \epsilon_{mnj}
\]  
(Eq. 3.5)

where:

\(\mu\) is a constant

\(\alpha_r\) are constants corresponding to the \(r\)th level of the mixing ratio

\(\beta_m\) is an independent normal random variable with expectation zero and variance \(\sigma^2_\beta\) corresponding to the effect for the \(m\)th dilution order

\(\gamma_{n(m)}\) is an independent normal random variable with expectation zero and variance \(\sigma^2_\gamma\) corresponding to the effect for the \(n\)th technical replicate nested within the \(m\)th dilution order

\(\delta_{j(mn)}\) is an independent normal random variable with expectation zero and variance \(\sigma^2_\delta\) corresponding to the effect for the \(j\)th CpG site nested within the \(m\)th dilution order and the \(n\)th technical replicate

\(r = 1,\ldots,6, \ m = 1,2, \ n = 1,2,3,4, \ j = 1,\ldots,17\)

(Kutner et al., 2005).

A likelihood ratio test for the significance of the CpG site nested effect

\((H_0 : \delta_{j(mn)} = 0 \ \text{vs. } H_A : \delta_{j(mn)} \neq 0)\) demonstrated that we rejected the null hypothesis and concluded that a nested random effect for the effect of the \(j\)th CpG site was valid in the
model fit (p-value=0.001). A plot of the standardized residuals versus the fitted values for the mixed ANOVA model is presented in Figure 3.7 below.

Figure 3.7 Standardized residuals versus fitted values in final mixed ANOVA model.

Using the mixed ANOVA model (Eq. 3.5), the previously listed pairwise comparisons for testing differences in the factor level means of the adjacent mixing ratios were conducted. Here, $H_0: \mu_{r_{-}} - \mu_{r_{+}} = 0$ versus $H_A: \mu_{r_{-}} - \mu_{r_{+}} \neq 0$. Testing was conducted at the 95% significance level and results for the five pairwise comparisons are summarized in Table 3.3.
<table>
<thead>
<tr>
<th>Pairwise Comparison</th>
<th>Expected Differential Methylation</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu(100% F:0% M)$ vs. $\mu(50% F:50% M)$</td>
<td>0.25</td>
<td>0.104</td>
<td>0.033</td>
<td>3.128</td>
<td>0.026</td>
</tr>
<tr>
<td>$\mu(50% F:50% M)$ vs. $\mu(20% F:80% M)$</td>
<td>0.15</td>
<td>0.170</td>
<td>0.033</td>
<td>5.116</td>
<td>0.004</td>
</tr>
<tr>
<td>$\mu(20% F:80% M)$ vs. $\mu(10% F:90% M)$</td>
<td>0.05</td>
<td>0.134</td>
<td>0.033</td>
<td>4.024</td>
<td>0.010</td>
</tr>
<tr>
<td>$\mu(10% F:90% M)$ vs. $\mu(5% F:95% M)$</td>
<td>0.025</td>
<td>0.088</td>
<td>0.033</td>
<td>2.641</td>
<td>0.046</td>
</tr>
<tr>
<td>$\mu(5% F:95% M)$ vs. $\mu(0% F:100% M)$</td>
<td>0.025</td>
<td>0.086</td>
<td>0.033</td>
<td>2.595</td>
<td>0.049</td>
</tr>
</tbody>
</table>

Table 3.3 Pairwise comparisons across adjacent mixing ratios.

Notice first the discrepancies between the values of the expected differential methylation level and the estimate of differential methylation from the fitted mixed ANOVA model. The difference in these two columns is likely due to the high variation in proportion of methylation across CpG site. Note from Figure 3.1, the boxplot of proportion methylation by percent of female genomic DNA in the hybridized sample, the amount of overlap among quantile boxes, especially between samples containing each 100% Female: 0% Male and 50% Female: 50% Male. Nevertheless, at each pairwise comparison, the p-value is less than 0.05, and so we conclude that the Illumina technology can, in fact, detect differential methylation at levels less than 0.17. Theoretically, this analysis demonstrates that evidence exists to support a hypothesis that the Illumina technology can detect differences in methylation at least as low as 0.025 since differential methylation was detected for two pairwise comparisons designed to query for this methylation level. However, given that the observed differential methylation levels were as low as 0.086, a
more conservative concluded estimate of the minimum detectable differential methylation would be closer to 0.086.

3.3 Verification of fixed asymptote assumption

In the constructing the nonlinear model with nested random components in section 3.2 above, it was assumed that as the percent of Female DNA in the sample approached 100%, the asymptote of predicted methylation was fixed at 0.5. This was motivated by the biological principle of methylation at CpG sites on the X-chromosome. We now seek to determine if the given data support this assumption. In doing so, we fit separate nonlinear models for each of the 17 CpG sites and used the predicted methylation level under the conditions of 100% Female DNA: 0% Male to predict site-specific asymptotes and associated confidence intervals. The description of the model construction follows.

CpG site-specific nonlinear mixed models were constructed under each the Logistic regression framework previously described (Eq. 3.2) and the Asymptotic regression framework found in Pinheiro and Bates (2000) using a random effect for dilution order (1 or 2). These two models were chosen as appropriate candidate models because the shape of the nonlinear curve described appeared to adhere well to the shape of the dose-response curves exhibited in the mixture data set.
For the Asymptotic Regression model, an initial model (Equation 3.6) was fit including independent random effects for each the dilution order level (1 or 2). Generally, the nonlinear mixed model employing one random effect is represented

\[
y = f(\phi_{mj}, x) + \epsilon
\]

\[
\epsilon \sim N(0, \sigma^2)
\]

\[
\phi_{mj} = A_{mj} \beta + B_{mj} b_{mj}
\]  \hspace{1cm} (Eq. 3.6)

\[
m = 1, 2, \ j = 1, ..., 17
\]

\[
b_{mj} \sim N(0, \Psi),
\]

where \( y \) summarizes the mean proportion of methylation as a function of \( x \), mixing ratio.

Here, \( \beta \) is a vector of fixed effects, with design matrix \( A_{mj} \). The random effects, \( b_{mj} \), those corresponding to the \( m \)th dilution order for the \( j \)th CpG site, are independently distributed vectors with variance-covariance matrix \( \Psi \) and a corresponding design matrix, \( B_{mj} \).

Using the framework of the Asymptotic regression model, this model is expressed

\[
f(\phi_{mj}, x) = \phi_{1mj} + (\phi_{2mj} - \phi_{1mj}) \exp(-\exp(\phi_{3mj})x),
\]  \hspace{1cm} (Eq. 3.7)

and the random coefficients are modeled

\[
\phi = \begin{bmatrix}
\phi_{1mj} \\
\phi_{2mj} \\
\phi_{3mj}
\end{bmatrix}, \quad \phi_{1mj} = \beta_{1mj} + b_{1mj} \\
\phi_{2mj} = \beta_{2mj} + b_{2mj} \\
\phi_{3mj} = \beta_{3mj} + b_{3mj}
\]  \hspace{1cm} (Eq. 3.8)

\( \phi_1 \) represents the horizontal asymptote on the right side, \( \phi_2 \) represents the proportion methylated when the percent of Female genomic DNA present in the mixture is zero, and
\( \phi \) represents the natural logarithm of the rate constant and so quantifies the dose response relation.

3.3(b) Nonlinear Mixed Model Formulation: Logistic Regression Model

Again, an initial model (Equation 3.9) was fit including independent random effects for the dilution order level. The same basic framework for the model stays the same in that for the mean proportion methylated, \( y \), for the \( m \)th dilution order group and the \( j \)th CpG site is written as

\[
y = f(\phi_{mj}, x) + \epsilon \\
\epsilon \sim N(0, \sigma^2) \\
\phi_{mj} = A_{mj} \beta + B_{mj} \beta_{mj} \\
m = 1, 2, \ldots \text{ and } j = 1, \ldots, 17 \\
\beta_{mj} \sim N(0, \Psi),
\]

where \( \epsilon \) is a normally distributed within-group error term. The difference in the Asymptotic Regression model and the Logistic Regression model lies in the link function. Again, \( f \) is the real-valued, differentiable logistic function of a group specific parameter vector \( \phi_{mj} \) containing information on dilution order and a covariate vector \( x_j \) containing the levels of mixing ratios. But now \( f(\phi_{mj}, x_j) \) is written as

\[
f(\phi_{mj}, x) = \frac{\phi_{1mj}^x}{1 + \exp\left(\frac{\phi_{2mj} - x}{\phi_{3mj}}\right)}
\]

(3.10)

and the random coefficients are modeled
\[ \varphi_{mj} = \begin{bmatrix} \varphi_{1mj} \\ \varphi_{2mj} \\ \varphi_{3mj} \end{bmatrix} \], \quad \varphi_{1mj} = \beta_{1mj} + b_{1mj}, \quad \varphi_{2mj} = \beta_{2mj} + b_{2mj}, \quad \varphi_{3mj} = \beta_{3mj} + b_{3mj}. \]  

(3.11)

\( \varphi_1 \) represents the horizontal asymptote, \( \varphi_2 \) represents the value of the percent of Female genomic DNA in the sample at the inflection point of the curve, and \( \varphi_3 \) represents the scale parameter on the x-axis.

3.3(c) Model Selection

Now that the methodology underlying both the candidate models has been outlined, both models were fit for each CpG site and model comparison was conducted according to the Akaike Information Criterion (AIC). The AIC is a measure of the goodness of fit of an estimated statistical model and is commonly used as a tool for model selection. A model with a lower AIC is considered better. The AIC is computed,

\[ AIC = 2\eta - 2\ln(L) \]  

(3.12)

where \( \eta \) is the number of parameters in the statistical model, and \( L \) is the maximized function of the likelihood for the estimated model. Results of the AIC analysis are summarized in Table 3.4.
Table 3.4 AIC comparison by CpG site between Asymptotic Regression and Logistic Regression models

<table>
<thead>
<tr>
<th>CpG site</th>
<th>Asymptotic Regression Model</th>
<th>Logistic Regression Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-214.24</td>
<td>-126.81</td>
</tr>
<tr>
<td>2</td>
<td>-214.86</td>
<td>-152.14</td>
</tr>
<tr>
<td>3</td>
<td>-190.65</td>
<td>-157.43</td>
</tr>
<tr>
<td>4</td>
<td>-195.82</td>
<td>-124.20</td>
</tr>
<tr>
<td>5</td>
<td>-177.35</td>
<td>-127.88</td>
</tr>
<tr>
<td>6</td>
<td>-192.56</td>
<td>-144.77</td>
</tr>
<tr>
<td>7</td>
<td>-198.72</td>
<td>-154.47</td>
</tr>
<tr>
<td>8</td>
<td>-205.08</td>
<td>-131.58</td>
</tr>
<tr>
<td>9</td>
<td>-203.98</td>
<td>-118.07</td>
</tr>
<tr>
<td>10</td>
<td>-195.46</td>
<td>-114.94</td>
</tr>
<tr>
<td>11</td>
<td>-201.70</td>
<td>-125.15</td>
</tr>
<tr>
<td>12</td>
<td>-196.10</td>
<td>-109.88</td>
</tr>
<tr>
<td>13</td>
<td>-168.16</td>
<td>-154.94</td>
</tr>
<tr>
<td>14</td>
<td>-229.57</td>
<td>-210.85</td>
</tr>
<tr>
<td>15</td>
<td>-189.13</td>
<td>-124.10</td>
</tr>
<tr>
<td>16</td>
<td>-206.59</td>
<td>-208.94</td>
</tr>
<tr>
<td>17</td>
<td>-203.94</td>
<td>-177.54</td>
</tr>
</tbody>
</table>

For each CpG site, the AIC demonstrates that the Asymptotic Regression model is more appropriate to our given dilution data. The one possible exception to this is site 16 where the AIC of the Logistic Regression model is only slightly more negative than that of the Asymptotic Regression model. However, since these values are so close, we conclude that the models are relatively equivalent in fit, and for consistency purposes, we choose to proceed with the Asymptotic Regression model for this site as well.

In the search for the most well-suited yet parsimonious model, we next explored and modifications to the variance-covariance structures of our 17 Asymptotic Regression models. Up to this point, we had assumed independent variance components. Given that
plots of residuals did not show any particular trend such as with technical replicate order, we next dropped the constraint of independent variance components and considered an unstructured variance matrix. However, in all cases, the original constraint of independence proved to be a better fit. A final modification explored on our models, was to determine if each of the three random effects for the three modeling parameters were necessary. This was assessed by fitting new Asymptotic Regression models for each CpG site where one of the random effects was dropped from the model. The resulting reduced models were compared to the original full model by means of a Likelihood Ratio Test, where the test statistic is computed

\[
D = -2 \ln \left( \frac{L(H_0)}{L(H_1)} \right)
\]

(3.13)

Here, \( L(H_0) \) is the likelihood of the of the Asymptotic Regression model under the null hypothesis of a full model containing a random term for each of the three parameters in the model. And, \( L(H_0) \) is the likelihood under the alternative hypothesis of a reduced model where either \( b_{1jm} = 0 \), \( b_{2jm} = 0 \), or \( b_{3jm} = 0 \). The likelihood ratio test statistic was compared to a chi-squared distribution with \( n_1 - n_2 \) degrees of freedom, where \( n_1 \) and \( n_2 \) are the degrees of freedom of the full and reduced models, respectively.

For each CpG site, it was found that \( \phi_{2jm} \) did not warrant a random component. Likelihood ratios and p-values for this test are found in Table 3.2.
<table>
<thead>
<tr>
<th>CpG site</th>
<th>Likelihood of Full Model</th>
<th>Likelihood of Reduced Model (no random term for $\phi_{2mj}$)</th>
<th>Likelihood Ratio</th>
<th>P-value of Likelihood Ratio Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: EFNB1-645</td>
<td>114.12</td>
<td>114.12</td>
<td>6.04E-06</td>
<td>0.998</td>
</tr>
<tr>
<td>2: EFNB1-672</td>
<td>114.41</td>
<td>113.05</td>
<td>2.77E+00</td>
<td>0.096</td>
</tr>
<tr>
<td>3: ELK1-1163</td>
<td>102.32</td>
<td>102.32</td>
<td>3.15E-09</td>
<td>1.000</td>
</tr>
<tr>
<td>4: ELK1-1306</td>
<td>104.91</td>
<td>104.91</td>
<td>1.46E-06</td>
<td>0.990</td>
</tr>
<tr>
<td>5: ELK1-1495</td>
<td>95.67</td>
<td>95.67</td>
<td>2.13E-08</td>
<td>1.000</td>
</tr>
<tr>
<td>6: ELK1-877</td>
<td>103.28</td>
<td>101.87</td>
<td>2.83E+00</td>
<td>0.092</td>
</tr>
<tr>
<td>7: FMR1-1182</td>
<td>106.36</td>
<td>106.36</td>
<td>1.04E-09</td>
<td>1.000</td>
</tr>
<tr>
<td>8: FMR1-1440</td>
<td>109.54</td>
<td>109.54</td>
<td>4.66E-08</td>
<td>1.000</td>
</tr>
<tr>
<td>9: G6PD-1076</td>
<td>108.99</td>
<td>108.99</td>
<td>3.94E-06</td>
<td>0.998</td>
</tr>
<tr>
<td>10: G6PD-1304</td>
<td>104.73</td>
<td>104.73</td>
<td>1.02E-08</td>
<td>1.000</td>
</tr>
<tr>
<td>11: G6PD-834</td>
<td>107.85</td>
<td>107.85</td>
<td>2.11E-05</td>
<td>0.996</td>
</tr>
<tr>
<td>12: GLA-1158</td>
<td>105.05</td>
<td>105.05</td>
<td>5.51E-07</td>
<td>0.999</td>
</tr>
<tr>
<td>13: GLA-1294</td>
<td>91.08</td>
<td>91.08</td>
<td>2.45E-07</td>
<td>1.000</td>
</tr>
<tr>
<td>14: GLA-1306</td>
<td>121.78</td>
<td>121.78</td>
<td>5.95E-07</td>
<td>0.999</td>
</tr>
<tr>
<td>15: GLA-1388</td>
<td>101.56</td>
<td>101.56</td>
<td>6.54E-06</td>
<td>0.998</td>
</tr>
<tr>
<td>16: GLA-881</td>
<td>110.30</td>
<td>110.30</td>
<td>6.33E-06</td>
<td>0.998</td>
</tr>
<tr>
<td>17: GPC3-1182</td>
<td>108.97</td>
<td>108.97</td>
<td>4.75E+00</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Table 3.5 Results of Likelihood Ratio test to determine the significance of a random effect for $\phi_{2mj}$ parameter in Asymptotic Regression model.

Owing to the large p-values for the Likelihood Ratio test at each CpG site, we failed to conclude that there was a difference between the full Asymptotic Regression Model containing a random effect for each parameter in the model and the reduced model, which lacks a random effect for $\phi_{2mj}$. Therefore, we moved forward with the simpler reduced model. More formally, this model can be written,

$$f(\phi_{mj}, x) = \phi_{1mj} + (\phi_{2mj} - \phi_{1mj}) \exp(-\exp(\phi_{3mj}) x),$$

(3.14)
where

$$
\varphi_{mj} = \begin{bmatrix}
\varphi_{1mj} \\
\varphi_{2mj} \\
\varphi_{3mj}
\end{bmatrix}, \quad \varphi_{1mj} = \beta_{1mj} + b_{1mj} \\
\varphi_{2mj} = \beta_{2mj} + b_{2mj} \\
\varphi_{3mj} = \beta_{3mj} + b_{3mj}
$$

(3.15)

The appropriateness of this resulting model in summarizing the dilution data was verified both by plots of standardized residuals and plots of the fitted values overlaid upon observed values.
Figure 3.8. Standardized residuals plotted as a function of fitted values for the 17 CpG site-specific models.
In examining the plots of the standardized residuals plotted as a function of the fitted values for the 17 CpG site-specific models, we observed that for each model, the residuals appear to be randomly and uniformly distributed about the horizontal zero line. This gives confidence of the appropriateness of the fit for our models. A further check of model fit was accomplished by examining plots of model fitted values overlaid upon the observed values upon which the models were estimated (see Figure 3.5). We saw that for each CpG site, the fitted curve of mean proportion methylated appeared to adhere very well to the shape of the observed data. Therefore, we are confident in the fit of our Asymptotic Regression models, and we moved to construct hypothesis tests for the detection of differential methylation between mixing ratio levels at each CpG site.
Figure 3.9 Mean fitted proportion methylated curve overlaid upon observed proportion methylated values for each mixing ratio level by CpG site.
3.3(d) Predicted CpG site specific response asymptotes

Upon the construction of the CpG site specific nonlinear mixed models, it was of interest to use the predicted proportion of methylation to estimate the asymptote as the percent of Female genomic DNA in the sample approaches 100%. Therefore, using the 17 CpG site specific models, point estimates of the predicted mean proportion of methylation were obtained where the percent of Female genomic DNA in the sample was 100%, and 95% confidence intervals were calculated using the delta method to estimate the variance and applying a Bonferroni adjustment, i.e.,

$$\hat{\mu}(100\%\text{Female} : 0\%\text{Male}) \pm z_{(1−0.05/17)} \times \begin{bmatrix} \frac{\partial \varphi_{1j}}{\partial f} & \frac{\partial \varphi_{2j}}{\partial f} & \frac{\partial \varphi_{3j}}{\partial f} \\ \sigma_{\varphi_{1j}}^2 & 0 & 0 \\ \sigma_{\varphi_{2j}}^2 & 0 & 0 \\ 0 & \sigma_{\varphi_{3j}}^2 \end{bmatrix} \begin{bmatrix} \frac{\partial \varphi_{1j}}{\partial f} & \frac{\partial \varphi_{2j}}{\partial f} & \frac{\partial \varphi_{3j}}{\partial f} \end{bmatrix}^{T} \begin{bmatrix} \frac{\partial \varphi_{1j}}{\partial f} & \frac{\partial \varphi_{2j}}{\partial f} & \frac{\partial \varphi_{3j}}{\partial f} \end{bmatrix}^{T} \begin{bmatrix} \frac{\partial \varphi_{1j}}{\partial f} & \frac{\partial \varphi_{2j}}{\partial f} & \frac{\partial \varphi_{3j}}{\partial f} \end{bmatrix} \times 1/2$$

Point estimates and resulting 95% confidence intervals are summarized in Table 3.6 below.
Table 3.6 Point estimates and 95% confidence intervals of predicted proportion methylated for mixtures demonstrating 100% Female DNA

<table>
<thead>
<tr>
<th>CpG site</th>
<th>Point Estimate</th>
<th>Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>EFNB1-645</td>
<td>0.719</td>
<td>(0.673,0.765)</td>
</tr>
<tr>
<td>EFNB1-672</td>
<td>0.644</td>
<td>(0.551,0.737)</td>
</tr>
<tr>
<td>ELK1-1163</td>
<td>0.588</td>
<td>(0.544,0.631)</td>
</tr>
<tr>
<td>ELK1-1306</td>
<td>0.545</td>
<td>(0.423,0.668)</td>
</tr>
<tr>
<td>ELK1-1495</td>
<td>0.860</td>
<td>(0.836,0.884)</td>
</tr>
<tr>
<td>ELK1-877</td>
<td>0.679</td>
<td>(0.636,0.723)</td>
</tr>
<tr>
<td>FMR1-1182</td>
<td>0.551</td>
<td>(0.517,0.585)</td>
</tr>
<tr>
<td>FMR1-1440</td>
<td>0.736</td>
<td>(0.677,0.794)</td>
</tr>
<tr>
<td>G6PD-1076</td>
<td>0.638</td>
<td>(0.559,0.718)</td>
</tr>
<tr>
<td>G6PD-1304</td>
<td>0.819</td>
<td>(0.796,0.842)</td>
</tr>
<tr>
<td>G6PD-834</td>
<td>0.535</td>
<td>(0.404,0.667)</td>
</tr>
<tr>
<td>GLA-1158</td>
<td>0.812</td>
<td>(0.778,0.845)</td>
</tr>
<tr>
<td>GLA-1294</td>
<td>0.456</td>
<td>(0.108,0.802)</td>
</tr>
<tr>
<td>GLA-1306</td>
<td>0.223</td>
<td>(-0.008,0.458)</td>
</tr>
<tr>
<td>GLA-1388</td>
<td>0.747</td>
<td>(0.716,0.779)</td>
</tr>
<tr>
<td>GLA-881</td>
<td>0.297</td>
<td>(0.073,0.521)</td>
</tr>
<tr>
<td>GPC3-1182</td>
<td>0.500</td>
<td>(0.472,0.518)</td>
</tr>
</tbody>
</table>

It is evident from the above table that the asymptote of 0.5 was not obtained by every CpG site. This is likely due to the fact that, similar to other high-throughput genomic platforms, probes have different hybridization affinities and therefore the theoretical 0.5 estimate was not attained by each CpG site. However, as established in Chapter 3.1(c), an appropriate overall model fit was achieved and a significant dose-response relationship was indeed established.
3.4 Concluding remarks

In this chapter, we sought to explore our hypothesis that the Illumina technology can differentiate between levels of methylation less than 0.17. It had been previously claimed that the Illumina technology could differentiate between levels of methylation by 0.17, however, we hypothesized that the technology could indeed detect lesser levels of differential methylation. To demonstrate the validity of our hypothesis, we used a dilution data set provided by the Illumina company, and we modeled the relationship between mixing ratio and proportion of methylation in X-chromosome linked genes. This dilution data set of X-chromosome linked genes was key because it essentially provided us with a clinical data set in which the true methylation status was known. By modeling the relationship between mixing ratio and proportion methylation according to nonlinear mixed models for each X-chromosome linked CpG site, we were able to accurately describe the nature of the relationship between the mixing ratio and proportion of methylation while concluding that a significant relationship exists between these two factors. Furthermore, a mixed ANOVA model accounting for random variation due to dilution order, technical replicate order, and CpG site effect enabled hypothesis testing for differential mean methylation between adjacent mixing ratio groups. In doing so, we showed that the Illumina technology can, in fact, discriminate between levels of differentiation at least as low as 0.086, rather than 0.17. This is an especially interesting finding given the propensity of many to test differential methylation using a Filtered Two Sample t-test with a filter greater than or equal to 0.17, as summarized in Chapter 1.
Indeed, a filter of this magnitude likely reduces power for smaller but truly differential CpG sites that the technology is very capable of discriminating. In Chapter 4, alternatives to a Two Sample t-test with and without a filter to analyze CpG site methylation data will be introduced.
4.1 Alternative Analysis Exploration

Alternatives to the currently employed Two Sample t-test with and without a filter were sought, and an initial alternative analysis was motivated by the well known distributional property supposing that if $R_{ij} \sim \text{Gamma}(\theta_1, \kappa)$ and $G_{ij} \sim \text{Gamma}(\theta_2, \kappa)$ then

$$\beta_{ij} = \frac{R_{ij}}{R_{ij} + G_{ij}} \sim \text{Beta}(\theta_1, \theta_2).$$

Given that the formula for the proportion methylated resembles this property, a natural initial step in the analysis was to suppose that the Red and Green channels each follow a Gamma distribution. This seems plausible as both $R_{ij}$ and $G_{ij}$ have a distributional range of $[0, \infty)$ and because of the skewed shape of the distributions of the Red and Green channels present using the Hepatitis C Virus methylation dataset described in Chapter 2.1 (illustrated for a representative sample, beadarray 28, in Figure 4.1). An additional artifact of the Red and Green channel data present in Figure 4.1 is that of dye bias. The green dye is more efficient than the red, causing it to be incorporated faster and resulting in the bulk of the Green channel signal intensities being distributed at a much higher value than the Red channel signal intensities. Again in Figure 4.1 this artifact is demonstrated for a representative sample, beadarray.
number 28, where it can be seen that the intensities from the Green channel are much higher than the intensities from the Red Channel:

For this reason, a constant scaling normalization method was applied. Specifically, a unique scaling factor $s_i$ for each array was introduced to multiply the Green channel signal intensities, thus scaling down closer to the red values, i.e.,

$$s_i = \frac{\mu_{R_i}}{\mu_{G_i}}$$ (4.4)

for $i = 1, 2, \ldots, 36$. Here, $\mu_{R_i}$ and $\mu_{G_i}$ are the arithmetic means of the Red and Green signal intensities across all CpG sites for a single array. This array specific scaling factor was multiplied by the Green signal intensities, thus resulting in the Red and Green Channels
residing on a more similar range, as evidenced by the following histogram, again for beadarray 28.

![Histogram of Green uncorrected (top panel), Green corrected, and Red (bottom panel) channel intensities for a representative beadarray. Green corrected intensities were post-constant-scale normalized.](image)

**Figure 4.2** Histogram of Green uncorrected (top panel), Green corrected, and Red (bottom panel) channel intensities for a representative beadarray. Green corrected intensities were post-constant-scale normalized.

Upon computation of this scaled normalized data, $\beta$ estimates were calculated for each CpG site according to Equation 2.2. Because each the Red and Green channels were hypothesized to be distributed as Gamma, then it follows that we assumed $\beta$ is distributed as Beta. The Beta distribution is a family of continuous probability distributions
constrained on the interval \((0, 1)\) and defined by two positive shape parameters, \(p\) and \(q\).

The density function of a variable \(z\), that is distributed according to a beta distribution, \(\mathcal{B}(p, q)\), is

\[
f(z; p, q) = \frac{\Gamma(p + q)}{\Gamma(p)\Gamma(q)} z^p (1-z)^q.
\]  

(4.1)

Because the Beta distribution is constrained on the interval \((0, 1)\), it is often a good distribution to model variables which are measured in proportions. Further evidence in support of the hypothesis of the distribution of the proportion methylation \(\beta\) according to a Beta distribution is given that by nature of the computation of the proportion methylated estimates, \(\beta\) is a continuous random variable both centered on and restricted to the range [0, 1]. Furthermore, for each CpG site in the HCV data set a Kolmogorov-Smirnov Test was employed to determine if evidence existed to refute this hypothesis.

4.2 Kolmogorov-Smirnov Test

A Kolmogorov-Smirnov Test was conducted for each CpG site to assess whether evidence existed to refute the hypothesis that the proportion of methylation was distributed according to a beta distribution, i.e. \(H_0 : [F(t) = H(t), \text{ for every } t]\) where \(F(t)\) is the empirical distribution function for the 36 computed \(\beta\) ratios at each CpG site \((\beta_{ij} \sim \text{Beta}(p_j, q_j) \text{ under } H_0)\), calculated using the scaled data, and \(H(t)\) is the empirical distribution function for the Beta distribution function (Hollander and Wolfe 1999).
Letting $Z_{(1)} \leq \cdots \leq Z_{(72)}$ denote the 72 = $(i + h) = (36 + 36)$ ordered values for the combined sample of the $\beta_{1j}, \ldots, \beta_{q_j}$ and the $Y_{1j}, \ldots, Y_{h_j}$ observations of the empirical distribution function for the Beta distribution, $H_{h_j}$, the Kolmogorov-Smirnov test statistic is computed as

$$\Omega_{j} = \max_{s=1, \ldots, 72} \left\{ F_{h_j}(Z_{(s)}) - H_{h_j}(Z_{(s)}) \right\} \tag{4.2}$$

For the empirical distribution function of the Beta distribution at the $j$th CpG site, the method of moments was used to estimate the parameters for the empirical distribution function of the $Beta(p_{j}, q_{j})$ distribution as a function of the data at each CpG site where:

$$\hat{p}_{j} = \bar{z}_{j} \left( \frac{z_{j} (1 - \bar{z}_{j})}{s_{zj}^2} - 1 \right) \tag{4.3}$$

and,

$$\hat{q}_{j} = (1 - \bar{z}_{j}) \left( \frac{z_{j} (1 - \bar{z}_{j})}{s_{zj}^2} - 1 \right). \tag{4.4}$$

Above, $\bar{z}_{j}$ is the sample mean and $s_{zj}^2$ is the population variance of the $\beta_{ij}$ ratios at each CpG site.

This yielded promising results: in 1438 (93%) of the ratio distributions computed for each CpG site, we failed to reject the null hypothesis, and so we do not have evidence to refute the claim that the distribution of the proportion of methylation at each CpG site follows a Beta distribution. Thus it was decided to move forward with designing a hypothesis test that is based on a Beta distributed random variable to test for the equivalence of the two groups. The hypothesis test based on a Beta distributed random
variable took two forms: that of a test based on the beta distribution and that of a large sample approximation. A description of these two forms follows.

### 4.3 Testing for Differentially Methylated CpG sites: Beta Distribution Motivated Two-Sample Test

#### 4.3(a) Test Based on a Beta Distributed Random Variable

To design a test for the difference between two beta distributed random variables, assume that \( X \sim \text{Beta}(p_1, q_1) \) and \( Y \sim \text{Beta}(p_2, q_2) \) are two given random variables. It is of interest to test \( H_0 : \mu_X - \mu_Y = 0 \) against the alternative \( H_A : \mu_X - \mu_Y \neq 0 \). The null hypothesis can be equivalently expressed as \( H_0 : \mu_X = \mu_Y \). In formulating the hypothesis test, it must be conducted under the conditions of the null hypothesis. Therefore, given the observed data \( \beta_{j1} = (\beta_{1j1}, \beta_{2j1}, \ldots, \beta_{n_jj1}) \) and \( \beta_{j2} = (\beta_{1j2}, \beta_{2j2}, \ldots, \beta_{n_jj2}) \) at the \( j \)th CpG site, under the conditions of \( H_0 \), we form \( Z_{j} = (\beta_{j1}, \beta_{j2}) \) which is used to estimate \( \hat{p} \) and \( \hat{q} \) according to the method of moments estimators described in Equations 4.3 and 4.4, and as follows:

\[
\hat{p}_{j} = \frac{\bar{z}_{j} (1 - \bar{z}_{j})}{s_{j}^2} - 1 \quad (4.3a)
\]

and
\[
\hat{q}_j = \left(1 - \bar{z}_j \right) \left( \bar{z}_j \left(1 - \bar{z}_j \right) \right) - 1. 
\]

(4.4a)

Here, \( s_j^2 \) is the population variance of \( Z_j \). Furthermore, note that

\[
\mu_j = \text{Mean}(z_{j}) = \frac{p_j}{p_j + q_j},
\]

(4.5)

be the population mean, and that

\[
v_j = \text{Var}(z_{j}) = \frac{p_j q_j}{(p_j + q_j + 1)(p_j + q_j)^2},
\]

(4.6)

be the population variance of \( Z_j \sim \text{Beta}(p_j, q_j) \).

These values can then be used to obtain a standardized quantile value, \( Q_{0.975, j} \), from the beta distribution for a two-sided 95% confidence level (from \( 1 - \alpha/2 \) letting \( \alpha = 0.05 \) ) as follows:

\[
Q_{0.975, j} = \frac{F^{-1}(0.975, \hat{p}_j, \hat{q}_j) - \mu_j}{\sqrt{v_j}}.
\]

(4.7)

The critical value, \( C_{0.975, j} \), for testing the test statistic \( D_j = |\mu_{j1} - \mu_{j2}| \) at the \( j \)th CpG site would be computed as

\[
C_{0.975, j} = 0 + Q_{0.975, j} \sqrt{\frac{v(\beta_{j1})}{n_1} + \frac{v(\beta_{j2})}{n_2}},
\]

(4.8)

and one would reject the null hypothesis of equal means between the two groups at the \( j \)th CpG site where \( D_j > C_{0.975, j} \).
Under the framework just described, it is assumed that the combined data between the two Beta distributed groups follows a Beta distribution, and it is this Beta distribution of the combined data that is used to compute a standardized quantile and, hence, a critical value to which a test statistic is compared. Therefore, this Beta test specifically tests if the mean proportion methylated is equivalent between two groups at a specified CpG site.

4.3(b) Likelihood Ratio Test

As a large sample counterpart to the Beta test, an alternative test was designed under the likelihood ratio testing framework. The likelihood ratio test was appropriate here since two nested distributions were present. Under the likelihood ratio test, the same question of equivalence between groups is addressed in a slightly different manner. Rather a two sample test for comparing the ratios for each CpG site between two groups, cirrhosis with and without concomitant HCC, reduced to testing if the ratio of the signal intensities adhered to identical Beta distributions at each site. This was determined by testing if the Beta distribution parameters are the same for each group at each CpG site, i.e.,

\[ H_0 = p_1 = p_2 = p \quad \text{and} \quad q_1 = q_2 = q, \quad H_1 = p_1 \neq p_2 \quad \text{and/or} \quad q_1 \neq q_2 \quad (4.9) \]

The methodology of this large sample approximation test follows.

To test the hypothesis of identically shaped Beta distributions at each CpG site \(( H_0 = p_1 = p_2 = p \quad \text{and} \quad q_1 = q_2 = q )\), a likelihood ratio test was employed. The first distribution, the distribution under the null hypothesis, assumed that both \(p\) and \(q\) parameters from the two groups were equal, and hence, the likelihood equation at the \(j\)th
CpG site under the null hypothesis is that of a Beta distribution with sample size \( n \), given by

\[
L_{0,j}(p_j, q_j; z_{1j}, \ldots, z_{nj}) = \frac{1}{B(p_j, q_j)} \prod_{i=1}^{n_j} z_{ij}^{p_j - 1}(1 - z_{ij})^{q_j - 1} \tag{4.10}
\]

where the \( z_{ij} \)'s are the computed \( \beta \) ratios and here, \( n = 36 \). Now, the logarithm of the likelihood is

\[
\log L_{0,j}(p_j, q_j; z_{1j}, \ldots, z_{nj}) = -n \log[B(p_j, q_j)] + (p_j - 1) \sum_{i=1}^{n_j} \log z_{ij} + (q_j - 1) \sum_{i=1}^{n_j} \log(1 - z_{ij}). \tag{4.11}
\]

The second likelihood formed under the conditions of the alternative hypothesis, assumes that either or both \( p \) and \( q \) parameters from the two groups are not equal, and hence the likelihood equation of this second group is given by

\[
L_{1,j}(p_{j1}, q_{j1}, p_{j2}, q_{j2}; z_{1j1}, \ldots, z_{nj1}, z_{1j2}, \ldots, z_{nj2}) = \frac{1}{B(p_{j1}, q_{j1})} \prod_{i=1}^{n_{j1}} z_{ij1}^{p_{j1} - 1}(1 - z_{ij1})^{q_{j1} - 1} \cdot \frac{1}{B(p_{j2}, q_{j2})} \prod_{i=1}^{n_{j2}} z_{ij2}^{p_{j2} - 1}(1 - z_{ij2})^{q_{j2} - 1} \tag{4.12}
\]

where \( z_{ij1}, z_{ij2} \) and \( n_{j1}, n_{j2} \) are the computed \( \beta \) ratios and sample sizes, respectively for the two groups in the \( j \)th CpG site, where \( n_{j1} + n_{j2} = n \). For the purposes of the data here, the \( z_{ij1} \) are the \( \beta \) ratios from the HCC positive group (\( n_{j1} = 20 \)), and the \( z_{ij2} \) are the \( \beta \) ratios from the HCC negative group (\( n_{j2} = 16 \)). The logarithm of this likelihood is

\[
\log L_{1}(p_{j1}, q_{j1}, p_{j2}, q_{j2}; z_{1j1}, \ldots, z_{nj1}, z_{1j2}, \ldots, z_{nj2}) = -n_{j1} B(p_{j1}, q_{j1}) - n_{j2} B(p_{j2}, q_{j2}) + (p_{j1} - 1) \sum_{i=1}^{n_{j1}} z_{ij1}^{p_{j1} - 1} \tag{4.13}
\]

\[
+ (q_{j1} - 1) \sum_{i=1}^{n_{j1}} (1 - z_{ij1})^{q_{j1} - 1} + (p_{j2} - 1) \sum_{i=1}^{n_{j2}} z_{ij2}^{p_{j2} - 1} + (q_{j2} - 1) \sum_{i=1}^{n_{j2}} (1 - z_{ij2})^{q_{j2} - 1}
\]
Consequently, the likelihood ratio test statistic for the \( j \)th CpG site, \( \Lambda_j \), is

\[
\Lambda_j = -2\left( \log L_0(\hat{p}_j, \hat{q}_j) - \log L_1(\hat{p}_{j1}, \hat{q}_{j1}, \hat{p}_{j2}, \hat{q}_{j2}) \right) \\
= 2\left( n_1 + n_2 \right) \log \left[ B(\hat{p}_j, \hat{q}_j) \right] - n_1 B(\hat{p}_{j1}, \hat{q}_{j1}) - n_2 B(\hat{p}_{j2}, \hat{q}_{j2}) \\
- \left( \hat{p}_j - 1 \right) \sum_{i=1}^{n} \log z_{ij} + \left( \hat{q}_j - 1 \right) \sum_{i=1}^{n} \log (1 - z_{ij}) - \left( \hat{p}_j - 1 \right) \sum_{i=1}^{n} \log z_{ij} \\
- \left( \hat{q}_j - 1 \right) \sum_{i=1}^{n} \log (1 - z_{ij}) - \left( \hat{p}_{j1} - 1 \right) \sum_{i=1}^{n} \log z_{qj1} + \left( \hat{q}_{j1} - 1 \right) \sum_{i=1}^{n} \log (1 - z_{qj1}) \\
+ \left( \hat{p}_{j2} - 1 \right) \sum_{i=1}^{n} \log z_{qj2} + \left( \hat{q}_{j2} - 1 \right) \sum_{i=1}^{n} \log (1 - z_{qj2})
\] (4.14)

where \( \hat{p}_j \) and \( \hat{q}_j \) are the maximum likelihood estimates computed under the null hypothesis and \( \hat{p}_{j1}, \hat{q}_{j1}, \hat{p}_{j2}, \hat{q}_{j2} \) are those computed under the alternative hypothesis.

Hence, the likelihood ratio test can be written under \( H_0 \) as: Reject \( H_0 \) if \( \Lambda_j > C \) where \( C \) is the \( 100 \times (1 - \alpha) \) percentile point of a Chi Square distribution with \( k \) degrees of freedom and the p-value is computed as \( p_j = 1 - F_\nu(\Lambda_j) \). Here, \( \nu \) is the difference in the number of parameters between the logarithms of the likelihood under the alternative and null hypotheses (\( \nu = 4 - 2 = 2 \)), and \( F_\nu \) is the cumulative distribution function of the \( \chi^2 \) distribution. For the purposes of this analysis, tests were conducted at an \( \alpha = 0.05 \) significance level.

It should be noted that in the computation of the test statistic of the likelihood ratio test, an iterative algorithm was employed to compute the maximum likelihood estimates. A closed form expression for the computation of the maximum likelihood estimated under a beta distribution does not exist, so the Nelder Mead simplex was used.
4.3(c) Nelder Mead Simplex

The Nelder Mead algorithm is a direct search method which operates on the assumption of a simplex. More specifically, where \( n \) parameters are to be estimated, a simplex \( \mathcal{S} \) in \( \mathbb{R}^n \) is formed by the convex hull of the \( n+1 \) vertices \( x_0, \ldots, x_n \in \mathbb{R}^n \). The corresponding objective function, \( f(x) \) is then evaluated at the vertices of the simplex, \( f(x_j) \), for \( j = 0, 1, \ldots, n \), and it moves in the direction of improved response. At each step, the transformation of the simplex is determined by computing the vertices, together with their function values, and by comparison of these function values with those at the vertices of the existing simplex. This process is terminated when the working simplex \( \mathcal{S} \) becomes sufficiently small as judged by pre-determined threshold value.

Now, for a random variable \( y \) such that \( a \leq y \leq b \), the family of beta distributions includes all probability density functions of the form

\[
p(y) = \frac{1}{B(p, q)} \frac{(y-a)^{p-1}(b-y)^{q-1}}{(b-a)^{p+q-1}}
\]  
(Eq. 4.15)

where \( p > 0 \) and \( q > 0 \). Johnson & Kotz (1970) describe method of moments estimators for the four parameters in Equation 4.15. Additionally, they describe an iterative method for approximating maximum likelihood estimates of \( p \) and \( q \) when \( a \) and \( b \) are known. Krishnamoorthy (2006) suggested that the moment estimates be used as initial values in the iterative procedure for obtaining maximum likelihood estimates. Additionally, if one of the values (\( p \) or \( q \)) are known, the equations simplify (Johnson & Kotz, 1970). For
example, if $q$ is known to be 1, the distribution actually becomes the standard power function. Because when performing hundreds of hypothesis tests for a high-throughput methylation dataset, it would not be practical to know $a$ or $b$ or make assumptions about either $p$ or $q$, when maximizing the likelihood we used the standard form of the probability density function,

$$p(x) = \frac{1}{B(p, q)} x^{p-1} (1 - x)^{q-1} \quad \text{(Eq. 4.16)}$$

in which $x = (y - a)/(b - a)$ which allowed us to solve for $p$ and $q$. In this manner, the maximum likelihood estimates for the above described likelihood ratio test were determined, where the objective function was the maximum likelihood equation of the standard form of the probability density function.

Since the methodology which underlies both our Beta test and our proposed likelihood ratio based test has been fully developed, a third alternative analysis method is described.

4.4 Testing for Differentially Methylated CpG sites: Bivariate Normal Distribution

Motivated Test

As another exploratory measure, it was hypothesized that each the Red and Green channels were distributed Normal. This hypothesis, similar to the previous hypothesis of the Gamma distribution, is a common assumption made for log$_2$ transformed channel data (Brody et al., 2002). One could argue that in principle, the normal distribution is not an
appropriate fit for the Red and Green Channel intensities, values that were constrained on $[0, \infty)$ by nature. However, given that the channel intensities were centered about a value that was much greater than zero and that by artifact of the background phenomenon, the intensities are clustered such that they never truly get to zero, it was considered appropriate to continue with this line of study. Again, to see if evidence exists to refute this hypothesis, Kolmogorov Smirnov tests were performed for the Red and the scaled Green channel signal intensities at each CpG site in the HCV data set. Thus the null hypothesis $H_0: [F(t) = H(t), \text{for every } t]$ was tested where $F(t)$ is the empirical distribution function for each the Red and Green channels, calculated using the aforementioned scaled data, and $H(t)$ is the empirical distribution function for the Normal distribution function. The sample mean and sample variance for each CpG site were used to estimate the parameters for the empirical distribution function of the $N(\mu_j, \sigma_j^2)$ distribution as a function of the data. Here, we failed to reject the null hypothesis in 1,421 (92%) CpG sites in the red channel and 1,527 (99%) CpG sites in the green channel out of a total 1,547 CpG sites per channel. Therefore, it was concluded that we did not have evidence to refute our assumption of Normality. It was furthermore known that the correlation between the Red and Green channels was approximately -0.62, so it was concluded that the Red and Green channels were distributed Bivariate Normal$(\mu_R, \mu_G, \sigma_R^2, \sigma_G^2, \rho)$. 
Again for testing purposes, the ratio $\beta = R/(R + G)$ was of particular interest. Given that Red and Green Channels were distributed Bivariate Normal($\mu_R, \mu_G, \sigma^2_R, \sigma^2_G, \rho$), it follows then that the ratio $\beta = R/(R + G)$ has a complex distributional form.

For this reason, bootstrap methods underlay the hypothesis testing designed to compare the ratios for each CpG site between the two groups. The bootstrap method is especially useful because it enables one to design a hypothesis test on a set of values by estimating all pertinent values from an approximate distribution, namely the empirical distribution of the observed data. This proves valuable in situations such as the one present, where the underlying distribution made the formulation of a hypothesis test difficult.

The test formulation was as follows: first a test statistic was computed as a function of the observed data. Here, the mean was used as the location parameter, and the statistic used for testing was the difference in the mean $\beta$ ratio between the two groups, i.e.,

$$\hat{\theta}_j = \text{mean} \left( \frac{R_{y1}}{R_{y1} + G_{y1}} \right) - \text{mean} \left( \frac{R_{y2}}{R_{y2} + G_{y2}} \right).$$

(4.15)

This test statistic was computed for each CpG site, letting the $n_1$ samples originating from patients presenting cirrhosis without HCC be called Group 1 and letting the remaining $n_2$ samples originating from patients presenting cirrhosis with HCC be called Group 2. Next let the empirical distribution function of Group 1 be denoted $F$ and of Group 2, $G$. Under the null hypothesis, any of the $n_1 + n_2$ observations could have come from either $F$ or $G$, i.e., $H_0: F = G$ which implies that $\Pr_F = \Pr_G$ for any event $A$. This hypothesis was tested by
computing the empirical distribution function of the test statistic by combining the $n_1$ observations from $F$ and the $n_2$ observations from $G$ to form a single set of values of size $n_1 + n_2$. A sample of size $n_1$ was drawn with replacement from the combined set of $n_1 + n_2$ values to represent the observations from $F$, and the latter $n_2$ observations represent those from $F$. The statistic $\hat{\theta}^*$ was computed using these artificially created samples. For each CpG site, this process of sampling and computing the test statistic was repeated 2,500 times, thereby creating an empirical distribution function of the bootstrap test statistic. A two-sided $100(1-\alpha)$% level bootstrap hypothesis test was constructed for each CpG site by comparing $\hat{\theta}^*_j$ to the $100(\alpha/2)$% upper and the $100(1-\alpha/2)$% lower percentiles, $\theta^b_j$ computed from the empirical distribution of the observed data. The p-value was calculated,

$$p_j = \#\{\theta^b_j > |\hat{\theta}^*_j| \text{ or } \theta^b_j < -|\hat{\theta}^*_j| \}.$$  \hspace{1cm} (4.16)

This chapter has described methodology for three proposed alternatives to the widely employed Two Sample t-test and the Filtered Two Sample t-test. These three alternatives were in the form of a test based on a Beta distributed random variable, a Likelihood ratio based test, and a Bootstrap-based test. In the next chapter, the properties of all five tests will be compared and contrasted using an extensive simulation study.
CHAPTER 5 Simulation Study

In this chapter we describe a comparison of the three proposed inferential methods with the commonly used Two-Sample t-test and the Filtered Two Sample t-test performed using simulated data. Specifically, for each simulation scenario we examined either the Type I or Type II error rates. A simulated data set was imperative so that knowledge as to a certain CpG site’s true methylation status (differentially methylated between two groups, yes or no) was certain. We chose to design our simulation study to preserve as many of the characteristics inherent to our application dataset as possible. For this reason, data were generated in two manners, the first of which generated proportion methylated data according to a Beta distribution, and the second of which generated Red and Green channel data according to a Bivariate Normal distribution with correlation coefficient equal to -0.6 upon which \( \beta \) proportion methylated data were computed. The Type I and Type II error rates were calculated and compared for each inferential method under both simulation scenarios. Under each the Beta distribution and the Bivariate Normal distribution scenario, data were generated for 3,152 CpG sites where there were 20 samples in Group 1 and 16 samples in Group 2. This sample size was chosen through an analysis in nQuery. All significance tests were conducted at the \( \alpha = 0.05 \) significance level; therefore, because we wished to estimate a Type I error rate which we expected to be 0.05 and accurate to the hundredths decimal (\( +/- 0.01 \)), the number of samples needed for a two-sided 99.0%
confidence interval for a single proportion using the large sample normal approximation extending 0.01 from the observed proportion for an expected proportion of 0.05 is 3,152.

For a more complete analysis, the number of samples per group was varied to further explore the behavior of the Type I and Type II error rates due to sample size. Methodology for testing the Type I and Type II error rates follows.

5.1 Testing Type I error rates

Type I error is the error of rejecting a null hypothesis when it indeed is true. Comparing the Type I error rates of the two proposed tests along with the Two Sample t-tests involved calculating the number of times a certain test found a particular CpG site differentially methylated between two classes of outcomes when the truth was that the CpG site, in fact, was not differentially methylated. Therefore, it was of interest to generate a data set for two groups having no difference in proportion methylated. For a given inferential method, the Type I error was then calculated as the proportion of CpG sites declared significantly differentially methylated between two treatment groups.

In an effort to ensure the generation of realistic data, values from the test data were used as a starting point. The CpG site with Illumina code 3923 was used to model non-significantly differentially methylated data. Note that while the Two Sample t-test found CpG site 3923 non-significantly differentiated (p-value= 0.763), any difference in beta values between the two outcome groups present at this CpG site would become null by means of the data simulation method now explained.
5.1(a) Type 1 error data simulation methods and results: Beta distribution

To study the Type I error rate, beta distributed data were generated using the same shape and scale, \( p \) and \( q \), parameters for both outcome groups. Thus, using all proportion methylated observations across the two outcome groups of the CpG site with Illumina code 3923, the sample mean and variance were calculated. These moments were then used to compute the method of moments estimators for \( p \) and \( q \) as outlined in Equations 4.6 and 4.7. Using these estimators, \( \hat{p} \) and \( \hat{q} \), beta distributed data were generated for 3,152 CpG sites for both a Group 1 random beta variable \( X \sim Beta(\hat{p}, \hat{q}) \) containing 20 observations and a Group 2 random variable \( Y \sim Beta(\hat{p}, \hat{q}) \) containing 16. In this manner, it was ensured that no true statistically significant difference existed for each CpG site between the two groups of generated data.

The simulated data set was analyzed under the assumptions of each of the five testing measures. For the beta test, as outlined in Chapter 4, and the Type I error rate was calculated as

\[
\frac{\# \{ \text{CpG sites where } |\bar{x}_j - \bar{y}_j| > Q_{0.975, j} \}}{3,152},
\]

and under the remaining tests, the Type I error rate was calculated as

\[
\frac{\# \{ \text{CpG sites whose p value < 0.05} \}}{3,152}.
\]

Results are presented in Table 5.1.

Note that data generated according to the beta distribution are single valued on the range \((0,1)\). Because the bootstrap test relies on sampling data from the individual Red and Green
channels, it cannot be performed on data that have been generated according to a beta
distribution, and hence results of the Type I error behavior of the Bootstrap test under the
assumption of Beta distributed data is not available.

<table>
<thead>
<tr>
<th>Number of Samples per Group</th>
<th>Two Sample t-test</th>
<th>Beta test</th>
<th>Likelihood ratio test</th>
</tr>
</thead>
<tbody>
<tr>
<td>20.16</td>
<td>0.048</td>
<td>0</td>
<td>0.0416</td>
</tr>
</tbody>
</table>

Table 5.1 Type I Error rate comparison where data has been computed under the assumption of a Beta distribution.

Since testing was conducted at the $\alpha = 0.05$ significance level, a Type I error rate
close to 0.05 was considered most desirable. Using this criterion, it was evident that the
Two Sample t-test, the Beta distribution test and the Likelihood Ratio test each performed
well. Meanwhile, the Type I error rate of the filtered Two Sample t-test was too low.

To gain a better understanding of the behavior of the Type I error under the four
available testing conditions, data were again generated in the aforementioned manner while
varying the number of samples per group. Results are presented below in Table 5.2:
| Number of Samples per Group | Two Sample t-test | Two Sample T test p-value <0.05 and $|\Delta \beta| > 0.17$ | Beta test | Likelihood ratio test |
|-----------------------------|------------------|------------------------------------------------------|-----------|----------------------|
| 20,16                       | 0.048            | 0                                                   | 0.0416    | 0.048                |
| 40,40                       | 0.055            | 0                                                   | 0.0495    | 0.063                |
| 80,80                       | 0.054            | 0                                                   | 0.0425    | 0.059                |
| 45,50                       | 0.057            | 0                                                   | 0.0498    | 0.056                |
| 15,15                       | 0.046            | 0                                                   | 0.0374    | 0.043                |

Table 5.2 Type I Error rate comparison across varying sample sizes where data have been computed under the assumption of a Beta distribution.

Across a wide range of sample sizes, the Two Sample t-test, the Beta distribution test, and the Likelihood ratio based test perform very well and very comparably. Note that the filtered Two Sample t-test does not accurately quantify the Type I error rate. As expected, the Type I Error Rate was furthest from the nominal 0.05 level for all tests when the sample sizes were reduced to 15 and 15 in each group.

5.1(b) Type 1 error data simulation methods and results: Bivariate Normal distribution

As an alternative to the assumption that the proportion methylated variable follows a Beta distribution, it was discussed in Chapter 4.4 that there is evidence to support a claim that the Red and Green channels follow a Bivariate Normal distribution. Thus, a thorough examination of the Type I error rate behavior for these five statistical measures involves a comparison of the error rates when each test has been performed using data simulated under the assumption of Bivariate Normality. Once again, the CpG site in our test data set
with Illumina code 3923 was used to model the data. For this CpG site, the means and variances of each the red and green channels were calculated across all samples, regardless of classification group. This resulted in one mean and standard deviation for the red channel data and one mean and standard deviation for the green channel data. Furthermore, as previously stated, the chip-wide mean correlation between the Red and Green channels was approximately -0.6, so this value, together with the computed mean and variance parameters from CpG site 3923 were used as the parameters of the bivariate normal distribution that generated our test set data. Upon the generation of the 36 samples for each of the 3,152 CpG sites, any negative values were changed to zero to better simulate the beta variable data. Kolmogorov Smirnov tests were performed to ensure that evidence did not exist to claim that the resulting data did not preserve the Bivariate Normal distribution, and in all cases, this proved true.

Type I error rates were computed and are presented in Table 5.3.

| Number of Samples per Group | Two Sample t-test | Two Sample t-test p value <0.05 and |Δβ| > 0.17 | Beta test | Likelihood ratio test | Bootstrap test |
|-----------------------------|-------------------|------------------------------------|-----------|---------------|-----------------------|----------------|
| 20,16                       | 0.044             | 0                                  | 0.0416    | 0.048         | 0.048                 |

Table 5.3 Type I Error rate comparison where data has been computed under the assumption of a Bivariate Normal distribution.

Once again, the Two Sample t-test, the Beta test and the Likelihood ratio test all three perform reasonably well under the assumption that the data are distributed Bivariate Normal. Furthermore, neither the Filtered Two Sample t-test nor the Bootstrap test
accurately quantify the Type I error here. With the exception of the Likelihood ratio based test, this is a trend that continues as sample size is varied, as illustrated in Table 5.4 below. The Likelihood ratio based test, rather, has an inflated estimate of the Type I error rate as sample size increases, demonstrating that similar to the Filtered Two Sample t-test, it is not an appropriate decision measure.

It is interesting to note that regardless of whether the data are assumed to follow a Beta distribution or a Bivariate Normal Distribution, the Type I error behavior of each the Two-Sample t-test and the Beta test is both quite comparable and accurate.

| Number of Samples per Group | Two Sample t-test | Two Sample t-test p value < 0.05 and |Δβ| > 0.17 | Beta test | Likelihood ratio test | Bootstrap test |
|-----------------------------|-------------------|-------------------------------------|----------|----------------------|-------------------|
| 20,16                       | 0.044             | 0                                   | 0.0416   | 0.048                | 0.048             |
| 40,40                       | 0.042             | 0                                   | 0.0495   | 0.075                | 0.042             |
| 80,80                       | 0.059             | 0                                   | 0.0425   | 0.076                | 0.058             |
| 45,50                       | 0.047             | 0                                   | 0.0498   | 0.070                | 0.045             |
| 15,15                       | 0.040             | 0                                   | 0.0374   | 0.040                | 0.043             |

Table 5.4 Type I Error rate comparison across varying sample sizes where data has been computed under the assumption of a Bivariate Normal distribution.

5.2 Testing Type II Error Rates

In addition to comparing the Type I error rates, a complete analysis of the five proposed tests also involved the comparison of their Type II error rates. Type II error is the error of failing to reject a null hypothesis when it is not true. The computation of the Type II error resulted from calculating the number of times a certain test found no
difference in the methylation status of particular CpG site between two classes of outcomes when the truth was that the sites were indeed differentially methylated. Therefore, it was of interest to generate a data set that for two outcome groups, a difference in methylation status existed for every CpG site. The Type II error was then calculated as the proportion of CpG sites that a certain test found non-significantly differentially methylated between two treatment groups. All significance tests were performed at the $\alpha = 0.05$ significance level.

Again, to ensure the generation of realistic data, values from the test data were used as a starting point. In Bibikova, et al. (2006b), it was estimated that the Illumina technology can discriminate levels of methylation ($\beta$-values) that differ by 0.17 or more. So to generate data, the means and variances of a CpG site that demonstrated significant differential methylation were used as a starting point for generating data. Because we were searching for a site with measurable methylation levels, a certain CpG site was considered differentially methylated if the mean proportion methylated in each of the treatment groups was at 0.2 or greater and if the difference in the mean proportion methylated between the two groups was 0.17 or more. When tested by the currently accepted method of the two-sample t-test, the CpG site with Illumina code 4092 had a difference in beta values between the two groups of 0.172 with a corresponding p-value of 0.005. Furthermore, this site demonstrated a mean methylation of 0.46 and 0.29 in the first and second outcome groups, respectively. Thus, this site was considered differentially methylated and was used initially to model data for simulations exploring the behavior of the Type II error. Again, the Type II error behavior of each test was explored under the assumption that data were
distributed according to each the Beta distribution and the Bivariate Normal distribution. Data generation and results under these two scenarios are as follows.

5.2(a) Type II error data simulation methods and results: Beta distribution

To study the Type II error rate, data from the two outcome groups were generated using different beta distributions. To generate data from differing beta distributions, the method of moments estimators were calculated for each of the treatment groups of CpG site 4092 resulting in $\hat{p}_1$ and $\hat{q}_1$ for the first treatment group and $\hat{p}_2$ and $\hat{q}_2$ for the second treatment group. These estimators were used to generate two random beta variables, $A \sim Beta(\hat{p}_1, \hat{q}_1)$ and $B \sim Beta(\hat{p}_2, \hat{q}_2)$ each of length 1,000. One thousand was chosen here as a large number to give stable estimates. The standardized quantile, $Q_{0.975}$ was computed according to Equation 4.10 as a function of this combined generated data. Next, two new beta random variables, $X \sim Beta(\hat{p}_1, \hat{q}_1)$ and $Y \sim Beta(\hat{p}_2, \hat{q}_2)$ were generated of length 20 and 16, respectively, for the 3,152 CpG sites to achieve the desired accuracy. In this manner, it was ensured that a difference in the mean proportion methylated between $X$ and $Y$ for the $j$th CpG site was equal to approximately 0.17, as presented in the test data CpG site 4092.

The simulated data set was analyzed under the assumptions of each of the five testing measures. For the beta test, as outlined in Chapter 3, and the Type II error rate was calculated as
Type II Error $= 1 - \text{power}$

$$= 1 - \# \{ \text{CpG sites where } |\bar{x}_j - \bar{y}_j| > C_{0.975,j} \} / 3,152,$$

and under the remaining tests, the Type II error rate was calculated as

$$\# \{ \text{CpG sites whose } p \text{ value } > 0.05 \} / 3,152.$$

Results are presented in Table 5.5.

Note again that data generated according to the beta distribution are single valued on the range (0,1). Because the bootstrap test relies on sampling data from the individual Red and Green channels, it cannot be performed on data that have been generated according to a beta distribution, and hence results of the Type II error behavior of the Bootstrap test under the assumption of Beta distributed data is not available.

Furthermore, it is important to note that the Type II error rates for the Filtered Two Sample t-test are not included in subsequent Type II error rate tables (Tables 5.5-5.12). In the identification of a uniformly most powerful test, power should be compared among tests at a fixed alpha-level, and the Filtered Two Sample t-test cannot attain a Type I error of 0.05 due to the imposed threshold. Because the (Unfiltered) Two Sample t-test, the Beta test, the Likelihood ratio-based test, and the Bootstrap-based test (where applicable) each demonstrated a roughly equivalent Type I error rate about $\alpha = 0.05$ across various sample sizes, the Type II error rates of these tests are reported in subsequent tables with the Type II error rate of the Filtered Two Sample t-test following in the text for informational purposes.
Table 5.5: Type II Error rate comparison where data has been computed under the assumption of a Beta distribution.

<table>
<thead>
<tr>
<th>Test Set Δβ</th>
<th>Observed Δβ</th>
<th>Illumina Code</th>
<th>X</th>
<th>Y</th>
<th>Test Set p-value (t-test)</th>
<th>Two Sample t-test</th>
<th>Beta Test</th>
<th>Likelihood ratio test</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.17</td>
<td>0.17</td>
<td>946</td>
<td>0.46</td>
<td>0.29</td>
<td>0.005</td>
<td>0.177</td>
<td>0.181</td>
<td>0.055</td>
</tr>
</tbody>
</table>

In Table 5.5, the Test Set Δβ column lists the difference in mean proportion methylated between the two outcome groups for which we sought to generate data. The Observed Δβ column lists the actual Δβ present in our generated data. Meanwhile, the columns denoted Illumina Code, X, Y, and Test Set p-value (t-test) describe the reference Illumina code of the CpG site used to generate our data, the mean proportion methylated present in our test data for the first and second outcome groups, and the p-value of this CpG site when a Two Sample t-test is performed to test for differential methylation, respectively. The aforementioned columns are listed to demonstrate that our data were successfully generated using a CpG site containing both positive methylation and significantly differentiated methylation, according to the currently accepted testing method. The following three columns show the Type II error rates of the three applicable proposed tests under the assumption that the data were generated according to a Beta distribution. When assessing Type II error, a smaller rate is better, where approximately 0.20 approaches the upper limit of a Type II error rate for a well-suited test. Here, we see that each the Two Sample t-test, the Beta Test, and the Likelihood ratio based test each perform well. The Type II error rate for the Filtered Two Sample t-test under the assumption of a sample size of 20 in outcome group 1 and 16 in outcome group 2 is 0.491.
Therefore, while we cannot directly compare this error rate to those of the other three applicable tests under the assumption of Beta distributed data, we can conclude that the Filtered Two Sample t-test is a highly conservative testing measure that in possessing a small Type I error rate, likely trades for it a Type II error rate that is too high. This translates to a loss in power. This is a trend that is more fully illustrated as we vary both the level of differential methylation (Table 5.6) and the sample size (Tables 5.7-5.8).

Table 5.6 details how the Type II error rate changes with varying levels of $\Delta \beta$. We observe that a similar trend for the Type II error rate continues across $\Delta \beta$. Namely, the Two Sample t-test, the Beta test and the Likelihood ratio based test each perform well concerning the preservation of an acceptable error rate. The Type II error rates of the Filtered Two Sample t-test for differential methylation levels are as follows: Type II Error rate($\Delta \beta = 0.20$ )=0.345, Type II Error rate($\Delta \beta = 0.10$, CpG site 2358 )=1.0, Type II Error rate($\Delta \beta = 0.10$, CpG site 2208 )=0.949, and Type II Error rate($\Delta \beta = 0.05$ )=1.0. These rates are clearly consistently high and even become severely inflated, as hypothesized, when the observed $\Delta \beta$ dips below the threshold.

An additional artifact of interest in Table 5.6 is the difference in Type II Error rates observed for each the Two Sample t-test, the Beta test and the Likelihood ratio based test between the two sites presenting a test set $\Delta \beta=0.10$. Note that while each CpG site 2358 and 2208 were used to successfully generated data with $\Delta \beta=0.10$, the Type II error for the three aforementioned tests is much higher when CpG site 2208 had been used to generate data, rather than when CpG site 2358. Two reasons have been identified to explain this
phenomenon. First, we noted that according to the currently accepted testing method, the
Two Sample t-test, site 2358 has a much smaller p-value than site 2208. Second, it can be
computed that the variances of the two outcome groups present in site 2358 is much
smaller than the variances present in the two outcome groups present in site 2208. This is
verified by noting that as estimated by the method of moments parameters for the Beta
distribution in Equations 4.6 and 4.7. \( \hat{p}_1 = 1.89, \hat{q}_1 = 2.24, \hat{p}_2 = 4.09, \) and \( \hat{q}_2 = 10.22. \) So,
when computing the variances of each outcome group for these two Beta-distributed sites
according to the equation,
\[
\text{var}(z) = \frac{pq}{(p + q)^2 (p + q + 1)},
\]
(5.5)
it was determined that for site 2358, \( \text{var}(z_1) = 0.003 \) and \( \text{var}(z_2) = 0.005, \) while for site
2208, \( \text{var}(z_1) = 0.024 \) and \( \text{var}(z_2) = 0.014. \) Thus, due to these two reasons, we can
conclude that while sites 2358 and 2208 both generated data to test for a \( \Delta \beta = 0.10, \) the
data of CpG site 2358 both demonstrated a more significant difference between the two
treatment groups and possessed less variation in the data, resulting in more precisely
significantly differentiated generated data on which to conduct hypothesis testing.
Therefore, it is understood that CpG site 2358 generates data that produces a more
acceptable Type II Error rate than CpG site 2208.
The next step in the analysis was to explore the behavior of the Type II error rate across varying levels of differential proportion methylated for each 40 samples in outcome groups 1 and 2 and 80 samples in outcome groups 1 and 2. Results are presented in Tables 5.7 and 5.8, respectively.

When data have been generated according to the Beta distribution with 40 samples in each of the treatment groups, the same trends seen previously with smaller sample sizes
continued and became even more defined. We see that as sample size increased, the Type II error rates of each the Two Sample t-test, the Beta distribution test, and the Likelihood ratio based test decreased, as to be expected. We also note that the increased sample size helped to overcome the increased error rate for the more marginally significant and highly dispersed data generated by CpG site 2208 among the three currently well-performing tests. The Type II error rates under the assumption of 40 samples in each outcome group are as follows: Type II Error rate($\Delta \beta = 0.17$)=0.494, Type II Error rate($\Delta \beta = 0.20$)=0.264, Type II Error rate($\Delta \beta = 0.10$, CpG site 2358)=1.0, Type II Error rate($\Delta \beta = 0.10$, CpG site 2208)=0.991, and Type II Error rate($\Delta \beta = 0.05$)=1.0. Because of the unbending threshold imposed upon the Filtered Two Sample t-test, the Type II error rate of this test still remained severely inflated for levels of $\Delta \beta$ at or below 0.17. As seen below, these trends become further defined as sample size increases to 80 samples in each outcome group. Therefore, when solely observing the Type II error rate across both varying levels of differential methylation and sample size, the Two Sample t-test, the Beta distribution test, and the Likelihood ratio based test each perform well.
We now turn to examine the behavior of the Type II error under both varying levels of differential methylation and sample size for the proposed tests where data have been generated according to the Bivariate Normal distribution.

5.2(b) Type II error data simulation methods and results: Bivariate Normal distribution

As in the examination of Type I error, the Type II error behavior of the five proposed tests was explored under the assumption that it was generated both under a Beta distribution and a Bivariate Normal distribution. Data generation under the Bivariate Normal distribution was as follows. Again, initially using test data CpG site 4092 to model realistic data with an approximate mean proportion methylated difference of 0.17 between the two outcome groups, the mean and variance of each the Red and Green channels for the Group 1 were found. Together with a correlation coefficient of $-0.6$, these values were used to generate bivariate normal random variables that became the Red

---

### Table 5.8 Type II Error rate comparison across varying levels of differential proportion methylated where data has been generated with 80 samples each in outcome groups 1 and 2 according to a Beta distribution.

<table>
<thead>
<tr>
<th>Test Set $\Delta \beta$</th>
<th>Observed $\Delta \beta$</th>
<th>Illumina Code</th>
<th>$\bar{X}$</th>
<th>$\bar{Y}$</th>
<th>Two Sample t-test p-value (t test)</th>
<th>Beta Test p-value (likelihood ratio test)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.17</td>
<td>0.17</td>
<td>4092</td>
<td>0.46</td>
<td>0.29</td>
<td>0.005</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>0.20</td>
<td>0.20</td>
<td>4297</td>
<td>0.52</td>
<td>0.32</td>
<td>0.008</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>0.10</td>
<td>0.10</td>
<td>2358</td>
<td>0.47</td>
<td>0.38</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>0.10</td>
<td>0.10</td>
<td>2208</td>
<td>0.47</td>
<td>0.37</td>
<td>0.039</td>
<td>0.006</td>
</tr>
<tr>
<td>-0.05</td>
<td>-0.05</td>
<td>4038</td>
<td>0.76</td>
<td>0.81</td>
<td>0.009</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

---
and Green channels for the first group of the simulated data. This process was repeated by calculating the mean and variance of each the Red and Green channels for Group 2 in selected the CpG site of the test data set. The generated bivariate normal random variables using these values became the Red and Green channels for the second group of the simulated data. Like when testing the Type I error rate, the generated data set was filtered so that any negative values were changed to zero, per the formula for the calculation of $\beta$, $(\beta = \max(R,0)/(\max(R,0) + \max(G,0)))$. This process was performed 3,152 times, meaning that all CpG sites were generated to be differentially methylated between the two groups. Then, with this fully prepared data set, the significance testing again commenced. Results of this analysis are presented below.

<table>
<thead>
<tr>
<th>Test Set $\Delta\beta$</th>
<th>Obs. $\Delta\beta$</th>
<th>Illum. Code</th>
<th>$\bar{X}$</th>
<th>$\bar{Y}$</th>
<th>Test Set p-value (t test)</th>
<th>Two Sample t-test</th>
<th>Beta Test</th>
<th>Likelihood ratio test</th>
<th>Bootstrap Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.17</td>
<td>0.15</td>
<td>946</td>
<td>0.46</td>
<td>0.29</td>
<td>0.005</td>
<td>0.192</td>
<td>0.251</td>
<td>0.135</td>
<td>0.213</td>
</tr>
</tbody>
</table>

**Table 5.9** Type II Error Rate comparison at minimum methylation discrimination level where data has been generated according to the Bivariate Normal distribution with 20 samples in outcome group 1 and 16 samples in outcome group 2.

To reiterate, in Table 5.9, the Test Set $\Delta\beta$ column lists the difference in mean proportion methylated between the two outcome groups for which we sought to generate data. The Observed $\Delta\beta$ column lists the actual $\Delta\beta$ present in our generated data. Meanwhile, the columns denoted Illumina Code, $\bar{X}$, $\bar{Y}$, and Test Set p-value (t-test) describe the reference Illumina code of the CpG site used to generate our data, the mean
proportion methylated present in our test data for the first and second outcome groups, and
the p-value of this CpG site when a two sample t-test is performed to test for differential
methylation, respectively. The following four columns illustrate the Type II error rate for
the four proposed tests when data have been generated under the assumptions of the given
CpG site. Because data generated under the Bivariate Normal distribution in a sense
creates two channel data, the Type II error of the Bootstrap test can now be assessed, as it
relies on this information.

Table 5.9 already demonstrates similar results to those seen when data were
generated according to the Beta distribution--each test preserved appropriate Type II error
rates. The Type II error rate of the Filtered Two Sample t-test when testing a differential
methylation of 0.17 is 0.616.

The effect of the magnitude of the difference in methylation levels on Type II error
rate was next explored. The process of generating data and running the five tests was
repeated where different significantly differentiated CpG sites from the test data were
chosen to generate the data. These CpG sites from the test data were chosen to achieve
varying levels in the \( \Delta \beta \) variable, as the CpG site with Illumina code 4092 was chosen to
achieve a \( \Delta \beta = 0.17 \). Results are presented in Table 5.10 below.
### Table 5.10

Type II Error rate comparison across varying levels of differential proportion methylated where data has been generated with 20 samples in outcome group 1 and 16 samples in outcome group 2 according to a Bivariate Normal distribution.

<table>
<thead>
<tr>
<th>Test Set ( \Delta \beta )</th>
<th>Obs. ( \Delta \beta )</th>
<th>Illum. Code</th>
<th>( \bar{X} )</th>
<th>( \bar{Y} )</th>
<th>Test Set p-value (t test)</th>
<th>Two Sample t-test</th>
<th>Beta Test</th>
<th>Likelihood ratio test</th>
<th>Bootstrap Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.17</td>
<td>0.15</td>
<td>4092</td>
<td>0.46</td>
<td>0.29</td>
<td>0.005</td>
<td>0.192</td>
<td>0.251</td>
<td>0.135</td>
<td>0.213</td>
</tr>
<tr>
<td>0.20</td>
<td>0.17</td>
<td>4297</td>
<td>0.52</td>
<td>0.32</td>
<td>0.008</td>
<td>0.176</td>
<td>0.208</td>
<td>0.171</td>
<td>0.156</td>
</tr>
<tr>
<td>0.10</td>
<td>0.09</td>
<td>2358</td>
<td>0.47</td>
<td>0.38</td>
<td>&lt;0.001</td>
<td>0.250</td>
<td>0.229</td>
<td>0.266</td>
<td>0.228</td>
</tr>
<tr>
<td>0.10</td>
<td>0.08</td>
<td>2208</td>
<td>0.47</td>
<td>0.37</td>
<td>0.039</td>
<td>0.539</td>
<td>0.576</td>
<td>0.441</td>
<td>0.563</td>
</tr>
<tr>
<td>-0.05</td>
<td>-0.04</td>
<td>4038</td>
<td>0.76</td>
<td>0.81</td>
<td>0.009</td>
<td>0.701</td>
<td>0.601</td>
<td>0.646</td>
<td>0.690</td>
</tr>
</tbody>
</table>

Here, each of the four tests included in Table 5.10 perform relatively well in respect to the Type II error rate across most varying levels of differential proportion methylated.

However, for small \( \Delta \beta \) (i.e., \( \Delta \beta = 0.05 \)), these tests all exhibit a high Type II error.

Alternatively, the Type II error rates of the Filtered Two Sample t-test are as follows: Type II Error rate(\( \Delta \beta = 0.17 \))=0.616, Type II Error rate(\( \Delta \beta = 0.20 \))=0.642, Type II Error rate(\( \Delta \beta = 0.10 \), CpG site 2358)=0.994, Type II Error rate(\( \Delta \beta = 0.10 \), CpG site 2208)=0.986, and Type II Error rate(\( \Delta \beta = 0.05 \))=1.0, which are, again, inflated, especially for \( \Delta \beta < 0.17 \), the imposed threshold level. Again, as observed in Chapter 5.2(a), the Type II error rates of all tests inflates when the data used to generate the data are both more marginally significant and more variable. This is seen when comparing the Type II error rates for data generated using the more highly significant and less variable CpG site 2358 to the less significant and more variable CpG site 2208 which are both used to test for a difference in proportion methylated of 0.10.
The next step in the analysis was to explore the behavior of the Type II error rate across varying levels of differential proportion methylated for each 40 samples in outcome groups 1 and 2 and 80 samples in outcome groups 1 and 2. Results are presented in Tables 5.11 and 5.12, respectively.

<table>
<thead>
<tr>
<th>Test Set Δβ</th>
<th>Obs. Δβ</th>
<th>Illum. Code</th>
<th>$\bar{X}$</th>
<th>$\bar{Y}$</th>
<th>Test Set p-value (t test)</th>
<th>Two Sample t-test</th>
<th>Beta Test</th>
<th>Likelihood ratio test</th>
<th>Bootstrap Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.17</td>
<td>0.15</td>
<td>4092</td>
<td>0.46</td>
<td>0.29</td>
<td>0.005</td>
<td>0.014</td>
<td>0.024</td>
<td>0.004</td>
<td>0.013</td>
</tr>
<tr>
<td>0.20</td>
<td>0.17</td>
<td>4297</td>
<td>0.47</td>
<td>0.38</td>
<td>&lt;0.001</td>
<td>0.023</td>
<td>0.015</td>
<td>0.028</td>
<td>0.023</td>
</tr>
<tr>
<td>0.10</td>
<td>0.09</td>
<td>2358</td>
<td>0.47</td>
<td>0.38</td>
<td>&lt;0.001</td>
<td>0.022</td>
<td>0.258</td>
<td>0.110</td>
<td>0.249</td>
</tr>
<tr>
<td>0.10</td>
<td>0.08</td>
<td>2208</td>
<td>0.47</td>
<td>0.37</td>
<td>0.039</td>
<td>0.418</td>
<td>0.314</td>
<td>0.347</td>
<td>0.411</td>
</tr>
<tr>
<td>-0.05</td>
<td>-0.04</td>
<td>4038</td>
<td>0.76</td>
<td>0.81</td>
<td>0.009</td>
<td>0.013</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
</tr>
</tbody>
</table>

**Table 5.11** Type II Error rate comparison across varying levels of differential proportion methylated where data has been generated with 40 samples in each outcome group 1 and group 2 according to a Bivariate Normal distribution.

<table>
<thead>
<tr>
<th>Test Set Δβ</th>
<th>Obs. Δβ</th>
<th>Illum. Code</th>
<th>$\bar{X}$</th>
<th>$\bar{Y}$</th>
<th>Test Set p-value (t test)</th>
<th>Two Sample t-test</th>
<th>Beta Test</th>
<th>Likelihood ratio test</th>
<th>Bootstrap Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.17</td>
<td>0.15</td>
<td>4092</td>
<td>0.46</td>
<td>0.29</td>
<td>&lt;0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>&lt;0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>0.20</td>
<td>0.17</td>
<td>4297</td>
<td>0.52</td>
<td>0.32</td>
<td>&lt;0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>0.10</td>
<td>0.09</td>
<td>2358</td>
<td>0.47</td>
<td>0.38</td>
<td>&lt;0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>0.10</td>
<td>0.08</td>
<td>2208</td>
<td>0.47</td>
<td>0.37</td>
<td>0.030</td>
<td>0.041</td>
<td>0.006</td>
<td>0.044</td>
<td>0.044</td>
</tr>
<tr>
<td>-0.05</td>
<td>-0.04</td>
<td>4038</td>
<td>0.76</td>
<td>0.81</td>
<td>0.009</td>
<td>0.132</td>
<td>0.089</td>
<td>0.064</td>
<td>0.130</td>
</tr>
</tbody>
</table>

**Table 5.12** Type II Error rate comparison across varying levels of differential proportion methylated where data has been generated with 80 samples in each outcome group 1 and group 2 according to a Bivariate Normal distribution.
Tables 5.11 and 5.12 illustrate that as sample size increases, the Type II error rate of each the Two Sample t-test, the Beta test, the Likelihood ratio based test, and the Bootstrap based test decreases. Even for testing a very small absolute difference in proportion methylated of 0.05 does the Type II error rate become acceptably small when sample size increases to 80 samples per outcome group. In fact, when there are 40 samples per outcome group, the Beta test, in particular, demonstrates a Type II error rate of 0.314, which could be considered an acceptable error rate under certain testing scenarios. The Type II error rates of the Filtered Two Sample t-test when there are 40 samples in each outcome groups 1 and 2 are as follows: Type II Error rate($\Delta \beta = 0.17$) = 0.671, Type II Error rate($\Delta \beta = 0.20$) = 0.725, Type II Error rate($\Delta \beta = 0.10, \text{CpG site } \text{CpG site } 2358$) = 1.0, Type II Error rate($\Delta \beta = 0.10, \text{CpG site } 2208$) = 0.999, and Type II Error rate($\Delta \beta = 0.05$) = 1.0.

Furthermore, the Type II error rates of the Filtered Two Sample t-test when there are 80 samples in each outcome groups 1 and 2 are as follows: Type II Error rate($\Delta \beta = 0.17$) = 0.746, Type II Error rate($\Delta \beta = 0.20$) = 0.803, Type II Error rate($\Delta \beta = 0.10, \text{CpG site } 2358$) = 1.0, Type II Error rate($\Delta \beta = 0.10, \text{CpG site } 2208$) = 1.0, and Type II Error rate($\Delta \beta = 0.05$) = 1.0.

5.3 Application of the Beta Test to Hepatitis C Virus data set

As demonstrated by both appropriate Type I and Type II error rates, the Beta test is a well-suited statistical measure for detecting differential methylation between two
outcome groups. As previously stated, methodologies for the development of alternative
testing procedures were motivated by a Hepatitis C Virus data set. (Archer et al., 2010)
The full dataset consists of methylation data from 1,547 CpG sites for analysis for each of
36 patients who present either HCV cirrhosis without concomitant Hepatocellular
Carcinoma (HCC) (N=16) or HCV cirrhosis with concomitant HCC (N=20). While results
have been published applying a Filtered Two Sample t-test to this data, a re-analysis was
conducted applying the Beta test here developed. At the $\alpha = 0.05$ level, it was determined
that 277 CpG sites were significantly differentially methylated, as compared with the 205
sites differentially methylated using a standard Two Sample t-test. As to be expected,
there was much overlap between sites found significantly differentially methylated under
the two testing scenarios. A table of these 277 CpG sites found differentially methylated is
found in Appendix A.

5.4 Conclusions: Determination of a well-defined statistical testing measure

In choosing an appropriate statistical testing measure for a certain class of data, it is
important to weigh both the Type I and Type II error rates. Here, we sought an appropriate
testing measure for the identification of differential CpG site methylation between two
groups of samples—a typical testing scenario concerning data collected with the Illumina
GoldenGate technology. Currently, both the Two Sample t-test and, increasingly, the
Filtered Two Sample t-test are widely used to analyze differential methylation patterns in
this class of data. After concern that both the distributional assumption of normality
underlying the Two Sample t-test might not be appropriate for testing on a proportion methylated variable and that the Filtered Two Sample t-test could reduce power for sites exhibiting truly differential methylation that was below the filter level, alternative testing measures were explored. These three alternative testing measures took the form of a test based on the Beta distribution and on two large sample tests, a likelihood ratio based test and a bootstrap based test. It had been hypothesized that the proportion methylated variable, $\beta$, upon which hypothesis testing was conducted, followed a Beta distribution, and it had also been hypothesized that the Red and Green two channel data followed a Bivariate Normal distribution. Therefore, data were generated under each of these distributions, and the Type I and Type II error rates were computed for each test when applied to these simulated data sets.

For data generated under each of the distributions, both the Two Sample t-test and the Beta distribution test accurately quantify Type I error rates across small to large sample sizes. The Bootstrap based test also accurately quantified the Type I error rate under the distribution assumption of Bivariate normality, but by nature of the test construction, the Type I error rates could not be applied to the single valued data generated here under a Beta distribution. Quite notably however, the Filtered Two Sample t-test demonstrates a Type I error rate that is too low, and so it does not accurately quantify the Type I error rate. Additionally, the Likelihood ratio based test exhibited a Type I error rate that was slightly inflated under the assumption of Bivariate Normal data. Concerning the Type II error rates, the Two Sample t-test, the Beta test and the Likelihood Ratio based test preserved an appropriate error, especially for large sample sizes, under both distributional assumptions.
The Bootstrap based test, performed well with respect to the Type II error rate, as well, under the testable assumption of Bivariate Normality. Owing to both its differential Type I error rate when compared to the four alternative testing measures and its inability for fixing this error, the Filtered Two Sample t-test could not be directly compared. Nevertheless, it was evident that the imposition of a filter likely decreases power for truly differential sites that exhibit $\Delta \beta$ less than the filtering level.

Because the Two Sample t-test and the Beta test demonstrate both appropriate Type I and Type II error rates, these two testing measures are decided to be the most preferable tests for CpG site methylation data. While the Bootstrap based test demonstrated both appropriate Type I and Type II error rates under the assumption of Bivariate Normal data, the Two Sample t-test and the Beta test have the distinct advantage that they can be performed on data presented as either two channel array data or computed proportion methylated data. Because the Bootstrap based test relies on sampling data from each the Red and Green channels, it cannot be applied to data that has already been summarized as proportion methylated data, as is frequently the case.

Concerning the Two Sample t-test and the Beta test, it cannot be claimed whether one of these two tests is superior to the other. Rather, the Type I and Type II error rates of both of these tests under each distributional assumption are quite comparable. A clinician might prefer to use the Two Sample t-test because it is a well known and straightforward testing measure that is already in use. However, despite its accuracy, it can still be claimed that the fundamental distributional assumption of data arising from the entire real number line is violated. It is for this reason, on the other hand, that the statistician might prefer to
use the Beta distribution test because it is more statistically rigorous in its distributional assumptions.

Two final points merit noting. First, one should note how poorly the Filtered Two Sample t-test performed. As hypothesized, the application of a filter severely reduces power for CpG sites demonstrating smaller but truly differential levels of proportion methylation. Therefore, the application of a Filtered Two Sample t test cannot be recommended for the analysis of CpG site methylation data. Second, one should note that especially for larger sample sizes, the Two Sample t-test and the Beta distribution test perform well across levels of differential methylation—even those levels of methylation less than 0.17. This gives evidence in support of a hypothesis that the Illumina technology can, in fact, detect differential methylation at levels less than $\Delta \beta = 0.17$. How small of a difference in methylation that the Illumina technology can detect was addressed in Chapter 4.
CHAPTER 6 Conclusions and Recommendations

Abnormal DNA methylation has emerged as an important area of study for understanding many clinical pathologies. As with any developing clinical theories, sound statistical methods must be devised to analyze the new types of emerging data to ensure accurate conclusions. In this thesis, the methods for identifying a change in methylation pattern were studied and modifications to these methods were proposed. Specifically, a study of abnormal DNA methylation as it relates in the progression from HCV-cirrhosis to HCC was employed as an impetus for our analysis. For this study, data were collected according to the GoldenGate Methylation BeadArray Cancer Panel I platform developed by Illumina for cancer-focused methylation analysis. This technology presents data in the form of two channel array data containing expression values for the methylation target sequence (Red) as well as expression values for the unmethylated target sequence (Green). Data are summarized in the form of a proportion methylated variable for each CpG site, and current analysis methods for the identification of differential methylation between two outcome groups at each CpG site include the performance of either a Two-Sample t-test or a Filtered Two-Sample t-test on the proportion methylated variable.

Each the Two Sample t-test and the Filtered Two Sample t-test elicit concerns over their statistical assumptions and thus the validity of their conclusions. The Two Sample t-test assumes that the data upon which it is being performed arise from a distribution
encompassing the entire real number line, which violates the fact that the proportion methylated variable is, in fact, a proportion on the range of 0 to 1. The Filtered Two Sample t-test imposes the additional constraint that a CpG site can be concluded differentially methylated between two outcome groups if the difference in the proportion methylated variables between the two groups is greater than some threshold. Any CpG sites that are truly differentially methylated but possess a difference in methylation that is less than the threshold would then present no statistical power. Therefore, the validity of the Filtered Two Sample t-test rests upon the accuracy of the imposed threshold. It was in response to these concerns that the aims of this thesis materialized. First, we sought statistically to demonstrate any deficiencies in the two existing testing methods while proposing appropriate alternative testing methods. Second, we sought to establish an accurate minimum level of discernable differential methylation for the Illumina technology, both for reference value and for statistical use if a filter was to be imposed upon a statistical test, as is increasingly commonly seen in literature.

In accomplishing our first aim, we proposed a test based on a Beta distributed random variable. Our motivation behind the Beta test was based on the assumption that the Red and Green channels followed a Gamma distribution and hence the ratio would be beta distributed. It is important to note that this distributional property supposes that the two Gamma distributed random variables were independent of each other. As expected from the lack of independence between the methylated and unmethylated states of a given CpG site, this was not the case with the given data, but rather the mean correlation of the Red and Green channels across all 36 chips was -0.62 (standard error = 0.03).
Nevertheless, by means of a simulation study comparing the Type I and Type II error rates of this proposed test to that of the Two Sample t-test, we concluded that both tests, in fact, performed equally well. Alternatively, due to such differing Type I error rates, the Type II error rates of the Filtered Two Sample t-test could not be directly compared to those of the remaining tests. However, we gained evidence in support of our hypothesis that the Filtered Two-Sample t-test is not a viable testing alternative in that it appears to inflate Type I error while diminishing Type II error, especially for the CpG sites mentioned—those with smaller but truly differential methylation that the filter allows. As a result of these findings, it is recommended that either the Two Sample t-test or the test based on a Beta distributed random variable be employed in future studies. In fact, both of these testing measures perform as well as any appropriately designed statistical test, so it is not recommended that any type of filter needs to be applied. As far as deciding between these two tests which testing measure to use, it is assumed that a clinician might find the Two Sample t-test preferable to use because of its simplicity and familiarity in implementation. On the other hand, it could be claimed that a statistician would prefer the use of the test based on a Beta distributed random variable, since it is statistically sound in its assumptions. For, as well as the Two Sample t-test performs, it still can be argued that it violates a basic distributional assumption of the proportion methylated variable.

Our findings into the poor performance of the Filtered t-test further supported the second aim—to establish an accurate minimum level of discernable differential methylation. Again, two alternative testing measures have been well established in the Two Sample t-test and the test based on a Beta distributed random variable to accurately
differentiate methylation patterns between two outcome groups, and so there is no need to apply any type of filter when simply looking for differentially methylated CpG sites. Nevertheless, if a minimum level of differentiation is to be employed, an accurate value must be established. While a previous estimate of 0.17 had been established based on a dilution study of female genomic DNA into male genomic DNA, a closer look into the analysis of the study revealed that not all known characteristics of the data had been properly accounted for, and so it was concluded that by re-analyzing the data, a more appropriate minimum discernable level could be established. By means of a mixed effects Logistic regression model, it was sought to model methylation levels as a function of the dilution ratio of Female to Male DNA while accounting for random differences among dilution order, technical replicate order, and CpG site. This nonlinear mixed effects model accounted for the true shape of the relationship between the percent of Female genomic DNA in a sample and the proportion of methylation while establishing the significance of this relation. This analysis was followed by the formulation of a mixed ANOVA model, accounting for the same variables and random differences, for the purpose of statistical comparisons of adjacent mixing ratios. In so doing, it was determined that the Illumina technology was able to differentiate between levels of methylation that differed by as little as 0.086 (comparing 5% Female: 0% Male versus 0% Female: 100% Male). Clearly, the Illumina technology can discern levels of methylation with more accuracy than anticipated, and so the use of filters on the order of 0.17 or greater that are seen in literature are missing much valuable information that is ready to be used.
In conclusion, one seeking to identify CpG sites which are differentially methylated between two outcome groups should employ either a Two Sample t-test or a test based on a Beta distribution. Furthermore, it has been concluded that the Illumina technology can differentiate methylation levels that differ at least as low as 0.086. Therefore, although it is not recommended in a traditional testing scenario, if a filter is to be applied to the Two Sample t-test procedure, it should be on the order of 0.086.
Literature Cited
Literature Cited


Bird, AP. Nucleic Acids Research, 8:1499-1504.


### APPENDIX A

Table of CpG sites found significant using test based on the Beta Distribution

<table>
<thead>
<tr>
<th>Gene Symbol</th>
<th>Illumina ID</th>
<th>Mean Proportion Methylated for Cirrhosis Group, $\beta_1$</th>
<th>Mean Proportion Methylated for HCC Group, $\beta_2$</th>
<th>$\Delta\beta$</th>
<th>Beta Test</th>
<th>Two Sample t-test</th>
<th>Filtered Two Sample t-test</th>
</tr>
</thead>
<tbody>
<tr>
<td>AATK</td>
<td>AATK_P709_R_01</td>
<td>0.021</td>
<td>0.031</td>
<td>0.010</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>ABCB4</td>
<td>ABCB4_E429_F_01</td>
<td>0.904</td>
<td>0.881</td>
<td>-0.023</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ABCB4</td>
<td>ABCB4_P51_F_01</td>
<td>0.587</td>
<td>0.670</td>
<td>0.083</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>ABL1</td>
<td>ABL1_P53_F_01</td>
<td>0.033</td>
<td>0.046</td>
<td>0.014</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>AFF3</td>
<td>AFF3_P808_F_01</td>
<td>0.156</td>
<td>0.236</td>
<td>0.079</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>AKT1</td>
<td>AKT1_P310_R_01</td>
<td>0.946</td>
<td>0.968</td>
<td>0.022</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>ALK</td>
<td>ALK_P28_F_01</td>
<td>0.826</td>
<td>0.776</td>
<td>-0.050</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ALPL</td>
<td>ALPL_P278_F_01</td>
<td>0.377</td>
<td>0.500</td>
<td>0.123</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>APP</td>
<td>APP_E8_F_01</td>
<td>0.975</td>
<td>0.963</td>
<td>-0.013</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ARAF</td>
<td>ARAF_E38_F_01</td>
<td>0.953</td>
<td>0.966</td>
<td>0.012</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ARHGDI B</td>
<td>ARHGDI B_P148_R_01</td>
<td>0.994</td>
<td>0.982</td>
<td>-0.012</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>ASCL1</td>
<td>ASCL1_E24_F_01</td>
<td>0.999</td>
<td>0.995</td>
<td>-0.003</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ATP10A</td>
<td>ATP10A_P147_F_01</td>
<td>0.486</td>
<td>0.599</td>
<td>0.113</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>AXL</td>
<td>AXL_E61_F_01</td>
<td>0.927</td>
<td>0.910</td>
<td>-0.017</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>BCAM</td>
<td>BCAM_E100_R_01</td>
<td>0.085</td>
<td>0.229</td>
<td>0.144</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>BCL2L2</td>
<td>BCL2L2_E172_F_01</td>
<td>0.939</td>
<td>0.915</td>
<td>-0.024</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>BCL2L2</td>
<td>BCL2L2_P280_F_01</td>
<td>0.054</td>
<td>0.046</td>
<td>-0.008</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>BCR</td>
<td>BCR_P422_F_01</td>
<td>1.000</td>
<td>0.998</td>
<td>-0.002</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>BMP3</td>
<td>BMP3_E147_F_01</td>
<td>0.009</td>
<td>0.020</td>
<td>0.011</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>BMP3</td>
<td>BMP3_P56_R_01</td>
<td>0.033</td>
<td>0.057</td>
<td>0.024</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>CAPG</td>
<td>CAPG_E228_F_01</td>
<td>0.344</td>
<td>0.287</td>
<td>-0.057</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

$H_0 : \Delta\beta = \beta_1 - \beta_2 = 0$

1 = reject $H_0$ 0 = fail to reject $H_0$
<table>
<thead>
<tr>
<th>Gene Symbol</th>
<th>Illumina ID</th>
<th>Mean Proportion Methylated for Cirrhosis Group, ( \beta_1 )</th>
<th>Mean Proportion Methylated for HCC Group, ( \beta_2 )</th>
<th>( \Delta \beta )</th>
<th>Beta Test</th>
<th>Two Sample t-test</th>
<th>Filtered Two Sample t-test</th>
</tr>
</thead>
<tbody>
<tr>
<td>CASP6</td>
<td>CASP6_P230_R_01</td>
<td>0.932</td>
<td>0.891</td>
<td>-0.042</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>CAV2</td>
<td>CAV2_E33_R_01</td>
<td>0.110</td>
<td>0.284</td>
<td>0.174</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>CCKAR</td>
<td>CCKAR_E79_F_01</td>
<td>0.041</td>
<td>0.092</td>
<td>0.050</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>CD81</td>
<td>CD81_P211_F_01</td>
<td>0.988</td>
<td>0.937</td>
<td>-0.051</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CD9</td>
<td>CD9_E14_R_01</td>
<td>0.979</td>
<td>0.972</td>
<td>-0.007</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CDH1</td>
<td>CDH1_P45_F_01</td>
<td>0.628</td>
<td>0.550</td>
<td>-0.077</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>CDH13</td>
<td>CDH13_E102_F_01</td>
<td>0.994</td>
<td>0.979</td>
<td>-0.015</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CDH17</td>
<td>CDH17_P532_F_01</td>
<td>0.409</td>
<td>0.356</td>
<td>-0.053</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>CDKN1A</td>
<td>CDKN1A_E101_F_01</td>
<td>0.048</td>
<td>0.119</td>
<td>0.071</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>CDKN1B</td>
<td>CDKN1B_P1161_F_01</td>
<td>0.972</td>
<td>0.958</td>
<td>-0.014</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CDKN2B</td>
<td>CDKN2B_E220_F_01</td>
<td>0.999</td>
<td>1.000</td>
<td>0.000</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CEBPA</td>
<td>CEBPA_P706_F_01</td>
<td>0.019</td>
<td>0.030</td>
<td>0.011</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>CHD2</td>
<td>CHD2_P667_F_01</td>
<td>0.990</td>
<td>0.997</td>
<td>0.006</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CLDN4</td>
<td>CLDN4_P1120_R_01</td>
<td>0.008</td>
<td>0.033</td>
<td>0.026</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>COL1A2</td>
<td>COL1A2_E299_F_01</td>
<td>0.935</td>
<td>0.899</td>
<td>-0.035</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>COL6A1</td>
<td>COL6A1_P425_F_01</td>
<td>0.973</td>
<td>0.979</td>
<td>0.007</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>COMT</td>
<td>COMT_E401_F_01</td>
<td>0.974</td>
<td>0.963</td>
<td>-0.011</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CPNE1</td>
<td>CPNE1_P138_F_01</td>
<td>0.017</td>
<td>0.024</td>
<td>0.007</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>CREBBP</td>
<td>CREBBP_P712_R_01</td>
<td>0.714</td>
<td>0.756</td>
<td>0.041</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>CRIP1</td>
<td>CRIP1_P274_F_01</td>
<td>0.600</td>
<td>0.544</td>
<td>-0.056</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>CSF1</td>
<td>CSF1_P339_F_01</td>
<td>0.314</td>
<td>0.443</td>
<td>0.129</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>CSF1R</td>
<td>CSF1R_P73_F_01</td>
<td>0.165</td>
<td>0.277</td>
<td>0.111</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>CSF2</td>
<td>CSF2_E248_R_01</td>
<td>0.944</td>
<td>0.954</td>
<td>0.010</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CSF2</td>
<td>CSF2_P605_F_01</td>
<td>0.987</td>
<td>0.981</td>
<td>-0.007</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>CSF3R</td>
<td>CSF3R_P8_F_01</td>
<td>0.938</td>
<td>0.924</td>
<td>-0.014</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CSK</td>
<td>CSK_P740_R_01</td>
<td>0.046</td>
<td>0.083</td>
<td>0.036</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>CTAG2</td>
<td>CTAG2_P1426_F_01</td>
<td>0.882</td>
<td>0.860</td>
<td>-0.022</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>CTHA</td>
<td>CTHA_P238_F_01</td>
<td>0.003</td>
<td>0.009</td>
<td>0.006</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>DAB2</td>
<td>DAB2_P35_F_01</td>
<td>0.993</td>
<td>0.998</td>
<td>0.005</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

\( H_0 : \Delta \beta = \beta_1 - \beta_2 = 0 \)

1 = reject \( H_0 \) 0 = fail to reject \( H_0 \)
<table>
<thead>
<tr>
<th>Gene Symbol</th>
<th>Illumina ID</th>
<th>Mean Proportion Methylated for Cirrhosis Group, $\beta_1$</th>
<th>Mean Proportion Methylated for HCC Group, $\beta_2$</th>
<th>$\Delta\beta$</th>
<th>Beta Test</th>
<th>Two Sample t-test</th>
<th>Filtered Two Sample t-test</th>
</tr>
</thead>
<tbody>
<tr>
<td>DAB2IP</td>
<td>DAB2IP_E18_R_01</td>
<td>0.995</td>
<td>0.992</td>
<td>-0.003</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>DAPK1</td>
<td>DAPK1_P345_R_01</td>
<td>0.974</td>
<td>0.981</td>
<td>0.007</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>DBC1</td>
<td>DBC1_E204_F_01</td>
<td>0.997</td>
<td>0.992</td>
<td>-0.006</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>DCC</td>
<td>DCC_E53_R_01</td>
<td>0.751</td>
<td>0.720</td>
<td>-0.031</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>DCC</td>
<td>DCC_P471_R_01</td>
<td>0.965</td>
<td>0.949</td>
<td>-0.016</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>DKC1</td>
<td>DKC1_E101_F_01</td>
<td>0.975</td>
<td>0.958</td>
<td>-0.017</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>DLG3</td>
<td>DLG3_P62_R_01</td>
<td>0.798</td>
<td>0.737</td>
<td>-0.062</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>DNAJC15</td>
<td>DNAJC15_E26_R_01</td>
<td>0.006</td>
<td>0.014</td>
<td>0.008</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>DNASE1L1</td>
<td>DNASE1L1_P39_R_01</td>
<td>0.626</td>
<td>0.553</td>
<td>-0.073</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>DST</td>
<td>DST_E31_F_01</td>
<td>0.989</td>
<td>0.930</td>
<td>-0.060</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>DST</td>
<td>DST_P262_R_01</td>
<td>1.000</td>
<td>0.997</td>
<td>-0.003</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>EDNRB</td>
<td>EDNRB_P148_R_01</td>
<td>0.037</td>
<td>0.072</td>
<td>0.034</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>EFNA1</td>
<td>EFNA1_P591_R_01</td>
<td>0.840</td>
<td>0.874</td>
<td>0.034</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>EFNB1</td>
<td>EFNB1_E69_F_01</td>
<td>0.276</td>
<td>0.421</td>
<td>0.145</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>EFNB3</td>
<td>EFNB3_P442_R_01</td>
<td>1.000</td>
<td>1.000</td>
<td>0.000</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>EGF</td>
<td>EGF_P413_F_01</td>
<td>0.958</td>
<td>0.949</td>
<td>-0.009</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ELK1</td>
<td>ELK1_E156_F_01</td>
<td>0.108</td>
<td>0.146</td>
<td>0.038</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>ELK3</td>
<td>ELK3_P514_F_01</td>
<td>0.955</td>
<td>0.937</td>
<td>-0.018</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>EMR3</td>
<td>EMR3_E61_F_01</td>
<td>0.073</td>
<td>0.120</td>
<td>0.047</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>ENC1</td>
<td>ENC1_P484_R_01</td>
<td>0.880</td>
<td>0.904</td>
<td>0.024</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>EPHA3</td>
<td>EPHA3_E156_R_01</td>
<td>0.897</td>
<td>0.810</td>
<td>-0.087</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>EPHA5</td>
<td>EPHA5_E158_R_01</td>
<td>0.821</td>
<td>0.780</td>
<td>-0.040</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>EPHA8</td>
<td>EPHA8_P256_F_01</td>
<td>0.371</td>
<td>0.468</td>
<td>0.097</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>EPHB4</td>
<td>EPHB4_P313_R_01</td>
<td>0.807</td>
<td>0.754</td>
<td>-0.053</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>EPHX1</td>
<td>EPHX1_P22_F_01</td>
<td>0.216</td>
<td>0.358</td>
<td>0.142</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>ESR1</td>
<td>ER_seq_a1_S60_F_01</td>
<td>0.959</td>
<td>0.923</td>
<td>-0.037</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>ERBB3</td>
<td>ERBB3_E331_F_01</td>
<td>0.982</td>
<td>0.971</td>
<td>-0.011</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ERG</td>
<td>ERG_E28_F_01</td>
<td>0.999</td>
<td>0.998</td>
<td>-0.001</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ERN1</td>
<td>ERN1_P809_R_01</td>
<td>0.689</td>
<td>0.625</td>
<td>-0.064</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Gene Symbol</td>
<td>Illumina ID</td>
<td>Mean Proportion Methylated for Cirrhosis Group, $\beta_1$</td>
<td>Mean Proportion Methylated for HCC Group, $\beta_2$</td>
<td>$\Delta\beta$</td>
<td>Beta Test</td>
<td>Two Sample t-test</td>
<td>Filtered Two Sample t-test</td>
</tr>
<tr>
<td>------------</td>
<td>----------------------</td>
<td>----------------------------------------------------------</td>
<td>--------------------------------------------------</td>
<td>--------------</td>
<td>-----------</td>
<td>------------------</td>
<td>---------------------------</td>
</tr>
<tr>
<td>F2R</td>
<td>F2R_P839_F_01</td>
<td>0.378</td>
<td>0.474</td>
<td>0.096</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>FABP3</td>
<td>FABP3_P598_F_01</td>
<td>0.680</td>
<td>0.751</td>
<td>0.072</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>FAT</td>
<td>FAT_P279_R_01</td>
<td>0.119</td>
<td>0.167</td>
<td>0.048</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>FES</td>
<td>FES_P223_R_01</td>
<td>0.961</td>
<td>0.934</td>
<td>-0.026</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>FGF2</td>
<td>FGF2_P229_F_01</td>
<td>0.079</td>
<td>0.118</td>
<td>0.039</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>FGF6</td>
<td>FGF6_E294_F_01</td>
<td>0.702</td>
<td>0.775</td>
<td>0.073</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>FGFR1</td>
<td>FGFR1_P204_F_01</td>
<td>0.873</td>
<td>0.844</td>
<td>-0.030</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>FGFR2</td>
<td>FGFR2_P266_R_01</td>
<td>0.290</td>
<td>0.533</td>
<td>0.243</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>FHT</td>
<td>FHT_P93_R_01</td>
<td>0.908</td>
<td>0.852</td>
<td>-0.056</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>FLI1</td>
<td>FLI1_P620_R_01</td>
<td>0.936</td>
<td>0.984</td>
<td>0.047</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>FMR1</td>
<td>FMR1_P484_R_01</td>
<td>0.178</td>
<td>0.208</td>
<td>0.030</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>FZD7</td>
<td>FZD7_E296_F_01</td>
<td>0.018</td>
<td>0.028</td>
<td>0.010</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>FZD9</td>
<td>FZD9_P15_R_01</td>
<td>1.000</td>
<td>0.999</td>
<td>-0.001</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>GABRA5</td>
<td>GABRA5_P862_R_01</td>
<td>0.841</td>
<td>0.800</td>
<td>-0.041</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>GABRG3</td>
<td>GABRG3_P75_F_01</td>
<td>0.137</td>
<td>0.100</td>
<td>-0.036</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>GAS1</td>
<td>GAS1_P754_R_01</td>
<td>0.042</td>
<td>0.078</td>
<td>0.036</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>GAS7</td>
<td>GAS7_P622_R_01</td>
<td>0.989</td>
<td>0.982</td>
<td>-0.007</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>GFI1</td>
<td>GFI1_P45_R_01</td>
<td>0.999</td>
<td>0.995</td>
<td>-0.003</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>GLI2</td>
<td>GLI2_E90_F_01</td>
<td>0.973</td>
<td>0.961</td>
<td>-0.012</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>GLI3</td>
<td>GLI3_P453_R_01</td>
<td>0.997</td>
<td>0.999</td>
<td>0.002</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>GNG7</td>
<td>GNG7_E310_R_01</td>
<td>0.967</td>
<td>0.946</td>
<td>-0.021</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>GPR116</td>
<td>GPR116_E328_R_01</td>
<td>0.178</td>
<td>0.137</td>
<td>-0.041</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>GSTM2</td>
<td>GSTM2_P453_R_01</td>
<td>0.092</td>
<td>0.123</td>
<td>0.031</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>GSTP1</td>
<td>GSTP1_P74_F_01</td>
<td>0.803</td>
<td>0.832</td>
<td>0.029</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>HBII-13</td>
<td>HBII-13_P991_R_01</td>
<td>0.010</td>
<td>0.025</td>
<td>0.015</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>HBII-52</td>
<td>HBII-52_P659_F_01</td>
<td>0.850</td>
<td>0.872</td>
<td>0.022</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>HCK</td>
<td>HCK_P46_R_01</td>
<td>0.637</td>
<td>0.572</td>
<td>-0.066</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>HDAC6</td>
<td>HDAC6_P153_F_01</td>
<td>0.289</td>
<td>0.247</td>
<td>-0.042</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>HDAC9</td>
<td>HDAC9_E38_F_01</td>
<td>0.995</td>
<td>0.984</td>
<td>-0.010</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

$H_0 : \Delta\beta = \beta_1 - \beta_2 = 0$

$1 = \text{reject } H_0 \quad 0 = \text{fail to reject } H_0$
<table>
<thead>
<tr>
<th>Gene Symbol</th>
<th>Illumina ID</th>
<th>Mean Proportion Methylated for Cirrhosis Group, $\beta_1$</th>
<th>Mean Proportion Methylated for HCC Group, $\beta_2$</th>
<th>$\Delta \beta$</th>
<th>Beta Test</th>
<th>Two Sample t-test</th>
<th>Filtered Two Sample t-test</th>
</tr>
</thead>
<tbody>
<tr>
<td>HGF</td>
<td>HGF_E102_R_01</td>
<td>0.990</td>
<td>0.985</td>
<td>-0.006</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>HIC2</td>
<td>HIC2_P498_F_01</td>
<td>0.797</td>
<td>0.956</td>
<td>-0.024</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>HLA-DOB</td>
<td>HLA-DOB_E432_R_01</td>
<td>0.046</td>
<td>0.092</td>
<td>0.046</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>HOXA9</td>
<td>HOXA9_P1141_R_01</td>
<td>0.744</td>
<td>0.812</td>
<td>0.069</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>HOXB2</td>
<td>HOXB2_P488_R_01</td>
<td>0.094</td>
<td>0.128</td>
<td>0.034</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>HPN</td>
<td>HPN_P823_F_01</td>
<td>0.999</td>
<td>1.000</td>
<td>0.001</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>HPSE</td>
<td>HPSE_P93_F_01</td>
<td>0.999</td>
<td>0.995</td>
<td>-0.004</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>HTR1B</td>
<td>HTR1B_P222_F_01</td>
<td>0.876</td>
<td>0.849</td>
<td>-0.027</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>HTR2A</td>
<td>HTR2A_P853_F_01</td>
<td>0.110</td>
<td>0.148</td>
<td>0.038</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>IAPP</td>
<td>IAPP_E280_F_01</td>
<td>0.986</td>
<td>0.966</td>
<td>-0.019</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>ICAM1</td>
<td>ICAM1_P119_R_01</td>
<td>0.455</td>
<td>0.593</td>
<td>0.139</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>IFNG</td>
<td>IFNG_P459_R_01</td>
<td>0.973</td>
<td>0.988</td>
<td>0.015</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>IFNGR2</td>
<td>IFNGR2_E164_F_01</td>
<td>0.364</td>
<td>0.522</td>
<td>0.157</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>IGF1R</td>
<td>IGF1R_E186_R_01</td>
<td>0.589</td>
<td>0.509</td>
<td>-0.080</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>IGF2R</td>
<td>IGF2R_P396_R_01</td>
<td>0.391</td>
<td>0.485</td>
<td>0.094</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>IGFBP1</td>
<td>IGFBP1_E48_R_01</td>
<td>0.367</td>
<td>0.451</td>
<td>0.084</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>IGFBP6</td>
<td>IGFBP6_P328_R_01</td>
<td>0.052</td>
<td>0.085</td>
<td>0.033</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>IGF4C</td>
<td>IGF4C_P533_R_01</td>
<td>0.552</td>
<td>0.470</td>
<td>-0.082</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>IHH</td>
<td>IHH_P246_R_01</td>
<td>0.387</td>
<td>0.343</td>
<td>-0.044</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>IL13</td>
<td>IL13_E75_R_01</td>
<td>0.168</td>
<td>0.351</td>
<td>0.183</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>IL16</td>
<td>IL16_P226_F_01</td>
<td>0.252</td>
<td>0.335</td>
<td>0.083</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>IL18BP</td>
<td>IL18BP_E285_F_01</td>
<td>0.410</td>
<td>0.465</td>
<td>0.055</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>IL4</td>
<td>IL4_P262_R_01</td>
<td>0.945</td>
<td>0.958</td>
<td>0.013</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>IL6</td>
<td>IL6_E168_F_01</td>
<td>0.439</td>
<td>0.411</td>
<td>-0.027</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>IL6</td>
<td>IL6_P213_R_01</td>
<td>0.994</td>
<td>0.988</td>
<td>-0.006</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>IMPACT</td>
<td>IMPACT_P186_F_01</td>
<td>0.801</td>
<td>0.766</td>
<td>-0.035</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>IMPACT</td>
<td>IMPACT_P234_R_01</td>
<td>0.999</td>
<td>0.995</td>
<td>-0.004</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>IRAK1</td>
<td>IRAK1_P312_F_01</td>
<td>0.039</td>
<td>0.044</td>
<td>0.005</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>IRAK1</td>
<td>IRAK1_P455_R_01</td>
<td>0.105</td>
<td>0.080</td>
<td>-0.025</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>IRF5</td>
<td>IRF5_E101_F_01</td>
<td>0.940</td>
<td>0.956</td>
<td>0.015</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

$H_0 : \Delta \beta = \beta_1 - \beta_2 = 0$

1 = reject $H_0$ 0 = fail to reject $H_0$
<table>
<thead>
<tr>
<th>Gene Symbol</th>
<th>Illumina ID</th>
<th>Mean Proportion Methylated for Cirrhosis Group, $\beta_1$</th>
<th>Mean Proportion Methylated for HCC Group, $\beta_2$</th>
<th>$\Delta\beta$</th>
<th>Beta Test</th>
<th>Two Sample t-test</th>
<th>Filtered Two Sample t-test</th>
</tr>
</thead>
<tbody>
<tr>
<td>IRF7</td>
<td>IRF7_E236_R_01</td>
<td>0.680</td>
<td>0.729</td>
<td>0.049</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>ITGA6</td>
<td>ITGA6_P238_R_01</td>
<td>0.133</td>
<td>0.195</td>
<td>0.062</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>ITGB1</td>
<td>ITGB1_P451_F_01</td>
<td>0.819</td>
<td>0.784</td>
<td>-0.036</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>KCNK4</td>
<td>KCNK4_P171_R_01</td>
<td>0.999</td>
<td>0.998</td>
<td>-0.001</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>KCNQ1</td>
<td>KCNQ1_E349_R_01</td>
<td>1.000</td>
<td>0.999</td>
<td>-0.001</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>KRAS</td>
<td>KRAS_P651_F_01</td>
<td>0.128</td>
<td>0.169</td>
<td>0.041</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>LAMB1</td>
<td>LAMB1_E144_R_01</td>
<td>0.195</td>
<td>0.243</td>
<td>0.048</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>LAMC1</td>
<td>LAMC1_E466_R_01</td>
<td>0.286</td>
<td>0.406</td>
<td>0.119</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>LAMC1</td>
<td>LAMC1_P808_F_01</td>
<td>0.412</td>
<td>0.459</td>
<td>0.047</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>LCN2</td>
<td>LCN2_P86_R_01</td>
<td>0.952</td>
<td>0.937</td>
<td>-0.014</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>LEFTY2</td>
<td>LEFTY2_P561_F_01</td>
<td>0.093</td>
<td>0.130</td>
<td>0.038</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>LIG4</td>
<td>LIG4_P194_F_01</td>
<td>0.992</td>
<td>0.903</td>
<td>-0.088</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>LOX</td>
<td>LOX_P71_F_01</td>
<td>0.448</td>
<td>0.515</td>
<td>0.067</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>MAF</td>
<td>MAF_P826_R_01</td>
<td>0.062</td>
<td>0.141</td>
<td>0.078</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>MAP3K1</td>
<td>MAP3K1_E81_F_01</td>
<td>0.283</td>
<td>0.391</td>
<td>0.108</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>MAP3K1</td>
<td>MAP3K1_P7_F_01</td>
<td>0.004</td>
<td>0.011</td>
<td>0.007</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>MAP3K9</td>
<td>MAP3K9_E17_R_01</td>
<td>0.767</td>
<td>0.706</td>
<td>-0.061</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MAPK12</td>
<td>MAPK12_P416_F_01</td>
<td>0.211</td>
<td>0.258</td>
<td>0.047</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>MAPK14</td>
<td>MAPK14_P327_R_01</td>
<td>0.998</td>
<td>0.997</td>
<td>-0.002</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MAS1</td>
<td>MAS1_P469_R_01</td>
<td>0.989</td>
<td>0.981</td>
<td>-0.008</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MAS1</td>
<td>MAS1_P657_R_01</td>
<td>0.966</td>
<td>0.953</td>
<td>-0.014</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>MCF2</td>
<td>MCF2_P445_F_01</td>
<td>0.954</td>
<td>0.942</td>
<td>-0.012</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ABCB1</td>
<td>ABCB1_seq_42_S300_R_01</td>
<td>0.996</td>
<td>0.997</td>
<td>0.001</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MFAP4</td>
<td>MFAP4_P10_R_01</td>
<td>0.013</td>
<td>0.023</td>
<td>0.010</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>MKRN3</td>
<td>MKRN3_P108_F_01</td>
<td>0.985</td>
<td>0.978</td>
<td>-0.007</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MLH3</td>
<td>MLH3_E72_F_01</td>
<td>0.218</td>
<td>0.275</td>
<td>0.057</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>MLLT4</td>
<td>MLLT4_P1400_F_01</td>
<td>0.130</td>
<td>0.188</td>
<td>0.058</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>MME</td>
<td>MME_E29_F_01</td>
<td>0.800</td>
<td>0.757</td>
<td>-0.043</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MMP1</td>
<td>MMP1_P397_R_01</td>
<td>0.739</td>
<td>0.779</td>
<td>0.040</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Gene Symbol</td>
<td>Illumina ID</td>
<td>Mean Proportion Methylated for Cirrhosis Group, $\beta_1$</td>
<td>Mean Proportion Methylated for HCC Group, $\beta_2$</td>
<td>$\Delta \beta$</td>
<td>Filtered Two Sample t-test</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------</td>
<td>-----------------</td>
<td>----------------------------------------------------------</td>
<td>--------------------------------------------------</td>
<td>---------------</td>
<td>-----------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MMP9</td>
<td>MMP9_E88_R_01</td>
<td>0.992</td>
<td>0.985</td>
<td>-0.007</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MMP9</td>
<td>MMP9_P237_R_01</td>
<td>0.814</td>
<td>0.764</td>
<td>-0.050</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MPO</td>
<td>MPO_E302_R_01</td>
<td>0.038</td>
<td>0.062</td>
<td>0.024</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MST1R</td>
<td>MST1R_P87_R_01</td>
<td>0.110</td>
<td>0.090</td>
<td>-0.021</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MT1A</td>
<td>MT1A_E13_R_01</td>
<td>0.791</td>
<td>0.832</td>
<td>0.042</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MYBL2</td>
<td>MYBL2_P354_F_01</td>
<td>0.286</td>
<td>0.457</td>
<td>0.172</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MYCL1</td>
<td>MYCL1_P502_R_01</td>
<td>0.946</td>
<td>0.964</td>
<td>0.018</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MYCN</td>
<td>MYCN_P464_R_01</td>
<td>0.393</td>
<td>0.582</td>
<td>0.188</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MYLK</td>
<td>MYLK_E132_R_01</td>
<td>0.551</td>
<td>0.800</td>
<td>0.249</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NAT2</td>
<td>NAT2_P11_F_01</td>
<td>0.999</td>
<td>1.000</td>
<td>0.001</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NCL</td>
<td>NCL_P1102_F_01</td>
<td>0.998</td>
<td>0.996</td>
<td>-0.002</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NES</td>
<td>NES_P239_R_01</td>
<td>0.017</td>
<td>0.023</td>
<td>0.006</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NOTCH2</td>
<td>NOTCH2_P312_R_01</td>
<td>0.753</td>
<td>0.709</td>
<td>-0.044</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NOTCH3</td>
<td>NOTCH3_P198_R_01</td>
<td>0.050</td>
<td>0.068</td>
<td>0.018</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NPR2</td>
<td>NPR2_P1093_F_01</td>
<td>0.079</td>
<td>0.066</td>
<td>-0.014</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NPR2</td>
<td>NPR2_P618_F_01</td>
<td>0.034</td>
<td>0.078</td>
<td>0.044</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NRG1</td>
<td>NRG1_P558_R_01</td>
<td>0.986</td>
<td>0.976</td>
<td>-0.010</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NTRK1</td>
<td>NTRK1_E74_F_01</td>
<td>0.015</td>
<td>0.020</td>
<td>0.005</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NTRK2</td>
<td>NTRK2_P395_R_01</td>
<td>0.943</td>
<td>0.908</td>
<td>-0.035</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OSM</td>
<td>OSM_P34_F_01</td>
<td>0.997</td>
<td>0.979</td>
<td>-0.018</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PAX6</td>
<td>PAX6_E129_F_01</td>
<td>0.911</td>
<td>0.868</td>
<td>-0.043</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PAX6</td>
<td>PAX6_P1121_F_01</td>
<td>0.315</td>
<td>0.220</td>
<td>-0.095</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PCTK1</td>
<td>PCTK1_E77_R_01</td>
<td>0.900</td>
<td>0.853</td>
<td>-0.047</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PDE1B</td>
<td>PDE1B_E141_F_01</td>
<td>0.488</td>
<td>0.444</td>
<td>-0.044</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PDGFB</td>
<td>PDGFB_P719_F_01</td>
<td>0.994</td>
<td>0.979</td>
<td>-0.014</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PECAM1</td>
<td>PECAM1_E32_R_01</td>
<td>0.317</td>
<td>0.515</td>
<td>0.199</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PEG3</td>
<td>PEG3_E496_F_01</td>
<td>0.901</td>
<td>0.919</td>
<td>0.018</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PENK</td>
<td>PENK_E26_F_01</td>
<td>0.816</td>
<td>0.864</td>
<td>0.048</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PENK</td>
<td>PENK_P447_R_01</td>
<td>1.000</td>
<td>1.000</td>
<td>0.000</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PGF</td>
<td>PGF_E33_F_01</td>
<td>0.940</td>
<td>0.958</td>
<td>0.018</td>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$H_0 : \Delta \beta = \beta_1 - \beta_2 = 0$

1 = reject $H_0$ 0 = fail to reject $H_0$
<table>
<thead>
<tr>
<th>Gene Symbol</th>
<th>Illumina ID</th>
<th>Mean Proportion Methylated for Cirrhosis Group, $\beta_1$</th>
<th>Mean Proportion Methylated for HCC Group, $\beta_2$</th>
<th>$\Delta \beta$</th>
<th>Beta Test</th>
<th>Two Sample t-test</th>
<th>Filtered Two Sample t-test</th>
</tr>
</thead>
<tbody>
<tr>
<td>PGF</td>
<td>PGF_P320_F_01</td>
<td>0.058</td>
<td>0.092</td>
<td>0.034</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>PGR</td>
<td>PGR_E183_R_01</td>
<td>0.925</td>
<td>0.900</td>
<td>-0.024</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>PLAGL1</td>
<td>PLAGL1_P236_R_01</td>
<td>0.993</td>
<td>0.986</td>
<td>-0.007</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PLAGL1</td>
<td>PLAGL1_P334_F_01</td>
<td>0.862</td>
<td>0.879</td>
<td>0.017</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>PLAUR</td>
<td>PLAUR_P82_F_01</td>
<td>0.266</td>
<td>0.301</td>
<td>0.034</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>POMC</td>
<td>POMC_E254_F_01</td>
<td>0.991</td>
<td>0.984</td>
<td>-0.006</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PSA</td>
<td>PSA_P135_F_01</td>
<td>0.908</td>
<td>0.882</td>
<td>-0.026</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PSIP1</td>
<td>PSIP1_P163_R_01</td>
<td>0.964</td>
<td>0.930</td>
<td>-0.034</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>PTPH2</td>
<td>PTPH2_P568_R_01</td>
<td>0.879</td>
<td>0.836</td>
<td>-0.043</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>PTGS2</td>
<td>PTGS2_P308_F_01</td>
<td>0.998</td>
<td>0.996</td>
<td>-0.002</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PTHLH</td>
<td>PTHLH_P15_R_01</td>
<td>0.966</td>
<td>0.934</td>
<td>-0.032</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>PTK2B</td>
<td>PTK2B_P673_R_01</td>
<td>0.898</td>
<td>0.879</td>
<td>-0.020</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>PTPN6</td>
<td>PTPN6_E171_R_01</td>
<td>0.998</td>
<td>0.997</td>
<td>-0.001</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PTPN6</td>
<td>PTPN6_P282_R_01</td>
<td>0.044</td>
<td>0.074</td>
<td>0.030</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>PTPNS1</td>
<td>PTPNS1_P301_R_01</td>
<td>0.039</td>
<td>0.050</td>
<td>0.011</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>PTNGR</td>
<td>PTNGR_E40_R_01</td>
<td>0.962</td>
<td>0.947</td>
<td>-0.014</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PTNGR</td>
<td>PTNGR_P476_F_01</td>
<td>0.942</td>
<td>0.859</td>
<td>-0.083</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PTPRH</td>
<td>PTPRH_P255_F_01</td>
<td>0.999</td>
<td>1.000</td>
<td>0.001</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PYCARD</td>
<td>PYCARD_P150_F_01</td>
<td>0.822</td>
<td>0.864</td>
<td>0.041</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>RAD50</td>
<td>RAD50_P191_F_01</td>
<td>0.810</td>
<td>0.766</td>
<td>-0.044</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>RAN</td>
<td>RAN_P581_R_01</td>
<td>0.959</td>
<td>0.919</td>
<td>-0.040</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>RET</td>
<td>RET_seq_53_S374_F_01</td>
<td>0.227</td>
<td>0.342</td>
<td>0.115</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>RYK</td>
<td>RYK_P493_F_01</td>
<td>0.978</td>
<td>0.945</td>
<td>-0.034</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>S100A2</td>
<td>S100A2_E36_R_01</td>
<td>0.171</td>
<td>0.238</td>
<td>0.067</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>S100A4</td>
<td>S100A4_P194_R_01</td>
<td>0.996</td>
<td>0.998</td>
<td>0.002</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SCGB3A1</td>
<td>SCGB3A1_P103_R_01</td>
<td>0.909</td>
<td>0.932</td>
<td>0.023</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SEMA3B</td>
<td>SEMA3B_P110_R_01</td>
<td>0.006</td>
<td>0.016</td>
<td>0.010</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>5-Sep</td>
<td>SEPT5_P464_R_01</td>
<td>0.117</td>
<td>0.159</td>
<td>0.042</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>SERPINE1</td>
<td>SERPINE1_E189_R_01</td>
<td>0.039</td>
<td>0.080</td>
<td>0.041</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

$H_0 : \Delta \beta = \beta_1 - \beta_2 = 0$

1 = reject $H_0$ 0 = fail to reject $H_0$
<table>
<thead>
<tr>
<th>Gene Symbol</th>
<th>Illumina ID</th>
<th>Mean Proportion Methylated for Cirrhosis Group, $\beta_1$</th>
<th>Mean Proportion Methylated for HCC Group, $\beta_2$</th>
<th>$\Delta \beta$</th>
<th>Beta Test</th>
<th>Two Sample t-test</th>
<th>Filtered Two Sample t-test</th>
</tr>
</thead>
<tbody>
<tr>
<td>SFN</td>
<td>SFN_P248_F_01</td>
<td>0.992</td>
<td>0.982</td>
<td>-0.010</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>SFTPB</td>
<td>SFTPB_P689_R_01</td>
<td>0.890</td>
<td>0.847</td>
<td>-0.043</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SGCE</td>
<td>SGCE_E149_F_01</td>
<td>0.163</td>
<td>0.185</td>
<td>0.022</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>SGCE</td>
<td>SGCE_P250_R_01</td>
<td>0.013</td>
<td>0.023</td>
<td>0.010</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>SH3BP2</td>
<td>SH3BP2_P771_R_01</td>
<td>0.080</td>
<td>0.110</td>
<td>0.030</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>SHB</td>
<td>SHB_P473_R_01</td>
<td>0.209</td>
<td>0.274</td>
<td>0.065</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>SIN3B</td>
<td>SIN3B_P607_F_01</td>
<td>0.999</td>
<td>0.999</td>
<td>0.000</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SLIT2</td>
<td>SLIT2_E111_R_01</td>
<td>0.950</td>
<td>0.963</td>
<td>0.012</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>SMO</td>
<td>SMO_E57_F_01</td>
<td>0.908</td>
<td>0.930</td>
<td>0.032</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>SNRPN</td>
<td>SNRPN_E14_F_01</td>
<td>0.999</td>
<td>0.999</td>
<td>0.000</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SNURF</td>
<td>SNURF_P2_R_01</td>
<td>0.650</td>
<td>0.599</td>
<td>-0.051</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SPARC</td>
<td>SPARC_P195_F_01</td>
<td>0.980</td>
<td>0.968</td>
<td>-0.012</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SPDEF</td>
<td>SPDEF_P6_R_01</td>
<td>0.696</td>
<td>0.749</td>
<td>0.053</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SPI1</td>
<td>SPI1_P48_F_01</td>
<td>1.000</td>
<td>0.999</td>
<td>0.000</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SPP1</td>
<td>SPP1_E140_R_01</td>
<td>0.998</td>
<td>0.996</td>
<td>-0.002</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SPP1</td>
<td>SPP1_P647_F_01</td>
<td>0.851</td>
<td>0.791</td>
<td>-0.060</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SRC</td>
<td>SRC_E100_R_01</td>
<td>0.602</td>
<td>0.530</td>
<td>-0.072</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>STAT5A</td>
<td>STAT5A_P704_R_01</td>
<td>0.950</td>
<td>0.933</td>
<td>-0.017</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TAL1</td>
<td>TAL1_P817_F_01</td>
<td>0.160</td>
<td>0.130</td>
<td>-0.030</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>TDG</td>
<td>TDG_E129_F_01</td>
<td>0.996</td>
<td>0.992</td>
<td>-0.004</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TFDP1</td>
<td>TFDP1_P543_R_01</td>
<td>0.916</td>
<td>0.935</td>
<td>0.019</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TFF2</td>
<td>TFF2_P178_F_01</td>
<td>0.997</td>
<td>0.999</td>
<td>0.002</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TFRC</td>
<td>TFRC_P414_R_01</td>
<td>0.942</td>
<td>0.920</td>
<td>-0.022</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TIMP1</td>
<td>TIMP1_E254_R_01</td>
<td>0.964</td>
<td>0.943</td>
<td>-0.021</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TIMP3</td>
<td>TIMP3_P1114_R_01</td>
<td>0.782</td>
<td>0.839</td>
<td>0.057</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>TJP2</td>
<td>TJP2_P518_F_01</td>
<td>0.505</td>
<td>0.649</td>
<td>0.144</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>TMEFF1</td>
<td>TMEFF1_E180_R_01</td>
<td>0.009</td>
<td>0.019</td>
<td>0.009</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>TMEFF1</td>
<td>TMEFF1_P234_F_01</td>
<td>0.030</td>
<td>0.090</td>
<td>0.059</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>TMEFF2</td>
<td>TMEFF2_P152_R_01</td>
<td>0.152</td>
<td>0.193</td>
<td>0.041</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
$H_0: \Delta \beta = \beta_1 - \beta_2 = 0$

1 = reject $H_0$  0 = fail to reject $H_0$

<table>
<thead>
<tr>
<th>Gene Symbol</th>
<th>Illumina ID</th>
<th>Mean Proportion Methylated for Cirrhosis Group, $\beta_1$</th>
<th>Mean Proportion Methylated for HCC Group, $\beta_2$</th>
<th>$\Delta \beta$</th>
<th>Beta Test</th>
<th>Two Sample t-test</th>
<th>Filtered Two Sample t-test</th>
</tr>
</thead>
<tbody>
<tr>
<td>TNFRSF10B</td>
<td>TNFRSF10B_E198_R_01</td>
<td>0.108</td>
<td>0.087</td>
<td>-0.021</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>TNFRSF1B</td>
<td>TNFRSF1B_E5_F_01</td>
<td>0.200</td>
<td>0.232</td>
<td>0.032</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>TNFSF10</td>
<td>TNFSF10_P2_R_01</td>
<td>0.855</td>
<td>0.831</td>
<td>-0.024</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>TP73</td>
<td>TP73_P496_F_01</td>
<td>0.988</td>
<td>0.992</td>
<td>0.005</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TSC2</td>
<td>TSC2_E140_F_01</td>
<td>0.753</td>
<td>0.685</td>
<td>-0.068</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TSG101</td>
<td>TSG101_P257_R_01</td>
<td>0.831</td>
<td>0.902</td>
<td>0.072</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>UGT1A1</td>
<td>UGT1A1_P564_R_01</td>
<td>1.000</td>
<td>1.000</td>
<td>0.000</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>UGT1A7</td>
<td>UGT1A7_P751_R_01</td>
<td>1.000</td>
<td>0.994</td>
<td>-0.006</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>VAMP8</td>
<td>VAMP8_P241_F_01</td>
<td>0.966</td>
<td>0.932</td>
<td>-0.034</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>VAV2</td>
<td>VAV2_E58_F_01</td>
<td>0.982</td>
<td>0.969</td>
<td>-0.013</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>VIM</td>
<td>VIM_P811_R_01</td>
<td>0.001</td>
<td>0.008</td>
<td>0.007</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>WNT1</td>
<td>WNT1_E157_F_01</td>
<td>0.683</td>
<td>0.732</td>
<td>0.049</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>YES1</td>
<td>YES1_P600_F_01</td>
<td>0.994</td>
<td>0.979</td>
<td>-0.015</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>ZAP70</td>
<td>ZAP70_P220_R_01</td>
<td>0.993</td>
<td>0.980</td>
<td>-0.013</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>ZIM3</td>
<td>ZIM3_E203_F_01</td>
<td>0.338</td>
<td>0.283</td>
<td>-0.055</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>ZNF264</td>
<td>ZNF264_P397_F_01</td>
<td>0.460</td>
<td>0.396</td>
<td>-0.065</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>ZNFN1A1</td>
<td>ZNFN1A1_E102_F_01</td>
<td>0.609</td>
<td>0.658</td>
<td>0.049</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CC 8 (3433)</td>
<td>PCR16_HK3_1321_F</td>
<td>1.000</td>
<td>0.999</td>
<td>-0.001</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Negative</td>
<td>Neg7_PPIH_9972_R</td>
<td>0.158</td>
<td>0.311</td>
<td>0.153</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Negative</td>
<td>Neg15_HK3_8743_R</td>
<td>0.029</td>
<td>0.049</td>
<td>0.021</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Negative</td>
<td>Neg16_PPIF_4992_R</td>
<td>0.942</td>
<td>0.895</td>
<td>-0.047</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Negative</td>
<td>Neg48_PPIG_45960_R</td>
<td>0.999</td>
<td>0.997</td>
<td>-0.002</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
APPENDIX B

Source Code for the construction of the nonlinear mixed model in Chapter 3

```r
############################################################
### # This file constructs the most appropriate hierarchical model and establishes a dose response relation. 
############################################################
```

```r
setwd("C://Documents and Settings//Laptop User//Desktop//Research")
xchrom.new<-read.csv('XchromBeta.new.csv')

library(nlme)
names(xchrom.new)
# [1] "index" "site" "TargetID" "meth" "dose" "sample" "trep"
"SampleRep" "trep2"
attach(xchrom.new)
site<-as.factor(site);dose<-as.factor(dose);sample<-as.factor(sample)
plot(meth~dose,xlab="Percent of female genomic DNA in sample",ylab="Proportion Methylated")

CpGdata<-xchrom.new

dilution<-groupedData(meth~dose|sample/data=CpGdata)
plot(dilution,display=1,collapse=1,ylab="Proportion Methylated",xlab="Percent of female genomic DNA in sample")

# Simple Logistic Model

model1<-nlme(meth~.5/(1+exp(-(dose-mid)/scal)),data =dilution,start=c(5,5),fixed = mid + scal ~ 1,random=list(sample=pdDiag(mid+scal~1),Rep= pdDiag(mid+scal~1),site= pdDiag(mid+scal~1)))

model1a<-update(model1, weights=varExp(form=~dose|site))  # Variance model in which the site variance increases exponentially with age
```
model1b <- update(model1a, random = list(sample = pdDiag(mid+scal~1), Rep = pdDiag(mid+scal~1), site = pdDiag(scal~1)))

model1c <- update(model1a, random = list(sample = pdDiag(mid+scal~1), Rep = pdDiag(mid+scal~1)))

#> anova(model1a, model1c)
#   Model df      AIC      BIC    logLik Test L.Ratio p-value
#model1a     1 26 -1222.222 -1100.454 637.1108
#model1c     2 24 -1231.735 -1119.334 639.8674 1 vs 2 5.513169  0.0635

p3.est <- summary(model1a)$tTable[2,]
#  Value     Std.Error    DF    t-value     p-value
#8.398883e+00 1.394291e+00 5.940000e+02 6.023765e+0 0 2.988831e-09

dimnames(p3.est)[[2]] <- dimnames(summary(model1a)$tTable)[[2]]

p3.est2 <- cbind(p3.est, lower.ci, upper.ci)
p3.est2[1,]
#  p3.est lower.ci upper.ci
#8.398883  5.610301 11.187466

fitted.mean <- numeric()
fitted.mean[1] <- mean(fitted(model1a)[CpGdata$dose == "100"])
fitted.mean[2] <- mean(fitted(model1a)[CpGdata$dose == "50"])
fitted.mean[3] <- mean(fitted(model1a)[CpGdata$dose == "20"])
fitted.mean[4] <- mean(fitted(model1a)[CpGdata$dose == "10"])
fitted.mean[5] <- mean(fitted(model1a)[CpGdata$dose == "5"])
fitted.mean[6] <- mean(fitted(model1a)[CpGdata$dose == "0"])
dose.mean <- c(100, 50, 20, 10, 5, 0)

plot(fitted.mean ~ dose.mean, type = 'l', ylab = "Proportion Methylated", xlab = "Percent of female genomic DNA in sample", ylim = c(0, 1), main = "")

points(CpGdata$meth ~ CpGdata$dose)

# Nonlinear method for testing significance of dose-response
dilution <- groupedData(meth ~ dose | sample/Rep/site, data = CpGdata)
p3.est[i,] <- summary(model2b)$tTable[3,]
}
dimnames(p3.est)[[2]] <- dimnames(summary(model1a)$tTable)[[2]]
upper.ci <- p3.est[,1] + 2 * p3.est[,2]
p3.est2 <- cbind(p3.est, lower.ci, upper.ci)
Source code for the establishment of a minimum level of detectable differential level of methylation

setwd("C://Documents and Settings//Laptop User//Desktop//Research")
xchrom.new<-read.csv('XchromBeta.new.csv')
xchrom.cor<-read.csv('XchromBeta.csv')

#Correlation matrix
xchrom.cor<-read.csv('XchromBeta.csv',row.names=1)
correlation<-round(cor(t(xchrom.cor)),3)
dimnames(correlation)[[1]]<-rownames(xchrom.cor)
dimnames(correlation)[[2]]<-rownames(xchrom.cor)
write.table(correlation,"Correlation.csv",sep=",")
library(nlme)

names(xchrom.new)
#[1] "index" "site" "TargetID" "meth" "dose" "sample" "trep"
"SampleRep" "trep2"
attach(xchrom.new)
site<-as.factor(site);dose<-as.factor(dose);sample<-as.factor(sample)
plot(meth~dose,xlab="Percent of female genomic DNA in sample",ylab="Proportion Methylated")

##Fit one anova model for all CpG sites using dose as fixed effect and site nested within sample as random effects

# Since model is function of categorical variable (dose), regression and anova same here
xchrom.new<-read.csv('XchromBeta.new.csv')
xchrom.new$site<-as.factor(xchrom.new$site)
xchrom.new$dose<-as.factor(xchrom.new$dose)
xchrom.new$sample<-as.factor(xchrom.new$sample)
xchrom.new$Rep<-as.factor(xchrom.new$Rep)
attach(xchrom.new)
plot(meth~dose,xlab="Percent of female genomic DNA in sample",ylab="Proportion Methylated")
detach()
library(nlme)
overall<-lme(meth~dose,random=~1|sample/Rep/site,data=xchrom.new)
overall.reduced<-lme(meth~dose,random=~1|sample/Rep,data=xchrom.new)
anova(overall,overall.reduced)
overall.reduced2<-lme(meth~dose,random=~1|sample,data=xchrom.new)
anova(overall.reduced,overall.reduced2)
anova(overall)
plot(overall)
##Linear Contrasts for pairwise comparisons
library(gmodels)
contrast.0.5<-c(-1,1,0,0,0)
contrast.5.10<-c(0,-1,1,0,0)
contrast.10.20<-c(0,0,-1,1,0)
contrast.20.50<-c(0,0,0,-1,1)
contrast.50.100<-c(0,0,0,0,-1)
fit<-lme(meth~dose,random=~1|sample/Rep/site,data=xchrom.new)
all.contrasts<-
rbind(contrast.0.5,contrast.5.10,contrast.10.20,contrast.20.50,contrast.50.100)
pairwise.result<-fit.contrast(model=fit, varname="dose", all.contrasts)

#Since pvalues<0.05, each contrast significant, so we reject H0:equal mean methylation in two groups
APPENDIX D

Source code for Chapter 3.3 CpG site specific hierarchical models

###############################################################
##### This file constructs hierarchical models per CpG site          #######
##### It creates diagnostic plots for these models and computes point #######
##### estimates at dose=100 and corresponding confidence intervals   #######
###############################################################

setwd("C://Documents and Settings//Laptop User//Desktop//Research")
xchrom.new<-read.csv('XchromBeta.new.csv')

library(nlme)
names(xchrom.new)
# [1] "index" "site" "TargetID" "meth" "dose" "sample" "trep"
"SampleRep" "trep2"
attach(xchrom.new)
site<-as.factor(site);dose<-as.factor(dose);sample<-as.factor(sample)
plot(meth~dose,xlab="Percent of female genomic DNA in sample",ylab="Proportion Methylated")

#Change site depending on for which CpG site the model is being formulated
CpGdata<-xchrom.new[site==17,]
dilution<-groupedData(meth~dose|sample,data=CpGdata)

#Simple Logistic Model
model1<-nlme(meth~Asym/(1+exp(-(dose-mid)/scal)),data=dilution,start=c(.8,5,5),fixed =
Asym + mid + scal ~ 1,random=list(sample=pdDiag(Asym+mid+scal~1)))

model1a<-nlme(meth~Asym/(1+exp(-(dose-mid)/scal)),data=dilution,start=c(.72,5,5),fixed =
Asym + mid + scal ~ 1,random=list(sample=~Asym+mid+scal~1))

#Asymptotic Regression Model
model2<-nlme(meth~p1+(p2-p1)*exp(-exp(p3)*dose),data=dilution,start=c(.72,.03,-3),fixed = p1 + p2 + p3~ 1,random=list(sample=pdDiag(p1+p2+p3~1)))

anova(model1,model2)

model2a<-nlme(meth~p1+(p2-p1)*exp(-exp(p3)*dose),data=dilution,start=c(.72,.03,-3),fixed = p1 + p2 + p3~ 1,random=list(sample=pdDiag(p1+p2~1)))

model2b<-nlme(meth~p1+(p2-p1)*exp(-exp(p3)*dose),data=dilution,start=c(.72,.03,-3),fixed = p1 + p2 + p3~ 1,random=list(sample=pdDiag(p1+p3~1)))

model2c<-nlme(meth~p1+(p2-p1)*exp(-exp(p3)*dose),data=dilution,start=c(.72,.03,-3),fixed = p1 + p2 + p3~ 1,random=list(sample=pdDiag(p2+p3~1)))

model2d<-nlme(meth~p1+(p2-p1)*exp(-exp(p3)*dose),data=dilution,start=c(.72,.03,-3),fixed = p1 + p2 + p3~ 1,random=list(sample=pdDiag(p1~1)))

model3<-nlme(meth~p1+(p2-p1)*exp(-exp(p3)*dose),data=dilution,start=c(.72,.03,-3),fixed = p1 + p2 + p3~ 1,random=list(sample=~p2+p3~1))

anova(model2,model2a,model2b,model2c,model2d)

#calculating predicted value and confidence interval at dose=100 for each CpG site
out<-summary(model2b)

### Extract Point estimate ###
mean(predict(model2b,newdata=dilution[dilution$dose ==100,]))

### Extract the estimates ###
a<-model2b$coefficients$fixed[1]
b<-model2b$coefficients$fixed[2]
c<-model2b$coefficients$fixed[3]

### Construct Jacobian (at x=100) ###
da<- 1-exp(-100*exp(c))
db<- exp(-100*exp(c))
dc<- b*exp(-100*exp(c))*(-100*exp(c))+a*exp(-100*exp(c))*(100*exp(c))
vec<-matrix(c(da,db,dc),ncol=3)

### Estimate CI with Bonferroni adjustment ###
### Using the delta method to estimate variance JV(t)J^T where J is Jacobian, V(t) is variance-covariance matrix. ^T is transpose.
### and use in constructing the CI with a Bonferroni adjustment
mean(predict(model2b,newdata=dilution[dilution$dose ==100,])) - qnorm(1-0.05/17)*sqrt(vec%*%out$varFix%*%t(vec))
mean(predict(model2b,newdata=dilution[dilution$dose==100,])) + qnorm(1-0.05/17)*sqrt(vec%*%out$varFix%*%t(vec))

#################
## Residual Plots ##
#################

par(mfrow=c(4,5))
ident<-rep(0,2)
other<-c(-.5,1.5)

# Residuals vs fitted values
#Change main depending on CpG site
#plot(resid(model2b)~fitted(model2b),xlab="Fitted Values",ylab="Standardized Residuals",main="CpG Site 1: EFNB1-645",ylim=c(-.08,.08),xlim=c(0,1))
# lines(ident~other)

# Fitted mean methylation vs dose
fitted.mean<-numeric()
fitted.mean[1]<-mean(fitted(model2b)[CpGdata$dose=="100"])
fitted.mean[2]<-mean(fitted(model2b)[CpGdata$dose=="50"])
fitted.mean[3]<-mean(fitted(model2b)[CpGdata$dose=="20"])
fitted.mean[4]<-mean(fitted(model2b)[CpGdata$dose=="10"])
fitted.mean[5]<-mean(fitted(model2b)[CpGdata$dose=="5"])
fitted.mean[6]<-mean(fitted(model2b)[CpGdata$dose=="0"])
dose.mean<-c(100,50,20,10,5,0)
plot(fitted.mean~dose.mean,type='l',ylab="Proportion Methylated",xlab="Percent of female genomic DNA in sample",ylim=c(0,1),main="CpG Site 1: EFNB1-645")

points(CpGdata$meth~CpGdata$dose)
APPENDIX E

Source code for Chapter 5 data simulations for the computation of Type I error under the assumption of Beta distributed data

###################################################
##### This file tests type I error according to the ttest, the filtered
##### ttest, the LRT, the bootstrap test and the beta test under the
##### condition that the data has been generated according to the
##### Beta distribution.
###################################################

library(beadarray)
setwd("D:/")

## Read in data with no normalization or background correction
targets <- read.csv("phenotype.csv", header=T)
liver<-readIllumina(arrayNames=targets$FileName,useImages=FALSE,singleChannel=FALSE,targets=targets,backgroundMethod="none")

## Bead Summary Data
R<-createBeadSummaryData(liver,log=FALSE,imagesPerArray=1,what="R")
G<-createBeadSummaryData(liver,log=FALSE,imagesPerArray=1,what="G")
dim(R) #there are 1624 beadtypes in the Red Channel
dim(G) #there are 1624 beadtypes in the Green Channel

##Merging Red Channel Data with associated gene information
codeR<-featureNames(R)
M<-exprs(R)
RedData<-data.frame(M,codeR)
dim(RedData)
OMA<-read.csv("GS0007005-OMA.csv",header=TRUE)
RedOMA<-
merge(RedData,OMA,by.x="codeR",by.y="IllumiCode.name",all.x=FALSE,all.y=TRUE)
dim(RedOMA)
#There are 1547 beadtypes in the red channel summaries that also have gene annotation
## Merging Green Channel Data with associated gene information

codeG <- featureNames(G)
U <- exprs(G)
GreenData <- data.frame(U, codeG)
dim(GreenData)
GreenOMA <- merge(GreenData, OMA, by.x = "codeG", by.y = "IllumiCode.name", all.x = FALSE, all.y = TRUE)
dim(GreenOMA)
# There are 1547 beadtypes in the green channel summaries that also have gene annotation info

# There are 22 negative control beads

# Red Channel
# create dataframe of negative control beads
indR <- grep("Negative", RedOMA$Gene_Symbol)
negR <- RedOMA[indR,]
# create vector of length 36, where each entry is the sample mean of the neg control beadtype in a particular array
RedMean <- numeric()
for (i in 1:36) {
    RedMean[i] <- mean(negR[, i + 1])
i <- i + 1
}

# Frequency histogram of the means
hist(RedMean, main = paste(" "), xlab = paste("mean of Red (Methylated) Channel"), col = "red")

# Green Channel
# create dataframe of negative control beads
indG <- grep("Negative", GreenOMA$Gene_Symbol)
negG <- GreenOMA[indG,]
# create vector of length 36, where each entry is the sample mean of the negative control beadtype in a particular array
GreenMean <- numeric()
for (i in 1:36) {
    GreenMean[i] <- mean(negG[, i + 1])
i <- i + 1
}
# Frequency histogram of the means

```r
x11()
hist(GreenMean, main=paste(" "), xlab=paste("mean of Green (Unmethylated) Channel"), col="green")
```

### Compute beta variable

# initialize beta data frame to store values
```r
beta<-data.frame()
```

# isolating average intensity for each beadtype for R and G channels
```r
RedAvg<-RedOMA[,2:37]
GreenAvg<-GreenOMA[,2:37]
```

# compute beta
```r
a<-0
for (i in 1:36) {
  j<-1
  for (j in 1:1547) {
    diff1<-RedAvg[j,i]-RedMean[i]
    diff2<-GreenAvg[j,i]-GreenMean[i]
    num<-max(diff1,0)
    den<-num+max(diff2,0)
    n<-num/den
    a<-ifelse(n=="NaN",a+1,a)
    beta[j,i]<-n
  }
  i<-i+1
}
```

# This produces NaN values where beta=0/(0+0)

## Apply Scaling factor
```r
RedMean<-apply(RedOMA[,2:37],2,mean)
GreenMean<-apply(GreenOMA[,2:37],2,mean)
scale2<-numeric()
for (i in 1:36) {
  scale2[i]<-RedMean[i]/GreenMean[i]  ### KJA changes MRed to RedMean[i]
}
```

CorrectedGreenOMA2<-matrix(nrow=1547,ncol=36)
```r
for (i in 1:36) {
  CorrectedGreenOMA2[,i]<-scale2[i]*GreenOMA[,i+1]
}
library(multtest)
# Compute t-test test statistic
liver_results<-mt.teststat(beta,targets$Diagnosis)
# Compute adjusted degrees of freedom for unequal sample size
rawp0.liver<-2*(1-pnorm(abs(liver_results)))
df.func<-function(beta,class=targets$Diagnosis) {
  var1<-var(beta[class=="Cirrhosis HCC"])
  var2<-var(beta[class=="Cirrhosis non-HCC"])
  n1<-length(beta[class=="Cirrhosis HCC"])
  n2<-length(beta[class=="Cirrhosis non-HCC"])
  df<-(var1+var2)^2/(var1^2/(n1-1)+var2^2/(n2-1))
  df
}
degrees.f<apply(beta,1,df.func)
## p values for t distn
rawp0.liver<-2*(1-pt(abs(liver_results),degrees.f))
code<-data.frame(GreenOMA$codeG)
class<-targets$Diagnosis
deltabeta<-numeric()
for (i in 1:1547){
deltabeta[i]<-mean(as.numeric(beta[i,class=="Cirrhosis HCC"]))-mean(as.numeric(beta[i,class=="Cirrhosis non-HCC"]))
}
data<-cbind(code,deltabeta,rawp0.liver)
code<-data.frame(GreenOMA$codeG)
class<-targets$Diagnosis
deltabeta<-numeric()
for (i in 1:1547){
deltabeta[i]<-mean(as.numeric(beta[i,class=="Cirrhosis HCC"]))-mean(as.numeric(beta[i,class=="Cirrhosis non-HCC"]))
}
data<-cbind(code,deltabeta,rawp0.liver)

############################
### Generating Data Null case data: Non-differentially methylated
############################
set.seed(123)
test<-as.character(deltabeta)
site<-1493
sig.beta<-beta[site,]
betapositive<-as.numeric(sig.beta)
mu.beta<-mean(betapositive)
s2.beta<-var(betapositive)
a<-mu.beta*(mu.beta*(1-mu.beta)/s2.beta-1)
p<-(1-mu.beta)*(mu.beta*(1-mu.beta)/s2.beta-1)

n1<-20
n2<-16
set.seed(234)
for (i in 1:3152){
  beta.new[i,1:20]<-rbeta(n1,a,p)
  beta.new[i,21:36]<-rbeta(n2,a,p)
}

library(multtest)
#Compute t-test test statistic
##0 is HCC pos group; 1 is HCC neg group
#Change these values depending on tested sample size
indicator<-rep(0:1,c(20,16))
#indicator<-rep(0:1,c(15,15))
liver.results <- mt.teststat(beta.new, indicator)

# Compute adjusted degrees of freedom for unequal sample size
rawp0.liver <- 2*(1 - pnorm(abs(liver.results)))
df.func <- function(beta.new, class = indicator) {
  var1 <- var(beta.new[class == 0])
  var2 <- var(beta.new[class == 1])
  n1 <- length(beta.new[class == 0])
  n2 <- length(beta.new[class == 1])
  df <- (var1 + var2)^2 / (var1^2 / (n1 - 1) + var2^2 / (n2 - 1))
  df
}
degrees.f <- apply(beta.new, 1, df.func)
p0.liver <- 2*(1 - pt(abs(liver.results), degrees.f)) ## p values for t distn

sig.level <- 0.05
type1 <- numeric()
not.sig.ttest <- p0.liver[1:3152]
for (i in 1:3152) {
  type1[i] <- ifelse(not.sig.ttest[i] < sig.level, 1, 0)
}
type1.error.ttest <- sum(type1)/3152

#################################
#### Bibikova ####
#################################
type1.bib <- numeric()
for (i in 1:3152) {
  delta.beta <- mean(beta.new[i, 1:20]) - mean(beta.new[i, 21:36])
  type1.bib[i] <- ifelse(not.sig.ttest[i] < sig.level & delta.beta > 0.17, 1, 0)
}
type1.error.bib <- sum(type1.bib)/3152
# derivative using R
null.expression<-expression(n*(log(gamma(p+q))-log(gamma(p))-log(gamma(q)))+(p-1)*
log(input)+(q-1)*log(1-input))
dx.null<-deriv(null.expression,c('p','q'))

# grr is the gradient for the null likelihood
grr<-function(par,input) {
p<-par[1]
q<-par[2]
n<-length(input)
c(n*(digamma(p+q)/gamma(p+q)-digamma(p)/gamma(p))+sum(log(input)),
n*(digamma(p+q)/gamma(p+q)-digamma(q)/gamma(q))+sum(log(1-input)))
}

# grr1 is the gradient for the alt likelihood
grr1<-function(par,input1,input2) {
p1<-par[1]
q1<-par[2]
n1<-length(input1)
p2<-par[3]
q2<-par[4]
n2<-length(input2)
c(n1*(digamma(p1+q1)/gamma(p1+q1)-digamma(p1)/gamma(p1))+sum(log(input1)),
n1*(digamma(p1+q1)/gamma(p1+q1)-digamma(q1)/gamma(q1))+sum(log(1-input1)),
n2*(digamma(p2+q2)/gamma(p2+q2)-digamma(p2)/gamma(p2))+sum(log(input2)),
n2*(digamma(p2+q2)/gamma(p2+q2)-digamma(q2)/gamma(q2))+sum(log(1-input2)))
}

# This is the null function to be minimized
null.function<-function(par,input) {
p<-par[1]
q<-par[2]
n<-length(input)
null.like<- n*(log(gamma(p+q))-log(gamma(p))-log(gamma(q)))+(p-1)*sum(log(input))+(q-1)*sum(log(1-input))
-null.like
}

# Alternative function
alt.function<-function(par,input1,input2) {

p1<-par[1]
p2<-par[3]
n1<-length(input1)
n2<-length(input2)

log.like<- n1*(log(gamma(p1+q1))-log(gamma(p1))-log(gamma(q1)))+(p1-1)*sum(log(input1))+(q1-1)*sum(log(1-input1)) + n2*(log(gamma(p2+q2))-log(gamma(p2))-log(gamma(q2)))+(p2-1)*sum(log(input2))+(q2-1)*sum(log(1-input2)) -log.like

null<-numeric()
alt<-numeric()

for (i in 1:3152){
  null[i]<-optim(par=c(.5,.5),fn=null.function,gr=grr,input=beta.new[i,],method='Nelder-Mead')$value
  alt[i]<-optim(par=rep(.5,4),fn=alt.function,input1=beta.new[i,1:20],input2=beta.new[i,21:36],method='Nelder-Mead')$value
}

like.test<--2*(alt-null)
pvals.like<-pchisq(like.test,df=2,lower.tail=FALSE)

type1.like<-numeric()
not.sig.like<-pvals.like[1:3152]

for (i in 1:3152){
  type1.like[i]<-ifelse(not.sig.like[i]<sig.level,1,0)
}

type1.error.like<-sum(type1.like)/3152

#########################

#### Bstrap###############

Red<-red.nonsig
Green<-green.nonsig
#Change limits depending on sample size
Red.pos<-Red[,1:20]
Red.neg<-Red[,21:36]
Green.pos<-Green[,1:20]
Green.neg<-Green[,21:36]
teststat<-numeric()
for (i in 1:3152){
teststat[i]<-mean((Red.pos[i,])/(Red.pos[i,]+Green.pos[i,]))-
mean((Red.neg[i,])/(Red.neg[i,]+Green.neg[i,]))
}

# Number of bootstrap samples
B<-2500

# Compute bootstrap test statistic
b.teststat<-matrix(ncol=B,nrow=3152)

for (b in 1:B) {
 bsample<-sample(1:36,replace=TRUE)
 red<-Red[,bsample]
 green<-Green[,bsample]
 #Change limits depending on sample size
 Red.positive<-red[,1:20]
 Green.positive<-green[,1:20]
 Red.negative<-red[,21:36]
 Green.negative<-green[,21:36]
 for (i in 1:3152){
 b.teststat[i,b]<-mean((Red.positive[i,])/(Red.positive[i,]+Green.positive[i,]))-
 mean((Red.negative[i,])/(Red.negative[i,]+Green.negative[i,]))
 }
}

b.teststat.sort<-matrix(ncol=B,nrow=3152)
for (i in 1:3152){
b.teststat.sort[i,]<-sort(b.teststat[i,])
}

result<-numeric()
for (i in 1:3152){
 result[i]<-ifelse(teststat[i]<b.teststat.sort[i,25]||teststat[i]>b.teststat.sort[i,975],1,0)
}
# Achieved significance level
asl<-numeric()
for (i in 1:3152){
  asl[i]<-sum(ifelse(b.teststat[i,]>abs(teststat[i]) | b.teststat[i,]< -abs(teststat[i]),1,0))/B
}
sig.level<-0.05

# Type 1: say not equal when are equal
type1.bstrap<-numeric()
not.sig.bstrap<-asl[1:3152]
for (i in 1:3152){
type1.bstrap[i]<-ifelse(not.sig.bstrap[i]<sig.level,1,0)
}
type1.error.bstrap<-sum(type1.bstrap)/3152

#############################
### Beta Test ###############
#############################
set.seed(123)
test<-as.character(deltabeta)
site<-1493
sig.beta<-beta[site,]
beta.positive<-as.numeric(sig.beta)
mu.beta<-mean(beta.positive)
s2.beta<-var(beta.positive)
a<-mu.beta*(mu.beta*(1-mu.beta)/s2.beta-1)
p<-(1-mu.beta)*(mu.beta*(1-mu.beta)/s2.beta-1)
mean <- a/(a + p)
variance <- function(a, b) a * b/((a + b + 1) * (a + b)^2)
Q <- (qbeta(0.975, a, p) - mean)/sqrt(variance(a, p))
```r
set.seed(123)
reject <- numeric()
#Change n1 and n2 depending on sample size
n1 <- 20
n2 <- 16
for (i in 1:10000) {
x <- rbeta(n1, a, p)
y <- rbeta(n2, a, p)
Critical <- 0 + Q * sqrt(variance(a, p)/n1 + variance(a,p)/n2)
reject[i] <- ifelse(abs(mean(x) - mean(y)) > Critical, 1,0)
}
#Type I error according to Beta Test
type1error.beta<-sum(reject)/10000
type1error.ttest
type1error.bib
type1error.like
type1error.bstrap
type1error.beta
```
APPENDIX F

Source code for Chapter 5 data simulations for the computation of Type I error under the assumption of Bivariate Normal distributed data

###########################
#### This file tests type I error according to the ttest, the filtered ttest, the LRT, the bootstrap test and the beta test under the condition that the data has been generated according to the Bivariate Normal distribution.####
###########################

library(beadarray)
setwd("D:/")

## Read in data with no normalization or background correction
targets <- read.csv("phenotype.csv", header=T)
liver<-readIllumina(arrayNames=targets$FileName,useImages=FALSE,singleChannel=FALSE,target=targets,backgroundMethod="none")

## Bead Summary Data
R<-createBeadSummaryData(liver,log=FALSE,imagesPerArray=1,what="R")
G<-createBeadSummaryData(liver,log=FALSE,imagesPerArray=1,what="G")
dim(R) #there are 1624 beadtypes in the Red Channel
dim(G) #there are 1624 beadtypes in the Green Channel

##Merging Red Channel Data with associated gene information
codeR<-featureNames(R)
M<-exprs(R)
RedData<-data.frame(M,codeR)
dim(RedData)
OMA<-read.csv("GS0007005-OMA.csv",header=TRUE)
RedOMA<-merge(RedData,OMA,by.x="codeR",by.y="IllumiCode.name",all.x=FALSE,all.y=TRUE)
dim(RedOMA)
#There are 1547 beadtypes in the red channel summaries that also have gene annotation
## Merging Green Channel Data with associated gene information

```
codeG <- featureNames(G)
U <- exprs(G)
GreenData <- data.frame(U, codeG)
dim(GreenData)
GreenOMA <-
  merge(GreenData, OMA, by.x = "codeG", by.y = "IllumiCode.name", all.x = FALSE, all.y = TRUE)
dim(GreenOMA)
# There are 1547 beadtypes in the green channel summaries that also have gene annotation info

# There are 22 negative control beads

# Red Channel
# create dataframe of negative control beads
indR <- grep("Negative", RedOMA$Gene_Symbol)
negR <- RedOMA[indR,]
# create vector of length 36, where each entry is the sample mean of the neg control beadtype in a particular array
RedMean <- numeric()
for (i in 1:36) {
  RedMean[i] <- mean(negR[, i+1])
i <- i+1
}

# Frequency histogram of the means
hist(RedMean, main = paste(" "), xlab = paste("mean of Red (Methylated) Channel"), col = "red")

# Green Channel
# create dataframe of negative control beads
indG <- grep("Negative", GreenOMA$Gene_Symbol)
negG <- GreenOMA[indG,]
# create vector of length 36, where each entry is the sample mean of the # negative control beadtype in a particular array
GreenMean <- numeric()
for (i in 1:36) {
  GreenMean[i] <- mean(negG[, i+1])
i <- i+1
}
# Frequency histogram of the means

```r
x11()
hist(GreenMean,main=paste(" "),xlab=paste("mean of Green (Unmethylated) Channel"),col="green")
```  

### Compute beta variable

```r
# initialize beta data frame to store values
beta<-data.frame()

# isolating average intensity for each beadtype for R and G channels
RedAvg<-RedOMA[,2:37]
GreenAvg<-GreenOMA[,2:37]

# compute beta
a<-0
for (i in 1:36) {
  j<-1
  for (j in 1:1547) {
    diff1<-RedAvg[j,i]-RedMean[i]
    diff2<-GreenAvg[j,i]-GreenMean[i]
    num<-max(diff1,0)
    den<-num+max(diff2,0)
    n<-num/den
    a<-ifelse(n="NaN",a+1,a)
    beta[j,i]<-n
    j<-j+1
  }
  i<-i+1
}
# This produces NaN values where beta=0/(0+0)

## Apply Scaling factor
RedMean<-apply(RedOMA[,2:37],2,mean)
GreenMean<-apply(GreenOMA[,2:37],2,mean)
scale2<-numeric()
for (i in 1:36) {
  scale2[i]<-RedMean[i]/GreenMean[i]  ### KJA changes MRed to RedMean[i]
}

CorrectedGreenOMA2<-matrix(nrow=1547,ncol=36)
for (i in 1:36) {
  CorrectedGreenOMA2[,i]<-scale2[i]*GreenOMA[,i+1]
}
### Two sample t-test to determine differentially methylated sites

```r
library(multtest)
# Compute t-test test statistic
liver.results<-mt.teststat(beta,targets$Diagnosis)
# Compute adjusted degrees of freedom for unequal sample size
rawp0.liver<-2*(1-pnorm(abs(liver.results)))
df.func<-function(beta,class=targets$Diagnosis) {
  var1<-var(beta[class=="Cirrhosis HCC"])
  var2<-var(beta[class=="Cirrhosis non-HCC"])
  n1<-length(beta[class=="Cirrhosis HCC"])
  n2<-length(beta[class=="Cirrhosis non-HCC"])
  df<-(var1+var2)^2/(var1^2/(n1-1)+var2^2/(n2-1))
  df
}
degrees.f<-apply(beta,1,df.func)
## p values for t distn
rawp0.liver<-2*(1-pt(abs(liver.results),degrees.f))

code<-data.frame(GreenOMA$codeG)
class<-targets$Diagnosis
deltabeta<-numeric()
for (i in 1:1547){
deltabeta[i]<-mean(as.numeric(beta[i,class=="Cirrhosis HCC"]))-mean(as.numeric(beta[i,class=="Cirrhosis non-HCC"]))
}
data<-cbind(code,deltabeta,rawp0.liver)
code<-data.frame(GreenOMA$codeG)
class<-targets$Diagnosis
deltabeta<-numeric()
for (i in 1:1547){
deltabeta[i]<-mean(as.numeric(beta[i,class=="Cirrhosis HCC"]))-mean(as.numeric(beta[i,class=="Cirrhosis non-HCC"]))
}
data<-cbind(code,deltabeta,rawp0.liver)

#################
#### Generating Data Null case data: Non-differentially methylated
#################
set.seed(123)
site<-703

library(MASS)
## Differentially expressed CpG sites
nreps<-3152
#Change ncol depending on desired sample size
red.nonsig<-matrix(nrow=nreps,ncol=36)
green.nonsig<-matrix(nrow=nreps,ncol=36)
nonsig.green<-CorrectedGreenOMA2[site,]
green<-as.numeric(nonsig.green)
mu.green.nonsig<-mean(green)
s2.green.nonsig<-var(green)
nonsig.red<-RedOMA[site,2:37]
red<-as.numeric(nonsig.red)
mu.red.nonsig<-mean(red)
s2.red.nonsig<-var(red)
covar<-(0.6)*sqrt(s2.green.nonsig)*sqrt(s2.red.nonsig)
for (i in 1:nreps){
  #Change length depending on desired sample size
  nonsig<-mvrnorm(36,mu=c(mu.red.nonsig,mu.green.nonsig),Sigma=matrix(c(s2.red.nonsig,covar,c
  covar,s2.green.nonsig),nrow=2))
  red.nonsig[i,]<-nonsig[,1]
green.nonsig[i,]<-nonsig[,2]
}
beta.new<-red.nonsig/(red.nonsig+green.nonsig)
library(multtest)
#Compute t-test test statistic
##0 is HCC pos group; 1 is HCC neg group

#Change these values depending on tested sample size
indicator<-rep(0:1,c(20,16))
#indicator<-rep(0:1,c(15,15))
liver.results<-mt.teststat(beta.new,indicator)

#Compute adjusted degrees of freedom for unequal sample size
rawp0.liver<-2*(1-pnorm(abs(liver.results)))
df.func<-function(beta.new,class=indicator) {
  var1<-var(beta.new[class==0])
  var2<-var(beta.new[class==1])
  n1<-length(beta.new[class==0])
  n2<-length(beta.new[class==1])
  df<-(var1+var2)^2/(var1^2/(n1-1)+var2^2/(n2-1))
  df
}
degrees.f<-apply(beta.new,1,df.func)
p0.liver<-2*(1-pt(abs(liver.results),degrees.f)) ##p values for t distn

sig.level<-0.05
type1<-numeric()

not.sig.ttest<-p0.liver[1:3152]
for (i in 1:3152){
type1[i]<-ifelse(not.sig.ttest[i]<sig.level,1,0)
}
type1.error.ttest<-sum(type1)/3152
type1.bib<-numeric()
for (i in 1:3152){
delta.beta<-mean(beta.new[i,1:20])-mean(beta.new[i,21:36])
type1.bib[i]<-ifelse(not.sig.ttest[i]<sig.level & delta.beta>0.17,1,0)
}
type1.error.bib<-sum(type1.bib)/3152

### LR Test###

# the derivative using R
null.expression<-expression(n*(log(gamma(p+q))-log(gamma(p))-log(gamma(q)))+(p-1)*log(input)+(q-1)*log(1-input))
dx.null<-deriv(null.expression,c('p','q'))
# grr is the gradient for the null likelihood
grr<-function(par,input) {
p<-par[1]
q<-par[2]
n<-length(input)
c(n*(digamma(p+q)/gamma(p+q)-digamma(p)/gamma(p))+sum(log(input)),
   n*(digamma(p+q)/gamma(p+q)-digamma(q)/gamma(q))+sum(log(1-input)))
}
# grr1 is the gradient for the alt likelihood
grr1<-function(par,input1,input2) {
p1<-par[1]
q1<-par[2]
n1<-length(input1)
p2<-par[3]
q2<-par[4]
n2 <- length(input2)
c(n1*(digamma(p1+q1)/gamma(p1+q1)-digamma(p1)/gamma(p1))+sum(log(input1)),
n1*(digamma(p1+q1)/gamma(p1+q1)-digamma(q1)/gamma(q1))+sum(log(1-input1)),
n2*(digamma(p2+q2)/gamma(p2+q2)-digamma(p2)/gamma(p2))+sum(log(input2)),
n2*(digamma(p2+q2)/gamma(p2+q2)-digamma(q2)/gamma(q2))+sum(log(1-input2))
)

# This is the null function to be minimized
null.function <- function(par,input) {
  p <- par[1]
  q <- par[2]
  n <- length(input)
  null.like <- n*(log(gamma(p+q))-log(gamma(p))-log(gamma(q)))+(p-1)*sum(log(input))+ (q-1)*sum(log(1-input))
  -null.like
}

# Alternative function
alt.function <- function(par,input1,input2) {
  p1 <- par[1]
  q1 <- par[2]
  n1 <- length(input1)
  p2 <- par[3]
  q2 <- par[4]
  n2 <- length(input2)
  log.like <- n1*(log(gamma(p1+q1))-log(gamma(p1))-log(gamma(q1)))+(p1-1)*sum(log(input1))+ (q1-1)*sum(log(1-input1)) + n2*(log(gamma(p2+q2))-log(gamma(p2))-log(gamma(q2)))+(p2-1)*sum(log(input2))+ (q2-1)*sum(log(1-input2))
  -log.like
}

null <- numeric()
alt <- numeric()
for (i in 1:3152) {
  null[i] <- optim(par=c(.5,.5), fn=null.function, gr=grr, input=beta.new[i,], method='Nelder-Mead')$value
  # Change limits depending on sample size
  alt[i] <- optim(par=rep(.5,4), fn=alt.function, input1=beta.new[i,1:20], input2=beta.new[i,21:36], method='Nelder-Mead')$value
}
like.test <- -2*(alt-null)
pvals.like<-pchisq(like.test,df=2,lower.tail=FALSE)
#Type 1: say not equal when are equal
type1.like<-numeric()
not.sig.like<-pvals.like[1:3152]
for (i in 1:3152){
type1.like[i]<-ifelse(not.sig.like[i]<sig.level,1,0)
}
type1.error.like<-sum(type1.like)/3152

############################
#### Bstrap#################
############################
Red<-red.nonsig
Green<-green.nonsig
#Change indices depending on sample size
Red.pos<-Red[,1:20]
Red.neg<-Red[,21:36]
Green.pos<-Green[,1:20]
Green.neg<-Green[,21:36]
teststat<-numeric()
for (i in 1:3152){
teststat[i]<-mean((Red.pos[i,])/(Red.pos[i,]+Green.pos[i,]))-
mean((Red.neg[i,])/(Red.neg[i,]+Green.neg[i,]))
}
# Number of bootstrap samples
B<-2500

# Compute bootstrap test statistic
b.teststat<-matrix(ncol=B,nrow=3152)
for (b in 1:B) {
bsample<-sample(1:36,replace=TRUE)
red<-Red[,bsample]
green<-Green[,bsample]
#Change indices depending on Sample size
Red.positive <- red[,1:20]
Green.positive <- green[,1:20]
Red.negative <- red[,21:36]
Green.negative <- green[,21:36]
for (i in 1:3152) {
b.teststat[i,b] <- mean((Red.positive[i,])/(Red.positive[i,]+Green.positive[i,]))-
mean((Red.negative[i,])/(Red.negative[i,]+Green.negative[i,]))
}

b.teststat.sort <- matrix(ncol=B,nrow=3152)
for (i in 1:3152) {
b.teststat.sort[i,] <- sort(b.teststat[i,])
}

result <- numeric()
for (i in 1:3152) {
result[i] <- ifelse(teststat[i]<b.teststat.sort[i,25] | teststat[i]>b.teststat.sort[i,975],1,0)
}

# Achieved significance level
asl <- numeric()
for (i in 1:3152) {
asl[i] <- sum(ifelse(b.teststat[i,]>abs(teststat[i]) | b.teststat[i,]< -abs(teststat[i]),1,0))/B
}
sig.level <- 0.05

# Type 1: say not equal when are equal
type1.bstrap <- numeric()
not.sig.bstrap <- asl[1:3152]
for (i in 1:3152) {
type1.bstrap[i] <- ifelse(not.sig.bstrap[i]<sig.level,1,0)
}
type1.error.bstrap <- sum(type1.bstrap)/3152

####################################################
####### Beta Test ###################################
####################################################
set.seed(123)

site<-1493

library(MASS)

nonsig.green<-CorrectedGreenOMA2[site,]

green<-as.numeric(nonsig.green)

mu.green.nonsig<-mean(green)

s2.green.nonsig<-var(green)

nonsig.red<-RedOMA[site,2:37]

red<-as.numeric(nonsig.red)

mu.red.nonsig<-mean(red)

s2.red.nonsig<-var(red)

covar<-(0.6)*sqrt(s2.green.nonsig)*sqrt(s2.red.nonsig)

nonsig<-mvrnorm(1000,mu=c(mu.red.nonsig,mu.green.nonsig),Sigma=matrix(c(s2.red.nonsig,covar,covar,s2.green.nonsig),nrow=2))

red.nonsig<-nonsig[,1]

green.nonsig<-nonsig[,2]

beta<-red.nonsig/(red.nonsig+green.nonsig)

mu.beta<-mean(beta)

s2.beta<-var(beta)

a<-mu.beta*(mu.beta*(1-mu.beta)/s2.beta-1)
p<-(1-mu.beta)*(mu.beta*(1-mu.beta)/s2.beta-1)

mean <- a/(a + p)

variance <- function(a, b) a * b/((a + b + 1) * (a + b)^2)

Q <- (qbeta(0.975, a, p) - mean)/sqrt(variance(a, p))

set.seed(123)
reject <- numeric()
# Change n1 and n2 depending on sample size
n1 <- 20
n2 <- 16
for (i in 1:10000) {
  x <- rbeta(n1, a, p)
  y <- rbeta(n2, a, p)
  Critical <- 0 + Q * sqrt(variance(a, p)/n1 + variance(a,p)/n2)
  reject[i] <- ifelse(abs(mean(x) - mean(y)) > Critical, 1, 0)
}

type1error.beta<-sum(reject)/10000
APPENDIX G

Source code for Chapter 5 data simulations for the computation of Type II error under the assumption of Beta distributed data

###################################################
##### This file tests type II error according to the ttest, the filtered ttest, the LRT, the bootstrap test and the beta test under the condition that the data has been generated according to the Beta distribution. Concludes with the application of the Beta test to HCV data set.
###################################################

library(beadarray)
setwd("D:/")

## Read in data with no normalization or background correction
targets <- read.csv("phenotype.csv", header=T)
liver<-readIllumina(arrayNames=targets$FileName,useImages=FALSE,singleChannel=FALSE,targets=targets,backgroundMethod="none")

## Bead Summary Data
R<-createBeadSummaryData(liver,log=FALSE,imagesPerArray=1,what="R")
G<-createBeadSummaryData(liver,log=FALSE,imagesPerArray=1,what="G")
dim(R) #there are 1624 beadtypes in the Red Channel
dim(G) #there are 1624 beadtypes in the Green Channel

##Merging Red Channel Data with associated gene information
codeR<-featureNames(R)
M<-exprs(R)
RedData<-data.frame(M,codeR)
dim(RedData)
OMA<-read.csv("GS0007005-OMA.csv",header=TRUE)
RedOMA<-
merge(RedData,OMA,by.x="codeR",by.y="IllumiCode.name",all.x=FALSE,all.y=TRUE)
dim(RedOMA) #There are 1547 beadtypes in the red channel summaries that also have gene annotation info
## Merging Green Channel Data with associated gene information

codeG <- featureNames(G)
U <- exprs(G)
GreenData <- data.frame(U, codeG)
dim(GreenData)

# OMA <- read.csv("GS0007005-OMA.csv", header=TRUE)
GreenOMA <-
merge(GreenData, OMA, by.x = "codeG", by.y = "IllumiCode.name", all.x = FALSE, all.y = TRUE)
dim(GreenOMA)  # There are 1547 beadtypes in the green channel summaries that also have gene annotation info

# There are 22 negative control beads

# Red Channel
# create dataframe of negative control beads
indR <- grep("Negative", RedOMA$Gene_Symbol)
negR <- RedOMA[indR,]
# create vector of length 36, where each entry is the sample mean of the negative control bead type in a particular array
RedMean <- numeric()
for (i in 1:36) {
    RedMean[i] <- mean(negR[, i + 1])
i <- i + 1
}

# Frequency histogram of the means
hist(RedMean, main = paste(" "), xlab = paste("mean of Red (Methylated) Channel"), col = "red")

# Green Channel
# create dataframe of negative control beads
indG <- grep("Negative", GreenOMA$Gene_Symbol)
negG <- GreenOMA[indG,]
# create vector of length 36, where each entry is the sample mean of the negative control bead type in a particular array
GreenMean <- numeric()
for (i in 1:36) {
    GreenMean[i] <- mean(negG[, i + 1])
i <- i + 1
}

# Frequency histogram of the means
# Compute beta

# initialize beta data frame to store values
beta<-data.frame()

# isolating average intensity for each beadtype for R and G channels
RedAvg<-RedOMA[,2:37]
GreenAvg<-GreenOMA[,2:37]

# compute beta
a<-0
for (i in 1:36) {
  j<-1
  for (j in 1:1547) {
    diff1<-RedAvg[j,i]-RedMean[i]
    diff2<-GreenAvg[j,i]-GreenMean[i]
    num<-max(diff1,0)
    den<-num+max(diff2,0)
    n<-num/den
    a<-ifelse(n=="NaN",a+1,a)
    beta[j,i]<-n
    j<-j+1
  }
  i<-i+1
}
# This produces NaN values where beta=0/(0+0)
# Set NaN values to 0 ??

### Scale

RedMean<-apply(RedOMA[,2:37],2,mean)
GreenMean<-apply(GreenOMA[,2:37],2,mean)
scale2<-numeric()
for (i in 1:36) {
  scale2[i]<-RedMean[i]/GreenMean[i]  ### KJA changes MRed to RedMean[i]
}

CorrectedGreenOMA2<-matrix(nrow=1547,ncol=36)
for (i in 1:36){
  CorrectedGreenOMA2[,i]<-scale2[i]*GreenOMA[,i+1]
}

#################
### Two sample t-test to determine differentially methylated sites
 #################

library(multtest)
#Compute t-test test statistic
liver.results<-mt.teststat(beta,targets$Diagnosis)
#Compute adjusted degrees of freedom for unequal sample size
rawp0.liver<-2*(1-pnorm(abs(liver.results)))
df.func<-function(beta,class=targets$Diagnosis) {
  var1<-var(beta[class=="Cirrhosis HCC"])
  var2<-var(beta[class=="Cirrhosis non-HCC"])
  n1<-length(beta[class=="Cirrhosis HCC"])
  n2<-length(beta[class=="Cirrhosis non-HCC"])
  df<-(var1+var2)^2/(var1^2/(n1-1)+var2^2/(n2-1))
  df
}
degrees.f<-apply(beta,1,df.func)
rawp0.liver<-2*(1-pt(abs(liver.results),degrees.f)) ##p values for t distn

code<-data.frame(GreenOMA$codeG)
class<-targets$Diagnosis
deltabeta<-numeric()
for (i in 1:1547){
deltabeta[i]<-mean(as.numeric(beta[i,class=="Cirrhosis HCC"]))-mean(as.numeric(beta[i,class=="Cirrhosis non-HCC"]))
}
data<-cbind(code,deltabeta,rawp0.liver)
code<-data.frame(GreenOMA$codeG)
class<-targets$Diagnosis
deltabeta<-numeric()
for (i in 1:1547){
deltabeta[i]<-mean(as.numeric(beta[i,class=="Cirrhosis HCC"]))-mean(as.numeric(beta[i,class=="Cirrhosis non-HCC"]))
}
data<-cbind(code,deltabeta,rawp0.liver)

#To identify sites with specific deltabetas
# Compute delta beta

delbeta<-numeric()
mu.HCCpos<-numeric()
mu.HCCneg<-numeric()
abs.delbeta<-numeric()
for (i in 1:1547){
  mu.HCCpos[i]<-mean(as.numeric(beta[i,targets$Diagnosis=="Cirrhosis HCC"]))
  mu.HCCneg[i]<-mean(as.numeric(beta[i,targets$Diagnosis=="Cirrhosis non-HCC"]))
  delbeta[i]<-mu.HCCpos[i]-mu.HCCneg[i]
  abs.delbeta[i]<-abs(delbeta[i])
}

index<-1:1547
new.data<-cbind(index,mu.HCCpos,mu.HCCneg,delbeta,abs.delbeta,rawp0.liver)

# sites of interest are those where both means > 0.2
ind<-numeric()
ind.a<-numeric()
for (i in 1:1547){
  ind[i]<-ifelse(mu.HCCpos[i]>0.2 & mu.HCCneg[i]>0.2,1,0)
}
of.interest<-new.data[ind==1,]

# sort by increasing values of absolute value of delta beta
of.interest<-of.interest[order(of.interest[,5]),]

# 843 sites where both means > 0.2
# for testing about 0.17
ind2<-numeric()
for (i in 1:843){
  ind2[i]<-ifelse(.15<of.interest[i,5] & of.interest[i,5]<.18,1,0)
}
sub2<-of.interest[ind2==1,]

# site 946

# for testing about 0.2
ind3<-numeric()
for (i in 1:843){
  ind3[i]<-ifelse(.18<of.interest[i,5],1,0)
}
sub3<-of.interest[ind3==1,]

# site 1051

# for testing about 0.1
ind4<-numeric()
for (i in 1:843){
  ind4[i]<-ifelse(.08<of.interest[i,5] & of.interest[i,5]<.12,1,0)
}
sub4<-of.interest[ind4==1,]
# site 441

# for testing about .5
ind5<-numeric()
for (i in 1:843){
  ind5[i]<-ifelse(.049<of.interest[i,5] && of.interest[i,5]<.055,1,0)
}
sub5<-of.interest[ind5==1,]
# site 920

##############################
# Generate beta data###
##############################

## Type II error
## Generate significantly differentially expressed data
nreps<-3152

set.seed(123)
test<as.character(deltabeta)
site<-946

sig.beta<-beta[site,]
beta.positive<-as.numeric(sig.beta[targets$Diagnosis=='Cirrhosis HCC'])
mu.beta.positive<-mean(beta.positive)
s2.beta.positive<-var(beta.positive)
a.positive<-mu.beta.positive*(mu.beta.positive*(1-mu.beta.positive)/s2.beta.positive-1)
p.positive<-(1-mu.beta.positive)*(mu.beta.positive* (1-mu.beta.positive)/s2.beta.positive-1)
beta.negative<-as.numeric(sig.beta[targets$Diagnosis=='Cirrhosis non-HCC'])
mu.beta.negative<-mean(beta.negative)
s2.beta.negative<-var(beta.negative)
a.negative<-mu.beta.negative*(mu.beta.negative* (1-mu.beta.negative)/s2.beta.negative-1)
p.negative<-(1-mu.beta.negative)*(mu.beta.negative*(1-mu.beta.negative)/s2.beta.negative-1)

#Change ncol depending on sample size

beta.new<-matrix(nrow=nreps,ncol=36)

set.seed(123)

for (i in 1:nreps)
{
  #change length of data depending on sample size
  beta.new[i,1:20]<-rbeta(20,a.positive,p.positive)
  beta.new[i,21:36]<-rbeta(16,a.negative,p.negative)
}

########################################################################
#### T Test####################
########################################################################

library(multtest)

#Compute t-test test statistic

##0 is HCC pos group; 1 is HCC neg group

#Change these numbers depending on sample size

indicator<-rep(0:1,c(20,16))

liver.results<-mt.teststat(beta.new,indicator)

#Compute adjusted degrees of freedom for unequal sample size

rawp0.liver<-2*(1-pnorm(abs(liver.results)))

df.func<-function(beta.new,class=indicator) {
  var1<-var(beta.new[class==0])
  var2<-var(beta.new[class==1])
  n1<-length(beta.new[class==0])
  n2<-length(beta.new[class==1])
  df<-(var1+var2)^2/(var1^2/(n1-1)+var2^2/(n2-1))
  df
degrees.f<-apply(beta.new,1,df.func)
p0.liver<-2*(1-pt(abs(liver.results),degrees.f)) ## p values for t distn

sig.level<-0.05
type2<-numeric()
not.sig.ttest<-p0.liver
type2<-numeric()
for (i in 1:nreps){
type2[i]<-ifelse(p0.liver[i]>sig.level,1,0)
}
type2error.ttest<-sum(type2)/nreps

###############################################################
#### Bibikova#######
#### Filtered ttest#####
###############################################################
type2.bib<-numeric()
for (i in 1:nreps){
#Change indices depending on sample size
delta.beta<-abs(mean(beta.new[i,1:20])-mean(beta.new[i,21:36]))
type2.bib[i]<-ifelse(not.sig.ttest[i]>sig.level | delta.beta<0.17,1,0)
}
type2error.bib<-sum(type2.bib)/nreps

###############################################################
#### LR Test########
###############################################################
# derivative using R
null.expression <- expression(n*(log(gamma(p+q))-log(gamma(p))-log(gamma(q)))+(p-1)*log(input)+(q-1)*log(1-input))
dx.null <- deriv(null.expression, c('p', 'q'))
# grr is the gradient for the null likelihood
grr <- function(par, input) {
p <- par[1]
q <- par[2]
n <- length(input)
c(n*(gamma(p+q)*digamma(p+q)/gamma(p+q)-gamma(p)*digamma(p)/gamma(p))+sum(log(input)),
n*(gamma(p+q)*digamma(p+q)/gamma(p+q)-gamma(q)*digamma(q)/gamma(q))+sum(log(1-input)))
}
# null function to be minimized
null.function <- function(par, input) {
p <- par[1]
q <- par[2]
n <- length(input)
null.like <- n*(log(gamma(p+q))-log(gamma(p))-log(gamma(q)))+(p-1)*sum(log(input))+(q-1)*sum(log(1-input)) - null.like
}
# Alternative function
alt.function <- function(par, input1, input2) {
p1 <- par[1]
q1 <- par[2]
n1 <- length(input1)
p2 <- par[3]
q2 <- par[4]
n2 <- length(input2)
log.like <- -n1*(log(gamma(p1+q1))-log(gamma(p1))-log(gamma(q1)))+(p1-1)*sum(log(input1))+(q1-1)*sum(log(1-input1)) + n2*(log(gamma(p2+q2))-log(gamma(p2))-log(gamma(q2)))+(p2-1)*sum(log(input2))+(q2-1)*sum(log(1-input2)) - log.like
}
null <- numeric()
alt <- numeric()
for (i in 1:nreps) {
null[i] <- optim(par = c(.5, .5), fn = null.function, gr = grr, input = beta.new[i, ], method = 'Nelder-Mead')$value
}
alt[i] <-
optim(par=rep(.5,4), fn=alt.function, input1=beta.new[i, 1:20], input2=beta.new[i, 21:36], method='Nelder-Mead')$value
}
like.test <- -2*(alt-null)

pvals.like <- pchisq(like.test, df=2, lower.tail=FALSE)

# Type 2: say equal when not equal

sig.like <- pvals.like

type2.like <- numeric()
for (i in 1:nreps){
type2.like[i] <- ifelse(sig.like[i] > sig.level, 1, 0)
}
type2.error.like <- sum(type2.like)/nreps

####################################################

#### Bootstrap
####################################################

Red <- red.data.new
Green <- green.data.new
# Change indices depending on sample size
Red.pos <- Red[, 1:20]
Red.neg <- Red[, 21:36]
Green.pos <- Green[, 1:20]
Green.neg <- Green[, 21:36]
teststat <- numeric()
for (i in 1:nreps){
teststat[i] <- mean((Red.pos[i,])/(Red.pos[i,] + Green.pos[i,]))-
mean((Red.neg[i,])/(Red.neg[i,] + Green.neg[i,]))
}

# Number of bootstrap samples
B <- 2500

b.teststat <- matrix(ncol=B, nrow=nreps)
set.seed(123)
for (b in 1:B) {
    bsample<-sample(1:36,replace=TRUE)
    red<-Red[,bsample]
    green<-Green[,bsample]
    Red.positive<-red[,1:20]
    Green.positive<-green[,1:20]
    Red.negative<-red[,21:36]
    Green.negative<-green[,21:36]
    for (i in 1:nreps){
        b.teststat[i,b]<-mean((Red.positive[i,])/ (Red.positive[i,]+Green.positive[i,]))-
        mean((Red.negative[i,])/ (Red.negative[i,]+Green.negative[i,]))
    }
}

b.teststat.sort<-matrix(ncol=B,nrow=nreps)
for (i in 1:nreps){
    b.teststat.sort[i,]<-sort(b.teststat[i,])
}

result<-numeric()
for (i in 1:nreps){
    result[i]<-ifelse(teststat[i]<b.teststat.sort[i,25] | teststat[i]>b.teststat.sort[i,975],1,0)
}

asl<-numeric()
for (i in 1:nreps){
    asl[i]<-sum(ifelse(b.teststat[i,]>abs(teststat[i]) | b.teststat[i,]<-abs(teststat[i]),1,0))/B
}
sig.level<-0.05

# Type 2: say equal when not equal

sig.bstrap<-asl

type2.bstrap<-numeric()
for (i in 1:nreps){
    type2.bstrap[i]<-ifelse(sig.bstrap[i]>sig.level,1,0)
}

type2error.bstrap<-sum(type2.bstrap)/nreps
variance <- function(a, b) a * b/((a + b + 1) * (a + b)^2)

mean.beta <- function(x,y) {
    #xbar <- (length(x)*mean(x)+length(y)*mean(y))/(length(x)+length(y))
    xbar<-mean(c(x,y))
    #v <-
    (length(x)*(var(x)+(mean(x))^2)+length(y)*(var(y)+(mean(y))^2))/(length(x++length(y))-
    xbar^2
    v<-var(c(x,y))*(length(c(x,y))-1)/length(c(x,y))
    a <- xbar * (xbar * (1 - xbar)/v - 1)
    b <- (1 - xbar) * (xbar * (1 - xbar)/v - 1)
    var <- variance(a, b)
    list(xbar = xbar, alpha = a, beta = b, var = var)
}

set.seed(123)
x.pos<-rbeta(3152,a.positive,p.positive)
set.seed(123)
x.neg<-rbeta(3152,a.negative,p.negative)

combined <- mean.beta(x.pos, x.neg)
Q <- (qbeta(0.975, combined$Alpha, combined$Beta) - combined$Alpha/(combined$Alpha + combined$Beta))/sqrt(variance(combined$Alpha, combined$Beta))

set.seed(123)
power <- numeric()
n1 <- 20
n2 <- 16
set.seed(123)
for (i in 1:3152) {
x <- rbeta(n1, a.positive, p.positive)
y <- rbeta(n2, a.negative, p.negative)
Critical <- 0 + Q * sqrt(variance(a.positive, p.positive)/n1 + variance(a.negative,p.negative)/n2)
power[i] <- ifelse(abs(mean(x) - mean(y)) > Critical, 1, 0)
}

# Applying Beta test to Archer data

```r
variance <- function(a, b) a * b/((a + b + 1) * (a + b)^2)
mean.beta <- function(x,y) {
  #xbar <- (length(x)*mean(x)+length(y)*mean(y))/(length(x)+length(y))
  xbar<-mean(c(x,y))
  #v <-
  (length(x)*(var(x)+(mean(x))^2)+length(y)*(var(y)+(mean(y))^2))/(length(x++)length(y))-xbar^2
  v<-var(c(x,y))*(length(c(x,y))-1)/length(c(x,y))
  a <- xbar * (xbar * (1 - xbar)/v - 1)
  b <- (1 - xbar) * (xbar * (1 - xbar)/v - 1)
  var <- variance(a, b)
  list(xbar = xbar, alpha = a, beta = b, var = var)
}

Q<-numeric()
for (i in 1:1547){
  x.pos<-as.numeric(beta[i,targets$Diagnosis=="Cirrhosis HCC"])
  x.neg<-as.numeric(beta[i,targets$Diagnosis=="Cirrhosis non-HCC"])
  combined <- mean.beta(x.pos, x.neg)
  Q [i]<- (qbeta(0.975, combined$alpha, combined$beta) -
  combined$alpha/(combined$alpha + combined$beta))/sqrt(variance(combined$alpha, combined$beta))
}

Critical<-numeric()
```
a.positive<-numeric()
p.positive<-numeric()
a.negative<-numeric()
p.negative<-numeric()
reject<-numeric()
index<-numeric()
beta.pval<-numeric()
for (i in 1:1547) {
  x <- as.numeric(beta[i,targets$Diagnosis=="Cirrhosis HCC"])
  y <- as.numeric(beta[i,targets$Diagnosis=="Cirrhosis non-HCC"])
  v.x<-var(x)*(length(x)-1)/length(x)
  v.y<-var(y)*(length(y)-1)/length(y)
  a.positive[i]<- mean(x)* (mean(x) * (1 - mean(x))/v.x - 1)
  p.positive[i]<- (1 - mean(x)) * (mean(x) * (1 - mean(x))/v.x - 1)
  a.negative[i]<-mean(y) * (mean(y) * (1 - mean(y))/v.y - 1)
  p.negative[i]<- (1 -mean(y)) * (mean(y) * (1 -mean(y))/v.y - 1)
  Critical[i]<- 0 + Q[i] * sqrt(variance(a.positive[i], p.positive[i])/length(x) + variance(a.negative[i],p.negative[i])/length(y))
  reject[i]<-ifelse(abs(mean(x)-mean(y))>Critical[i], 1,0)
}
sum(reject,na.rm=TRUE) #277 significant sites for two sided test at 0.05 significance level versus 227 with ttest
ttest.sig<-ifelse(rawp0.liver<0.05,1,0) #205 significant
filt.sig<-ifelse(rawp0.liver>0.05|deltabeta<0.17,0,1) #7 significant
index<-1:1547
compare<-cbind(index,ttest.sig,reject) #see that for the most part, the two tests find the same CpG sites significantly differentially methylated
results<-cbind(as.numeric(mu.HCCpos),as.numeric(mu.HCCneg),as.numeric(delbeta),as.numeric(reject),as.numeric(ttest.sig),as.numeric(filt.sig),as.vector(OMA$Gene_Symbol),as.vector(OMA$Ilmn_ID))
keep<-results[results[,4]==1,]
Source code for Chapter 5 data simulations for the computation of Type II error under the assumption of Bivariate Normal distributed data

library(beadarray)
setwd("D:/")

## Read in data with no normalization or background correction
targets <- read.csv("phenotype.csv", header=T)
liver<-readIllumina(arrayNames=targets$FileName,useImages=FALSE,singleChannel=FALSE,targets=targets,backgroundMethod="none")

## Bead Summary Data
R<-createBeadSummaryData(liver,log=FALSE,imagesPerArray=1,what="R")
G<-createBeadSummaryData(liver,log=FALSE,imagesPerArray=1,what="G")
dim(R) #there are 1624 beadtypes in the Red Channel
dim(G) #there are 1624 beadtypes in the Green Channel

##Merging Red Channel Data with associated gene information
codeR<-featureNames(R)
M<exprs(R)
RedData<-data.frame(M,codeR)
dim(RedData)
OMA<-read.csv("GS0007005-OMA.csv",header=TRUE)
RedOMA<merge(RedData,OMA,by.x="codeR",by.y="IllumiCode.name",all.x=FALSE,all.y=TRUE)
dim(RedOMA)
# There are 1547 beadtypes in the red channel summaries that also have gene annotation info

## Merging Green Channel Data with associated gene information

codeG <- featureNames(G)
U <- exprs(G)
GreenData <- data.frame(U, codeG)
dim(GreenData)

GreenOMA <- merge(GreenData, OMA, by.x = "codeG", by.y = "IllumiCode.name", all.x = FALSE, all.y = TRUE)
dim(GreenOMA)

# There are 1547 beadtypes in the green channel summaries that also have gene annotation info

# There are 22 negative control beads

# Red Channel
# create dataframe of negative control beads
indR <- grep("Negative", RedOMA$Gene_Symbol)
negR <- RedOMA[indR,]
# create vector of length 36, where each entry is the sample mean of the neg control beadtype in a particular array
RedMean <- numeric()
for (i in 1:36) {
    RedMean[i] <- mean(negR[, i + 1])
i <- i + 1
}

# Frequency histogram of the means
hist(RedMean, main = paste(" "), xlab = paste(" mean of Red (Methylated) Channel"), col = "red")

# Green Channel
# create dataframe of negative control beads
indG <- grep("Negative", GreenOMA$Gene_Symbol)
negG <- GreenOMA[indG,]
# create vector of length 36, where each entry is the sample mean of the negative control beadtype in a particular array
GreenMean <- numeric()
for (i in 1:36) {
    GreenMean[i] <- mean(negG[, i + 1])
i <- i + 1
}
# Frequency histogram of the means

x11()
hist(GreenMean, main=paste(" " ), xlab=paste("mean of Green (Unmethylated) Channel"), col="green")

### Compute beta variable

# Initialize beta data frame to store values
beta<-data.frame()
# Isolating average intensity for each beadtype for R and G channels
RedAvg<-RedOMA[,2:37]
GreenAvg<-GreenOMA[,2:37]
# Compute beta
a<-0
for (i in 1:36) {
  j<-1
  for (j in 1:1547) {
    diff1<-RedAvg[j,i]-RedMean[i]
    diff2<-GreenAvg[j,i]-GreenMean[i]
    num<-max(diff1,0)
    den<-num+max(diff2,0)
    n<-num/den
    a<-ifelse(n=="NaN", a+1, a)
    beta[j,i]<-n
    j<-j+1
  }
  i<-i+1
}
# This produces NaN values where beta=0/(0+0)

## Apply scaling factor
RedMean<-apply(RedOMA[,2:37],2,mean)
GreenMean<-apply(GreenOMA[,2:37],2,mean)
scale2<-numeric()
for (i in 1:36) {
  scale2[i]<-RedMean[i]/GreenMean[i]  ### KJA changes MRed to RedMean[i]
}
CorrectedGreenOMA2<-matrix(nrow=1547, ncol=36)
for (i in 1:36){
  CorrectedGreenOMA2[,i]<-scale2[i]*GreenOMA[,i+1]
### Two sample t-test to determine differentially methylated sites

```r
library(multtest)
#Compute t-test test statistic
liver.results<-mt.teststat(beta,targets$Diagnosis)
#Compute adjusted degrees of freedom for unequal sample size
rawp0.liver<-2*(1-pnorm(abs(liver.results)))
df.func<-function(beta,class=targets$Diagnosis) {
  var1<-var(beta[class=="Cirrhosis HCC"])
  var2<-var(beta[class=="Cirrhosis non-HCC"])
  n1<-length(beta[class=="Cirrhosis HCC"])
  n2<-length(beta[class=="Cirrhosis non-HCC"])
  df<-(var1+var2)^2/(var1^2/(n1-1)+var2^2/(n2-1))
  df
}
degrees.f<-apply(beta,1,df.func)
##p values for t distn
rawp0.liver<-2*(1-pt(abs(liver.results),degrees.f))
code<-data.frame(GreenOMA$codeG)
class<-targets$Diagnosis
deltabeta<-numeric()
for (i in 1:1547){
deltabeta[i]<-mean(as.numeric(beta[i,class=="Cirrhosis HCC"]))-
  mean(as.numeric(beta[i,class=="Cirrhosis non-HCC"]))
}
data<-cbind(code,deltabeta,rawp0.liver)
code<-data.frame(GreenOMA$codeG)
class<-targets$Diagnosis
deltabeta<-numeric()
for (i in 1:1547){
deltabeta[i]<-mean(as.numeric(beta[i,class=="Cirrhosis HCC"]))-
  mean(as.numeric(beta[i,class=="Cirrhosis non-HCC"]))
}
data<-cbind(code,deltabeta,rawp0.liver)
```

### To identify sites with specific delta betas

#Compute delta beta
delbeta<-numeric()
mu.HCCpos<-numeric()
mu.HCCneg<-numeric()
abs.deltabeta<-numeric()
```
for (i in 1:1547) {
  mu.HCCpos[i] <- mean(as.numeric(beta[i, targets$Diagnosis == "Cirrhosis HCC"]))
  mu.HCCneg[i] <- mean(as.numeric(beta[i, targets$Diagnosis == "Cirrhosis non-HCC"]))
  delbeta[i] <- mu.HCCpos[i] - mu.HCCneg[i]
  abs.delbeta[i] <- abs(delbeta[i])
}

index <- 1:1547
new.data <- cbind(index, mu.HCCpos, mu.HCCneg, delbeta, abs.delbeta, rawp0.liver)
# sites of interest are those where both means > 0.2
ind <- numeric()
ind.a <- numeric()
for (i in 1:1547) {
  ind[i] <- ifelse(mu.HCCpos[i] > 0.2 & mu.HCCneg[i] > 0.2, 1, 0)
}
of.interest <- new.data[ind == 1,]
# sort by increasing values of absolute value of delta beta
of.interest <- of.interest[order(of.interest[,5]),]

# 843 sites where both means > 0.2
# for testing about 0.17
ind2 <- numeric()
for (i in 1:843) {
  ind2[i] <- ifelse(.15 < of.interest[i,5] & of.interest[i,5] < .18, 1, 0)
}
sub2 <- of.interest[ind2 == 1,]
# site 946

# for testing about 0.2
ind3 <- numeric()
for (i in 1:843) {
  ind3[i] <- ifelse(.18 < of.interest[i,5], 1, 0)
}
sub3 <- of.interest[ind3 == 1,]
# site 1051

# for testing about 0.1
ind4 <- numeric()
for (i in 1:843) {
  ind4[i] <- ifelse(.08 < of.interest[i,5] & of.interest[i,5] < .12, 1, 0)
}
sub4 <- of.interest[ind4 == 1,]
# site 441

# for testing about 0.5
ind5 <- numeric()
for (i in 1:843) {
ind5[i]<-ifelse(.049<of.interest[i,5] && of.interest[i,5]<.055,1,0)
sub5<-of.interest[ind5==1,]
#site 920

#########################################
#########################################
#########################################

## Generating Data that's significantly differently expressed: bivariate normal

#Set seed so data is replicable
set.seed(123)
test<-as.character(deltabeta)

### Change this site depending on site that data is being modeled after
site<-946
sig.green<-CorrectedGreenOMA2[site,]
green.positive<-as.numeric(sig.green[targets$Diagnosis=="Cirrhosis HCC"])
mu.green.positive<-mean(green.positive)
green.negative<-as.numeric(sig.green[targets$Diagnosis=="Cirrhosis non-HCC"])
mu.green.negative<-mean(green.negative)
s2.green.positive<-var(green.positive)
s2.green.negative<-var(green.negative)

sig.red<-RedOMA[site,2:37]
red.positive<-as.numeric(sig.red[targets$Diagnosis=="Cirrhosis HCC"])
mu.red.positive<-mean(red.positive)
red.negative<-as.numeric(sig.red[targets$Diagnosis=="Cirrhosis non-HCC"])
mu.red.negative<-mean(red.negative)
s2.red.positive<-var(red.positive)
s2.red.negative <- var(red.negative)
cov.positive <- (-.6)*sqrt(s2.red.positive)*sqrt(s2.green.positive)
cov.negative <- (-.6)*sqrt(s2.red.negative)*sqrt(s2.green.positive)

library(MASS)

## Differentially expressed CpG sites
nreps <- 3152
set.seed(123)

## Change ncol depending on sample size being tested
red.pos.sig <- matrix(nrow=nreps,ncol=40)
green.pos.sig <- matrix(nrow=nreps,ncol=40)
red.neg.sig <- matrix(nrow=nreps,ncol=40)
green.neg.sig <- matrix(nrow=nreps,ncol=40)

for (i in 1:nreps) {
  ## Change length of generated data depending on sample size
  positive.sig <- mvrnorm(40, mu=c(mu.red.positive,mu.green.positive), Sigma=matrix(c(s2.red.positive,cov.positive,cov.positive,s2.green.positive),nrow=2))
  red.pos.sig[i,] <- positive.sig[,1]
green.pos.sig[i,] <- positive.sig[,2]

  negative.sig <- mvrnorm(40, mu=c(mu.red.negative,mu.green.negative), Sigma=matrix(c(s2.red.negative,cov.negative,cov.negative,s2.green.negative),nrow=2))
  red.neg.sig[i,] <- negative.sig[,1]
green.neg.sig[i,] <- negative.sig[,2]
}

beta1.positive <- red.pos.sig / (red.pos.sig + green.pos.sig)
beta2.negative <- red.neg.sig / (red.neg.sig + green.neg.sig)
delta.beta<-numeric()
sd.beta1.positive<-numeric()
sd.beta2.negative<-numeric()
for (i in 1:nreps){
delta.beta[i]<-mean(beta1.positive[i,])-mean(beta2.negative[i,])
sd.beta1.positive[i]<-sqrt(var(beta1.positive[i,]))
sd.beta2.negative[i]<-sqrt(var(beta2.negative[i,]))
}
summary(delta.beta)
summary(sd.beta1.positive)
summary(sd.beta2.negative)

###Combine data
red.data<-cbind(red.pos.sig,red.neg.sig)
green.data<-cbind(green.pos.sig,green.neg.sig)
beta<-red.data/(red.data+green.data)

###Change negative values to positive
#Change ncol and limit of loop depending on sample size
red.data.new<-matrix(nrow=nreps,ncol=80)
green.data.new<-matrix(nrow=nreps,ncol=80)
for (i in 1:nreps){
  for (j in 1:80){
    red.data.new[i,j]<-ifelse(red.data[i,j]<0,(-red.data[i,j]),red.data[i,j])
    green.data.new[i,j]<-ifelse(green.data[i,j]<0,(-green.data[i,j]),green.data[i,j])
  }
}
beta.new <- red.data.new / (red.data.new + green.data.new)

## Conduct KS test to ensure still follow biv norm
mu.green <- apply(green.data.new, 1, mean)
s2.green <- apply(green.data.new, 1, var)
mu.red <- apply(red.data.new, 1, mean)
s2.red <- apply(red.data.new, 1, var)
red.data.norm <- numeric()
green.data.norm <- numeric()
for (i in 1:nreps) {
  red.data.norm[i] <- ks.test(red.data.new[i, ], "pnorm", mean = mu.red[i], sd = sqrt(s2.red[i]))$p.value
  green.data.norm[i] <- ks.test(green.data.new[i, ], "pnorm", mean = mu.green[i], sd = sqrt(s2.green[i]))$p.value
}
sum(ifelse(green.data.norm > 0.05, 1, 0))
sum(ifelse(red.data.norm > 0.05, 1, 0))
## This demonstrates that vast majority of generated CpG site still follow normal distribution

## Conduct subsequent analysis on red.data.new, green.data.new, beta.new

##############################
### T Test####################

library(multtest)
# Compute t-test test statistic
##0 is HCC pos group; 1 is HCC neg group

#Change these numbers depending on sample size
#indicator<-rep(0:1,c(20,16))
indicator<-rep(0:1,c(40,40))
liver.results<-mt.teststat(beta.new,indicator)

#Compute adjusted degrees of freedom for unequal sample size
rawp0.liver<-2*(1-pnorm(abs(liver.results)))
df.func<-function(beta.new,class=indicator) {
  var1<-var(beta.new[class==0])
  var2<-var(beta.new[class==1])
  n1<-length(beta.new[class==0])
  n2<-length(beta.new[class==1])
  df<-(var1+var2)^2/(var1^2/(n1-1)+var2^2/(n2-1))
  df
}
degrees.f<-apply(beta.new,1,df.func)
p0.liver<-2*(1-pt(abs(liver.results),degrees.f)) ##p values for t distn

sig.level<-0.05

not.sig.ttest<-p0.liver

for (i in 1:nreps){
type2[i]<-ifelse(p0.liver[i]>sig.level,1,0)
}
type2error.ttest<-sum(type2)/nreps

#################################

####Bibikova####

####Filtered ttest####

#################################
type2.bib<-numeric()

for (i in 1:nreps){

#Change limits depending on sample size

delta.beta<-abs(mean(beta.new[i,1:40])-mean(beta.new[i,41:80]))

type2.bib[i]<-ifelse(not.sig.ttest[i]>sig.level | delta.beta<0.17,1,0)
}

type2error.bib<-sum(type2.bib)/nreps

##############################
#### LR Test######
##############################

# Getting the derivative using R

null.expression<-expression(n*(log(gamma(p+q))-log(gamma(p))-log(gamma(q)))+(p-1)*log(input)+(q-1)*log(1-input))

dx.null<-deriv(null.expression,c('p','q'))

# grr is the gradient for the null likelihood

grr<-function(par,input) {
p<-par[1]
q<-par[2]
n<-length(input)
c(n*(gamma(p+q)*digamma(p+q)/gamma(p+q)-gamma(p)*digamma(p)/gamma(p))+sum(log(input)),
   n*(gamma(p+q)*digamma(p+q)/gamma(p+q)-gamma(q)*digamma(q)/gamma(q))+sum(log(1-input)))
}

# null function to be minimized

null.function<-function(par,input) {
p<-par[1]
q<-par[2]
n<-length(input)
null.like<- n*(log(gamma(p+q))-log(gamma(p))-log(gamma(q)))+(p-1)*sum(log(input))+(q-1)*sum(log(1-input))
-null.like
}

# Alternative function
alt.function<-function(par,input1,input2) {
  p1<-par[1]
  q1<-par[2]
  n1<-length(input1)
  p2<-par[3]
  q2<-par[4]
  n2<-length(input2)
  log_like<- n1*(log(gamma(p1+q1))-log(gamma(p1))-log(gamma(q1)))+(p1-1)*sum(log(input1))+(q1-1)*sum(log(1-input1)) + n2* (log(gamma(p2+q2))-log(gamma(p2))-log(gamma(q2)))+(p2-1)*sum(log(input 2))+(q2-1)*sum(log(1-input2)) -log_like
}

null<-numeric()
alt<-numeric()
for (i in 1:nreps){
  null[i]<-optim(par=c(.5,.5),fn=null.function,gr=grr,input=beta.new[i,],method='Nelder-Mead')$value
  #Change limits depending on sample size
  alt[i]<-optim(par=rep(.5,4),fn=alt.function,input1=beta.new[i,1:40],input2=beta.new[i,41:80],method='Nelder-Mead')$value
}
like.test<--2*(alt-null)

pvals.like<-pchisq(like.test,df=2,lower.tail=FALSE)
#Type 2: say equal when not equal

sig.like<-pvals.like
type2.like<-numeric()
for (i in 1:nreps){
type2.like[i]<-ifelse(sig.like[i]>sig.level,1,0)
}
type2error.like<-sum(type2.like)/nreps

########################################################################
#### Bstrap##############
########################################################################
Red<-red.data.new
Green<-green.data.new
#Change limits depending on sample size
Red.pos<-Red[,1:40]
Red.neg<-Red[,41:80]
Green.pos<-Green[,1:40]
Green.neg<-Green[,41:80]
teststat<-numeric()
for (i in 1:nreps){
teststat[i]<-mean((Red.pos[i,])/(Red.pos[i,]+Green.pos[i,]))-
mean((Red.neg[i,])/(Red.neg[i,]+Green.neg[i,]))
}
#Number of bootstrap samples
B<-2500

b.teststat<-matrix(ncol=B,nrow=nreps)
set.seed(123)
for (b in 1:B) {
#Change limits depending on sample size
bsample<-sample(1:80,replace=TRUE)
red<-Red[,bsample]
green<-Green[,bsample]
Red.positive<-red[,1:40]
Green.positive<-green[,1:40]
Red.negative<-red[,41:80]
Green.negative<-green[,41:80]
for (i in 1:nreps){
b.teststat[i,b]<-mean((Red.positive[i,])/(Red.positive[i,]+Green.positive[i,]))-
mean((Red.negative[i,])/(Red.negative[i,]+Green.negative[i,]))
}
}
b.teststat.sort<-matrix(ncol=B,nrow=nreps)
for (i in 1:nreps){
b.teststat.sort[i,]<-sort(b.teststat[i,])
}

result<-numeric()
for (i in 1:nreps){
result[i]<-ifelse(teststat[i]<b.teststat.sort[i,25]|teststat[i]>b.teststat.sort[i,975],1,0)
}
asl<-numeric()
for (i in 1:nreps){
asl[i]<-sum(ifelse(b.teststat[i,]>abs(teststat[i]) | b.teststat[i,]< -abs(teststat[i]),1,0))/B
} 
sig.level<-0.05

#Type 2: say equal when not equal

sig.bstrap<-asl
type2.bstrap<-numeric()
for (i in 1:nreps){
type2.bstrap[i]<-ifelse(sig.bstrap[i]>sig.level,1,0)
}
type2error.bstrap<-sum(type2.bstrap)/nreps

########################
##### Beta Test##########
########################

## Differentially expressed CpG sites

#Generate many observations to get more stable estimate

set.seed(123)

positive.sig<-mvrnorm(3152,mu=c(mu.red.positive,mu.green.positive),Sigma=matrix(c(s2.red.positive,cov.positive,cov.positive,s2.green.positive),nrow=2))

red.positive<-positive.sig[,1]
green.positive<-positive.sig[,2]

negative.sig<-mvrnorm(3152,mu=c(mu.red.negative,mu.green.negative),Sigma=matrix(c(s2.red.negative,cov.negative,cov.negative,s2.green.negative),nrow=2))

red.negative<-negative.sig[,1]
green.negative<-negative.sig[,2]

beta.positive<-red.positive/(red.positive+green.positive)

beta.negative<-red.negative/(red.negative+green.negative)

mu.beta.positive<-mean(beta.positive)

s2.beta.positive<-var(beta.positive)

a.positive<-mu.beta.positive*(mu.beta.positive*(1-mu.beta.positive)/s2.beta.positive-1)

p.positive<-(1-mu.beta.positive)*(mu.beta.positive*(1-mu.beta.positive)/s2.beta.positive-1)

mu.beta.negative<-mean(beta.negative)

s2.beta.negative<-var(beta.negative)

a.negative<-mu.beta.negative*(mu.beta.negative*(1-mu.beta.negative)/s2.beta.negative-1)

p.negative<-(1-mu.beta.negative)*(mu.beta.negative*(1-mu.beta.negative)/s2.beta.negative-1)

variance <- function(a, b) a * b/((a + b + 1) * (a + b)^2)

mean.beta <- function(x,y) {
  xbar<-mean(c(x,y))
  v<-var(c(x,y))*(length(c(x,y))-1)/length(c(x,y))
  a <- xbar * (xbar * (1 - xbar)/v - 1)
  b <- (1 - xbar) * (xbar * (1 - xbar)/v - 1)
  var <- variance(a, b)
  list(xbar = xbar, alpha = a, beta = b, var = var)
}

set.seed(123)

x.pos<-rbeta(3152,a.positive,p.positive)
set.seed(123)
x.neg<-rbeta(3152,a.negative,p.negative)

combined <- mean.beta(x.pos, x.neg)

#Compute Quantile value under null hypothesis of equality
Q <- (qbeta(0.975, combined$alpha, combined$beta) - combined$alpha/(combined$alpha + combined$beta))/sqrt(variance(combined$alpha, combined$beta))

set.seed(123)

power <- numeric()
#Change n1 and n2 based on sample size
n1 <- 40
n2 <- 40
set.seed(123)
for (i in 1:3152) {
  x <- rbeta(n1, a.positive, p.positive)
y <- rbeta(n2, a.negative, p.negative)
Critical <- 0 + Q * sqrt(variance(a.positive, p.positive)/n1 + variance(a.negative, p.negative)/n2)
power[i] <- ifelse(abs(mean(x) - mean(y)) > Critical, 1, 0)
}
type2error.beta<-1-sum(power)/3152

#Type 2 error rates and parameter estimates (for use in calculating variance of beta distribution)
type2error.ttest
type2error.bib
type2error.beta
type2error.like
type2error.bstrap
a.positive
p.positive
a.negative
p.negative
Maria Irene Agapis Capparuccini was born on September 14, 1983 in Richmond, Virginia, and is an American citizen. She received her Bachelor of Arts in Mathematics from the University of Virginia, Charlottesville, Virginia in 2005. She completed her first semester of graduate training at the University of North Carolina at Chapel Hill School of Public Health, Chapel Hill, NC in 2005 before transferring to the Virginia Commonwealth University School of Medicine. Her graduate studies have been supported by NIEHS training grants “Biostatistics for Research in Environmental Health” (UNC) and “Integration of Chemical Mixtures Toxicology and Statistics” (VCU).